ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΧΡΗΣΗ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ ΓΙΑ ΠΕΡΙΓΡΑΦΗ ΒΙΝΤΕΟ

ΤΣΙΑΚΙΡΗΣ ΑΝΤΩΝΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία, 5 Φεβρουαρίου 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΧΡΗΣΗ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ ΓΙΑ ΠΕΡΙΓΡΑΦΗ ΒΙΝΤΕΟ

ΤΣΙΑΚΙΡΗΣ ΑΝΤΩΝΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία, 5 Φεβρουαρίου 2026

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

Deep Learning for Video Description

TSIAKIRIS ANTONIOS

FINAL THESIS

ADVISOR

Kolomvatsos Konstantinos
Associate Professor

Lamia, 5 February 2026

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 05/02/2026

Ο – Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία εξετάζει το Dense Video Captioning, την αυτόματη παραγωγή περιγραφικών λεζάντων για πολλαπλά χρονικά εντοπισμένα τμήματα ενός video, συνδυάζοντας τον χρονικό εντοπισμό συμβάντων με την παραγωγή φυσικής γλώσσας. Κεντρικός στόχος της εργασίας είναι η συγκριτική αξιολόγηση τριών διαφορετικών αρχιτεκτονικών παραγωγής λεζάντων, οι οποίες αντιπροσωπεύουν διακριτές προσεγγίσεις στην κατανόηση video. Συγκεκριμένα, το BLIP αποτελεί ένα frame-based μοντέλο που παράγει περιγραφές από μεμονωμένα frames, το GIT-VATEX αποτελεί ένα video-based μοντέλο που επεξεργάζεται πολλαπλά frames για την κατανόηση της χρονικής δυναμικής και προκύπτει από fine-tuning του GIT στο σύνολο δεδομένων VATEX, ενώ το Qwen2-VL ανήκει στην κατηγορία των σύγχρονων Vision-Language Models με αυξημένες δυνατότητες πολυτροπικής και σημασιολογικής κατανόησης. Για την ολοκληρωμένη επεξεργασία video αναπτύχθηκε ένα ενιαίο και παραμετροποιήσιμο σύστημα Dense Video Captioning, το οποίο επιτρέπει την επιλογή του εκάστοτε μοντέλου και υλοποιεί μια διαδοχική διαδικασία επεξεργασίας. Η διαδικασία αυτή περιλαμβάνει προσαρμοστική ανίχνευση σκηνών, εξαγωγή και επιλογή frames προσαρμοσμένη στις απαιτήσεις κάθε μοντέλου, καθώς και σημασιολογική συγχώνευση παρόμοιων σκηνών με στόχο τη βελτίωση της συνοχής των τελικών αποτελεσμάτων. Η αξιολόγηση του προτεινόμενου συστήματος πραγματοποιήθηκε στο σύνολο δεδομένων ActivityNet Captions με τη χρήση μετρικών που αφορούν τόσο την ακρίβεια χρονικού εντοπισμού (Recall, Precision) όσο και την ποιότητα των παραγόμενων λεζάντων (BLEU, METEOR, ROUGE-L). Επιπλέον, διεξήχθησαν ablation studies με σκοπό την αποτίμηση της συνεισφοράς των επιμέρους συνιστωσών του συστήματος στη συνολική απόδοση.

ABSTRACT

This thesis examines Dense Video Captioning, the automatic generation of descriptive captions for multiple temporally localized segments within a video, combining temporal event localization with natural language generation. The primary objective is the comparative evaluation of three different caption generation architectures representing distinct approaches to video understanding. Specifically, BLIP is a frame-based model that generates descriptions from individual frames, GIT-VATEX is a video-based model that processes multiple frames to capture temporal dynamics and is derived from fine-tuning GIT on the VATEX dataset, while Qwen2-VL belongs to the category of modern Vision-Language Models with enhanced multimodal and semantic understanding capabilities. For comprehensive video processing, a unified and parameterizable Dense Video Captioning system was developed, enabling model selection and implementing a sequential processing pipeline. This pipeline includes adaptive scene detection, frame extraction and selection tailored to each model's requirements, as well as semantic merging of similar scenes to improve the coherence of final results. The evaluation of the proposed system was conducted on the ActivityNet Captions dataset using metrics addressing both temporal localization accuracy (Recall, Precision) and caption quality (BLEU, METEOR, ROUGE-L). Additionally, ablation studies were performed to assess the contribution of individual system components to overall performance.

Table of Contents

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

1.1 Κίνητρο και Σκοπός της Εργασίας

Τα τελευταία χρόνια παρατηρείται εκθετική αύξηση του όγκου video που παράγεται και διακινείται καθημερινά σε διαδικτυακές πλατφόρμες, υπηρεσίες streaming, μέσα κοινωνικής δικτύωσης και εκπαιδευτικά περιβάλλοντα. Η χειροκίνητη περιγραφή και ευρετηρίαση αυτού του περιεχομένου είναι πρακτικά αδύνατη, γεγονός που καθιστά αναγκαία την ανάπτυξη αυτοματοποιημένων μεθόδων κατανόησης και περιγραφής video [1], [2].

Η κλασική μορφή video captioning στοχεύει στην παραγωγή μίας συνοπτικής περιγραφής για ολόκληρο το video, προσφέροντας μια γενική εικόνα του περιεχομένου. Ωστόσο, πολλά πραγματικά video αποτελούνται από διακριτές σκηνές με διαφορετικές δραστηριότητες, γεγονότα ή θεματικές ενότητες, τις οποίες μία και μόνο λεζάντα δεν μπορεί να αποδώσει επαρκώς. Σε τέτοιες περιπτώσεις, απαιτούνται μέθοδοι που λαμβάνουν υπόψη τη χρονική δομή του video και αποδίδουν λεπτομερείς περιγραφές σε επίπεδο επιμέρους τμημάτων [1].

Στο πλαίσιο αυτών των αναγκών, το Dense Video Captioning αναδεικνύεται ως ένα από τα πλέον απαιτητικά και ταυτόχρονα υποσχόμενα πεδία της αυτόματης ανάλυσης video. Στόχος του είναι η πιο λεπτομερής κατανόηση του περιεχομένου, μέσω της απόδοσης περιγραφών στα επιμέρους γεγονότα που εκτυλίσσονται μέσα στο video και όχι απλώς μιας συνολικής περίληψης. Η αυξημένη πολυπλοκότητά του το καθιστά τόσο επιστημονικά ενδιαφέρον όσο και ιδιαίτερα χρήσιμο σε πρακτικές εφαρμογές, ενώ παράλληλα αναδεικνύει και τις βασικές προκλήσεις που σχετίζονται με την κατανόηση της χρονικής εξέλιξης και της σημασιολογίας του οπτικοακουστικού περιεχομένου [2], [3].

Η παρούσα εργασία αποσκοπεί στη συγκριτική αξιολόγηση τριών διαφορετικών προσεγγίσεων στην παραγωγή λεζάντων για video, οι οποίες περιλαμβάνουν μοντέλα που βασίζονται σε στατικές εικόνες, μοντέλα που λαμβάνουν υπόψη τη χρονική δυναμική, καθώς και προηγμένα Vision-Language μοντέλα. Παρότι τα περισσότερα συστήματα Dense Video Captioning απαιτούν εκπαίδευση από την αρχή σε εξειδικευμένα σύνολα δεδομένων με μεγάλο όγκο επισημειωμένου υλικού, τα σύγχρονα προ-εκπαιδευμένα μοντέλα, όπως τα BLIP, GIT-VATEX και Qwen2-VL, προσφέρουν τη δυνατότητα άμεσης αξιοποίησης χωρίς επιπλέον εκπαίδευση. Στόχος της εργασίας είναι να διερευνηθεί κατά πόσο τα μοντέλα αυτά μπορούν να αξιοποιηθούν αποτελεσματικά σε ένα σύστημα Dense Video Captioning, ποιες διαφορές παρουσιάζουν ως προς την ποιότητα των παραγόμενων περιγραφών, την ταχύτητα επεξεργασίας και τις υπολογιστικές απαιτήσεις, καθώς και ποια αρχιτεκτονική είναι καταλληλότερη για διαφορετικά σενάρια χρήσης [4].

1.2 Περιγραφή και Προκλήσεις

Το Dense Video Captioning αποτελεί ένα σύνθετο πρόβλημα που βρίσκεται στο σημείο συνάντησης της Υπολογιστικής Όρασης και της Επεξεργασίας Φυσικής Γλώσσας. Η βασική πρόκληση είναι να αναλυθεί σωστά η χρονική εξέλιξη ενός video, να εντοπιστούν τα σημαντικά γεγονότα και να παραχθούν περιγραφές που αποδίδουν με ακρίβεια τόσο το οπτικό περιεχόμενο όσο και το νόημα των ενεργειών που παρουσιάζονται. Πιο συγκεκριμένα, το πρόβλημα διατυπώνεται ως εξής, έχοντας ένα video συνολικής διάρκειας T δευτερολέπτων, στόχος είναι να εντοπιστούν αυτόματα N σημαντικά γεγονότα και να παραχθεί για το καθένα μια φυσική γλωσσική περιγραφή. Κάθε γεγονός χαρακτηρίζεται από τη χρονική του έναρξη, τη χρονική του λήξη, και την αντίστοιχη λεζάντα που περιγράφει τι συμβαίνει σε εκείνο το διάστημα του video [1], [3].

Η επίλυση αυτού του προβλήματος συνεπάγεται την αντιμετώπιση πολλαπλών προκλήσεων:

- **Χρονικός Εντοπισμός:** Ο προσδιορισμός των χρονικών ορίων των σημαντικών γεγονότων σε ένα video είναι μια μη-τετριμμένη διαδικασία. Τα παραδοσιακά όρια σκηνών που βασίζονται σε απότομες αλλαγές δεν επαρκούν, καθώς πολλά γεγονότα εκτείνονται σε πολλαπλές λήψεις. Απαιτούνται πιο εξελιγμένες τεχνικές που κατανοούν το σημασιολογικό περιεχόμενο και όχι μόνο τις οπτικές ασυνέχειες.
- **Αναπαράσταση Χρονικής Πληροφορίας:** Σε αντίθεση με τις στατικές εικόνες, τα video περιέχουν πλούσια χρονική δυναμική. Η επιλογή των κατάλληλων frame για ανάλυση, η κωδικοποίηση της χρονικής εξέλιξης και η διατήρηση της πληροφορίας κίνησης αποτελούν κρίσιμες αποφάσεις σχεδιασμού.
- **Παραγωγή Φυσικής Γλώσσας:** Οι παραγόμενες λεζάντες πρέπει να είναι όχι μόνο ακριβείς ως προς το οπτικό περιεχόμενο, αλλά και γλωσσικά φυσικές, συνεκτικές και περιεκτικές. Αυτό απαιτεί μοντέλα που κατανοούν βαθιά τόσο την οπτική όσο και τη γλωσσική μορφολογία.
- **Υπολογιστικό Κόστος:** Η επεξεργασία video είναι υπολογιστικά απαιτητική. Τα video υψηλής ανάλυσης μπορεί να περιέχουν δεκάδες χιλιάδες frames, κάνοντας απαγορευτική την πλήρη επεξεργασία κάθε frame με βαριά μοντέλα βαθιάς μάθησης. Απαιτούνται έξυπνες στρατηγικές δειγματοληψίας και αποδοτικές αρχιτεκτονικές.
- **Ποιότητα και Συνέπεια:** Οι λεζάντες διαδοχικών σκηνών πρέπει να είναι σημασιολογικά συνεπείς και να αποφεύγουν περιττή επανάληψη. Παρόμοια γεγονότα θα πρέπει να περιγράφονται με παρόμοιο τρόπο, ενώ διαφορετικά γεγονότα να διακρίνονται ξεκάθαρα.

Στις παλαιότερες προσεγγίσεις, το πρόβλημα αντιμετωπιζόταν με δύο διακριτά στάδια: αρχικά ο εντοπισμός των χρονικών τμημάτων του video και στη συνέχεια η παραγωγή των αντίστοιχων λεζάντων. Η προσέγγιση αυτή παρουσιάζει περιορισμούς, καθώς η έλλειψη ουσιαστικής αλληλεπίδρασης μεταξύ των δύο σταδίων συχνά οδηγεί σε λιγότερο ακριβή αποτελέσματα. Για τον λόγο αυτό, οι πιο σύγχρονες μέθοδοι επιδιώκουν μια ενιαία διαδικασία μάθησης, όπου ο εντοπισμός των γεγονότων και η παραγωγή περιγραφών πραγματοποιούνται ταυτόχρονα, επιτυγχάνοντας καλύτερη συνολική απόδοση [4].

1.3 Συνεισφορά της Εργασίας

Παρότι οι σύγχρονες ερευνητικές προσεγγίσεις στρέφονται όλο και περισσότερο προς ενιαία (end-to-end) μοντέλα, ο στόχος της παρούσας εργασίας δεν είναι η εκπαίδευση ενός νέου εξειδικευμένου μοντέλου, αλλά η συστηματική διερεύνηση και σύγκριση της απόδοσης τριών υπαρχόντων, προ-εκπαιδευμένων μοντέλων που αντιπροσωπεύουν διαφορετικές προσεγγίσεις στην παραγωγή λεζάντων. Για τον σκοπό αυτό, υιοθετείται συνειδητά μια αρχιτεκτονική διακριτών σταδίων, η οποία λειτουργεί ως ελεγχόμενο πειραματικό περιβάλλον και εξασφαλίζει ότι όλα τα μοντέλα αξιολογούνται υπό τις ίδιες ακριβώς συνθήκες. Στο πλαίσιο αυτό, το προτεινόμενο σύστημα λειτουργεί ως πλατφόρμα συγκριτικής αξιολόγησης που επιτρέπει την εμπειρική μελέτη των δυνατοτήτων, των περιορισμών και των ανταλλαγών (trade-offs) μεταξύ διαφορετικών μοντέλων σε πραγματικές συνθήκες εφαρμογής.

Οι κύριες συνεισφορές της εργασίας συνοψίζονται ως εξής:

- **Συστηματική Σύγκριση Τριών Μοντέλων Παραγωγής Λεζάντων:** Το κεντρικό στοιχείο της εργασίας είναι η λεπτομερής αξιολόγηση τριών διαφορετικών προσεγγίσεων. Το BLIP αναλύει μεμονωμένες εικόνες και παράγει περιγραφές με βάση στατικό περιεχόμενο, εξασφαλίζοντας υψηλή ταχύτητα

επεξεργασίας [5]. Το GIT-VATEX επεξεργάζεται ταυτόχρονα πολλαπλές εικόνες από κάθε σκηνή και κατανοεί τη χρονική ροή των γεγονότων [39], [6]. Το Qwen2-VL είναι ένα προηγμένο Vision-Language Model που συνδυάζει βαθιά οπτική και γλωσσική κατανόηση για την παραγωγή πλούσιων, σημασιολογικά ακριβών περιγραφών [7]. Η σύγκριση εξετάζει την ποιότητα των περιγραφών, την ταχύτητα, και τις υπολογιστικές απαιτήσεις.

- **Ενιαίο Σύστημα Επεξεργασίας για Δίκαιη Σύγκριση:** Σχεδίαση και υλοποίηση ενός ενιαίου, modular συστήματος dense video captioning, το οποίο διασφαλίζει δίκαιη και συγκρίσιμη αξιολόγηση διαφορετικών μοντέλων, μέσω κοινής ροής επεξεργασίας. Το σύστημα περιλαμβάνει αυτόματη ανίχνευση σκηνών, εξαγωγή frame προσαρμοσμένη στις απαιτήσεις κάθε μοντέλου, παραγωγή λεζάντων και συγχώνευση διαδοχικών σκηνών, εξασφαλίζοντας ότι όλα τα μοντέλα αξιολογούνται υπό ακριβώς τις ίδιες συνθήκες εισόδου.
- **Αυτόματη Ανίχνευση Σκηνών:** Υλοποιήθηκε μηχανισμός αυτόματου εντοπισμού σκηνών χρησιμοποιώντας τον PySceneDetect με ContentDetector, ο οποίος ανιχνεύει μεταβάσεις βάσει του οπτικού περιεχομένου. Το σύστημα επιτρέπει τη ρύθμιση της ευαισθησίας ανίχνευσης, ώστε να μπορεί να προσαρμόζεται σε διαφορετικούς τύπους video, από δυναμικό περιεχόμενο με συχνές αλλαγές έως video με πιο ομαλές μεταβάσεις [8].
- **Διαφοροποιημένες Στρατηγικές Εξαγωγής Frame:** Για το BLIP, εφαρμόζεται έξυπνη επιλογή του βέλτιστου frame βάσει μετρικών ποιότητας (sharpness και entropy), επιλέγοντας το πιο αντιπροσωπευτικό από πολλαπλά υποψήφια. Για το GIT-VATEX και το Qwen2-VL, χρησιμοποιείται ομοιόμορφη δειγματοληψία πολλαπλών frames (6 και 5 αντίστοιχα) με ισαπέχουσες χρονικές θέσεις εντός κάθε σκηνής, εξασφαλίζοντας πλήρη κάλυψη της χρονικής δυναμικής.
- **Αυτόματη Σημασιολογική Συγχώνευση Σκηνών:** Υλοποιήθηκε αλγόριθμος που αναλύει τις παραγόμενες λεζάντες και εντοπίζει αυτόματα διαδοχικές σκηνές με παρόμοιο περιεχόμενο χρησιμοποιώντας το μοντέλο Sentence Transformers. Οι παρόμοιες σκηνές συγχωνεύονται σε μία ενιαία περιγραφή, επιλέγοντας τη βέλτιστη λεζάντα με βάση συνδυασμό κριτηρίων ομοιότητας και συνοχής [9].
- **Πολυεπίπεδο Σύστημα Αξιολόγησης με Ablation Studies:** Αναπτύχθηκε ολοκληρωμένο σύστημα αξιολόγησης που υπολογίζει τόσο την ακρίβεια χρονικού εντοπισμού (IoU, Precision, Recall) όσο και την ποιότητα των παραγόμενων περιγραφών (BLEU, METEOR, ROUGE-L). Πραγματοποιήθηκαν ablation studies για την αποτίμηση της συνεισφοράς της σημασιολογικής συγχώνευσης και της έξυπνης επιλογής keyframes. Τα αποτελέσματα οπτικοποιούνται με γραφήματα απόδοσης, καμπύλες IoU, κατανομές scores και ανάλυση μήκους λεζάντων [4].
- **Διαδραστική Εφαρμογή με Streamlit:** Δημιουργήθηκε web interface που επιτρέπει σε χρήστες χωρίς τεχνικές γνώσεις να χρησιμοποιήσουν το σύστημα. Υποστηρίζει ανέβασμα video, επιλογή μοντέλου, ρύθμιση παραμέτρων, real-time επεξεργασία με οπτικοποίηση αποτελεσμάτων, και εξαγωγή σε πολλαπλές μορφές (JSON, CSV, SRT, subtitled video).

ΚΕΦΑΛΑΙΟ 2 Μοντέλα Βαθιάς Μάθησης για Επεξεργασία video

2.1 Εισαγωγή στην Επεξεργασία video με Βαθιά Μάθηση

Η αυτόματη κατανόηση και περιγραφή video αποτελεί ένα από τα πιο απαιτητικά προβλήματα στην τομή της Όρασης Υπολογιστών (Computer Vision) και της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Σε αντίθεση με τις στατικές εικόνες, τα video εισάγουν την επιπλέον διάσταση του χρόνου, απαιτώντας μοντέλα που μπορούν να συλλάβουν τόσο χωρικά (spatial) όσο και χρονικά (temporal) μοτίβα. Η αυτόματη περιγραφή video και, ειδικότερα, το Dense Video Captioning προϋποθέτουν την ικανότητα ενός συστήματος να εξάγει αναπαραστάσεις που είναι ταυτόχρονα πλούσιες ως προς τη χωρική πληροφορία (αντικείμενα, σκηνικό, ιδιότητες) και ευαίσθητες στη χρονική εξέλιξη (κινήσεις, γεγονότα, αλληλουχίες ενεργειών). Στην πράξη, το πεδίο έχει αναπτυχθεί σε στενή συνάφεια με τις εξελίξεις σε video understanding / action recognition, όπου η κεντρική πρόκληση είναι η μοντελοποίηση χωροχρονικών προτύπων σε μεγάλη κλίμακα. Για τον λόγο αυτό, η βιβλιογραφία του Dense Video Captioning «κληρονομεί» σε σημαντικό βαθμό τις αρχιτεκτονικές επιλογές και τα διλήμματα που διαμορφώθηκαν αρχικά για αναγνώριση δράσεων και γεγονότων [10].

Το παρόν κεφάλαιο παρουσιάζει μια επισκόπηση των βασικών αρχιτεκτονικών βαθιάς μάθησης που έχουν αναπτυχθεί για την επεξεργασία video, από τα πρώιμα συνελικτικά δίκτυα έως τα σύγχρονα προ-εκπαιδευμένα vision–language μοντέλα.

2.2 Συνελικτικά Νευρωνικά Δίκτυα για Εξαγωγή Χωρικών Χαρακτηριστικών

Η εξαγωγή νοηματικών οπτικών χαρακτηριστικών αποτελεί το πρώτο καθοριστικό βήμα σε κάθε σύστημα κατανόησης video. Ένας visual encoder μετατρέπει την ακατέργαστη οπτική είσοδο (frames) σε συμπαγείς αναπαραστάσεις (embeddings) που κωδικοποιούν πληροφορία για αντικείμενα, σκηνές, αλληλεπιδράσεις και όπου είναι δυνατό στοιχεία κίνησης. Οι προσεγγίσεις διακρίνονται κυρίως με βάση το αν και πώς ενσωματώνουν τη χρονική διάσταση, δηλαδή αν δουλεύουν αποκλειστικά χωρικά (ανά frame) ή αν εξάγουν εξαρχής spatiotemporal αναπαραστάσεις [10].

2.2.1 Δισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα (2D CNNs)

Τα παραδοσιακά δισδιάστατα Συνελικτικά Νευρωνικά Δίκτυα (2D CNNs), όπως τα VGG, ResNet και Inception, σχεδιάστηκαν αρχικά για την ταξινόμηση στατικών εικόνων. Στο πλαίσιο της επεξεργασίας video, τα 2D CNNs εφαρμόζονται συνήθως σε επίπεδο μεμονωμένων frames, εξάγοντας χωρικά χαρακτηριστικά από κάθε στιγμιότυπο ανεξάρτητα. Κάθε frame επεξεργάζεται παράλληλα ή σειριακά, και τα προκύπτοντα χαρακτηριστικά είτε συναθροίζονται (pooling) είτε τροφοδοτούνται σε χρονικά μοντέλα για τη σύλληψη των εξαρτήσεων μεταξύ των frames [11].

Η κύρια αδυναμία των 2D CNNs στο video είναι ότι αδυνατούν να συλλάβουν άμεσα την κίνηση και τις χρονικές μεταβάσεις μεταξύ διαδοχικών frames, καθώς κάθε frame αντιμετωπίζεται ως ανεξάρτητη εικόνα. Αυτό σημαίνει ότι τα 2D CNNs από μόνα τους είναι κατάλληλα για ισχυρή χωρική κατανόηση (π.χ. ποια αντικείμενα υπάρχουν και ποιο είναι το περιβάλλον), αλλά ανεπαρκή όταν απαιτείται λεπτομερής χρονική κατανόηση, όπως η

αναγνώριση δράσεων (action recognition). Για τον λόγο αυτό, συχνά συνδυάζονται με ξεχωριστά χρονικά μοντέλα ή μηχανισμούς προσοχής [11].

2.2.2 Τριοδιάστατα Συνελικτικά Δίκτυα (3D CNNs)

Για να αντιμετωπιστεί ο περιορισμός των 2D CNNs, αναπτύχθηκαν τα τριοδιάστατα Συνελικτικά Νευρωνικά Δίκτυα (3D CNNs), τα οποία επεκτείνουν τη συνέλιξη στη χρονική διάσταση. Ένα 3D CNN εφαρμόζει φίλτρα που κινούνται ταυτόχρονα στο χώρο (ύψος και πλάτος καρέ) και στο χρόνο (διαδοχικά καρέ), επιτρέποντας την εξαγωγή spatiotemporal χαρακτηριστικών [12]. Με αυτόν τον τρόπο, το μοντέλο μπορεί να μάθει μοτίβα κίνησης, μεταβολές σε στάσεις σωμάτων, καθώς και τοπικές χρονικές μεταβάσεις.

Ιστορικά, αρχιτεκτονικές όπως το C3D ανέδειξαν ότι η άμεση μοντελοποίηση μικρών video clips οδηγεί σε πλουσιότερες αναπαραστάσεις για εργασίες όπως η αναγνώριση δράσεων. Στη συνέχεια, προσεγγίσεις όπως το I3D αξιοποίησαν την ιδέα της μεταφοράς γνώσης από ισχυρά 2D backbones, «φουσκώνοντας» (inflating) 2D kernels σε 3D, ώστε το μοντέλο να ξεκινά από καλά χωρικά χαρακτηριστικά και να μαθαίνει συμπληρωματικά χρονικές δυναμικές. Παράλληλα, προτάθηκαν παραλλαγές όπως οι R(2+1)D και P3D, οι οποίες διαχωρίζουν ρητά τη χωρική και χρονική συνέλιξη, επιδιώκοντας καλύτερο συμβιβασμό μεταξύ απόδοσης και κόστους. Παρά τα πλεονεκτήματα, τα 3D CNNs είναι υπολογιστικά απαιτητικά, η προσθήκη της χρονικής διάστασης αυξάνει τον αριθμό των παραμέτρων και το κόστος σε μνήμη, ιδιαίτερα όταν αυξάνονται η ανάλυση και η διάρκεια του clip. Ως αποτέλεσμα, σε πολλές πρακτικές εφαρμογές επιλέγονται είτε πιο ελαφριές spatiotemporal παραλλαγές είτε εντελώς διαφορετικές αρχιτεκτονικές, όπως οι Transformers [12].

2.3 Temporal Modeling: RNNs, LSTMs και Self-Attention Mechanisms

Η αποτελεσματική μοντελοποίηση της χρονικής εξέλιξης σε ακολουθίες video αποτελεί μία από τις βασικές προκλήσεις στην αυτόματη κατανόηση οπτικοακουστικού περιεχομένου. Σε αντίθεση με τις στατικές εικόνες, όπου η εξαγωγή χωρικών χαρακτηριστικών μέσω συνελικτικών νευρωνικών δικτύων είναι συνήθως επαρκής, τα video απαιτούν την ταυτόχρονη επεξεργασία της χωρικής πληροφορίας που αφορά την κατανομή αντικειμένων και γεγονότων σε κάθε frame, καθώς και της χρονικής πληροφορίας που περιγράφει την εξέλιξή τους σε διαδοχικές χρονικές στιγμές. Η ικανότητα ενός μοντέλου να αποτυπώνει με συνέπεια αυτές τις χρονικές εξαρτήσεις επηρεάζει καθοριστικά την απόδοσή του σε εργασίες όπως η αναγνώριση δράσεων, η πρόβλεψη γεγονότων και η παραγωγή περιγραφικών λεζάντων σε video.

2.3.1 Recurrent Neural Networks και Long Short-Term Memory

Τα Recurrent Neural Networks (RNNs) εισήχθησαν ως μια φυσική επέκταση των feedforward νευρωνικών δικτύων για την επεξεργασία ακολουθιακών δεδομένων. Η θεμελιώδης καινοτομία τους έγκειται στη διατήρηση μιας εσωτερικής κατάστασης μνήμης (hidden state), η οποία ενημερώνεται επαναληπτικά καθώς το δίκτυο επεξεργάζεται διαδοχικά στοιχεία της εισόδου. Αυτή η επαναληπτική δομή επιτρέπει στο RNN να «θυμάται» πληροφορίες από προηγούμενες χρονικές στιγμές και να τις ενσωματώνει στην επεξεργασία της τρέχουσας εισόδου. Τυπικά, για μια ακολουθία εισόδων $x1$, $x2$,… , xt , το RNN υπολογίζει την κρυφή κατάσταση ht στη χρονική στιγμή t μέσω της σχέσης:

$$h_t = \sigma(W_{hh} \cdot h_{t-1} + W_{xh} \cdot x_t + b_h)$$

Τα Whh και Wxh αντιπροσωπεύουν τους πίνακες βαρών για την προηγούμενη κρυφή κατάσταση και την τρέχουσα είσοδο αντίστοιχα, bh είναι το διάνυσμα bias, και σ μια μη-γραμμική συνάρτηση ενεργοποίησης (συνήθως tanh ή sigmoid).

Παρά την αρχική τους επιτυχία, τα παραδοσιακά RNNs αντιμετώπισαν σοβαρούς περιορισμούς κατά την εκπαίδευση σε μακρές ακολουθίες. Το φαινόμενο του vanishing gradient, αναλυτικά τεκμηριωμένο από τους Bengio et al. τη δεκαετία του 1990 [13], περιγράφει την εκθετική μείωση της παραγώγου καθώς διαδίδεται προς τα πίσω μέσω πολλών χρονικών βημάτων. Αυτό καθιστά πρακτικά αδύνατη την εκμάθηση μακρινών χρονικών εξαρτήσεων, καθώς οι παράμετροι που επηρεάζουν τα πρώιμα στάδια της ακολουθίας δεν λαμβάνουν επαρκές training signal.

Για την αντιμετώπιση αυτού του θεμελιώδους περιορισμού, οι Hochreiter και Schmidhuber πρότειναν το 1997 την αρχιτεκτονική Long Short-Term Memory (LSTM) [14]. Το LSTM εισάγει έναν πολυπλοκότερο μηχανισμό διαχείρισης μνήμης μέσω της έννοιας των gates (πυλών), οι οποίες ρυθμίζουν τη ροή πληροφορίας εντός του δικτύου. Κάθε LSTM cell περιλαμβάνει τρία βασικά gates, το forget gate, το input gate, και το output gate, καθώς και το cell Ct. Το cell state λειτουργεί ως μια εσωτερική «λωρίδα μεταφοράς» πληροφορίας που διατρέχει ολόκληρη την ακολουθία με ελάχιστες γραμμικές αλληλεπιδράσεις, επιτρέποντας την απρόσκοπτη διάδοση gradients σε μεγάλες χρονικές αποστάσεις.

Οι λειτουργίες των τριών gates ορίζονται ως εξής.

- Το **forget gate** ft καθορίζει ποιο μέρος της προηγούμενης πληροφορίας θα διατηρηθεί:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$$

- Το input gate it ελέγχει πόση νέα πληροφορία θα ενσωματωθεί στο cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- Η ενημέρωση του cell state πραγματοποιείται μέσω γραμμικής παρεμβολής μεταξύ της παλιάς και της νέας πληροφορίας:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t}$$

όπου το σύμβολο $\odot$ υποδηλώνει στοιχειακό (element-wise) πολλαπλασιασμό.

- Τέλος, το output gate ot προσδιορίζει ποιο τμήμα του cell state θα εκτεθεί ως έξοδος:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$

Η αρχιτεκτονική LSTM έφερε επανάσταση στην επεξεργασία ακολουθιών και εφαρμόστηκε με μεγάλη επιτυχία σε πλήθος εργασιών, από τη μηχανική μετάφραση (Machine Translation - MT) και την αναγνώριση ομιλίας (Speech Recognition ή Automatic Speech Recognition - ASR) έως την παραγωγή κειμένου (Text Generation ή Natural Language Generation - NLG). Στο πλαίσιο της επεξεργασίας video, τα LSTMs χρησιμοποιήθηκαν εκτεναμένα ως temporal encoders ή decoders σε συνδυασμό με συνελικτικά δίκτυα για την εξαγωγή χωρικών χαρακτηριστικών.

Ένα χαρακτηριστικό παράδειγμα αποτελεί η πρωτοποριακή εργασία των Venugopalan et al. το 2015 [15], η οποία εισήγαγε την αρχιτεκτονική **Sequence-to-Sequence** για video captioning. Στην προσέγγιση αυτή, ένα συνελικτικό δίκτυο (συνήθως VGG ή ResNet) εξάγει χωρικά χαρακτηριστικά από κάθε frame του video ανεξάρτητα. Αυτά τα χαρακτηριστικά τροφοδοτούνται στη συνέχεια σε έναν LSTM encoder, ο οποίος κωδικοποιεί την ακολουθία των frames σε μια πυκνή αναπαράσταση σταθερού μήκους. Ένας δεύτερος LSTM decoder χρησιμοποιεί αυτή την αναπαράσταση ως αρχική κατάσταση και παράγει

αυτοπαλίνδρομα μια ακολουθία λέξεων που αποτελούν την περιγραφή του video. Η μοντελοποίηση και των δύο ροών πληροφορίας (οπτικής και γλωσσικής) μέσω LSTMs επέτρεψε στο μοντέλο να μαθαίνει τόσο τη χρονική δομή των γεγονότων στο video όσο και τη γλωσσική δομή των παραγόμενων προτάσεων.

Οι Yao et al. επέκτειναν αυτή την προσέγγιση εισάγοντας μηχανισμούς attention σε συνδυασμό με LSTMs [16], επιτρέποντας στον decoder να εστιάζει επιλεκτικά σε συγκεκριμένα frames κατά την παραγωγή κάθε λέξης. Ο συνδυασμός αυτός αύξησε σημαντικά την ποιότητα των παραγόμενων περιγραφών, ιδιαίτερα για video με πολύπλοκες ακολουθίες δράσεων. Παράλληλα, προτάθηκαν και **bidirectional LSTMs** για video description [17], τα οποία επεξεργάζονται την ακολουθία των frames τόσο προς τα εμπρός όσο και προς τα πίσω, συλλαμβάνοντας έτσι πλουσιότερη πληροφορία για το συνολικό πλαίσιο του video.

Παρά την ευρεία τους υιοθέτηση, τα RNNs και τα LSTMs παρουσιάζουν εγγενείς περιορισμούς που γίνονται ιδιαίτερα προβληματικοί στην επεξεργασία μακρών video sequences. Ο κυριότερος περιορισμός αφορά την αναγκαστικά **σειριακή φύση της επεξεργασίας τους**. Κάθε χρονική στιγμή t εξαρτάται άμεσα από τον υπολογισμό της προηγούμενης κατάστασης $ht-1$, γεγονός που καθιστά αδύνατη την παραλληλοποίηση των υπολογισμών κατά τη διάσταση του χρόνου. Για ένα video με T frames, η υπολογιστική πολυπλοκότητα είναι $O(T)$, αλλά ο πραγματικός χρόνος επεξεργασίας είναι γραμμικά εξαρτώμενος από το T λόγω της αδυναμίας παραλληλοποίησης. Αυτό καθιστά τα LSTMs σημαντικά πιο αργά από τα συνελικτικά δίκτυα κατά την εκπαίδευση σε σύγχρονες GPU accelerators, οι οποίες είναι βελτιστοποιημένες για ταυτόχρονους υπολογισμούς.

Δεύτερον, παρότι τα LSTMs αντιμετωπίζουν μερικώς το πρόβλημα του vanishing gradient, εξακολουθούν να **δυσκολεύονται με πολύ μακρινές εξαρτήσεις**. Η πληροφορία που κωδικοποιείται στο cell state υφίσταται διαδοχικές μη-γραμμικές μετατροπές σε κάθε χρονικό βήμα, και αν και οι πύλες επιτρέπουν στην πληροφορία να διατηρείται για μεγαλύτερα χρονικά διαστήματα, η ικανότητα αυτή εξασθενεί καθώς η απόσταση αυξάνεται. Σε πρακτικό επίπεδο, τα LSTMs τείνουν να έχουν αποτελεσματική μνήμη περιορισμένη σε μερικές δεκάδες ή εκατοντάδες χρονικά βήματα, ανάλογα με την εργασία και τα δεδομένα εκπαίδευσης.

Τρίτον, η **επαγωγική μεροληψία (inductive bias)** των RNNs προς τη χρονική σειρά, αν και ωφέλιμη σε πολλές περιπτώσεις, μπορεί να αποβεί περιοριστική. Τα μοντέλα αυτά δομούνται έτσι ώστε να δίνουν εξ ορισμού μεγαλύτερη σημασία στο recent context παρά σε απομακρυσμένες χρονικές στιγμές, ακόμη και όταν αυτό δεν ανταποκρίνεται στην πραγματική δομή των δεδομένων. Για παράδειγμα, σε ένα video όπου ένα σημαντικό γεγονός εμφανίζεται στην αρχή και επηρεάζει την ερμηνεία μιας πολύ μεταγενέστερης σκηνής, το LSTM θα πρέπει να «μεταφέρει» αυτή την πληροφορία μέσω πολλών ενδιάμεσων καταστάσεων, κάτι που αυξάνει τον κίνδυνο απώλειας ή παραμόρφωσης του σήματος.

Οι περιορισμοί αυτοί οδήγησαν τη ερευνητική κοινότητα στην αναζήτηση εναλλακτικών μηχανισμών χρονικής μοντελοποίησης που θα μπορούσαν να παραλληλοποιηθούν πιο αποτελεσματικά, να χειρίζονται μακρινές εξαρτήσεις με μεγαλύτερη ακρίβεια, και να προσφέρουν μεγαλύτερη ευελιξία στην κατανομή της προσοχής σε διαφορετικά τμήματα της ακολουθίας. Η λύση προέκυψε από τομείς εκτός της υπολογιστικής όρασης, ειδικά από την επεξεργασία φυσικής γλώσσας, όπου οι μηχανισμοί attention είχαν ήδη αρχίσει να αντικαθιστούν ή να συμπληρώνουν τα recurrent networks.

2.3.2 Η Μετάβαση στα Self-Attention Mechanisms

Οι περιορισμοί των RNNs και LSTMs που αναλύθηκαν προηγουμένως (η αδυναμία παραλληλοποίησης, η εξασθένηση του σήματος σε πολύ μακρές ακολουθίες, και η επαγωγική μεροληψία προς την τοπική χρονική συνέχεια) οδήγησαν την ερευνητική κοινότητα στην αναζήτηση εναλλακτικών μηχανισμών που θα μπορούσαν να υπερβούν αυτά τα θεμελιώδη εμπόδια. Η λύση προέκυψε από την εξέλιξη των μηχανισμών προσοχής (attention mechanisms), οι οποίοι εισήχθησαν αρχικά ως συμπληρωματικά στοιχεία στα sequence-to-sequence μοντέλα.

Οι Bahdanau et al. πρότειναν το 2015 έναν μηχανισμό [18] που επιτρέπει στον decoder να εστιάζει επιλεκτικά σε διαφορετικά τμήματα της εισόδου κατά την παραγωγή κάθε στοιχείου της εξόδου, αντί να βασίζεται αποκλειστικά σε μια σταθερή αναπαράσταση. Στο πλαίσιο του video captioning, αυτή η ιδέα υιοθετήθηκε γρήγορα, επιτρέποντας στα μοντέλα να εστιάζουν σε συγκεκριμένα frames κατά την παραγωγή κάθε λέξης της περιγραφής [16]. Αυτή η προσέγγιση βελτίωσε σημαντικά την ποιότητα των παραγόμενων λεζάντων, αλλά ο υποκείμενος μηχανισμός ακολουθίας εξακολουθούσε να βασίζεται σε LSTMs.

Το καθοριστικό βήμα προς την πλήρη απομάκρυνση από τις recurrent δομές ήρθε το 2017 με την εισαγωγή της αρχιτεκτονικής Transformer από τους Vaswani et al. [19]. Η θεμελιώδης καινοτομία έγκειται στην κατάργηση κάθε μορφής επαναληπτικής ή συνελικτικής επεξεργασίας υπέρ ενός μοντέλου που βασίζεται αποκλειστικά σε self-attention mechanisms. Σε αντίθεση με τα RNNs όπου η πληροφορία πρέπει να διαδοθεί μέσω διαδοχικών κρυφών καταστάσεων, στους Transformers κάθε θέση της ακολουθίας μπορεί να αλληλεπιδράσει άμεσα με κάθε άλλη θέση μέσω του μηχανισμού προσοχής.

Αυτή η αρχιτεκτονική μετατόπιση προσφέρει τρία κρίσιμα πλεονεκτήματα για την επεξεργασία video. Πρώτον, επιτρέπει την πλήρη παραλληλοποίηση των υπολογισμών κατά την εκπαίδευση, καθώς όλες οι αλληλεπιδράσεις μεταξύ frames υπολογίζονται ταυτόχρονα αντί διαδοχικά. Δεύτερον, εξαλείφει το πρόβλημα των μακρινών εξαρτήσεων (long-range dependencies), αφού κάθε frame μπορεί να «δει» και να επηρεαστεί άμεσα από οποιοδήποτε άλλο frame ανεξαρτήτως απόστασης. Τρίτον, προσφέρει μεγαλύτερη ευελιξία στην εκμάθηση χρονικών μοτίβων που δεν ακολουθούν απαραίτητα τη γραμμική χρονική σειρά.

Η μετάβαση από τα RNNs/LSTMs στους Transformers αντιπροσωπεύει μια θεμελιώδη αλλαγή παραδείγματος στη χρονική μοντελοποίηση. Από σειριακή επεξεργασία με temporal bias προς την τοπική χρονική συνέχεια, σε παράλληλη επεξεργασία με άμεση μοντελοποίηση όλων των δυνατών αλληλεπιδράσεων. Αυτή η εξέλιξη έθεσε τα θεμέλια για τα σύγχρονα Vision-Language Models που αξιολογούνται στην παρούσα εργασία, τα οποία εφαρμόζουν Transformer αρχιτεκτονικές σε όλα τα στάδια της επεξεργασίας, από την κωδικοποίηση των frames έως την παραγωγή των γλωσσικών περιγραφών.

2.4 Transformers και Μηχανισμοί Attention

Η αρχιτεκτονική Transformer, που εισήχθη αρχικά το 2017 για εφαρμογές επεξεργασίας φυσικής γλώσσας (Vaswani et al., "Attention Is All You Need") [19], έχει φέρει επανάσταση και στον τομέα της υπολογιστικής όρασης. Η βασική καινοτομία των Transformers είναι ο μηχανισμός self-attention, ο οποίος επιτρέπει στο μοντέλο να υπολογίζει τη σχέση μεταξύ όλων των στοιχείων της εισόδου ταυτόχρονα, ανεξάρτητα από την απόστασή τους. Αυτό επιτρέπει την κατανόηση τόσο τοπικών όσο και μακρινών εξαρτήσεων (long-range dependencies) στα δεδομένα, χωρίς τους περιορισμούς της σειριακής επεξεργασίας των RNNs. Σε αντίθεση με τα CNNs που εφαρμόζουν τα ίδια φίλτρα τοπικά σε όλη την εικόνα (επαγωγική μεροληψία locality), οι Transformers υπολογίζουν σχέσεις μεταξύ όλων των στοιχείων μέσω self-attention, επιτρέποντας τη σύλληψη μακρινών εξαρτήσεων χωρίς γεωμετρικούς περιορισμούς.

2.4.1 Self-Attention: Ο Θεμελιώδης Μηχανισμός

Ο πυρήνας της αρχιτεκτονικής Transformer είναι ο μηχανισμός **self-attention**, ο οποίος επιτρέπει σε κάθε στοιχείο μιας ακολουθίας να σταθμίζει δυναμικά τη σημασία όλων των υπόλοιπων στοιχείων της ίδιας ακολουθίας. Με τον τρόπο αυτό, το μοντέλο μπορεί να συλλαμβάνει τόσο τοπικές όσο και μακρινές εξαρτήσεις χωρίς την ανάγκη σειριακής επεξεργασίας [19]. Τυπικά, ο μηχανισμός **self-attention** ορίζεται ως:

$$Attention(Q, K, V) = soft\max\left(QK^T/\sqrt{d_k}\right)V$$

Τα διανύσματα Q (Queries), Keys (K) και Values (V) προκύπτουν από γραμμικές προβολές της ίδιας εισόδου, ενώ το dk αντιστοιχεί στη διάσταση των διανυσμάτων των keys και χρησιμοποιείται για την κανονικοποίηση των βαθμολογιών ομοιότητας, ώστε να διασφαλίζεται σταθερότητα κατά την εκπαίδευση.

Ο μηχανισμός self-attention υπολογίζει, για κάθε στοιχείο της ακολουθίας, μια κατανομή προσοχής που καθορίζει τον βαθμό στον οποίο αυτό «εστιάζει» σε κάθε άλλο στοιχείο. Στο πλαίσιο της υπολογιστικής όρασης, η διαδικασία αυτή επιτρέπει στο μοντέλο να προσδιορίζει ποια patches μιας εικόνας είναι πιο σημαντικά για την αναπαράσταση ενός συγκεκριμένου patch. Αντίστοιχα, στην επεξεργασία video, ο self-attention μηχανισμός μπορεί να αναδείξει κρίσιμες χρονικές στιγμές ή frames που συμβάλλουν ουσιαστικά στην κατανόηση μιας σκηνής ή μιας δράσης. Με αυτόν τον τρόπο, ο self-attention αποτελεί τον θεμελιώδη μηχανισμό που επιτρέπει στους Transformers να μοντελοποιούν σύνθετες χωρικές και χρονικές σχέσεις με ενιαίο τρόπο [19].

2.4.2 Multi-Head Attention

Για την αύξηση της εκφραστικής ικανότητας του μηχανισμού self-attention, οι Transformers αξιοποιούν τον μηχανισμό Multi-Head Attention. Αντί να υπολογίζεται μία ενιαία συνάρτηση attention σε ολόκληρο τον χώρο αναπαράστασης, η είσοδος προβάλλεται σε πολλαπλούς υποχώρους, γνωστούς ως heads, μέσω ανεξάρτητων γραμμικών μετασχηματισμών. Για κάθε head, παράγεται ένα ξεχωριστό σύνολο διανυσμάτων ερωτημάτων, κλειδιών και τιμών (Qi ,Ki ,Vi) πάνω στα οποία εφαρμόζεται αυτοτελώς ο μηχανισμός self-attention. Με τον τρόπο αυτό, κάθε head μπορεί να εστιάσει σε διαφορετικές πτυχές της ακολουθίας εισόδου, όπως τοπικές ή μακρινές εξαρτήσεις, καθώς και διαφορετικά είδη σημασιολογικών ή δομικών συσχετίσεων.

Τα αποτελέσματα που προκύπτουν από τα επιμέρους heads συνενώνονται μέσω concatenation και στη συνέχεια προβάλλονται εκ νέου σε έναν ενιαίο χώρο αναπαράστασης μέσω ενός γραμμικού επιπέδου. Με αυτόν τον τρόπο, το Multi-Head Attention επιτρέπει στο μοντέλο να συλλαμβάνει ταυτόχρονα διαφορετικά είδη συσχετίσεων μέσα στην ίδια είσοδο, όπως τοπικές και μακρινές εξαρτήσεις ή επαναλαμβανόμενα δομικά μοτίβα. Η παράλληλη αυτή επεξεργασία πολλαπλών «οπτικών» ενισχύει σημαντικά την ικανότητα του Transformer να μαθαίνει πλούσιες και πολύπλευρες αναπαραστάσεις εντός της ακολουθίας.

2.4.3 Αρχιτεκτονική Transformer Encoder–Decoder

Ο μηχανισμός της προσοχής (attention) αποτελεί το θεμέλιο πάνω στο οποίο βασίζεται η αρχιτεκτονική Encoder–Decoder του Transformer. Η αρχιτεκτονική αυτή επέτρεψε την αποτελεσματική μοντελοποίηση πολύπλοκων εξαρτήσεων σε ακολουθιακά δεδομένα και έχει καθιερωθεί ως βασικό δομικό στοιχείο σε σύγχρονα μοντέλα επεξεργασίας φυσικής γλώσσας και πολυτροπικών δεδομένων.

Ο **encoder** αποτελείται από μια ακολουθία πανομοιότυπων blocks, όπου κάθε block περιλαμβάνει έναν μηχανισμό multi-head self-attention που επιτρέπει την ταυτόχρονη μοντελοποίηση διαφορετικών συσχετίσεων μεταξύ των στοιχείων της εισόδου, καθώς και ένα πλήρως συνδεδεμένο feed-forward δίκτυο για μη γραμμικό μετασχηματισμό των χαρακτηριστικών. Η σταθερότητα και η αποδοτικότητα της εκπαίδευσης διασφαλίζονται μέσω residual συνδέσεων και layer normalization, οι οποίες εφαρμόζονται σε κάθε υπομονάδα.

Ο **decoder** ακολουθεί παρόμοια δομή, επεκτείνοντας ωστόσο τη λειτουργικότητά του με δύο διακριτούς μηχανισμούς προσοχής. Αρχικά, το masked multi-head self-attention διασφαλίζει ότι κατά την αυτοπαλίνδρομη παραγωγή κειμένου το μοντέλο δεν έχει πρόσβαση σε μελλοντικά tokens. Στη συνέχεια, ο μηχανισμός cross-attention επιτρέπει στον decoder να εστιάζει επιλεκτικά στις εξόδους του encoder, συνδέοντας την παραγόμενη ακολουθία με τις κωδικοποιημένες αναπαραστάσεις της εισόδου.

Στο πλαίσιο εργασιών όρασης–γλώσσας, ο encoder μπορεί να επεξεργάζεται οπτικά tokens, όπως patches εικόνας ή frames video, ενώ ο decoder λειτουργεί ως γλωσσικό μοντέλο που παράγει λεζάντες ή περιγραφές. Η διασύνδεση των δύο modalities επιτυγχάνεται μέσω μηχανισμών cross-attention, οι οποίοι επιτρέπουν στο γλωσσικό μοντέλο να αντλεί σχετική οπτική πληροφορία κατά την παραγωγή κειμένου.

2.4.4 Εξειδικευμένοι Μηχανισμοί Attention: Spatial, Temporal, Channel

Οι βασικοί μηχανισμοί attention που εισήχθησαν με την αρχιτεκτονική Transformer μπορούν να εξειδικευτούν περαιτέρω, ανάλογα με τον τύπο της πληροφορίας που επιθυμούμε να αναδειχθεί. Στα οπτικά και πολυτροπικά μοντέλα, η εξειδίκευση αυτή οδηγεί σε μηχανισμούς attention που εστιάζουν στη χωρική και χρονική δομή των δεδομένων, καθώς και στη σχετική σημασία των επιμέρους χαρακτηριστικών (feature channels), επιτρέποντας πιο στοχευμένη και αποδοτική επεξεργασία της πληροφορίας.

Το **spatial attention** εστιάζει στη χωρική διάσταση της εικόνας, με στόχο τον εντοπισμό των περιοχών που είναι πιο σημαντικές για την εκάστοτε εργασία, όπως η ανίχνευση αντικειμένων ή η περιγραφή σκηνών. Αντί το μοντέλο να αντιμετωπίζει ολόκληρη την εικόνα ισοδύναμα, μαθαίνει να δίνει μεγαλύτερη έμφαση σε περιοχές που περιέχουν σημασιολογικά κρίσιμη πληροφορία. Στα Vision Transformers, το spatial attention υλοποιείται φυσικά μέσω self-attention μεταξύ των patch tokens μιας εικόνας. Ο μηχανισμός αυτός επιτρέπει στο μοντέλο να αναθέτει υψηλότερα βάρη σε patches που περιέχουν αντικείμενα, πρόσωπα ή χαρακτηριστικές κινήσεις, ενώ περιοχές φόντου λαμβάνουν μικρότερη σημασία. Αντίστοιχα, σε CNN-based αρχιτεκτονικές, το spatial attention μπορεί να εφαρμοστεί μέσω επιπλέον modules που παράγουν χωρικούς χάρτες βαρών πάνω στα feature maps, τροποποιώντας την ενεργοποίηση κάθε χωρικής θέσης.

Το **temporal attention** αφορά τη χρονική διάσταση και είναι ιδιαίτερα σημαντικό σε βιντεο-βασισμένες εφαρμογές. Ο μηχανισμός αυτός επιτρέπει στο μοντέλο να επιλέγει ποια frames ή χρονικές στιγμές ενός video είναι πιο πληροφοριακές, αγνοώντας τμήματα με μικρή σημασιολογική αξία. Στα Video Transformers, ο μηχανισμός temporal attention μπορεί να υλοποιηθεί με διαφορετικές αρχιτεκτονικές στρατηγικές. Μία προσέγγιση βασίζεται σε διαδοχική επεξεργασία, κατά την οποία εφαρμόζεται αρχικά spatial attention εντός κάθε frame, ώστε να εξαχθούν χωρικές συσχετίσεις, και ακολούθως temporal attention μεταξύ των frames, με σκοπό τη μοντελοποίηση της χρονικής εξέλιξης.

Εναλλακτικά, η χρονική και χωρική πληροφορία μπορεί να ενσωματωθεί μέσω ενιαίου space–time attention, όπου τα tokens από όλα τα frames αλληλεπιδρούν ταυτόχρονα μέσω ενός κοινού μηχανισμού self-attention, επιτρέποντας την άμεση σύλληψη χωροχρονικών εξαρτήσεων.

Με αυτόν τον τρόπο, το μοντέλο μπορεί να συσχετίσει χρονικά απομακρυσμένα γεγονότα, όπως την αρχική εμφάνιση ενός αντικειμένου και τις επακόλουθες κινήσεις του, αναγνωρίζοντας ακολουθίες δράσης που είναι κρίσιμες για την κατανόηση του video.

Το **channel attention** λειτουργεί σε επίπεδο καναλιών χαρακτηριστικών (feature channels) και συναντάται κυρίως σε CNN-based αρχιτεκτονικές. Η βασική ιδέα είναι ότι δεν συμβάλλουν όλα τα feature channels εξίσου στην τελική απόφαση του μοντέλου. Μηχανισμοί όπως τα Squeeze-and-Excitation Networks επιτρέπουν στο δίκτυο να εκτιμά τη σχετική σημασία κάθε καναλιού και να το ενισχύει ή να το καταστέλλει ανάλογα με τη χρησιμότητά του. Η διαδικασία αυτή βασίζεται στη δημιουργία μιας συνοπτικής αναπαράστασης της πληροφορίας κάθε feature channel μέσω global pooling, η οποία στη συνέχεια τροφοδοτεί έναν μηχανισμό επαναβαρύτητας (re-weighting). Με τον τρόπο αυτό, το μοντέλο μαθαίνει να δίνει μεγαλύτερη έμφαση σε κανάλια που αντιστοιχούν σε διακριτικά χαρακτηριστικά, όπως ανθρώπινα σώματα, αντικείμενα ή άκρα, και να μειώνει την επίδραση καναλιών που συνεισφέρουν λιγότερο στην αναγνώριση της σκηνής.

Σε πιο σύνθετες multimodal εργασίες, όπως η περιγραφή video, οι μηχανισμοί attention δεν λειτουργούν μεμονωμένα, αλλά συνδυάζονται ώστε το μοντέλο να αξιοποιεί ταυτόχρονα διαφορετικές πτυχές της διαθέσιμης πληροφορίας. Στο πλαίσιο αυτό, το **spatial attention** συμβάλλει στον εντοπισμό των σημαντικών αντικειμένων και χωρικών περιοχών εντός κάθε frame, το **temporal attention** επιτρέπει την επιλογή των κρίσιμων χρονικών στιγμών και τη μοντελοποίηση της εξέλιξης των γεγονότων στον χρόνο, ενώ το **channel attention** ενισχύει τα πιο διακριτικά feature channels και περιορίζει την επίδραση λιγότερο χρήσιμων αναπαραστάσεων. Ο συνδυασμός αυτών των μηχανισμών επιτρέπει στο μοντέλο να εστιάζει ταυτόχρονα στο «πού», στο «πότε» και στο «ποια χαρακτηριστικά» είναι ουσιώδη για την παραγωγή συνεκτικών και σημασιολογικά πλούσιων περιγραφών video.

2.5 Vision Transformers (ViT)

Η εμφάνιση των Vision Transformers (ViT) αποτέλεσε ένα σημαντικό ορόσημο στην εξέλιξη των μεθόδων υπολογιστικής όρασης, καθώς εισήγαγε την απευθείας εφαρμογή της αρχιτεκτονικής Transformer στην επεξεργασία εικόνων. Σε αντίθεση με τις κλασικές συνελικτικές αρχιτεκτονικές, οι οποίες βασίζονται στη χρήση τοπικών φίλτρων και στην ιεραρχική εξαγωγή χαρακτηριστικών, τα ViT υιοθετούν έναν μηχανισμό επεξεργασίας που βασίζεται αποκλειστικά στο attention. Στο πλαίσιο αυτό, η εικόνα αναπαρίσταται ως ακολουθία οπτικών tokens, πάνω στα οποία εφαρμόζεται self-attention, επιτρέποντας τη μοντελοποίηση χωρικών συσχετίσεων μεγάλης κλίμακας ήδη από τα πρώτα επίπεδα του δικτύου. Με τον τρόπο αυτό, τα ViT μπορούν να αποτυπώνουν παγκόσμιες χωρικές σχέσεις χωρίς να απαιτείται η σταδιακή διεύρυνση του receptive field μέσω διαδοχικών συνελικτικών επιπέδων [20].

2.5.1 Από CNNs στα Attention-based Μοντέλα

Οι συνελικτικές νευρωνικές αρχιτεκτονικές (Convolutional Neural Networks – CNNs) αποτέλεσαν το κυρίαρχο πρότυπο στην υπολογιστική όραση για περισσότερο από μία δεκαετία, με εμβληματικά μοντέλα όπως τα AlexNet, VGG, ResNet και EfficientNet να επιτυγχάνουν κορυφαίες επιδόσεις σε σύνολα αναφοράς όπως το ImageNet και συναφή benchmarks. Η ευρεία υιοθέτησή τους αποδίδεται στην παρουσία ισχυρών επαγωγικών μεροληψιών (inductive biases), όπως η χρήση τοπικών φίλτρων με weight sharing και η

ιεραρχική σύνθεση χαρακτηριστικών, η οποία επιτρέπει τη σταδιακή μετάβαση από χαμηλού επιπέδου οπτικά μοτίβα, όπως ακμές και υφές, σε υψηλού επιπέδου σημασιολογικές έννοιες. Αν και οι ιδιότητες αυτές καθιστούν τα CNNs ιδιαίτερα αποδοτικά για την επεξεργασία εικόνων, περιορίζουν ταυτόχρονα την ικανότητά τους να αποτυπώνουν μακρινές χωρικές συσχετίσεις, οι οποίες συνήθως απαιτούν βαθιά στοίβαξη επιπέδων και αυξημένο αριθμό παραμέτρων.

Η επιτυχία των αρχιτεκτονικών Transformer στην επεξεργασία φυσικής γλώσσας ανέδειξε το ερώτημα κατά πόσο παρόμοιοι μηχανισμοί, βασισμένοι αποκλειστικά στο attention, θα μπορούσαν να εφαρμοστούν αποτελεσματικά και σε δεδομένα εικόνας. Μία βασική πρόκληση σε αυτή τη μετάβαση είναι το γεγονός ότι οι εικόνες δεν είναι εγγενώς ακολουθιακές δομές και ότι η απευθείας γραμμικοποίηση όλων των pixels οδηγεί σε ακολουθίες εξαιρετικά μεγάλου μήκους, καθιστώντας τον υπολογισμό του self-attention υπολογιστικά μη βιώσιμο. Η κεντρική ιδέα πίσω από τον Vision Transformer έγκειται στην αναπαράσταση της εικόνας ως ακολουθίας από patches σταθερού μεγέθους, καθένα από τα οποία αντιμετωπίζεται ως διακριτό token. Με την προσέγγιση αυτή, το μήκος της ακολουθίας μειώνεται σημαντικά, ενώ παράλληλα διατηρείται επαρκής χωρική πληροφορία για την αποτελεσματική μοντελοποίηση του οπτικού περιεχομένου.

2.5.2 Βασική Αρχιτεκτονική Vision Transformer (ViT)

Ο Vision Transformer (ViT), όπως προτάθηκε από τους Dosovitskiy et al. [20], προσαρμόζει την κλασική αρχιτεκτονική Transformer encoder στην επεξεργασία εικόνων, αντιμετωπίζοντας την εικόνα ως ακολουθία από patches αντί για μεμονωμένα pixels. Με τον τρόπο αυτό καθίσταται δυνατή η άμεση εφαρμογή του μηχανισμού self-attention στην υπολογιστική όραση, κατά αναλογία με την επεξεργασία ακολουθιών tokens στη φυσική γλώσσα. Η αρχιτεκτονική του ViT βασίζεται σε μια διαδοχική διαδικασία κατά την οποία η εικόνα αρχικά **διαμερίζεται σε μη επικαλυπτόμενα patches (Patch Partitioning)**, τα οποία **προβάλλονται σε έναν κοινό χώρο embeddings (Patch embedding)**, συνδυάζονται με **positional encodings** ώστε να διατηρείται η χωρική πληροφορία και στη συνέχεια επεξεργάζονται από μια στοίβα **Transformer encoder blocks**. Η τελική αναπαράσταση που προκύπτει μπορεί να αξιοποιηθεί είτε για εργασίες classification είτε ως γενικού σκοπού visual representation για μεταγενέστερα στάδια επεξεργασίας.

Ο Διαμερισμός σε patches (Patch Partitioning), αποτελεί το πρώτο στάδιο της αρχιτεκτονικής**,** κατά το οποίο μια εικόνα εισόδου διαστάσεων $H \times W \times 3$, όπου H και W είναι το ύψος και το πλάτος, και 3 αντιστοιχεί στα RGB channels, διασπάται σε μικρότερα χωρικά τμήματα σταθερού μεγέθους. Η εικόνα χωρίζεται σε μη επικαλυπτόμενα τετράγωνα patches διαστάσεων $P \times P$. Στις συνήθεις υλοποιήσεις, το P είναι συνήθως 16 (ViT-B/16, ViT-L/16) ή 32 (ViTL/32), με μικρότερα patches να επιτρέπουν λεπτομερέστερη ανάλυση αλλά με αυξημένο υπολογιστικό κόστος λόγω του μεγαλύτερου αριθμού tokens. Ο συνολικός αριθμός των patches δίνεται από την παρακάτω σχέση:

$$N = \frac{H}{P} \cdot \frac{W}{P}$$

Το **Patch Embedding** ακολουθεί τον διαμερισμό της εικόνας, όπου κάθε τμήμα (patch) υπόκειται σε διαδικασία γραμμικοποίησης (flattening), κατά την οποία η τρισδιάστατη πληροφορία του αναδιατάσσεται σε ένα μονοδιάστατο διάνυσμα μεγέθους $P^2 \cdot 3$. Ακολούθως, το εν λόγω διάνυσμα προβάλλεται σε έναν χώρο χαρακτηριστικών (feature space) σταθερής διάστασης D , μέσω ενός εκπαιδεύσιμου γραμμικού μετασχηματισμού (Learnable Linear Projection). Το αποτέλεσμα της διαδικασίας είναι η παραγωγή μιας ακολουθίας N διανυσμάτων (embeddings). Ειδικότερα για τις πρότυπες αρχιτεκτονικές, η διάσταση του χώρου χαρακτηριστικών ορίζεται ως $D = 768$ για το ViT-Base και ως $D = 1024$ για το ViT-Large.

Τα **Class Token** και **Positional Encoding** συμπληρώνουν τη διαμόρφωση της ακολουθίας εισόδου, όπου, ακολουθώντας την προσέγγιση του μοντέλου BERT, ενσωματώνεται στην αρχή της ακολουθίας ένα ειδικό learnable token [**CLS**]. Η τελική αναπαράσταση του εν λόγω token, μετά την επεξεργασία από τον Transformer, αξιοποιείται ως η συνολική αναπαράσταση (global embedding) της εικόνας. Επιπροσθέτως, σε όλα τα tokens της ακολουθίας (συμπεριλαμβανομένου του [**CLS**]) αθροίζονται learnable positional embeddings, με στόχο την κωδικοποίηση της χωρικής πληροφορίας των patches. Σε αντιδιαστολή με τα sinusoidal positional encodings της αρχικής αρχιτεκτονικής Transformer, το ViT αξιοποιεί learnable 1D positional embeddings, οι παράμετροι των οποίων βελτιστοποιούνται κατά τη διαδικασία της εκπαίδευσης [21]. Ως εκ τούτου, η ακολουθία εισόδου στον Transformer διαμορφώνεται ως εξής:

$$z_0 = \left[x_{\{class\}};\ x_1^p E;\ x_2^p E;\ \ldots;\ x_N^p E\right] +\ E_{\{pos\}}$$

όπου E είναι το patch embedding matrix, E_{pos} τα positional embeddings, x_i^p το i -οστό patch.

Transformer Encoder. Η ακολουθία των tokens τροφοδοτείται σε μία στοίβα από L Transformer encoder blocks (όπου τυπικά $L = 12$ για το ViT-Base και $L = 24$ για το ViT-Large). Κάθε block απαρτίζεται από ένα Multi-Head Self-Attention (MSA) layer, το οποίο επιτρέπει σε κάθε token να εφαρμόσει attention σε όλα τα υπόλοιπα tokens, μοντελοποιώντας με αυτόν τον τρόπο τα global dependencies. Επιπλέον, το block περιλαμβάνει ένα positionwise Feed-Forward Network (MLP), το οποίο αποτελείται από δύο linear layers με GELU activation. Η αρχιτεκτονική ενσωματώνει residual connections και Layer Normalization (LN), το οποίο εφαρμόζεται πριν από κάθε υπο-στρώμα (pre-norm configuration). Η επεξεργασία στο πλαίσιο του block l περιγράφεται από τις ακόλουθες σχέσεις:

$$z'_\ell = MSA\big(LN(z_{\ell-1})\big) + z_{\ell-1}$$
$$z_\ell = MLP\big(LN(z'_\ell)\big) + z'_\ell$$

Τέλος, η **έξοδος και χρήση των αναπαραστάσεων** ολοκληρώνει τη διαδικασία κωδικοποίησης. Η τελική αναπαράσταση του [**CLS**] token, συμβολιζόμενη ως z_L^0 , ή εναλλακτικά, το αποτέλεσμα του global average pooling επί των patch tokens, αξιοποιείται ως το global embedding της εικόνας. Στο πλαίσιο εργασιών image classification, έπεται ένα MLP classification head το οποίο πραγματοποιεί την πρόβλεψη της κλάσης. Όταν το ViT χρησιμοποιείται ως γενικού σκοπού visual encoder, όπως παρατηρείται στα Vision-Language Models (π.χ. CLIP, BLIP), τα patch embeddings ή το [**CLS**] embedding διοχετεύονται σε επόμενα modules, όπως projectors, cross-attention layers, ή τροφοδοτούνται απευθείας σε Large Language Models (LLMs).

2.5.3 Προεκπαίδευση και Κλιμάκωση των Vision Transformers

Ένα από τα βασικά ευρήματα της αρχικής μελέτης για το ViT είναι ότι τα καθαρά attention-based μοντέλα για εικόνες απαιτούν πολύ μεγαλύτερα σύνολα δεδομένων για να ξεπεράσουν τα CNNs, όταν εκπαιδεύονται από την αρχή. Συγκεκριμένα, όταν το ViT εκπαιδεύτηκε απευθείας στο ImageNet (~1.3M εικόνες), η απόδοσή του υστερούσε σε σχέση με αντίστοιχου μεγέθους CNNs. Ωστόσο, όταν χρησιμοποιήθηκε pre-training σε πολύ μεγαλύτερα σύνολα δεδομένων, όπως το JFT-300M, ακολουθούμενη από fine-tuning στο ImageNet, τα ViT μοντέλα πέτυχαν state-of-the-art αποτελέσματα.

Στη συνέχεια, προτάθηκαν ποικίλες προσεγγίσεις self-supervised και masked pre-training για Vision Transformers, με χαρακτηριστικά παραδείγματα τα **MAE (Masked Autoencoders)** και **DINO**, οι οποίες επιτρέπουν την εκμάθηση πλούσιων οπτικών αναπαραστάσεων χωρίς την ανάγκη πλήρως επισημασμένων δεδομένων [22]. Οι τεχνικές

αυτές αποκτούν ιδιαίτερη σημασία σε εφαρμογές που αφορούν video, όπου η συλλογή και επισήμανση δεδομένων είναι ακόμη πιο απαιτητική, ενώ η αξιοποίηση μεγάλων ποσοτήτων μη επισημασμένου οπτικού υλικού μπορεί να προσφέρει ουσιαστικό πλεονέκτημα [23].

Η ικανότητα των ViT να κλιμακώνονται τόσο σε βάθος όσο και σε πλάτος, σε συνδυασμό με ισχυρά σχήματα pre-training, τα κατέστησε το κυρίαρχο backbone για πολλές σύγχρονες αρχιτεκτονικές στην υπολογιστική όραση. Η εξέλιξη αυτή αποτέλεσε το φυσικό υπόβαθρο για την ανάπτυξη των Video Vision Transformers, οι οποίοι επεκτείνουν την ίδια σχεδιαστική φιλοσοφία στη μοντελοποίηση της τρισδιάστατης χωροχρονικής πληροφορίας που χαρακτηρίζει τα δεδομένα video.

2.6 Επεκτάσεις των Vision Transformers σε video

Η επέκταση των Vision Transformers από στατικές εικόνες σε video απαιτεί την ταυτόχρονη αντιμετώπιση της χωρικής και της χρονικής διάστασης. Σε αντίθεση με τις εικόνες, τα video αποτελούνται από ακολουθίες frames, δημιουργώντας ένα τρισδιάστατο χωροχρονικό tensor διαστάσεων $T \times H \times W \times C$. Η άμεση εφαρμογή ενός ViT που θα αντιμετώπιζε κάθε patch σε κάθε frame ως ξεχωριστό token οδηγεί σε ακολουθίες τεράστιου μήκους και καθιστά τον υπολογισμό του self-attention με πολυπλοκότητα $O((T \cdot H \cdot W)^2)$ πρακτικά ανέφικτο για ρεαλιστικά video.

Για τον λόγο αυτό, προτάθηκαν εξειδικευμένες παραλλαγές όπως τα ViViT, TimeSformer και Video Swin Transformer, οι οποίες επιδιώκουν να μοντελοποιήσουν αποτελεσματικά τη χωροχρονική πληροφορία μέσω δομημένων attention μηχανισμών.

2.6.1 Η Πρόκληση της Χρονικής Διάστασης

Η επέκταση των ViT σε video εισάγει δύο βασικές προκλήσεις. Πρώτον, ο αριθμός των tokens αυξάνεται γραμμικά με τον αριθμό των frames T , πολλαπλασιάζοντας τον αριθμό των patch tokens που παράγονται από κάθε frame. Αν μια εικόνα 224 × 224 με patches 16 × 16 παράγει 196 tokens, ένα video 32 frames με τις ίδιες χωρικές διαστάσεις και patch size παράγει32 × 196 = 6,272 tokens, καθιστώντας το self-attention τετραγωνικής πολυπλοκότητας απαγορευτικά δαπανηρό. Δεύτερον, η χρονική διάσταση εισάγει νέους τύπους συσχετίσεων: τα μοντέλα πρέπει να συλλάβουν τόσο τις χωρικές σχέσεις εντός κάθε frame όσο και τις χρονικές σχέσεις μεταξύ διαφορετικών frames. Για να αντιμετωπιστούν αυτές οι προκλήσεις, οι σύγχρονοι Video Vision Transformers υιοθετούν στρατηγικές factorization του attention (διαχωρίζοντας χωρικό και χρονικό attention), ιεραρχικές δομές με τοπικά παράθυρα (local attention windows) , ή συνδυασμούς αυτών των προσεγγίσεων. Οι τεχνικές αυτές αποσκοπούν στη μείωση της πολυπλοκότητας από το πλήρες $O((T \cdot H \cdot W)^2)$ σε πιο διαχειρίσιμες μορφές όπως $O(T^2 \cdot H \cdot W + T \cdot H^2 \cdot W^2)$ ή ακόμη και σε περίπου γραμμική πολυπλοκότητα ως προς T .

2.6.2 Factorized Spatiotemporal Attention: ViViT και TimeSformer

Η βασική πρόκληση στη μετάβαση από εικόνες σε video αφορά την υπολογιστική πολυπλοκότητα του μηχανισμού self-attention, η οποία αυξάνεται τετραγωνικά ως προς τον αριθμό των tokens. Στην περίπτωση του video, ο αριθμός των tokens προκύπτει από το γινόμενο του αριθμού των spatial patches ανά frame με τον αριθμό των frames $N_{\text{spatial}} \times T$. Για παράδειγμα, σε ένα video με 8 frames και 196 spatial patches ανά frame, ο συνολικός αριθμός tokens ανέρχεται σε 1,568, καθιστώντας την πολυπλοκότητα $O(N^2)$ ιδιαίτερα υψηλή και συχνά υπολογιστικά απαγορευτική. Για την αντιμετώπιση αυτής της πρόκλησης, δύο σημαντικές αρχιτεκτονικές, το ViViT και το TimeSformer, προτείνουν

διαφορετικές στρατηγικές factorization του attention, διαχωρίζοντας ρητά τη χωρική και τη χρονική επεξεργασία. Οι προσεγγίσεις αυτές μειώνουν την πολυπλοκότητα από $O((T \cdot N)^2)$ σε $O(T \cdot N^2 + N \cdot T^2)$, επιτυγχάνοντας ουσιαστική βελτίωση χωρίς σημαντική απώλεια εκφραστικότητας.

Το **ViViT (Video Vision Transformer)**, προτεινόμενο από τους Arnab et al. [24], διερευνά συστηματικά τον σχεδιαστικό χώρο των Video Vision Transformers μέσω πολλαπλών παραλλαγών. Η κεντρική καινοτομία του ViViT έγκειται στην έννοια των **tubelet embeddings**, η οποία αποτελεί φυσική επέκταση του patch embedding concept από το ViT σε τρεις διαστάσεις. Αντί να αντιμετωπίζεται κάθε 2D patch ξεχωριστά, το video χωρίζεται σε χωροχρονικά tubelets διαστάσεων $t \times P \times P$, τα οποία εκτείνονται σε t διαδοχικά frames και καλύπτουν χωρική περιοχή $P \times P$ pixels σε κάθε frame. Κάθε tubelet αναδιαμορφώνεται σε διάνυσμα διάστασης $t \cdot P^2 \cdot 3$ και προβάλλεται γραμμικά σε embedding space διάστασης D μέσω ενός learnable linear projection, παράγοντας tokens που ενσωματώνουν εγγενώς πληροφορία κίνησης και χρονικής συνέχειας. Η προσέγγιση αυτή μειώνει τον αριθμό των tokens κατά παράγοντα t , καθώς πολλά frames συμπτύσσονται σε κάθε tubelet. Για $t = 2$ και video 16 frames, ο αριθμός temporal positions μειώνεται από 16 σε 8, μειώνοντας ανάλογα το κόστος του attention. Ωστόσο, αυτή η πρώιμη temporal fusion εισάγει έναν trade-off, μπορεί να περιορίσει την ικανότητα του μοντέλου να συλλαμβάνει λεπτές χρονικές αλλαγές και γρήγορες κινήσεις, ιδίως σε δράσεις υψηλής δυναμικής όπου τα διαδοχικά frames διαφέρουν σημαντικά.

Για να ισορροπήσει αυτό το trade-off, το ViViT προτείνει **factorized encoder architectures**, όπου η χωρική και χρονική επεξεργασία διαχωρίζονται ρητά σε διαδοχικά στάδια. Στην πιο συνηθισμένη παραλλαγή (Model 2), εφαρμόζεται αρχικά ένας spatial encoder που επεξεργάζεται ανεξάρτητα κάθε frame, παράγοντας χωρικά πλούσιες αναπαραστάσεις. Στη συνέχεια, ένας temporal encoder λαμβάνει ως είσοδο την ακολουθία των χωρικών αναπαραστάσεων (συνήθως τα [**CLS**] tokens από κάθε frame) και μοντελοποιεί τις χρονικές εξαρτήσεις μέσω temporal self-attention, όπου κάθε frame υπολογίζει attention με όλα τα άλλα frames. Αυτός ο διαχωρισμός επιτρέπει στο μοντέλο να εκμεταλλευτεί pretrained ViT weights για τον spatial encoder, αξιοποιώντας έτσι τη γνώση που έχει αποκτηθεί από εκατοντάδες εκατομμύρια εικόνες, ενώ ο temporal encoder εκπαιδεύεται εξ αρχής για την εργασία video, προσαρμόζοντας το δίκτυο στα ιδιαίτερα χαρακτηριστικά της χρονικής διάστασης.

Το **TimeSformer**, το οποίο προτάθηκε από τους Bertasius et al. [25], υιοθετεί μια πιο αυστηρά factorized προσέγγιση, γνωστή ως divided space-time attention. Σε αντίθεση με τη διαδοχική εφαρμογή spatial και temporal encoders που χρησιμοποιείται στο ViViT, ο TimeSformer ενσωματώνει τη factorization απευθείας μέσα σε κάθε Transformer layer. Συγκεκριμένα, σε κάθε layer, το self-attention διαχωρίζεται σε δύο διακριτά στάδια που εκτελούνται διαδοχικά. Κατά το πρώτο στάδιο, εφαρμόζεται spatial attention block, όπου το self-attention υπολογίζεται μόνο μεταξύ patch tokens του ίδιου frame. Κάθε token στη θέση (t, i) (δηλαδή frame t και spatial position i υπολογίζει attention scores μόνο με tokens (t,j) για όλα τα j, επιτρέποντας τη μοντελοποίηση χωρικών σχέσεων εντός του frame. Κάθε frame επεξεργάζεται ανεξάρτητα, γεγονός που μειώνει την πολυπλοκότητα αυτού του βήματος σε $O(T \cdot N^2)$ όπου N ο αριθμός patches ανά frame και T ο αριθμός frames. Κατά το δεύτερο στάδιο, εφαρμόζεται temporal attention block, όπου το self-attention υπολογίζεται μόνο μεταξύ tokens της ίδιας χωρικής θέσης σε διαφορετικά frames. Κάθε token (t, i) υπολογίζει attention με tokens (t', i) για όλα τα $t' \in \{1, \dots, T\}$, συλλαμβάνοντας τη χρονική εξέλιξη σε κάθε χωρική περιοχή. Επιπλέον, η αρχιτεκτονική του TimeSformer επιτρέπει την άμεση μεταφορά pre-trained ViT weights, τα spatial attention layers αρχικοποιούνται

από το pretrained ViT, ενώ τα temporal attention layers αρχικοποιούνται τυχαία ή με μηδενικά weights ώστε να μην επηρεάζουν αρχικά την έξοδο. Αυτή η στρατηγική weight transfer επιταχύνει σημαντικά τη σύγκλιση σε video tasks.

Τα εμπειρικά αποτελέσματα επιδεικνύουν την αποτελεσματικότητα και των δύο προσεγγίσεων. Το TimeSformer-L πετυχαίνει 80.7% top-1 accuracy στο Kinetics-400, ξεπερνώντας σημαντικά προηγούμενα 3D CNN μοντέλα [25]. Ιδιαίτερα αξιοσημείωτη είναι η απόδοση στο Something-Something V2 dataset, το οποίο απαιτεί λεπτομερή temporal reasoning, αποδεικνύοντας ότι το divided attention συλλαμβάνει αποτελεσματικά χρονικές εξαρτήσεις και λεπτές κινήσεις. Ως προς την υπολογιστική αποδοτικότητα, το TimeSformer επιτυγχάνει περίπου τριπλάσια ταχύτερη εκπαίδευση σε σχέση με joint space-time attention για το ίδιο μέγεθος μοντέλου, ενώ διατηρεί συγκρίσιμη ή ανώτερη ακρίβεια. Παράλληλα, το ViViT με tubelet embeddings και factorized architectures προσφέρει συμπληρωματικές δυνατότητες, παρέχοντας επιπλέον ευελιξία στον σχεδιασμό της αρχιτεκτονικής. Όταν προεκπαιδεύεται σε μεγάλα video datasets, το ViViT-L/16 (factorized) πετυχαίνει 81.7% top-1 accuracy στο Kinetics-400 με pre-training στο JFT [24], σημαντικά υπερβαίνοντας το I3D (71.1%) και αποδεικνύοντας την ισχύ της attention-based χωροχρονικής μοντελοποίησης.

Και οι δύο προσεγγίσεις αποτελούν σημαντικές εξελίξεις στην κατανόηση της αποδοτικής χωροχρονικής επεξεργασίας μέσω attention mechanisms, παρέχοντας τη θεωρητική και πρακτική βάση για τα σύγχρονα Video Vision Transformers που λειτουργούν ως visual encoders σε μεγάλα Vision-Language Models. Με την εφαρμογή factorized attention, τα ViViT και TimeSformer καταδεικνύουν πώς είναι δυνατή η αποδοτική μοντελοποίηση του video ως ακολουθίας οπτικών tokens, συλλαμβάνοντας ταυτόχρονα χωρικές σχέσεις εντός κάθε frame και χρονικές σχέσεις μεταξύ διαδοχικών frames, παράγοντας πλούσιες spatiotemporal αναπαραστάσεις που ενσωματώνουν πληροφορία κίνησης, αλλαγών σκηνής και αλληλουχιών δράσης.

2.6.3 Video Swin Transformer: Ιεραρχική Spatiotemporal Attention

Ο **Video Swin Transformer** αποτελεί μία σημαντική επέκταση των Vision Transformers στο πεδίο του video [26]. Βασίζεται στις αρχές του **Swin Transformer** και τις γενικεύει για την περίπτωση των χωροχρονικών δεδομένων. Σε αντίθεση με το κλασικό ViT, όπου το selfattention εφαρμόζεται σε όλα τα patches ταυτόχρονα με τετραγωνική πολυπλοκότητα, ο Swin Transformer εισάγει την έννοια των local attention windows. Τα παράθυρα αυτά μετατοπίζονται (shifted) μεταξύ διαδοχικών επιπέδων του δικτύου, δημιουργώντας σταδιακή global επικοινωνία με γραμμικό υπολογιστικό κόστος ως προς τον αριθμό των patches. Αυτή η ιεραρχική προσέγγιση με shifted windows επιτρέπει την ανταλλαγή πληροφορίας μεταξύ απομακρυσμένων περιοχών μέσω πολλαπλών επιπέδων, διατηρώντας παράλληλα την υπολογιστική αποδοτικότητα.

Στην περίπτωση του Video Swin Transformer, η μεθοδολογία αυτή επεκτείνεται σε τρεις διαστάσεις: ύψος, πλάτος και χρόνο. Ο χωροχρονικός όγκος διαμερίζεται σε μικρά τρισδιάστατα παράθυρα (spatiotemporal windows), εντός των οποίων το self-attention υπολογίζεται τοπικά. Η προσέγγιση αυτή αποφεύγει το απαγορευτικό υπολογιστικό κόστος της εφαρμογής joint attention στο σύνολο του video. Σε επόμενα επίπεδα της αρχιτεκτονικής, τα παράθυρα μετατοπίζονται τόσο χωρικά όσο και χρονικά, επιτρέποντας στις αναπαραστάσεις να ανταλλάσσουν πληροφορία και πέρα από τα αρχικά τους όρια. Κατά συνέπεια, το δίκτυο συνδυάζει σταδιακά πληροφορία από διαφορετικές χρονικές στιγμές και χωρικές περιοχές. Το αποτέλεσμα είναι μία αρχιτεκτονική με γραμμική πολυπλοκότητα, ανάλογη των CNNs, η οποία όμως διατηρεί το global receptive field των Transformers.

Επιπροσθέτως, ο Video Swin υιοθετεί ιεραρχική δομή (hierarchical architecture), παρόμοια με αυτή των CNNs (π.χ. ResNet). Στα αρχικά επίπεδα, η επεξεργασία πραγματοποιείται σε υψηλή χωρική και χρονική ανάλυση με περιορισμένο αριθμό καναλιών (feature channels), ενώ σταδιακά, μέσω των patch merging layers, μειώνεται η ανάλυση και αυξάνεται η διάσταση των χαρακτηριστικών. Ειδικότερα, το δίκτυο εκκινεί με patches μεγέθους 2 × 4 × 4 (χρόνος × ύψος × πλάτος) και εξελίσσεται μέσω τεσσάρων διακριτών σταδίων (stages). Σε κάθε στάδιο, η ανάλυση μειώνεται κατά συντελεστή 2, ενώ το πλήθος των καναλιών διπλασιάζεται. Η ιεραρχική αυτή αναπαράσταση αποδεικνύεται ιδιαίτερα χρήσιμη για εργασίες όπως action recognition, temporal action localization και γενικό video understanding, καθώς επιτρέπει στο μοντέλο να συνδυάζει λεπτομερή τοπική πληροφορία (π.χ. μικρές κινήσεις, λεπτομέρειες αντικειμένων) με πιο αφηρημένες, global αναπαραστάσεις (π.χ. συνολικές δράσεις, σκηνές).

Η σημασία αρχιτεκτονικών όπως ο Video Swin Transformer έγκειται στην αποδοτική προσαρμογή της λογικής των Transformers στα video. Τα μοντέλα αυτά διατηρούν την ικανότητα global προσοχής και spatiotemporal μοντελοποίησης, σεβόμενα παράλληλα τους υπολογιστικούς περιορισμούς που επιβάλλει η υψηλή διαστασιμότητα των δεδομένων. Αν και τα μοντέλα που εξετάζονται στην παρούσα εργασία (BLIP, GIT-VATEX, Qwen2-VL) δεν βασίζονται αποκλειστικά στο Video Swin, μοιράζονται κοινές αρχές. Το μεθοδολογικό πλαίσιο που διέπεται από το spatiotemporal attention, τη χρήση ιεραρχικών visual encoders και τη βελτιστοποιημένη διαχείριση των χρονικών συσχετίσεων, συγκροτεί τη θεμελιώδη βάση για την ανάπτυξη των σύγχρονων αρχιτεκτονικών Vision-Language.

2.6.4 Vision Transformers ως Visual Encoders σε Vision-Language Models

Η επιρροή των Vision Transformers (ViTs) στα σύγχρονα Vision-Language Models (VLMs) είναι καθοριστική, καθώς αποτελούν το βασικό υποσύστημα που επιτρέπει την αποτελεσματική αντίληψη και κωδικοποίηση του οπτικού περιεχομένου. Η γενική αρχιτεκτονική των περισσότερων VLMs ακολουθεί ένα κοινό πρότυπο, το οποίο περιλαμβάνει έναν οπτικό κωδικοποιητή (Vision Encoder), έναν μηχανισμό διασύνδεσης μεταξύ των δύο τροπικοτήτων (cross-modal connector) και ένα Μεγάλο Γλωσσικό Μοντέλο (Large Language Model). Στο πλαίσιο αυτό, οι Vision Transformers έχουν σε μεγάλο βαθμό αντικαταστήσει τις παραδοσιακές αρχιτεκτονικές CNN, αναλαμβάνοντας την εξαγωγή σημασιολογικά πλούσιων οπτικών αναπαραστάσεων (visual embeddings) από εικόνες ή video. Οι αναπαραστάσεις αυτές, βασισμένες σε μηχανισμούς self-attention, κλιμακώνονται αποδοτικά σε μεγάλα σύνολα δεδομένων και υποστηρίζουν ομοιόμορφη μοντελοποίηση διαφορετικών τύπων οπτικής εισόδου, διευκολύνοντας την ευθυγράμμισή τους με τον διανυσματικό χώρο του κειμένου. Στη συνέχεια, μέσω μηχανισμών ευθυγράμμισης ή μετασχηματισμού, όπως projection layers, adapters ή query-based connectors οι οπτικές αυτές αναπαραστάσεις συνδέονται με το γλωσσικό μοντέλο.

Η αρχιτεκτονική σύζευξη των Vision Transformers με LLMs δεν ακολουθεί μία ενιαία ή μονοσήμαντη προσέγγιση, αλλά υλοποιείται μέσω διαφορετικών στρατηγικών που επιδιώκουν να εξισορροπήσουν την υπολογιστική αποδοτικότητα με την αποτελεσματική σημασιολογική ευθυγράμμιση (semantic alignment) μεταξύ οπτικής και γλωσσικής πληροφορίας. Μία κυρίαρχη κατηγορία περιλαμβάνει τη χρήση **frozen visual encoders**, όπως προεκπαιδευμένα μοντέλα τύπου CLIP ή SIGLIP, όπου τα βάρη του οπτικού δικτύου παραμένουν αμετάβλητα κατά την εκπαίδευση του VLM. Η προσέγγιση αυτή αξιοποιεί τις γενικεύσιμες αναπαραστάσεις που έχουν αποκτηθεί κατά την προ-εκπαίδευση σε μεγάλης κλίμακας πολυτροπικά δεδομένα, μειώνοντας σημαντικά το υπολογιστικό κόστος και τις απαιτήσεις μνήμης. Αντιθέτως, η στρατηγική των **fine-tuned encoders**, η οποία υιοθετείται από μοντέλα όπως το BLIP, επιτρέπει την προσαρμογή των παραμέτρων του Vision Transformer στο πλαίσιο της εκπαίδευσης του VLM, οδηγώντας σε καλύτερη ευθυγράμμιση με τις απαιτήσεις της εκάστοτε πολυτροπικής εργασίας, αν και με αυξημένο

υπολογιστικό κόστος και κίνδυνο υπερπροσαρμογής (overfitting). Παράλληλα, αναδύεται μία τρίτη κατηγορία, αυτή των **generative visual encoders**, οι οποίοι έχουν προ-εκπαιδευτεί με παραγωγικούς στόχους και προσφέρουν αναπαραστάσεις διαφορετικής δομής και σημασιολογικού πλούτου.

Παράλληλα με τη στρατηγική εκπαίδευσης, κρίσιμο ρόλο στην απόδοση των VLMs διαδραματίζει και η διαχείριση της χωρικής ανάλυσης της οπτικής εισόδου. Οι παραδοσιακές υλοποιήσεις Vision Transformer βασίζονται σε σταθερή ανάλυση εισόδου (fixed resolution), συνήθως της τάξης των 224 × 224 ή 384 × 384 , γεγονός που καθιστά αναγκαία τη διαδικασία αλλαγής μεγέθους (resizing) ή συμπλήρωσης (padding) των εικόνων. Οι πρακτικές αυτές οδηγούν συχνά είτε σε απώλεια λεπτής χωρικής πληροφορίας είτε στην εισαγωγή τεχνητού θορύβου. Για την αντιμετώπιση αυτού του περιορισμού, σύγχρονες αρχιτεκτονικές, όπως το Qwen2-VL, εισάγουν την έννοια της δυναμικής ανάλυσης (Naive Dynamic Resolution), κατά την οποία η εικόνα μετατρέπεται σε ακολουθία tokens μεταβλητού μήκους και επεξεργάζεται στην εγγενή της ανάλυση. Η δυνατότητα αυτή υποστηρίζεται από προηγμένες τεχνικές κωδικοποίησης θέσης, όπως το 2D-Rotary Position Embedding (2DRoPE), το οποίο διατηρεί τις χωρικές συσχετίσεις ανεξάρτητα από το απόλυτο μήκος της ακολουθίας, γεφυρώνοντας αποτελεσματικά το χάσμα μεταξύ στατικής εικόνας και δυναμικής, υψηλής ανάλυσης οπτικής πληροφορίας.

Στην παρούσα εργασία, τα τρία μοντέλα που αξιολογούνται για το πρόβλημα του Dense Video Captioning, συγκεκριμένα τα **BLIP**, **GIT-VATEX** και **Qwen2-VL**, ενσωματώνουν Vision Transformers με διαφορετικούς τρόπους, αποτυπώνοντας με σαφήνεια τη διαχρονική εξέλιξη των αρχιτεκτονικών από την επεξεργασία στατικών εικόνων έως την προηγμένη χωροχρονική κατανόηση. Παρότι τα μοντέλα αυτά διαφοροποιούνται ως προς τον στόχο προεκπαίδευσης, τον τρόπο σύνδεσης με το γλωσσικό σκέλος και τον βαθμό ενσωμάτωσης της χρονικής πληροφορίας, μοιράζονται έναν κοινό αρχιτεκτονικό πυρήνα, ο οποίος βασίζεται στην αξιοποίηση Transformer-based οπτικών αναπαραστάσεων. Οι διαφοροποιήσεις τους εντοπίζονται κυρίως στον τρόπο με τον οποίο οι αναπαραστάσεις αυτές παράγονται και αξιοποιούνται, αντανακλώντας διαφορετικές σχεδιαστικές επιλογές ως προς την επεξεργασία των frames, την εισαγωγή ρητής χρονικής πληροφορίας και τον βαθμό ενοποίησης χωρικών και χρονικών χαρακτηριστικών. Το **BLIP** υιοθετεί μια κλασική προσέγγιση, αξιοποιώντας έναν τυπικό Vision Transformer (ViT-B ή ViT-L) ο οποίος επεξεργάζεται τα frames του video ως ανεξάρτητες στατικές εικόνες σταθερής ανάλυσης, χωρίς ρητή μοντελοποίηση της χρονικής συνέχειας στο επίπεδο του οπτικού encoder. Το **GIT-VATEX** επεκτείνει αυτή τη λογική μέσω της χρήσης ενός contrastive προεκπαιδευμένου vision encoder, ο οποίος, παρότι εξακολουθεί να επεξεργάζεται τα frames μεμονωμένα, ενισχύεται με την εισαγωγή εκπαιδεύσιμων temporal embeddings, επιτρέποντας τη βασική κωδικοποίηση της χρονικής αλληλουχίας πριν από τη σύνδεση με το γλωσσικό μοντέλο. Αντίθετα, το **Qwen2-VL** αντιπροσωπεύει μια πιο εξελιγμένη προσέγγιση, ενσωματώνοντας έναν εκτεταμένο Vision Transformer με δυνατότητες δυναμικής ανάλυσης και χωροχρονικής επεξεργασίας. Αυτό επιτυγχάνεται μέσω του μηχανισμού Multimodal Rotary Position Embedding (M-RoPE), ο οπτικός encoder του Qwen2-VL πραγματοποιεί ενιαία μοντελοποίηση χώρου και χρόνου, αντιμετωπίζοντας το video ως συνεκτικό χωροχρονικό σήμα και προσαρμόζοντας δυναμικά τόσο την ανάλυση όσο και τη χρονική δειγματοληψία.

2.7 Vision-Language Models

Η σύγκλιση της υπολογιστικής όρασης και της επεξεργασίας φυσικής γλώσσας αποτελεί ένα από τα πιο δυναμικά ερευνητικά πεδία της τελευταίας δεκαετίας. Τα Vision-Language Models έχουν αναδειχθεί ως μια κρίσιμη αρχιτεκτονική επιλογή για εφαρμογές που απαιτούν την κατανόηση και την ερμηνεία οπτικών δεδομένων μέσω της γλώσσας. Η ανάγκη για Vision-Language Models προέκυψε από τους περιορισμούς των εξειδικευμένων μοντέλων που σχεδιάζονταν για μεμονωμένες εργασίες. Παραδοσιακά, κάθε εργασία όρασης υπολογιστών (π.χ. image classification, object detection, semantic segmentation) απαιτούσε την εκπαίδευση ενός ξεχωριστού μοντέλου σε χειροκίνητα επισημειωμένα δεδομένα. Επιπλέον, τα μοντέλα αυτά δυσκολεύονταν να συνδέσουν το οπτικό περιεχόμενο με το γλωσσικό του πλαίσιο, περιορίζοντας την ικανότητά τους να κατανοούν την πλούσια σημασιολογία των σκηνών του πραγματικού κόσμου. Σε αντίθεση με τα παραδοσιακά μοντέλα που εξειδικεύονται είτε στην ανάλυση εικόνων είτε στην επεξεργασία κειμένου, τα Vision-Language Models επιδιώκουν την ενιαία μοντελοποίηση και των δύο πληροφοριακών διαστάσεων εντός ενός κοινού αναπαραστασιακού χώρου.

Η θεμελιώδης πρόκληση που αντιμετωπίζουν τα Vision-Language Models είναι η γεφύρωση του σημασιολογικού χάσματος μεταξύ της συνεχούς και πολυδιάστατης φύσης των οπτικών δεδομένων και της διακριτής, γραμμικής δομής της φυσικής γλώσσας. Οι εικόνες και τα video περιέχουν πλούσια και πολύπλοκη πληροφορία που εκτείνεται σε χωρικές και χρονικές διαστάσεις, ενώ η γλώσσα οργανώνει την πληροφορία σε ιεραρχικές δομές λέξεων, φράσεων και προτάσεων με αυστηρούς συντακτικούς και σημασιολογικούς κανόνες. Η επιτυχημένη ευθυγράμμιση αυτών των δύο τρόπων αναπαράστασης απαιτεί αρχιτεκτονικές που μπορούν να εξάγουν αφηρημένες σημασιολογικές έννοιες από την οπτική είσοδο και να τις αντιστοιχίσουν σε γλωσσικές περιγραφές με τρόπο που διατηρεί τη σημασιολογική συνοχή και την εκφραστικότητα.

Η ανάπτυξη των Vision-Language Models έχει επωφεληθεί σημαντικά από τρεις συγκλίνουσες τάσεις. Πρώτον, η διαθεσιμότητα τεράστιων συνόλων δεδομένων που περιέχουν ζεύγη εικόνων και κειμένου από διαδικτυακές πηγές έχει καταστήσει δυνατή την προεκπαίδευση μοντέλων σε πρωτοφανή κλίμακα. Δεύτερον, οι αρχιτεκτονικές Transformer, με τους μηχανισμούς cross-attention και self-attention, έχουν αποδειχθεί ιδανικές για τη μοντελοποίηση των πολύπλοκων αλληλεπιδράσεων μεταξύ οπτικών και γλωσσικών αναπαραστάσεων. Τρίτον, η εμφάνιση των μεγάλων γλωσσικών μοντέλων με εκατοντάδες δισεκατομμύρια παραμέτρων έχει δημιουργήσει νέες ευκαιρίες για την ενσωμάτωση οπτικής πληροφορίας σε μοντέλα που διαθέτουν ήδη βαθιά γλωσσική κατανόηση και ικανότητες αιτιολόγησης.

Τα τελευταία χρόνια, παρατηρείται μια δραστική μετατόπιση προς vision-language foundation models που προεκπαιδεύονται σε μεγάλης κλίμακας multimodal σύνολα δεδομένων εικόνας-κειμένου, συχνά αποτελούμενα από εκατοντάδες εκατομμύρια ή και δισεκατομμύρια ζεύγη. Η προσέγγιση αυτή επιτρέπει την εκμάθηση γενικών και ισχυρών multimodal αναπαραστάσεων, προσφέροντας δυνατότητες zero-shot και few-shot γενίκευσης σε ποικίλες downstream εργασίες χωρίς την ανάγκη εξειδικευμένης επαναεκπαίδευσης. Η ευρεία διαθεσιμότητα multimodal δεδομένων σε τομείς όπως η υγεία (ιατρικές εικόνες με διαγνωστικές αναφορές), τα αυτόνομα συστήματα (camera feeds με εντολές πλοήγησης), και τα μέσα κοινωνικής δικτύωσης (εικόνες με λεζάντες) ανέδειξε περαιτέρω τη σημασία των VLMs ως βασικό δομικό στοιχείο σύγχρονων ευφυών συστημάτων. Ως αποτέλεσμα, τα Vision-Language Models έχουν αναδειχθεί σε έναν ενοποιητικό πυλώνα μεταξύ όρασης υπολογιστών και επεξεργασίας φυσικής γλώσσας, διαμορφώνοντας το σύγχρονο τοπίο της multimodal τεχνητής νοημοσύνης.

2.7.1 Βασικά Συστατικά Αρχιτεκτονικής VLMs

Τα σύγχρονα Vision-Language Models αποτελούνται από τρεις βασικές αρχιτεκτονικές συνιστώσες που συνεργάζονται για την επίτευξη της πολυτροπικής κατανόησης. Η πρώτη συνιστώσα είναι ο **οπτικός κωδικοποιητής (Vision Encoder)**, ο οποίος αναλαμβάνει την εξαγωγή αναπαραστάσεων από την εικόνα ή το video. Η δεύτερη κρίσιμη συνιστώσα είναι ο **γλωσσικός κωδικοποιητής (text encoder/decoder)**, ο οποίος επεξεργάζεται το κείμενο και εξάγει γλωσσικές αναπαραστάσεις. Η τρίτη και ίσως η πιο καθοριστική συνιστώσα είναι ο **μηχανισμός ευθυγράμμισης και σύνδεσης (fusion mechanism)** των πολυτροπικών αναπαραστάσεων.

Ο **Vision Encoder** είναι υπεύθυνος για την εξαγωγή οπτικών χαρακτηριστικών από εικόνες ή video, μετατρέποντας την οπτική είσοδο σε μια ακολουθία feature vectors που αναπαριστούν το περιεχόμενο σε έναν συμπαγή χώρο αναπαράστασης. Οι σύγχρονες αρχιτεκτονικές χρησιμοποιούν κυρίως Vision Transformers (ViTs) που έχουν εκπαιδευτεί με contrastive learning (π.χ. CLIP ViT-L/14), αν και χρησιμοποιούνται επίσης CNNs (π.χ. ResNet, ConvNeXt) ή υβριδικοί encoders. Πολλά σύγχρονα VLMs χρησιμοποιούν προεκπαιδευμένους vision encoders από μοντέλα όπως το CLIP, τα οποία έχουν ήδη μάθει να ευθυγραμμίζουν οπτικές και γλωσσικές αναπαραστάσεις μέσω contrastive pre-training σε μεγάλης κλίμακας σύνολα δεδομένων εικόνας-κειμένου. Αυτό επιτρέπει στα VLMs να ξεκινούν από ισχυρές οπτικές αναπαραστάσεις που ήδη κωδικοποιούν σημασιολογική πληροφορία. Για παράδειγμα, το LLaVA χρησιμοποιεί το openai/clip-vit-large-patch14-336 με ανάλυση 336×336, όπου οι εικόνες χωρίζονται σε patches 14×14 pixels και επεξεργάζονται από τον transformer. Συνήθως, ο vision encoder παραμένει frozen (με σταθερά βάρη) κατά την εκπαίδευση για να διατηρηθούν οι ισχυρές ικανότητες γενίκευσης που αποκτήθηκαν στο στάδιο της προεκπαίδευσης.

Ο **Text Encoder/Decoder** είναι υπεύθυνος για την επεξεργασία και παραγωγή κειμένου, τα VLMs χρησιμοποιούν Transformer-based μοντέλα που λειτουργούν είτε ως encoders (για εργασίες κατανόησης) είτε ως decoders (για εργασίες παραγωγής). Στα πρώιμα VLMs, όπως το CLIP, χρησιμοποιείται ξεχωριστός text encoder βασισμένος σε αρχιτεκτονικές τύπου BERT ή GPT όπου μετατρέπει το κείμενο σε embeddings. Αντίθετα, τα σύγχρονα VLMs βασίζονται σε Large Language Models (LLMs), όπως τα LLaMA, Vicuna ή Qwen, τα οποία λειτουργούν ως αυτοπαλίνδρομοι decoders και παράγουν κείμενο token-by-token. Η ενσωμάτωση προεκπαιδευμένων LLMs επιτρέπει στα VLMs να αξιοποιούν προηγμένες ικανότητες reasoning, common-sense γνώσης και instruction-following.

Ο **Fusion Mechanism** καθορίζει τον τρόπο με τον οποίο οι οπτικές αναπαραστάσεις που παράγονται από τον vision encoder ενσωματώνονται στη γλωσσική επεξεργασία του μοντέλου και αποτελεί κρίσιμο παράγοντα για την απόδοση των Vision-Language Models. Οι σύγχρονες προσεγγίσεις fusion μπορούν να κατηγοριοποιηθούν με βάση τη χρονική στιγμή της σύνδεσης (early vs. late), τη θέση της fusion στην αρχιτεκτονική του μοντέλου (external vs. internal), καθώς και τον τρόπο ενσωμάτωσης της οπτικής πληροφορίας (modular vs. direct). Στη συνέχεια παρουσιάζονται συνοπτικά οι βασικές κατηγορίες μηχανισμών fusion, με έμφαση στα βασικά χαρακτηριστικά και τις σχεδιαστικές επιλογές που τις διαφοροποιούν.

Μία διαδεδομένη κατηγορία προσεγγίσεων είναι η λεγόμενη **Early External Fusion**, όπου η οπτική πληροφορία ενσωματώνεται σε πρώιμο στάδιο της αρχιτεκτονικής, πριν από την είσοδό της στο γλωσσικό μοντέλο, μέσω ενός εξωτερικού ενδιάμεσου μηχανισμού που λειτουργεί ως γέφυρα μεταξύ του vision encoder και του language model. Χαρακτηριστικό παράδειγμα αποτελεί το BLIP-2, το οποίο εισάγει τον Q-Former, έναν ελαφρύ querying transformer σχεδιασμένο να συνδέει έναν frozen vision encoder με έναν frozen LLM. Ο QFormer χρησιμοποιεί ένα σύνολο learnable query embeddings που αλληλεπιδρούν με τα οπτικά χαρακτηριστικά μέσω μηχανισμών cross-attention, επιτυγχάνοντας συμπίεση της οπτικής πληροφορίας από μεταβλητό αριθμό visual tokens σε έναν σταθερό και αποδοτικό

αριθμό αναπαραστάσεων. Οι παραγόμενες αναπαραστάσεις προβάλλονται στο embedding space του LLM και εισάγονται ως soft prompts, επιτρέποντας αποτελεσματική multimodal κατανόηση με περιορισμένο αριθμό εκπαιδεύσιμων παραμέτρων και χωρίς τροποποίηση της βασικής αρχιτεκτονικής του language model.

Μία εναλλακτική στρατηγική είναι η **late fusion μέσω gated cross-attention** μηχανισμών, στην οποία η οπτική πληροφορία ενσωματώνεται κατά τη διάρκεια της γλωσσικής επεξεργασίας και όχι εκ των προτέρων. Με τον τρόπο αυτό, το μοντέλο αποκτά τη δυνατότητα να αξιοποιεί το οπτικό περιεχόμενο επιλεκτικά, μόνο όταν αυτό κρίνεται απαραίτητο. Αντιπροσωπευτικό παράδειγμα αποτελεί το **Flamingo**, το οποίο εισάγει gated cross-attention layers παρεμβαλλόμενα ανάμεσα στα υπάρχοντα self-attention layers του language model. Κάθε μηχανισμός cross-attention ελέγχεται από έναν learnable gating παράγοντα, ο οποίος αρχικοποιείται κατά τρόπο ώστε το μοντέλο να συμπεριφέρεται αρχικά ως καθαρό LLM, επιτρέποντας τη σταδιακή ενσωμάτωση της οπτικής πληροφορίας κατά την εκπαίδευση. Με αυτή τη σχεδίαση, το οπτικό context ενσωματώνεται μόνο όπου είναι λειτουργικά χρήσιμο, χωρίς να διαταράσσεται η γλωσσική συνοχή. Για τη διαχείριση μεταβλητού αριθμού εικόνων ή video frames, το Flamingo αξιοποιεί τον Perceiver Resampler, ο οποίος συμπιέζει τα visual feature grids σε σταθερό αριθμό visual tokens, διασφαλίζοντας κλιμακωσιμότητα και ελεγχόμενο υπολογιστικό κόστος.

Μία πιο απλή αλλά ιδιαίτερα αποτελεσματική προσέγγιση είναι η **direct internal fusion μέσω Simple MLP projector**, όπου η ενσωμάτωση της οπτικής πληροφορίας πραγματοποιείται άμεσα στο εσωτερικό του language model χωρίς τη χρήση ενδιάμεσων μηχανισμών cross-attention ή bridging modules. Το **LLaVA** αποτελεί χαρακτηριστικό παράδειγμα αυτής της στρατηγικής, καθώς χρησιμοποιεί έναν MLP projector, αποτελούμενο από ένα ή περισσότερα fully-connected layers, για να προβάλλει τα patch embeddings του CLIP vision encoder στον χώρο embeddings του LLM, όπως το Vicuna. Τα προβαλλόμενα visual embeddings συνενώνονται με τα text token embeddings και τροφοδοτούνται απευθείας στο γλωσσικό μοντέλο ως ενιαία ακολουθία tokens. Παρά την αρχιτεκτονική της απλότητα, η συγκεκριμένη προσέγγιση έχει αποδειχθεί ιδιαίτερα αποδοτική όταν συνδυάζεται με ισχυρά pretrained components και υψηλής ποιότητας instruction-tuning δεδομένα, γεγονός που την καθιστά μία από τις πλέον διαδεδομένες μορφές direct fusion σε σύγχρονα Vision-Language Models.

Τέλος, πιο πρόσφατες ερευνητικές κατευθύνσεις διερευνούν μορφές **internal ή layerwise fusion**, στις οποίες τα οπτικά χαρακτηριστικά δεν ενσωματώνονται αποκλειστικά στην είσοδο του language model, αλλά συνδυάζονται με ενδιάμεσα layers της γλωσσικής αρχιτεκτονικής. Σε αυτές τις προσεγγίσεις, visual features που εξάγονται από συγκεκριμένα layers του vision encoder ευθυγραμμίζονται και συγχωνεύονται με αντίστοιχα layers του LLM, επιτρέποντας θεωρητικά βαθύτερη και πλουσιότερη αλληλεπίδραση μεταξύ όρασης και γλώσσας. Ωστόσο, η στρατηγική αυτή συνεπάγεται αυξημένη αρχιτεκτονική πολυπλοκότητα, καθώς απαιτεί ξεχωριστούς projectors ή μηχανισμούς σύνδεσης για κάθε layer, καθώς και πιο απαιτητικές διαδικασίες εκπαίδευσης. Πειραματικές μελέτες δείχνουν ότι, παρά το θεωρητικό τους πλεονέκτημα, οι layer-wise internal fusion προσεγγίσεις συχνά δεν υπερτερούν πρακτικά των απλούστερων external ή direct fusion στρατηγικών, οι οποίες εμφανίζουν πιο σταθερή και αποδοτική απόδοση.

2.7.2 Στόχοι Εκπαίδευσης (Pre-training Objectives)

Η εκπαίδευση των Vision-Language Models (VLMs) βασίζεται σε ένα σύνολο στόχων που εκμεταλλεύονται τη συσχέτιση μεταξύ εικόνων και κειμένου, με σκοπό την εκμάθηση πλούσιων και γενικών multimodal αναπαραστάσεων. Οι στόχοι αυτοί έχουν σχεδιαστεί ώστε να ενισχύουν τόσο τη συνολική (global) ευθυγράμμιση μεταξύ των δύο modalities όσο και τη λεπτομερή, τοπική αντιστοίχιση οπτικού και γλωσσικού περιεχομένου. Οι κυριότεροι στόχοι εκπαίδευσης συνοψίζονται παρακάτω.

Το **Image-Text Contrastive learning (ITC)** αποτελεί έναν από τους πιο θεμελιώδεις στόχους προεκπαίδευσης. Η μέθοδος αυτή εκπαιδεύει το μοντέλο να μεγιστοποιεί την ομοιότητα μεταξύ των embeddings σωστά αντιστοιχισμένων ζευγών εικόνας-κειμένου, ενώ ταυτόχρονα ελαχιστοποιεί την ομοιότητα μεταξύ μη αντιστοιχισμένων ζευγών. Το **CLIP** [27] αποτελεί χαρακτηριστικό παράδειγμα αυτής της προσέγγισης, όπου για κάθε δέσμη (batch) N ζεύγων, κατασκευάζεται ένας πίνακας ομοιότητας (matrix) διαστάσεων N x N, επί του οποίου εφαρμόζεται η συμμετρική συνάρτηση απώλειας cross-entropy (InfoNCE) [31]. Η συνάρτηση απώλειας έχει τη μορφή:

$$L_{ITC} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \quad \frac{\exp(sim(v_i, t_i)/\tau)}{\sum_{j=1}^{N}\exp\left(sim\left(v_i, t_j\right)/\tau\right)}\right]$$

όπου vi και ti αναπαριστούν τα normalized embeddings της εικόνας και του κειμένου αντίστοιχα, sim(·,·) υποδηλώνει συνήθως την cosine similarity, και τ αποτελεί παράμετρο θερμοκρασίας που ελέγχει την κλίμακα των logits. Η συνάρτηση κόστους είναι συμμετρική και εφαρμόζεται τόσο στην κατεύθυνση image-to-text όσο και στην text-to-image, επιτρέποντας zero-shot generalization σε διάφορα downstream tasks [27].

Το **Image-Text Matching (ITM)** συμπληρώνει τον contrastive στόχο με μια δυαδική εργασία ταξινόμησης (binary classification) που απαιτεί από το μοντέλο να προβλέπει αν ένα δεδομένο ζεύγος εικόνας-κειμένου είναι σωστά αντιστοιχισμένο. Σε αντίθεση με το ITC που βασίζεται σε συγκρίσεις global embeddings σε επίπεδο batch, το ITM απαιτεί βαθύτερη αλληλεπίδραση μεταξύ οπτικών και γλωσσικών αναπαραστάσεων μέσω cross-attention μηχανισμών. Αυτό επιτρέπει fine-grained alignment μεταξύ τμημάτων της εικόνας και τμημάτων του κειμένου, ενισχύοντας την ικανότητα του μοντέλου να εντοπίζει λεπτές σημασιολογικές ασυμφωνίες. Τα negative samples συνήθως δημιουργούνται είτε με τυχαία αντικατάσταση εικόνων (hard negatives από το batch) είτε με τροποποίηση του κειμένου, καθιστώντας την εργασία αρκετά απαιτητική για το μοντέλο. Η εκπαίδευση βασίζεται σε binary cross-entropy loss που ταξινομεί ζευγάρια ως matched (θετικά) ή unmatched (αρνητικά):

$$L_{ITM} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)]$$

όπου ∈ {0,1} είναι η ετικέτα matching (1 για matched pairs, 0 για unmatched), $p(v,t)$ είναι η πιθανότητα που προβλέπει το μοντέλο ότι το ζευγάρι είναι matched, και (v,t) αντιπροσωπεύει τις multimodal αναπαραστάσεις εικόνας-κειμένου.

Ο στόχος **Masked Language Modeling (MLM)**, εμπνευσμένος από την αρχιτεκτονική BERT [30], καλύπτει τυχαία tokens από το κείμενο και απαιτεί από το μοντέλο να τα προβλέψει αξιοποιώντας τόσο το γλωσσικό περιβάλλον όσο και την αντίστοιχη εικόνα. Τυπικά, ένα ποσοστό των tokens (συνήθως 15%) αντικαθίσταται με ένα ειδικό [MASK] token, και το μοντέλο εκπαιδεύεται να ανακατασκευάσει τα αρχικά tokens. Ο στόχος αυτός ενισχύει την εκμάθηση grounded language representations, δηλαδή γλωσσικών αναπαραστάσεων που είναι άμεσα συνδεδεμένες με την οπτική πληροφορία, και βελτιώνει σημαντικά τη multimodal reasoning ικανότητα του μοντέλου [32]. Η απώλεια υπολογίζεται μόνο για τα masked tokens:

$$L_{MLM} = -\sum_{i \in M}\log P(w_i | w_M, I)$$

όπου M είναι το σύνολο των θέσεων των masked tokens, wi το σωστό token στη θέση i, $w_{\setminus M}$ όλα τα μη masked tokens και v οπτική πληροφορία.

Για VLMs που στοχεύουν στην παραγωγή κειμένου, η εκπαίδευση βασίζεται επιπλέον σε αυτοπαλίνδρομους στόχους **Language Modeling (LM)**, όπου το μοντέλο προβλέπει το επόμενο token δεδομένης της εικόνας και του προηγούμενου κειμένου. Η διαδικασία αυτή, γνωστή και ως **Image-grounded Text Generation (ITG)**, επιτρέπει την παραγωγή φυσικών γλωσσικών περιγραφών όπως image captions, καθώς και τη δημιουργία multimodal διαλόγων [29], [19]. Η συνάρτηση απώλειας είναι η standard cross-entropy:

$$L_{LM} = -\frac{1}{T}\sum_{t=1}^{T} \log P(w_t | w_{<t}, I)$$

όπου wt είναι το token στην θέση t , $w<t$ είναι τα προηγούμενα tokens και I είναι η εικόνα. Αυτός ο στόχος επιτρέπει τη φυσική παραγωγή captions και την υλοποίηση multimodal conversations, καθώς το μοντέλο μαθαίνει να παράγει συνεκτικό κείμενο που βασίζεται στην οπτική πληροφορία [5].

Σύγχρονα Vision-Language models συνδυάζουν συχνά πολλαπλά pre-training objectives για την επίτευξη ισορροπίας μεταξύ understanding και generation capabilities. Η συνολική συνάρτηση κόστους ορίζεται ως γραμμικός συνδυασμός των επιμέρους objectives:

$$L_{total} = \lambda_{ITC} L_{ITC} + \lambda_{ITM} L_{ITM} + \lambda_{MLM} L_{MLM} + \lambda_{LM} L_{LM}$$

όπου τα βάρη λi ελέγχουν τη σχετική σπουδαιότητα κάθε objective κατά την εκπαίδευση. Η επιλογή των objectives και των αντίστοιχων βαρών εξαρτάται από τους στόχους του μοντέλου. Για παράδειγμα, το **BLIP** [5] χρησιμοποιεί τον συνδυασμό $L_{ITC} + L_{ITM} + L_{LM}$, επιτυγχάνοντας ισορροπία μεταξύ image-text understanding (μέσω ITC και ITM) και caption generation (μέσω LM). Η επιλογή του κατάλληλου συνδυασμού objectives εξαρτάται από τη φύση των downstream tasks που στοχεύει το pre-trained μοντέλο να υποστηρίξει.

2.7.3 Αντιπροσωπευτικές Αρχιτεκτονικές Vision-Language Models

Η τελευταία δεκαετία χαρακτηρίζεται από σημαντική πρόοδο στην ανάπτυξη Vision-Language Models που συνδυάζουν διαφορετικές αρχιτεκτονικές προσεγγίσεις και καινοτόμες στρατηγικές εκπαίδευσης. Το CLIP, που παρουσιάστηκε από την OpenAI το 2021, αποτελεί σημείο αναφοράς στο πεδίο των πολυτροπικών μοντέλων και έχει επηρεάσει καθοριστικά την κατεύθυνση της έρευνας. Η αρχιτεκτονική του CLIP βασίζεται σε δύο ξεχωριστούς κωδικοποιητές, έναν οπτικό κωδικοποιητή που μπορεί να υλοποιηθεί είτε ως ResNet είτε ως Vision Transformer, και έναν γλωσσικό κωδικοποιητή βασισμένο σε αρχιτεκτονική Transformer. Το κρίσιμο χαρακτηριστικό της επιτυχίας του CLIP έγκειται στην προεκπαίδευσή του σε 400 εκατομμύρια ζεύγη εικόνας-κειμένου που συλλέχθηκαν από το διαδίκτυο, χρησιμοποιώντας αποκλειστικά contrastive learning ως μέθοδο εκπαίδευσης [27]. Αυτή η προσέγγιση, παρά την απλότητά της, επιτρέπει στο CLIP να επιτυγχάνει εντυπωσιακές zero-shot ικανότητες σε εργασίες ταξινόμησης εικόνων και cross-modal retrieval χωρίς καμία εξειδικευμένη fine-tuning.

Το **BLIP**, που εισήχθη το 2022 [5], προσέφερε μια πιο ολοκληρωμένη αρχιτεκτονική που υποστηρίζει τόσο εργασίες κατανόησης όσο και εργασίες παραγωγής. Η αρχιτεκτονική του BLIP περιλαμβάνει τρία θεμελιώδη στοιχεία, πρώτον, έναν unimodal encoder για την εξαγωγή οπτικών χαρακτηριστικών, δεύτερον, έναν image-grounded text encoder που συνδυάζει οπτική και γλωσσική πληροφορία μέσω cross-attention μηχανισμών για image-text matching, και τρίτον, έναν image-grounded text decoder που παράγει κείμενο conditioned στην εικόνα. Η σημαντικότερη συνεισφορά του BLIP είναι η τεχνική Captioning and Filtering (CapFilt), που χρησιμοποιεί το ίδιο το μοντέλο για τη δημιουργία και το φιλτράρισμα synthetic captions, βελτιώνοντας σημαντικά την ποιότητα των

δεδομένων εκπαίδευσης [5]. Αυτή η bootstrapping στρατηγική επιτρέπει στο BLIP να εκπαιδεύεται αποδοτικά σε noisy web data και να επιτυγχάνει state-of-the-art απόδοση σε image captioning, visual question answering και image-text retrieval [33].

Το **BLIP-2**, που ακολούθησε το 2023 [33], προχώρησε ένα βήμα παραπέρα εισάγοντας την αρχιτεκτονική Q-Former. Ο Q-Former αποτελεί ένα ελαφρύ Transformer module που λειτουργεί ως γέφυρα μεταξύ ενός frozen pre-trained image encoder και ενός frozen large language model. Αντί να εκπαιδεύει ολόκληρα τα μεγάλα προ-εκπαιδευμένα μοντέλα, το BLIP-2 εκπαιδεύει μόνο τον Q-Former να εξάγει την πιο σχετική οπτική πληροφορία και να τη μετατρέπει σε query embeddings που μπορούν να ερμηνευτούν από το language model. Αυτή η προσέγγιση είναι εξαιρετικά αποδοτική υπολογιστικά και επιτρέπει στο BLIP-2 να αξιοποιήσει τις ισχυρές γλωσσικές ικανότητες μεγάλων LLMs όπως το OPT και το FlanT5, επιτυγχάνοντας εντυπωσιακή απόδοση σε zero-shot vision-language tasks με ελάχιστο υπολογιστικό κόστος [33].

Το **Flamingo**, που αναπτύχθηκε από την DeepMind [34], εστιάζει στην in-context few-shot learning ικανότητα. Η αρχιτεκτονική του Flamingo συνδυάζει έναν frozen vision encoder με έναν μεγάλο autoregressive language model, χρησιμοποιώντας Perceiver Resampler modules για τη συμπίεση των οπτικών features σε σταθερό αριθμό tokens. Το κλειδί του Flamingo είναι η ικανότητά του να επεξεργάζεται arbitrarily interleaved sequences of images and text, επιτρέποντάς του να μαθαίνει από few-shot examples που παρέχονται ως prompts [34]. Αυτό το καθιστά ιδιαίτερα ευέλικτο για rapid adaptation σε νέες εργασίες χωρίς gradient updates.

Το **LLaVA**, που παρουσιάστηκε το 2023 [35], αποτελεί σημαντική προσπάθεια σύνδεσης ισχυρών vision encoders με μεγάλα γλωσσικά μοντέλα όπως το LLaMA. Το LLaVA χρησιμοποιεί ένα απλό linear projection layer για τη μετατροπή των CLIP visual features σε embeddings που μπορούν να επεξεργαστούν από το LLaMA. Η διαδικασία εκπαίδευσής του περιλαμβάνει δύο στάδια: αρχικά την προ-εκπαίδευση του projection layer σε image-caption pairs, και στη συνέχεια instruction tuning σε πολυτροπικά instruction-following data [35]. Αυτή η προσέγγιση επιτρέπει στο LLaVA να επιτύχει εντυπωσιακές διαλογικές ικανότητες και να ακολουθεί σύνθετες οδηγίες που περιλαμβάνουν οπτική πληροφορία.

Το **InstructBLIP** επεκτείνει το BLIP-2 με instruction tuning [36], εκπαιδεύοντας τον QFormer να εξάγει instruction-aware visual features. Αντί να εξάγει generic representations από την εικόνα, ο instruction-aware Q-Former προσαρμόζει τα query embeddings βάσει της συγκεκριμένης οδηγίας ή ερώτησης του χρήστη, οδηγώντας σε πιο στοχευμένη και accurate απόκριση [36]. Αυτή η προσέγγιση βελτιώνει σημαντικά την απόδοση σε εργασίες που απαιτούν fine-grained visual reasoning και multi-hop reasoning.

2.7.4 Επέκταση των Vision-Language Models στην Επεξεργασία video

Η επέκταση των Vision-Language Models από στατικές εικόνες σε δυναμικό οπτικοακουστικό περιεχόμενο εισάγει σημαντικές τεχνικές προκλήσεις που αφορούν την αποτελεσματική κωδικοποίηση της χρονικής δυναμικής και την αποδοτική επεξεργασία μεγάλου όγκου οπτικών δεδομένων. Ενώ οι εικόνες μπορούν να αναπαρασταθούν με έναν σταθερό αριθμό visual tokens, τα video αποτελούνται από δεκάδες ή εκατοντάδες frames, κάθε ένα από τα οποία θα μπορούσε να επεξεργαστεί ένας image encoder, οδηγώντας σε εκθετική αύξηση των υπολογιστικών απαιτήσεων και της ακολουθίας εισόδου [10].

Ένα βασικό ζήτημα στα Video Vision-Language Models είναι ο τρόπος με τον οποίο κωδικοποιείται και συμπυκνώνεται η χρονική πληροφορία. Η ανεξάρτητη επεξεργασία κάθε frame ως στατικής εικόνας, αν και απλή, αποτυγχάνει να συλλάβει τη δυναμική των γεγονότων και τις χρονικές εξαρτήσεις που είναι κρίσιμες για την κατανόηση δράσεων, αλλαγών σκηνής και αλληλουχιών γεγονότων. Ως εκ τούτου, τα σύγχρονα μοντέλα υιοθετούν στρατηγικές που επιτρέπουν την επιλογή, τη σύνοψη και τη χρονική ευθυγράμμιση της οπτικής πληροφορίας πριν από τη σύνδεσή της με το γλωσσικό σκέλος.

Μία ευρέως χρησιμοποιούμενη προσέγγιση είναι το sparse frame sampling, κατά το οποίο επιλέγεται ένας περιορισμένος αριθμός αντιπροσωπευτικών frames από το video, με στόχο τη μείωση της υπολογιστικής επιβάρυνσης. Η επιλογή αυτή μπορεί να βασίζεται σε ομοιόμορφη δειγματοληψία, σε heuristics βασισμένες σε αλλαγές σκηνής ή σε learned μηχανισμούς επιλογής. Παρότι η στρατηγική αυτή μειώνει σημαντικά το κόστος, ενδέχεται να παραλείψει κρίσιμες χρονικές πληροφορίες, ιδίως σε video με σύνθετη δυναμική. Για την αποτελεσματικότερη σύνοψη της χρονικής πληροφορίας, έχουν προταθεί αρχιτεκτονικές που χρησιμοποιούν ειδικά modules συμπίεσης. Το **Video-LLaMA**, για παράδειγμα, εισάγει έναν **Video Q-Former**, ο οποίος επεκτείνει τη λογική του image QFormer του BLIP-2 στην περίπτωση ακολουθιών frames. Ο Video Q-Former λαμβάνει ως είσοδο visual features από πολλαπλά frames και χρησιμοποιεί learnable query tokens για να εξαγάγει τις πιο σημαντικές χρονικές και σημασιολογικές πληροφορίες, συμπιέζοντας ολόκληρο το video σε έναν μικρό, σταθερό αριθμό video tokens. Η συμπιεσμένη αυτή αναπαράσταση μπορεί στη συνέχεια να τροφοδοτηθεί σε ένα large language model, διατηρώντας κρίσιμες πληροφορίες για τη χρονική εξέλιξη των γεγονότων. Επιπλέον, το Video-LLaMA ενσωματώνει audio information μέσω του ImageBind audio encoder, επιτρέποντας πολυτροπική audio-visual κατανόηση video [37].

Το **Video-ChatGPT** αποτελεί εναλλακτική προσέγγιση που συνδυάζει video encoder με GPT language model [38], επιτρέποντας την παραγωγή απαντήσεων σε φυσική γλώσσα σχετικά με το περιεχόμενο ενός video. Η αρχιτεκτονική του χρησιμοποιεί spatiotemporal pooling για τη συμπίεση των frame-level features σε video-level representations και εκπαιδεύεται σε instruction-tuning data που περιλαμβάνουν video descriptions, question-answering pairs, και conversational interactions [38].

Συνολικά, τα Video Vision-Language Models αποτελούν μια φυσική και αναγκαία εξέλιξη των image-based VLMs, επεκτείνοντας τις δυνατότητές τους ώστε να καλύψουν τη δυναμική φύση του οπτικού κόσμου. Παρά τις σημαντικές προκλήσεις που εισάγει η χρονική διάσταση, οι σύγχρονες αρχιτεκτονικές καταδεικνύουν ότι η αποτελεσματική σύνοψη και ευθυγράμμιση της χωροχρονικής πληροφορίας με τη γλώσσα είναι εφικτή, ανοίγοντας τον δρόμο για προηγμένες εφαρμογές όπως dense video captioning, video question answering και πολυτροπική κατανόηση μακράς διάρκειας video.

2.7.5 BLIP: Bootstrapping Language-Image Pre-training

Το **BLIP** αποτελεί σημαντικό ορόσημο στην εξέλιξη των Vision-Language Models, εισάγοντας ενοποιημένη αρχιτεκτονική που αντιμετωπίζει τόσο εργασίες κατανόησης όσο και παραγωγής κειμένου από οπτικό περιεχόμενο. Η θεμελιώδης καινοτομία του BLIP έγκειται στον συνδυασμό ευέλικτης αρχιτεκτονικής που ονομάζεται Multimodal Mixture of EncoderDecoder και καινοτόμου μεθοδολογίας βελτίωσης δεδομένων που ονομάζεται CapFilt, η οποία επιτρέπει την αποδοτική εκμετάλλευση θορυβωδών δεδομένων από το διαδίκτυο [5].

Η αρχιτεκτονική του BLIP σχεδιάστηκε με γνώμονα την ευελιξία και την αποδοτικότητα. Το μοντέλο περιλαμβάνει οπτικό κωδικοποιητή βασισμένο σε Vision Transformer, ο οποίος διαχωρίζει την εικόνα σε patches και τα επεξεργάζεται ως ακολουθία οπτικών tokens [5], [20]. Το κρίσιμο στοιχείο της αρχιτεκτονικής είναι ο Multimodal Mixture of Encoder-Decoder, ο οποίος λειτουργεί σε τρεις διαφορετικές λειτουργικότητες. Στην πρώτη λειτουργικότητα, το μοντέλο λειτουργεί ως unimodal encoder, κωδικοποιώντας ξεχωριστά την εικόνα και το κείμενο. Ο κωδικοποιητής κειμένου βασίζεται στην αρχιτεκτονική BERT, προσθέτοντας ειδικό token [CLS] στην αρχή της ακολουθίας για την περίληψη της πρότασης [5], [30]. Στη δεύτερη λειτουργικότητα, το μοντέλο λειτουργεί ως image-grounded text encoder, εισάγοντας οπτική πληροφορία μέσω επιπλέον στρωμάτων cross-attention μεταξύ των self-attention layers και των feed-forward δικτύων. Ειδικό token [Encode] προστίθεται στο κείμενο και το αντίστοιχο output embedding

χρησιμοποιείται ως η πολυτροπική αναπαράσταση του ζεύγους εικόνας-κειμένου. Στην τρίτη λειτουργικότητα, το μοντέλο λειτουργεί ως image-grounded text decoder, αντικαθιστώντας τα bidirectional self-attention layers με causal self-attention layers για την αυτοπαλίνδρομη παραγωγή κειμένου [5].

Κρίσιμη αρχιτεκτονική επιλογή αποτελεί η κοινή χρήση παραμέτρων μεταξύ του encoder και του decoder. Συγκεκριμένα, ο text encoder και ο text decoder μοιράζονται όλες τις παραμέτρους εκτός από τα self-attention layers. Αυτή η στρατηγική βασίζεται στην παρατήρηση ότι οι διαφορές μεταξύ των εργασιών encoding και decoding αποτυπώνονται καλύτερα από τα self-attention layers, ενώ τα embedding layers, τα cross-attention layers και τα feed-forward δίκτυα λειτουργούν παρόμοια και στις δύο εργασίες. Η κοινή χρήση αυτών των στρωμάτων βελτιώνει την αποδοτικότητα της εκπαίδευσης ενώ επωφελείται από το multi-task learning, μειώνοντας τον συνολικό αριθμό παραμέτρων του μοντέλου [5].

Η εκπαίδευση του BLIP βασίζεται σε τρεις συμπληρωματικούς στόχους που βελτιστοποιούνται από κοινού. Ο πρώτος στόχος είναι το Image-Text Contrastive Learning, το οποίο ενεργοποιεί τον unimodal encoder και στοχεύει στην ευθυγράμμιση των χώρων χαρακτηριστικών της εικόνας και του κειμένου [5], [31]. Ο στόχος αυτός ακολουθεί την προσέγγιση που εισήγαγε το CLIP, χρησιμοποιώντας momentum encoder για την παραγωγή χαρακτηριστικών και soft labels για να λάβει υπόψη τα πιθανά θετικά ζεύγη μεταξύ των αρνητικών δειγμάτων [27]. Ο δεύτερος στόχος είναι το Image-Text Matching, το οποίο ενεργοποιεί τον image-grounded text encoder και διατυπώνεται ως πρόβλημα binary classification. Το μοντέλο πρέπει να προβλέψει αν ένα ζεύγος εικόνας-κειμένου είναι θετικό ή αρνητικό με βάση την πολυτροπική του αναπαράσταση. Για την εύρεση πιο πληροφοριακών αρνητικών δειγμάτων, το BLIP υιοθετεί στρατηγική hard negative mining, όπου τα αρνητικά ζεύγη με υψηλότερη contrastive ομοιότητα έχουν μεγαλύτερη πιθανότητα να επιλεγούν για τον υπολογισμό της απώλειας. Ο τρίτος στόχος είναι το Language Modeling, το οποίο ενεργοποιεί τον image-grounded text decoder και εκπαιδεύει το μοντέλο να παράγει περιγραφές κειμένου δοθείσης μιας εικόνας. Ο στόχος αυτός βελτιστοποιεί cross-entropy loss με label smoothing, εκπαιδεύοντας το μοντέλο να μεγιστοποιεί την πιθανότητα του κειμένου με αυτοπαλίνδρομο τρόπο.

Η μεθοδολογία CapFilt αποτελεί τη δεύτερη σημαντική συνεισφορά του BLIP και αντιμετωπίζει θεμελιώδες πρόβλημα στην προ-εκπαίδευση Vision-Language Models. Τα περισσότερα μοντέλα βασίζονται σε ζεύγη εικόνας-κειμένου που συλλέγονται αυτόματα από το διαδίκτυο, όπου τα alt-texts συχνά δεν περιγράφουν με ακρίβεια το οπτικό περιεχόμενο των εικόνων. Το CapFilt εισάγει δύο modules για τη βελτίωση της ποιότητας των δεδομένων, έναν captioner για τη δημιουργία συνθετικών λεζαντών και έναν filter για την αφαίρεση θορυβωδών λεζαντών [5]. Και τα δύο modules αρχικοποιούνται από το ίδιο προ-εκπαιδευμένο μοντέλο MED και fine-tuned ξεχωριστά στο dataset COCO.

Ο captioner είναι image-grounded text decoder που fine-tuned με τον στόχο Language Modeling για να αποκωδικοποιεί κείμενο δοθείσης εικόνων [5]. Κρίσιμη επιλογή σχεδιασμού είναι η χρήση nucleus sampling αντί για beam search για την παραγωγή των συνθετικών λεζαντών. Το nucleus sampling είναι στοχαστική μέθοδος αποκωδικοποίησης όπου κάθε token δειγματοληπτείται από σύνολο tokens των οποίων η αθροιστική πιθανότητα υπερβαίνει ένα κατώφλι. Παρόλο που το nucleus sampling παράγει πιο θορυβώδεις λεζάντες σε σύγκριση με το beam search, οι πειραματικές αξιολογήσεις δείχνουν ότι οδηγεί σε σημαντικά καλύτερη απόδοση στις downstream εργασίες. Η υπόθεση είναι ότι το nucleus sampling παράγει πιο ποικιλόμορφες και εκπλήσσουσες λεζάντες, οι οποίες περιέχουν περισσότερες νέες πληροφορίες από τις οποίες το μοντέλο μπορεί να επωφεληθεί [5].

Ο filter είναι image-grounded text encoder που fine-tuned με τους στόχους Image-Text Contrastive και Image-Text Matching για να μάθει αν ένα κείμενο ταιριάζει με μια εικόνα. Ο filter αφαιρεί θορυβώδη κείμενα τόσο από τις αρχικές λεζάντες του διαδικτύου όσο και από τις συνθετικές λεζάντες, θεωρώντας ένα κείμενο θορυβώδες αν το ITM head το

προβλέπει ως μη ταιριαστό με την εικόνα. Τελικά, τα φιλτραρισμένα ζεύγη εικόνας-κειμένου συνδυάζονται με τα ζεύγη με ανθρώπινες σημειώσεις για να σχηματίσουν νέο dataset, το οποίο χρησιμοποιείται για την προ-εκπαίδευση νέου μοντέλου [5].

Σημαντική παρατήρηση είναι ότι ο captioner και ο filter πρέπει να αποσυνδεθούν κατά τη διάρκεια του fine-tuning. Αν μοιράζονται παραμέτρους με τον ίδιο τρόπο όπως κατά την προεκπαίδευση, η απόδοση μειώνεται λόγω confirmation bias. Λόγω της κοινής χρήσης παραμέτρων, οι θορυβώδεις λεζάντες που παράγονται από τον captioner είναι λιγότερο πιθανό να φιλτραριστούν από τον filter. Η αποσύνδεση των δύο modules επιτρέπει στον filter να κρίνει πιο αντικειμενικά την ποιότητα των συνθετικών λεζαντών [5].

Το BLIP προ-εκπαιδεύεται σε σύνολο δεδομένων 14 εκατομμυρίων εικόνων που περιλαμβάνει δύο datasets με ανθρώπινες σημειώσεις και τρία web datasets. Η εκπαίδευση πραγματοποιείται για 20 epochs με batch size 2880 για το ViT-B και 2400 για το ViT-L. Το BLIP επιτυγχάνει state-of-the-art αποτελέσματα σε ευρύ φάσμα vision-language benchmarks. Στο image-text retrieval στο COCO dataset, το BLIP ξεπερνά το προηγούμενο καλύτερο μοντέλο ALBEF κατά 2.7% σε average recall@1 [5], [29]. Στο image captioning, το BLIP επιτυγχάνει 129.7 CIDEr στο COCO και 105.1 CIDEr στο NoCaps validation set. Στο Visual Question Answering, το BLIP επιτυγχάνει 77.54% στο VQA test-dev set [5]. Αξιοσημείωτα, το BLIP επιδεικνύει ισχυρές zero-shot ικανότητες μεταφοράς σε videolanguage tasks, επιτυγχάνοντας 43.3% recall@1 στο text-to-video retrieval στο MSRVTT dataset χωρίς καμία fine-tuning σε video data, ξεπερνώντας ακόμη και μοντέλα που έχουν fine-tuned στο target video dataset.

Η επιτυχία του BLIP οδήγησε σε περαιτέρω αναπτύξεις όπως το BLIP-2, το οποίο εισήγαγε τον Q-Former ως γέφυρα μεταξύ frozen pre-trained image encoders και frozen large language models [33]. Ο Q-Former αποτελεί ελαφρύ Transformer module που εκπαιδεύεται να εξάγει την πιο σχετική οπτική πληροφορία και να τη μετατρέπει σε query embeddings που μπορούν να κατανοηθούν από το language model. Αυτή η προσέγγιση είναι εξαιρετικά αποδοτική υπολογιστικά και επιτρέπει στο BLIP-2 να αξιοποιήσει τις ισχυρές γλωσσικές ικανότητες μεγάλων LLMs επιτυγχάνοντας εντυπωσιακή απόδοση σε zero-shot vision-language tasks με ελάχιστο υπολογιστικό κόστος [33].

2.7.6 GIT-VATEX: Ενοποιημένη Προσέγγιση για Video Captioning

Το GIT αποτελεί σημαντική προσέγγιση στην απλοποίηση της αρχιτεκτονικής των Vision-Language Models, εισάγοντας μινιμαλιστικό σχεδιασμό που συνδυάζει image encoder και text decoder υπό ενιαίο στόχο language modeling [39]. Η φιλοσοφία σχεδιασμού του GIT αντιτίθεται στην τάση πολυπλοκότητας που χαρακτηρίζει πολλά σύγχρονα Vision-Language Models, στοχεύοντας σε καθαρή και κλιμακούμενη αρχιτεκτονική χωρίς εξωτερικές εξαρτήσεις όπως object detectors ή OCR modules [39].

Η αρχιτεκτονική του GIT αποτελείται από δύο βασικά συστατικά [39]. Το πρώτο συστατικό είναι image encoder που βασίζεται σε Vision Transformer ή σε convolutional neural networks, ο οποίος αναλαμβάνει την εξαγωγή οπτικών αναπαραστάσεων από την εικόνα ή το video [20], [39]. Ο image encoder επεξεργάζεται την εισερχόμενη εικόνα και παράγει ακολουθία οπτικών features που αποτυπώνουν τα χαρακτηριστικά της εικόνας. Το δεύτερο συστατικό είναι text decoder που βασίζεται στην αρχιτεκτονική Transformer, ο οποίος λαμβάνει τα οπτικά features από τον encoder και παράγει αυτοπαλίνδρομα την ακολουθία κειμένου [19], [39]. Ο text decoder χρησιμοποιεί cross-attention layers για να ενσωματώσει την οπτική πληροφορία κατά τη διάρκεια της παραγωγής κειμένου, επιτρέποντας στο μοντέλο να συνδέσει άμεσα τα οπτικά χαρακτηριστικά με τις γλωσσικές δομές.

Το GIT εκπαιδεύεται με ενιαίο στόχο language modeling, όπου το μοντέλο μαθαίνει να μεγιστοποιεί την πιθανότητα της ακολουθίας κειμένου δοθείσης της εικόνας [39]. Αυτός ο απλός στόχος εκπαίδευσης επιτρέπει στο GIT να διατηρεί συνέπεια μεταξύ της

προεκπαίδευσης και του fine-tuning, καθώς και οι δύο φάσεις χρησιμοποιούν την ίδια αρχιτεκτονική και τον ίδιο στόχο. Σε αντίθεση με μοντέλα όπως το CLIP που βασίζονται κυρίως σε contrastive learning [27], το GIT εστιάζει στη generative προσέγγιση, κάτι που το καθιστά ιδιαίτερα κατάλληλο για εργασίες παραγωγής κειμένου όπως το image captioning και το video captioning [39].

Κρίσιμη συνεισφορά του GIT είναι η διερεύνηση των scaling laws για Vision-Language Models [39]. Τα πειραματικά αποτελέσματα δείχνουν ότι η κλιμάκωση τόσο του μεγέθους του μοντέλου όσο και του όγκου των δεδομένων προ-εκπαίδευσης οδηγεί σε σημαντικές βελτιώσεις της απόδοσης. Το GIT προ-εκπαιδεύεται σε τεράστια σύνολα δεδομένων που περιλαμβάνουν εκατοντάδες εκατομμύρια ζεύγη εικόνας-κειμένου από διάφορες πηγές του διαδικτύου [39]. Η μεγάλη κλίμακα των δεδομένων προ-εκπαίδευσης επιτρέπει στο μοντέλο να μάθει πλούσιες και ισχυρές αναπαραστάσεις που γενικεύουν καλά σε ποικίλες downstream εργασίες.

Το GIT επιτυγχάνει εντυπωσιακά αποτελέσματα σε ευρύ φάσμα benchmarks [39]. Στο TextCaps dataset, το GIT ξεπερνά για πρώτη φορά την ανθρώπινη απόδοση, επιτυγχάνοντας 138.2 CIDEr έναντι 125.5 της ανθρώπινης απόδοσης. Αυτό το αποτέλεσμα υπογραμμίζει την ικανότητα του μοντέλου να κατανοεί και να περιγράφει κείμενο μέσα σε εικόνες, δύσκολη εργασία που απαιτεί τόσο οπτική κατανόηση όσο και αναγνώριση κειμένου [39]. Το GIT επίσης καθιερώνει νέα state-of-the-art αποτελέσματα σε 12 challenging benchmarks, συμπεριλαμβανομένων του COCO Captioning, VQA, και άλλων vision-language tasks.

Η επέκταση του GIT στο video captioning με τη χρήση του VATEX dataset αντιπροσωπεύει φυσική εξέλιξη της αρχιτεκτονικής [6], [39]. Το VATEX αποτελεί μεγάλης κλίμακας πολυγλωσσικό dataset που περιέχει πάνω από 41.250 videos και 825.000 λεζάντες στα Αγγλικά και τα Κινέζικα [6]. Μεταξύ των λεζαντών, υπάρχουν πάνω από 206.000 παράλληλα ζεύγη μετάφρασης Αγγλικών-Κινέζικων. Σε σύγκριση με το ευρέως χρησιμοποιούμενο MSR-VTT dataset, το VATEX είναι πολυγλωσσικό, μεγαλύτερο, γλωσσικά πιο πολύπλοκο και πιο ποικιλόμορφο τόσο από την πλευρά του video όσο και από την πλευρά των περιγραφών φυσικής γλώσσας [6].

Για την επεξεργασία video εισόδου, το GIT-VATEX υιοθετεί απλή αλλά αποδοτική προσέγγιση [39]. Τα videos επεξεργάζονται ως ακολουθίες frames, όπου περιορισμένος αριθμός frames δειγματοληπτείται ομοιόμορφα από κάθε video. Κάθε frame επεξεργάζεται ξεχωριστά από τον image encoder, και τα προκύπτοντα οπτικά χαρακτηριστικά συνενώνονται σε ενιαία ακολουθία που τροφοδοτείται στον text decoder [39]. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να συλλάβει τη χρονική δυναμική του video χωρίς να απαιτεί εξειδικευμένες αρχιτεκτονικές για temporal modeling. Η απλότητα αυτής της προσέγγισης αποτελεί ένα από τα κύρια πλεονεκτήματα του GIT-VATEX, καθώς καθιστά το μοντέλο ευκολότερο να εκπαιδευτεί και να κλιμακωθεί.

Το fine-tuning του GIT στο VATEX dataset για την εργασία του multilingual video captioning επιδεικνύει την ικανότητα του μοντέλου να προσαρμόζεται σε πολυγλωσσικά περιβάλλοντα [6], [39]. Το μοντέλο εκπαιδεύεται να παράγει λεζάντες τόσο στα Αγγλικά όσο και στα Κινέζικα, μαθαίνοντας να χειρίζεται τις γλωσσικές ιδιαιτερότητες και των δύο γλωσσών ενώ διατηρεί την ικανότητά του να κατανοεί το οπτικό περιεχόμενο. Η πολυγλωσσική φύση του VATEX καθιστά το dataset ιδανικό για την αξιολόγηση της ικανότητας των Vision-Language Models να γενικεύουν σε διαφορετικές γλώσσες και πολιτισμικά πλαίσια [6].

Οι πειραματικές αξιολογήσεις του GIT-VATEX δείχνουν ανταγωνιστική απόδοση στις εργασίες video captioning [39]. Το μοντέλο επιτυγχάνει υψηλά scores σε metrics όπως το CIDEr, το BLEU και το METEOR, υποδεικνύοντας ότι οι παραγόμενες λεζάντες είναι τόσο ακριβείς όσο και ρέουσες. Η απλότητα της αρχιτεκτονικής του GIT-VATEX, σε συνδυασμό με τη μεγάλη κλίμακα προ-εκπαίδευσης, επιτρέπει στο μοντέλο να επιτύχει αυτά τα αποτελέσματα χωρίς την ανάγκη πολύπλοκων τεχνικών ή εξωτερικών modules.

2.7.7 Qwen2-VL: Προηγμένη Πολυτροπική Κατανόηση

Το Qwen2-VL αντιπροσωπεύει σημαντική εξέλιξη στην αρχιτεκτονική των Vision-Language Models, εισάγοντας καινοτομίες που επιτρέπουν την αποδοτική επεξεργασία εικόνων και videos οποιασδήποτε ανάλυσης. Η σειρά Qwen2-VL περιλαμβάνει τρία μοντέλα με 2, 8 και 72 δισεκατομμύρια παραμέτρους αντίστοιχα, προσφέροντας επιλογές για διαφορετικές υπολογιστικές απαιτήσεις και περιπτώσεις χρήσης. Η θεμελιώδης καινοτομία του Qwen2-VL έγκειται σε δύο κρίσιμες αρχιτεκτονικές βελτιώσεις, συγκεκριμένα τον μηχανισμό **Naive Dynamic Resolution** και το **Multimodal Rotary Position Embedding**.

Ο μηχανισμός **Naive Dynamic Resolution** επαναπροσδιορίζει τον τρόπο με τον οποίο τα Vision-Language Models επεξεργάζονται εικόνες διαφορετικών αναλύσεων [7]. Τα περισσότερα παραδοσιακά μοντέλα περιορίζονται σε σταθερό μέγεθος εισόδου εικόνας, όπως 224×224 pixels, και εφαρμόζουν downsampling, upsampling ή scale-then-padding προσεγγίσεις για να προσαρμόσουν τις εισερχόμενες εικόνες σε αυτό το μέγεθος [20]. Αυτή η one-size-fits-all στρατηγική οδηγεί σε απώλεια λεπτομερειών, ειδικά σε εικόνες υψηλής ανάλυσης, και περιορίζει την ικανότητα του μοντέλου να συλλαμβάνει πληροφορίες σε διαφορετικές κλίμακες.

Το Qwen2-VL ξεπερνά αυτόν τον περιορισμό επιτρέποντας στο μοντέλο να επεξεργάζεται δυναμικά εικόνες οποιασδήποτε ανάλυσης, μετατρέποντάς τες σε μεταβλητό αριθμό οπτικών tokens. Για την υποστήριξη αυτής της λειτουργίας, το μοντέλο τροποποιεί τον Vision Transformer αφαιρώντας τα αρχικά absolute position embeddings και εισάγοντας 2D Rotary Position Embedding για την αποτύπωση της δισδιάστατης χωρικής πληροφορίας των εικόνων [7], [21]. Κατά τη φάση inference, εικόνες διαφορετικών αναλύσεων συσκευάζονται σε ενιαία ακολουθία, με το μήκος της ακολουθίας να ελέγχεται για να περιοριστεί η χρήση μνήμης GPU. Για τη μείωση των οπτικών tokens κάθε εικόνας, απλό MLP layer χρησιμοποιείται μετά τον Vision Transformer για να συμπιέσει γειτονικά 2×2 tokens σε ένα μόνο token [7].

Το **Multimodal Rotary Position Embedding** αποτελεί τη δεύτερη σημαντική αρχιτεκτονική καινοτομία του Qwen2-VL [7]. Σε αντίθεση με το παραδοσιακό 1D-RoPE που χρησιμοποιείται στα Large Language Models και περιορίζεται στην κωδικοποίηση μονοδιάστατης πληροφορίας θέσης [21], το M-RoPE μοντελοποιεί αποτελεσματικά την πληροφορία θέσης πολυτροπικών εισόδων. Αυτό επιτυγχάνεται διαχωρίζοντας το αρχικό rotary embedding σε τρία συστατικά: temporal, height και width. Για τις εισόδους κειμένου, αυτά τα συστατικά χρησιμοποιούν ταυτόσημα position IDs, καθιστώντας το M-RoPE λειτουργικά ισοδύναμο με το 1D-RoPE. Κατά την επεξεργασία εικόνων, τα temporal IDs κάθε οπτικού token παραμένουν σταθερά, ενώ διακριτά IDs αντιστοιχίζονται στα συστατικά height και width με βάση τη θέση του token στην εικόνα [7].

Το **M-RoPE** όχι μόνο βελτιώνει τη μοντελοποίηση της πληροφορίας θέσης αλλά επίσης μειώνει την τιμή των position IDs για εικόνες και videos, επιτρέποντας στο μοντέλο να εκτείνεται σε μεγαλύτερες ακολουθίες κατά την inference. Αυτή η ικανότητα είναι ιδιαίτερα σημαντική για την επεξεργασία μεγάλων videos, όπου το μήκος της ακολουθίας μπορεί να γίνει σημαντικός περιορισμός. Οι πειραματικές αξιολογήσεις δείχνουν ότι το Qwen2-VL μπορεί να χειριστεί ακολουθίες έως και 80.000 tokens κατά την inference, παρόλο που εκπαιδεύτηκε με μέγιστο μήκος ακολουθίας 16.384 tokens [7].

Το Qwen2-VL υιοθετεί ενοποιημένη προσέγγιση για την επεξεργασία τόσο εικόνων όσο και videos, βελτιώνοντας τις ικανότητες οπτικής αντίληψης του μοντέλου. Για τη διατήρηση της πληροφορίας του video όσο το δυνατόν πληρέστερα, το μοντέλο δειγματοληπτεί κάθε video με δύο frames ανά δευτερόλεπτο. Επιπλέον, ενσωματώνει 3D convolutions με βάθος δύο για την επεξεργασία video εισόδων, επιτρέποντας στο μοντέλο να χειρίζεται 3D tubes αντί για 2D patches, και έτσι να επεξεργάζεται περισσότερα video frames χωρίς να αυξάνει το μήκος της ακολουθίας [7], [12].

Η εκπαίδευση του Qwen2-VL ακολουθεί τριφασική μεθοδολογία. Στην πρώτη φάση, εστιάζεται αποκλειστικά στην εκπαίδευση του Vision Transformer component, χρησιμοποιώντας τεράστιο corpus ζευγών εικόνας-κειμένου για τη βελτίωση της σημασιολογικής κατανόησης εντός του Large Language Model. Αυτή η φάση προεκπαίδευσης εστιάζει κυρίως στη μάθηση των σχέσεων εικόνας-κειμένου, στην αναγνώριση κειμενικού περιεχομένου μέσα σε εικόνες μέσω OCR, και σε εργασίες ταξινόμησης εικόνων. Η πρώτη φάση περιλαμβάνει περίπου 600 δισεκατομμύρια tokens.

Στη δεύτερη φάση προ-εκπαίδευσης, όλες οι παράμετροι ξεκλειδώνονται και το μοντέλο εκπαιδεύεται με ευρύτερο φάσμα δεδομένων για πιο ολοκληρωμένη μάθηση. Αυτή η φάση περιλαμβάνει επιπλέον 800 δισεκατομμύρια tokens δεδομένων σχετικών με εικόνες. Η φάση εισάγει μεγαλύτερο όγκο μικτού περιεχομένου εικόνας-κειμένου, διευκολύνοντας πιο λεπτή κατανόηση της αλληλεπίδρασης μεταξύ οπτικής και κειμενικής πληροφορίας. Η ενσωμάτωση datasets visual question answering βελτιώνει την ικανότητα του μοντέλου να ανταποκρίνεται σε ερωτήσεις σχετικές με εικόνες [7].

Η τρίτη φάση είναι το instruction fine-tuning, όπου οι παράμετροι του Vision Transformer κλειδώνονται και πραγματοποιείται αποκλειστικό fine-tuning του Large Language Model χρησιμοποιώντας instructional datasets. Τα πολυτροπικά συστατικά περιλαμβάνουν image question-answering, document parsing, multi-image comparison, video comprehension, video stream dialogue και agent-based interactions. Συνολικά, το Qwen2-VL επεξεργάζεται σωρευτικά 1.4 τρισεκατομμύρια tokens κατά τις φάσεις προ-εκπαίδευσης [7].

Οι πειραματικές αξιολογήσεις του Qwen2-VL επιδεικνύουν εξαιρετική απόδοση σε ευρύ φάσμα benchmarks. Το Qwen2-VL-72B επιτυγχάνει αποτελέσματα συγκρίσιμα με κορυφαία μοντέλα όπως το GPT-4o και το Claude 3.5-Sonnet σε διάφορα πολυτροπικά benchmarks. Στο DocVQA test set, το Qwen2-VL-72B επιτυγχάνει 96.5%, ξεπερνώντας τόσο το GPT-4o όσο και το προηγούμενο state-of-the-art. Στο OCRBench, το μοντέλο επιτυγχάνει score 877, ξεπερνώντας σημαντικά το προηγούμενο καλύτερο αποτέλεσμα [7].

Στις εργασίες video understanding, το Qwen2-VL επιδεικνύει ισχυρές επιδόσεις σε benchmarks που καλύπτουν από σύντομα videos μερικών δευτερολέπτων έως μεγάλα videos έως και μία ώρα. Στο MVBench, το Qwen2-VL-72B επιτυγχάνει 73.6%, ενώ στο EgoSchema test set επιτυγχάνει εντυπωσιακό 77.9%, ξεπερνώντας ακόμη και το GPT-4o. Η ικανότητα του μοντέλου να επεξεργάζεται videos διάρκειας πάνω από 20 λεπτά το καθιστά ιδιαίτερα κατάλληλο για εφαρμογές που απαιτούν κατανόηση εκτεταμένου χρονικού πλαισίου [7].

Το Qwen2-VL επίσης επιδεικνύει ισχυρές ικανότητες agent σε διάφορα benchmarks. Στην εργασία function calling, το μοντέλο επιτυγχάνει 93.1% type match και 53.2% exact match, ξεπερνώντας το GPT-4o. Στις εργασίες UI operations, το Qwen2-VL-72B επιτυγχάνει 89.6% type match στο AITZ benchmark, ξεπερνώντας σημαντικά το προηγούμενο state-of-the-art και το GPT-4o. Αυτές οι ικανότητες καθιστούν το Qwen2-VL κατάλληλο για ενσωμάτωση με συσκευές όπως smartphones και robots, επιτρέποντας αυτόνομη λειτουργία βασισμένη σε οπτικές εισόδους και κειμενικές οδηγίες [7].

2.7.8 Προκλήσεις και Αρχιτεκτονικοί Συμβιβασμοί

Η εφαρμογή των Vision-Language Models στο Dense Video Captioning συνοδεύεται από συγκεκριμένες προκλήσεις που σχετίζονται κυρίως με τη χωροχρονική φύση του video και τις αυξημένες απαιτήσεις σε σημασιολογική συνοχή [1]. Κεντρικό ζήτημα αποτελεί η αποδοτική επεξεργασία μακρών video sequences, γεγονός που καθιστά αναγκαίο τον προσεκτικό έλεγχο του αριθμού των visual tokens, ώστε να διατηρείται η κρίσιμη πληροφορία που σχετίζεται με γεγονότα και μεταβάσεις σκηνών χωρίς υπέρμετρο υπολογιστικό κόστος. Ιδιαίτερη σημασία έχει η κωδικοποίηση της χρονικής πληροφορίας, καθώς το μοντέλο καλείται να αναγνωρίσει γεγονότα που εξελίσσονται σε διαφορετικές

χρονικές κλίμακες και να διακρίνει σκηνές με παρόμοια χωρικά χαρακτηριστικά αλλά διαφορετικό χρονικό πλαίσιο. Παράλληλα, η παραγωγή πολλαπλών διαδοχικών περιγραφών προϋποθέτει μηχανισμούς που διασφαλίζουν σημασιολογική συνέπεια (semantic consistency) μεταξύ των παραγόμενων captions [3]. Το μοντέλο οφείλει να αναφέρεται με συνεπή τρόπο σε αντικείμενα και οντότητες που επανεμφανίζονται στο video, αποφεύγοντας τόσο αντιφάσεις όσο και πλεοναστικές επαναλήψεις.

Οι σύγχρονες προσεγγίσεις διαφοροποιούνται ως προς τον τρόπο δειγματοληψίας των frames, τη στρατηγική χρονικής μοντελοποίησης και τον βαθμό πολυπλοκότητας της αρχιτεκτονικής, εισάγοντας έναν σαφή συμβιβασμό (trade-off) μεταξύ ποιότητας περιγραφής και υπολογιστικής αποδοτικότητας. Απλούστερες προσεγγίσεις προσφέρουν υψηλή αποδοτικότητα αλλά περιορισμένη χρονική κατανόηση, ενώ πιο σύνθετοι μηχανισμοί temporal encoding βελτιώνουν το semantic και temporal reasoning με αυξημένες απαιτήσεις σε πόρους. Στο πλαίσιο αυτό, η επεξεργασία πολλών frames ανά σκηνή, σε συνδυασμό με μεγάλης κλίμακας Vision-Language Models, οδηγεί σε ανώτερη χρονική κατανόηση και πλουσιότερο semantic reasoning, αυξάνει όμως σημαντικά το latency κατά την inference και τις απαιτήσεις μνήμης [10]. Τεχνικές όπως η συγχώνευση tokens μέσω spatial pooling, καθώς και η χρήση frozen vision encoders αντί πλήρους end-to-end εκπαίδευσης, προσφέρουν πρακτικούς συμβιβασμούς, μειώνοντας τον αριθμό των visual tokens χωρίς αναγκαστική απώλεια κρίσιμης πληροφορίας [21].

Τα τρία μοντέλα που αξιολογούνται στην παρούσα εργασία αντιπροσωπεύουν διαφορετικές αρχιτεκτονικές φιλοσοφίες. Το BLIP υιοθετεί μια frame-based προσέγγιση, επεξεργαζόμενο ένα αντιπροσωπευτικό frame ανά σκηνή, γεγονός που προσφέρει υψηλή υπολογιστική αποδοτικότητα και το καθιστά κατάλληλο για σενάρια περιορισμένων πόρων, εις βάρος όμως της ρητής χρονικής μοντελοποίησης [5]. Το GIT-VATEX επεξεργάζεται πολλαπλά, ομοιόμορφα δειγματοληπτημένα frames με ενοποιημένη language modeling προσέγγιση, επιτυγχάνοντας πιο ισορροπημένο συνδυασμό temporal reasoning και αποδοτικότητας [39]. Τέλος, το Qwen2-VL ενσωματώνει προηγμένους μηχανισμούς χρονικής κωδικοποίησης, όπως το Multimodal Rotary Position Embedding (M-RoPE) και 3D convolutions, προσφέροντας ανώτερη χωροχρονική κατανόηση με τίμημα αυξημένες υπολογιστικές απαιτήσεις [7]. Η επιλογή του κατάλληλου μοντέλου εξαρτάται επομένως άμεσα από τις απαιτήσεις της εκάστοτε εφαρμογής και τη διαθέσιμη υπολογιστική υποδομή.

ΚΕΦΑΛΑΙΟ 3 Αλγόριθμοι εξαγωγής περιγραφής video

3.1 Κύριες Αρχιτεκτονικές Προσεγγίσεις για Dense Video Captioning

Το Dense Video Captioning αποτελεί μία από τις πλέον απαιτητικές εργασίες στη σύγχρονη έρευνα πολυτροπικών συστημάτων, καθώς συνδυάζει την πολυπλοκότητα του χρονικού εντοπισμού γεγονότων με την πρόκληση της παραγωγής φυσικών γλωσσικών περιγραφών. Σε αντίθεση με το παραδοσιακό video captioning που αποσκοπεί στη δημιουργία μίας ενιαίας λεζάντας για ολόκληρο το video, το Dense Video Captioning απαιτεί την αναγνώριση πολλαπλών γεγονότων εντός ενός video, τον ακριβή χρονικό προσδιορισμό των ορίων τους και την παραγωγή σημασιολογικά πλούσιων περιγραφών για κάθε ανιχνευμένο γεγονός. Η εργασία αυτή εισήχθη επίσημα το 2017 μέσω του ActivityNet Challenge και του πρωτοποριακού έργου των Krishna και συνεργατών [1], το οποίο όρισε το πρόβλημα και παρουσίασε το πρώτο σύνολο δεδομένων μεγάλης κλίμακας για την αξιολόγηση μεθόδων Dense Video Captioning. Από τότε, το πεδίο έχει γνωρίσει ραγδαία ανάπτυξη με την εμφάνιση διαφορετικών μεθοδολογικών προσεγγίσεων που κυμαίνονται από παραδοσιακές two-stage αρχιτεκτονικές έως ενοποιημένα end-toend μοντέλα και σύγχρονες προσεγγίσεις που αξιοποιούν προ-εκπαιδευμένα Vision-Language Models. Στο παρόν κεφάλαιο παρουσιάζονται οι βασικές αρχιτεκτονικές προσεγγίσεις που έχουν αναπτυχθεί για το Dense Video Captioning, με έμφαση στους μηχανισμούς παραγωγής λεζάντων και τις στρατηγικές που χρησιμοποιούνται για τη γέφυρα μεταξύ οπτικής αναπαράστασης και γλωσσικής έκφρασης. Οι μέθοδοι κατηγοριοποιούνται με βάση το αρχιτεκτονικό τους paradigm σε proposal-based (δύο σταδίων) και end-to-end προσεγγίσεις, ενώ εξετάζονται αναλυτικά οι μηχανισμοί που διέπουν την παραγωγή των τελικών περιγραφών και η συμβολή των σύγχρονων Vision-Language Models στην εξέλιξη του πεδίου.

3.2 Proposal-Based Μέθοδοι (Δύο Σταδίων)

Οι proposal-based μέθοδοι αποτελούν την πρώτη και πιο διαδεδομένη κατηγορία προσεγγίσεων για το Dense Video Captioning, ακολουθώντας τη φιλοσοφία της διαίρεσης του προβλήματος σε δύο ξεχωριστά στάδια. Στο πρώτο στάδιο, ένα event proposal module εντοπίζει υποψήφια χρονικά τμήματα που ενδεχομένως περιέχουν σημαντικά γεγονότα, ενώ στο δεύτερο στάδιο ένα captioning module παράγει φυσικές γλωσσικές περιγραφές για κάθε προτεινόμενο τμήμα [40]. Αυτή η modularity παρέχει σημαντικά πλεονεκτήματα όπως η δυνατότητα ανεξάρτητης βελτιστοποίησης κάθε συνιστώσας, η ευκολότερη ερμηνεία των ενδιάμεσων αποτελεσμάτων και η δυνατότητα επαναχρησιμοποίησης προεκπαιδευμένων μοντέλων για κάθε υποεργασία.

3.2.1 Θεμελιώδεις Αρχιτεκτονικές και Κλασικές Προσεγγίσεις

Το πρωτοποριακό έργο των Krishna και συνεργατών [1] εισήγαγε ένα ολοκληρωμένο σύστημα Dense Video Captioning που αποτελείται από ένα **proposal module** βασισμένο σε Deep Action Proposals και ένα **captioning module** που αξιοποιεί πληροφορίες χρονικού πλαισίου από γειτονικά γεγονότα. Το proposal module χρησιμοποιεί sliding temporal windows σε πολλαπλές χρονικές κλίμακες, επιτρέποντας την ανίχνευση γεγονότων τόσο μικρής όσο και μεγάλης διάρκειας. Κάθε παραγόμενο proposal ορίζεται από την αρχική και τελική χρονική στιγμή του, καθώς και από μια κρυφή αναπαράσταση που

κωδικοποιεί το οπτικό περιεχόμενο του αντίστοιχου τμήματος video. Το captioning module εισάγει έναν καινοτόμο μηχανισμό αξιοποίησης χρονικού πλαισίου μέσω ενός bidirectional LSTM, επιτρέποντας τη σύνθεση πιο συνεκτικών και σημασιολογικά πλούσιων περιγραφών. Η εργασία αυτή συνοδεύτηκε από τη δημιουργία του ActivityNet Captions dataset, το οποίο περιλαμβάνει περίπου 20.000 video και 100.000 χρονικά 32 προσδιορισμένες λεζάντες, καθιερώνοντας ένα από τα σημαντικότερα benchmarks για την αξιολόγηση συστημάτων Dense Video Captioning [1].

Μεταγενέστερες προσεγγίσεις επέκτειναν το αρχικό αυτό πλαίσιο με βελτιώσεις τόσο στο στάδιο της ανίχνευσης γεγονότων όσο και στην παραγωγή λεζαντών. Οι Wang και συνεργάτες εισήγαγαν bidirectional temporal context mechanisms που λαμβάνουν υπόψη τόσο το παρελθόν όσο και το μέλλον κάθε γεγονότος κατά τον προσδιορισμό των χρονικών του ορίων, βελτιώνοντας αισθητά την ακρίβεια του temporal localization [41]. Παράλληλα, η χρήση context gating mechanisms επιτρέπει την προσαρμοστική συγχώνευση πληροφορίας από διαφορετικά γεγονότα, ενισχύοντας τη συνοχή και τη σημασιολογική συνέπεια των παραγόμενων περιγραφών [41].

Μια επιπλέον κατεύθυνση αποτελεί η ανάπτυξη hierarchical αρχιτεκτονικών, όπου η αναπαράσταση των γεγονότων οργανώνεται σε πολλαπλά επίπεδα αφαίρεσης, επιτρέποντας την αποτελεσματικότερη μοντελοποίηση γεγονότων διαφορετικών χρονικών κλιμάκων [42].

3.2.2 Περιορισμοί και Προκλήσεις των Proposal-Based Προσεγγίσεων

Παρά τα πλεονεκτήματά τους, οι proposal-based μέθοδοι παρουσιάζουν σημαντικούς εγγενείς περιορισμούς. Η έλλειψη ουσιαστικής αλληλεπίδρασης μεταξύ των δύο σταδίων οδηγεί συχνά στο φαινόμενο της διάδοσης σφαλμάτων (error propagation), όπου ανακρίβειες στο στάδιο ανίχνευσης των proposals επηρεάζουν άμεσα την ποιότητα των τελικών λεζαντών [43].

Επιπλέον, η ανεξάρτητη βελτιστοποίηση κάθε module δεν εγγυάται τη βέλτιστη απόδοση του συνολικού συστήματος, καθώς απουσιάζει μηχανισμός άμεσης ανατροφοδότησης από το captioning module προς το proposal module [43]. Τέλος, οι προσεγγίσεις αυτές δυσκολεύονται να χειριστούν γεγονότα με ασαφή, συνεχόμενα ή επικαλυπτόμενα χρονικά όρια, δεδομένου ότι η διακριτή φύση των proposals δεν επιτρέπει την ευέλικτη αναπαράσταση τέτοιων περιπτώσεων [44].

3.3 Τεχνικές Χρονικού Εντοπισμού Γεγονότων (Temporal Localization)

Ο χρονικός εντοπισμός γεγονότων αποτελεί κρίσιμο συστατικό του Dense Video Captioning, καθώς ο ακριβής προσδιορισμός των temporal boundaries κάθε γεγονότος επηρεάζει άμεσα την ποιότητα των παραγόμενων περιγραφών. Οι μέθοδοι temporal localization έχουν εξελιχθεί σημαντικά τα τελευταία χρόνια, μεταβαίνοντας από απλές sliding window προσεγγίσεις σε εξελιγμένες τεχνικές που βασίζονται σε deep neural networks.

3.3.1 Anchor-Based Methods

Οι anchor-based μέθοδοι αντλούν έμπνευση από το object detection στον χώρο των εικόνων, χρησιμοποιώντας προκαθορισμένα temporal anchors διαφόρων διαρκειών ως σημεία αναφοράς για τον εντοπισμό γεγονότων [45]. Το **Single Shot Temporal Action Detection (SST)** αποτελεί μία από τις πρώτες anchor-based προσεγγίσεις, χρησιμοποιώντας 1D convolutional layers για την επεξεργασία χρονικών ακολουθιών χαρακτηριστικών και την πρόβλεψη τόσο των χρονικών ορίων όσο και των scores

εμπιστοσύνης για κάθε anchor [46]. Το SST ενσωματώνει μηχανισμούς multi-scale detection που επιτρέπουν την ανίχνευση γεγονότων με ποικίλες χρονικές διάρκειες μέσω της παράλληλης επεξεργασίας χαρακτηριστικών σε διαφορετικές χρονικές αναλύσεις [46].

Οι **Temporal Region Networks** επεκτείνουν την ιδέα των anchor-based μεθόδων με την εισαγωγή region proposal networks προσαρμοσμένων για την χρονική διάσταση [47]. Η αρχιτεκτονική τους περιλαμβάνει έναν temporal feature extractor που επεξεργάζεται ακολουθίες frames, ένα region proposal module που προτείνει υποψήφια 33 χρονικά τμήματα, και ένα classification module που αξιολογεί την ποιότητα κάθε πρότασης. Η χρήση διαφορετικών anchor shapes και scales επιτρέπει την αποτελεσματική κάλυψη γεγονότων με μεγάλο εύρος διαρκειών, από λίγα δευτερόλεπτα έως αρκετά λεπτά [47].

3.3.2 Boundary-Based Methods

Οι boundary-based μέθοδοι αντιπροσωπεύουν μία διαφορετική φιλοσοφία, εστιάζοντας στην ακριβή ανίχνευση των αρχών και των τελών γεγονότων αντί της πρόβλεψης ολόκληρων temporal segments. Το **Boundary-Sensitive Network (BSN)** εισάγει μία τριμερή αρχιτεκτονική που αποτελείται από τα temporal evaluation, proposal generation και proposal evaluation modules [48]. Το temporal evaluation module χρησιμοποιεί τριστρωματικά temporal convolutional networks για να υπολογίσει πιθανότητες για κάθε χρονική θέση του video σχετικά με το εάν αποτελεί αρχή ή τέλος ενός γεγονότος, καθώς και την πιθανότητα να ανήκει εντός ενός γεγονότος (actionness probability) [48]. Το proposal generation module συνδυάζει χρονικές θέσεις με υψηλές πιθανότητες αρχής και τέλους για τη δημιουργία candidate proposals, κατασκευάζοντας παράλληλα ένα Boundary-Sensitive Proposal feature για κάθε υποψήφια πρόταση βασισμένο στην ακολουθία actionness probabilities [48]. Τέλος, το proposal evaluation module, που υλοποιείται ως multilayer perceptron με ένα κρυφό επίπεδο, αξιολογεί το confidence score κάθε πρότασης χρησιμοποιώντας το BSP feature [48]. Η local-to-global φιλοσοφία του BSN επιτρέπει την ανίχνευση γεγονότων με ευέλικτα χρονικά όρια χωρίς την εξάρτηση από προκαθορισμένες διάρκειες, οδηγώντας σε υψηλή ακρίβεια και recall στα benchmarks ActivityNet-1.3 και THUMOS14 [48].

Το **Boundary-Matching Network (BMN)** επεκτείνει την προσέγγιση του BSN εισάγοντας τον Boundary-Matching μηχανισμό για την πυκνή αξιολόγηση των confidence scores όλων των πιθανών proposals [49]. Ο μηχανισμός BM αναπαριστά κάθε πρόταση ως ένα ζεύγος αρχικού και τελικού ορίου, συνδυάζοντας όλα τα πυκνά κατανεμημένα BM ζεύγη σε έναν δισδιάστατο BM confidence map [49]. Η αρχιτεκτονική του BMN περιλαμβάνει δύο παράλληλα branches που εκπαιδεύονται από κοινού σε ένα ενοποιημένο πλαίσιο. Το Temporal Evaluation Module (TEM) παράγει ακολουθίες πιθανοτήτων για τα temporal boundaries, ενώ το Proposal Evaluation Module (PEM) δημιουργεί τον BM confidence map που περιέχει confidence scores για πυκνά κατανεμημένες προτάσεις [49]. Ο BM layer, ένα κεντρικό συστατικό του PEM, παράγει BM feature maps από την χρονική ακολουθία χαρακτηριστικών, δειγματοληπτώντας ομοιόμορφα N σημεία εντός κάθε πρότασης για την κατασκευή του proposal-level feature representation [49]. Τα πειραματικά αποτελέσματα δείχνουν ότι το BMN επιτυγχάνει σημαντική βελτίωση απόδοσης με αξιοσημείωτη αποδοτικότητα και γενίκευση σε challenging datasets [49].

3.3.3 Anchor-Free Approaches

Οι anchor-free προσεγγίσεις αποτελούν την πιο πρόσφατη εξέλιξη στο temporal localization, αποφεύγοντας εντελώς τη χρήση προκαθορισμένων anchors ή boundaries [50]. Το **Anchor-Free Single-shot Detector (AFSD)** προβλέπει απευθείας τα κέντρα και τις διαρκείες των γεγονότων χωρίς να βασίζεται σε προκαθορισμένες παραμέτρους [50]. Η μέθοδος χρησιμοποιεί coarse-to-fine στρατηγική όπου αρχικά εντοπίζει περιοχές με υψηλή πιθανότητα να περιέχουν γεγονότα και στη συνέχεια εκλεπτύνει τον εντοπισμό σε αυτές τις περιοχές [50]. Οι anchor-free μέθοδοι προσφέρουν μεγαλύτερη ευελιξία στην αναπαράσταση γεγονότων με ασυνήθιστες διαρκείες και μειώνουν τον αριθμό των hyperparameters που απαιτούνται για tuning [50].

3.3.4 Scene Detection Techniques

Η ανίχνευση σκηνών αποτελεί μία εναλλακτική προσέγγιση για το temporal localization που εστιάζει στον εντοπισμό οπτικών μεταβάσεων μεταξύ διαφορετικών σκηνών ενός video. Το **PySceneDetect** αποτελεί ένα ευρέως διαδεδομένο εργαλείο που παρέχει πολλαπλούς αλγορίθμους για την ανίχνευση scene transitions [8], [75]. Ο **ContentDetector** υπολογίζει τη διαφορά στο περιεχόμενο μεταξύ διαδοχικών frames χρησιμοποιώντας ιστογράμματα του Y channel στον YCbCr χρωματικό χώρο, ανιχνεύοντας αλλαγές σκηνής όταν η διαφορά υπερβαίνει ένα προκαθορισμένο threshold [8]. Ο **AdaptiveDetector** επεκτείνει αυτή την προσέγγιση χρησιμοποιώντας προσαρμοστικό threshold βασισμένο στο rolling average των αλλαγών μεταξύ γειτονικών frames, βελτιώνοντας την απόδοση σε περιπτώσεις ταχείας κίνησης της κάμερας [8]. Ο **ThresholdDetector** εστιάζει στην ανίχνευση fade-in και fade-out transitions συγκρίνοντας τη μέση φωτεινότητα των frames με ένα καθορισμένο threshold [8]. Παρότι οι scene detection μέθοδοι είναι υπολογιστικά αποδοτικές και εύκολες στην υλοποίηση, συχνά δεν επιτυγχάνουν την ακρίβεια των μεθόδων που βασίζονται σε deep learning για τον εντοπισμό σημασιολογικά σημαντικών γεγονότων [8], [75].

3.4 End-to-End Αρχιτεκτονικές για Dense Video Captioning

Η μετάβαση από τις proposal-based (two-stage) προσεγγίσεις σε end-to-end αρχιτεκτονικές αποτελεί μία καθοριστική εξέλιξη στο πεδίο του Dense Video Captioning. Σε αντίθεση με τις αρθρωτές μεθόδους, όπου ο χρονικός εντοπισμός γεγονότων και η παραγωγή λεξαντών υλοποιούνται ως ανεξάρτητα στάδια, τα end-to-end μοντέλα επιτρέπουν την από κοινού βελτιστοποίηση των δύο υποεργασιών εντός ενός ενιαίου πλαισίου εκπαίδευσης [52]. Η ενοποίηση αυτή αντιμετωπίζει αποτελεσματικά το φαινόμενο της διάδοσης σφαλμάτων (error propagation) που χαρακτηρίζει τις proposal-based μεθόδους και επιτρέπει την άμεση αλληλεπίδραση μεταξύ temporal localization και caption generation μέσω κοινών ενδιάμεσων αναπαραστάσεων [52].

3.4.1 Joint Event Detection and Captioning

Η βασική ιδέα πίσω από τις end-to-end μεθόδους είναι η ταυτόχρονη εκπαίδευση για τον εντοπισμό γεγονότων και την παραγωγή περιγραφών, επιτρέποντας στο μοντέλο να μαθαίνει αναπαραστάσεις που είναι κατάλληλες και για τα δύο tasks. Πρωτοποριακές εργασίες όπως το JEDDi-Net (Joint Event Detection and Description Network) έδειξαν ότι είναι δυνατή η ενοποίηση των δύο σταδίων σε μία ενιαία αρχιτεκτονική που επεξεργάζεται συνεχώς το video stream με 3D συνελικτικά επίπεδα, προτείνει variable-length temporal events βασισμένο σε pooled features, και παράγει τις αντίστοιχες λεξάντες [73].

Μια κεντρική καινοτομία στις end-to-end προσεγγίσεις είναι η εισαγωγή μηχανισμών που συνδέουν άμεσα την ποιότητα της λεζάντας με τον χρονικό εντοπισμό. Για παράδειγμα, ορισμένες μέθοδοι προτείνουν την έννοια της descriptiveness regression, όπου το μοντέλο μαθαίνει να εκτιμά την περιγραφική πολυπλοκότητα κάθε detected proposal μέσω της διαδικασίας παραγωγής πρότασης (sentence generation). Αυτή η εκτίμηση χρησιμοποιείται στη συνέχεια για να προσαρμόσει τα χρονικά όρια κάθε event proposal, ευνοώντας proposals που μπορούν να περιγραφούν με πιο πλούσιες και λεπτομερείς λεζάντες. Με αυτόν τον τρόπο, το detection module λαμβάνει άμεσο feedback από το captioning module, δημιουργώντας μια αμφίδρομη ροή πληροφορίας που απουσιάζει στις παραδοσιακές two-stage αρχιτεκτονικές [73].

3.4.2 Transformer-Based End-to-End Αρχιτεκτονικές

Οι Transformer-based αρχιτεκτονικές έχουν επανασχεδιάσει τον τρόπο με τον οποίο προσεγγίζεται το Dense Video Captioning, αξιοποιώντας τους μηχανισμούς self-attention και cross-attention για την αποτελεσματική μοντελοποίηση των χρονικών εξαρτήσεων και την ταυτόχρονη παραγωγή event proposals και captions [52].

Το **Parallel Decoding for Dense Video Captioning (PDVC)** αποτελεί ένα ορόσημο στην κατηγορία αυτή, διατυπώνοντας το Dense Video Captioning ως πρόβλημα set prediction παρόμοιο με το DETR στο object detection. Το PDVC εισάγει μια καινοτόμο αρχιτεκτονική όπου ένας Transformer encoder επεξεργάζεται την ακολουθία των video features, ενώ ένας Transformer decoder με learnable event queries παράγει ταυτόχρονα τα temporal boundaries και τις λεζάντες για κάθε γεγονός [43].

Ένα κεντρικό χαρακτηριστικό του PDVC είναι ο event counter που προβλέπει τον αριθμό των γεγονότων στο video, επιτρέποντας στο μοντέλο να παράγει έναν event set κατάλληλου μεγέθους χωρίς να απαιτείται εφαρμογή heuristic non-maximum suppression. Τα event queries λειτουργούν ως learnable embeddings που αλληλεπιδρούν με τα video features μέσω cross-attention, κωδικοποιώντας πληροφορίες τόσο για τον χρονικό εντοπισμό όσο και για το σημασιολογικό περιεχόμενο κάθε γεγονότος. Οι enhanced representations των event queries τροφοδοτούνται παράλληλα σε δύο prediction heads, το localization head για την πρόβλεψη των temporal boundaries και confidence scores, και το captioning head για την παραγωγή των λεζαντών. Αυτή η παράλληλη αρχιτεκτονική επιτρέπει τη βαθιά αλληλεπίδραση και την αμοιβαία προώθηση των δύο υποεργασιών κατά τη διάρκεια της βελτιστοποίησης, οδηγώντας σε proposals με υψηλή descriptiveness και discriminative internal representations. Τα πειραματικά αποτελέσματα στα ActivityNet Captions και YouCook2 benchmarks δείχνουν ότι το PDVC ξεπερνά τις state-of-the-art two-stage μεθόδους όταν η ακρίβεια localization είναι συγκρίσιμη [43].

Το **Vid2Seq** αντιπροσωπεύει μία ακόμη πιο ενοποιημένη προσέγγιση, διατυπώνοντας το Dense Video Captioning ως sequence-to-sequence πρόβλημα όπου το μοντέλο παράγει μία ενιαία ακολουθία tokens που περιέχει τόσο χρονικές πληροφορίες όσο και κειμενικές περιγραφές [53]. Η αρχιτεκτονική του Vid2Seq επεκτείνει ένα pre-trained language model με ειδικά time tokens που αναπαριστούν διακριτοποιημένα timestamps στο video, παρόμοια με την προσέγγιση του Pix2Seq στο spatial domain. Ο visual encoder επεξεργάζεται την ακολουθία των video frames, ενώ ένας text encoder κωδικοποιεί προαιρετικά διαθέσιμο κείμενο όπως transcribed speech. Οι ενοποιημένες αναπαραστάσεις τροφοδοτούνται σε έναν text decoder που παράγει αυτορεγρεσσιακά την ακολουθία εξόδου, η οποία περιλαμβάνει εναλλασσόμενα time tokens και caption tokens. Αυτή η ενοποίηση επιτρέπει στο μοντέλο να μαθαίνει σύνθετες πολυτροπικές εξαρτήσεις μεταξύ των διαφορετικών γεγονότων μέσω του attention mechanism [53].

Ένα κρίσιμο χαρακτηριστικό του Vid2Seq είναι η δυνατότητά του για large-scale pretraining σε unlabeled narrated videos από το YT-Temporal-1B dataset που περιλαμβάνει 18 εκατομμύρια video. Η μεθοδολογία pretraining αναδιατυπώνει τα χρονικά

όρια των προτάσεων στο transcribed speech ως pseudo event boundaries και χρησιμοποιεί τις προτάσεις του speech ως pseudo event captions. Η εκπαίδευση συνδυάζει έναν generative objective που διδάσκει τον decoder να προβλέπει την ακολουθία του transcribed speech δεδομένων μόνο των οπτικών εισόδων, και έναν denoising objective που ενθαρρύνει την πολυτροπική μάθηση απαιτώντας από το μοντέλο να προβλέπει masked tokens δεδομένης μιας θορυβώδους ακολουθίας speech και οπτικών εισόδων. Το προ-εκπαιδευμένο Vid2Seq μοντέλο βελτιώνει το state of the art σε ποικίλα benchmarks όπως τα YouCook2, ViTT και ActivityNet Captions, και γενικεύει καλά σε few-shot settings, video paragraph captioning και standard video captioning [53].

3.4.3 Recurrent, Υβριδικές και Graph-Based Αρχιτεκτονικές

Παράλληλα με τις Transformer-based προσεγγίσεις, οι recurrent αρχιτεκτονικές συνεχίζουν να διαδραματίζουν σημαντικό ρόλο στο end-to-end Dense Video Captioning, ιδιαίτερα όταν συνδυάζονται με attention mechanisms. Τα LSTM-based μοντέλα εκμεταλλεύονται τη φυσική ικανότητα των recurrent networks να μοντελοποιούν χρονικές εξαρτήσεις, επεξεργαζόμενα το video ως μια ακολουθία γεγονότων όπου κάθε γεγονός εξαρτάται από τα προηγούμενα. Bidirectional LSTM αρχιτεκτονικές επιτρέπουν την ενσωμάτωση πληροφοριών τόσο από το παρελθόν όσο και από το μέλλον κάθε γεγονότος, βελτιώνοντας την ακρίβεια του χρονικού εντοπισμού και την ποιότητα των περιγραφών [56].

Οι **Graph Neural Networks (GNNs)** έχουν εισαχθεί πρόσφατα για τη μοντελοποίηση των σχέσεων μεταξύ γεγονότων σε ένα video [54]. Σε αυτές τις προσεγγίσεις, κάθε ανιχνευμένο γεγονός αναπαρίσταται ως ένας κόμβος στο γράφο, ενώ οι ακμές κωδικοποιούν χρονικές και σημασιολογικές σχέσεις μεταξύ των γεγονότων. Τα Graph Convolutional Networks επεξεργάζονται αυτή την αναπαράσταση για να εξάγουν refined event representations που λαμβάνουν υπόψη τις αλληλεπιδράσεις μεταξύ πολλαπλών γεγονότων [54]. Hybrid αρχιτεκτονικές που συνδυάζουν RNNs, CNNs και attention mechanisms έχουν επίσης προταθεί, προσπαθώντας να αξιοποιήσουν τα πλεονεκτήματα κάθε συνιστώσας για την ολοκληρωμένη κατανόηση του video content [55].

3.4.4 Multi-Task Learning Frameworks and Training Strategies

Τα **multi-task learning frameworks** προσεγγίζουν το Dense Video Captioning ως μέρος ενός ευρύτερου συνόλου συσχετιζόμενων εργασιών, επιτρέποντας στο μοντέλο να μαθαίνει πιο πλούσιες αναπαραστάσεις μέσω της κοινής εκπαίδευσης [59]. Τυπικές βοηθητικές εργασίες περιλαμβάνουν το video classification για την αναγνώριση των κύριων κατηγοριών δράσεων, το action recognition για τον εντοπισμό συγκεκριμένων actions εντός του video, και το temporal action localization για τον ακριβή προσδιορισμό των temporal boundaries των actions. Οι shared representations που μαθαίνονται μέσω αυτών των πολλαπλών εργασιών τείνουν να είναι πιο ανθεκτικές και να γενικεύουν καλύτερα σε unseen data. Η joint optimization με κατάλληλα σταθμισμένες loss functions για κάθε εργασία οδηγεί σε βελτιωμένη απόδοση τόσο στον χρονικό εντοπισμό όσο και στην παραγωγή λεξαντών [59].

Επιπλέον, στρατηγικές **curriculum learning** έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές για την εκπαίδευση σύνθετων end-to-end μοντέλων. Η βασική ιδέα είναι η σταδιακή αύξηση της πολυπλοκότητας των training examples κατά τη διάρκεια της εκπαίδευσης, ξεκινώντας από απλά video με λίγα και σαφώς διακριτά γεγονότα και προχωρώντας σταδιακά σε πιο σύνθετα video με πολλαπλά επικαλυπτόμενα γεγονότα. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να μαθαίνει σταδιακά τις απαραίτητες αναπαραστάσεις χωρίς να υπερφορτωθεί από την πλήρη πολυπλοκότητα του προβλήματος από την αρχή. Εναλλακτικά, η πολυπλοκότητα μπορεί να ελέγχεται μέσω της σταδιακής αύξησης του

μέγιστου αριθμού γεγονότων που το μοντέλο επιτρέπεται να προβλέψει, ή μέσω της προσαρμογής των βαρών των διαφόρων loss terms κατά τη διάρκεια της εκπαίδευσης [59].

3.4.5 Πλεονεκτήματα και Ανοιχτές Προκλήσεις

Οι end-to-end μέθοδοι προσφέρουν σημαντικά οφέλη σε σχέση με τις two-stage προσεγγίσεις. Η από κοινού βελτιστοποίηση και των δύο tasks επιτρέπει στο μοντέλο να μαθαίνει αναπαραστάσεις που είναι ταυτόχρονα κατάλληλες για detection και captioning, αποφεύγοντας την υπο-βελτιστοποίηση που προκύπτει από την ανεξάρτητη εκπαίδευση. Επιπλέον, η άμεση μετάδοση gradients από το captioning loss στο detection module δημιουργεί έναν feedback mechanism που βελτιώνει την ποιότητα των proposals ώστε να διευκολύνουν την παραγωγή συνεκτικών λεξάντων [73].

Παρά τα θεωρητικά τους πλεονεκτήματα, οι end-to-end μέθοδοι αντιμετωπίζουν και σημαντικές προκλήσεις. Η εκπαίδευση ενιαίων μοντέλων που αποδίδουν ικανοποιητικά και στα δύο tasks ταυτόχρονα είναι συχνά δύσκολη, καθώς τα δύο objectives μπορεί να έχουν διαφορετικές κλίμακες και να αντιμάχονται κατά την βελτιστοποίηση. Απαιτείται προσεκτική ρύθμιση των loss weights και συχνά multi-stage training strategies για την αποφυγή dominance του ενός task έναντι του άλλου. Επιπλέον, οι end-to-end μέθοδοι τείνουν να έχουν υψηλότερες απαιτήσεις σε δεδομένα εκπαίδευσης, καθώς η ταυτόχρονη εκμάθηση και των δύο tasks απαιτεί πλούσια και ποικιλόμορφα παραδείγματα για να γενικεύσει αποτελεσματικά [3].

3.5 Caption Generation: Αλγόριθμοι και Αρχιτεκτονικές

Η παραγωγή φυσικών γλωσσικών περιγραφών για video events αποτελεί τον πυρήνα του Dense Video Captioning, απαιτώντας την αποτελεσματική σύνδεση της οπτικής αναπαράστασης με τη γλωσσική έκφραση. Η εξέλιξη των caption generation μεθόδων έχει ακολουθήσει μια διαδρομή από απλά sequence-to-sequence μοντέλα έως σύνθετες Transformer αρχιτεκτονικές και τελικά σε πλούσια Vision-Language Models που αξιοποιούν τεράστιες ποσότητες προ-εκπαιδευμένων πολυτροπικών αναπαραστάσεων.

3.5.1 Sequence-to-Sequence Architectures και Attention Mechanisms

Τα sequence-to-sequence μοντέλα αποτέλεσαν την πρώτη επιτυχημένη προσέγγιση για την αυτόματη παραγωγή video captions, υιοθετώντας την encoder-decoder αρχιτεκτονική που είχε αποδειχθεί αποτελεσματική στη μηχανική μετάφραση [56]. Στην κλασική CNN-RNN προσέγγιση, ένα convolutional neural network εξάγει χωρικά χαρακτηριστικά από κάθε frame ή από αντιπροσωπευτικά frames του video event, ενώ ένα recurrent neural network λειτουργεί ως decoder για την παραγωγή της ακολουθίας λέξεων που συνθέτουν την περιγραφή [57]. Το LSTM χρησιμοποιήθηκε ευρέως ως decoder λόγω της ικανότητάς του να μοντελοποιεί μακροπρόθεσμες εξαρτήσεις στη γλώσσα, αποφεύγοντας το πρόβλημα του vanishing gradient που χαρακτηρίζει τα απλά RNNs [57].

Η εισαγωγή των attention mechanisms αποτέλεσε σημαντική καινοτομία που επέτρεψε στα μοντέλα να εστιάζουν επιλεκτικά σε διαφορετικά μέρη του οπτικού περιεχομένου κατά την παραγωγή κάθε λέξης. Ο Bahdanau attention mechanism υπολογίζει alignment scores μεταξύ του τρέχοντος hidden state του decoder και κάθε θέσης της encoder output, επιτρέποντας την προσαρμοστική συγκέντρωση πληροφοριών από το σύνολο των frames [18]. Ο Luong attention προσφέρει εναλλακτικές μεθόδους υπολογισμού των attention scores, όπως dot product, general bilinear form, και concatbased approaches, παρέχοντας διαφορετικά trade-offs μεταξύ υπολογιστικής πολυπλοκότητας και εκφραστικότητας [58]. Ο temporal attention επεκτείνει αυτές τις ιδέες για να λάβει υπόψη τη χρονική δομή του

video, επιτρέποντας στο μοντέλο να δίνει διαφορετικά βάρη σε frames ανάλογα με τη χρονική τους απόσταση από το γεγονός που περιγράφεται [65].

Οι hierarchical sequence models εισάγουν πολλαπλά επίπεδα αφαίρεσης στην αρχιτεκτονική, επεξεργαζόμενα πρώτα τα frames σε επίπεδο shot ή scene και στη συνέχεια συνθέτοντας τις αναπαραστάσεις αυτών των υψηλότερου επιπέδου μονάδων για την παραγωγή της τελικής περιγραφής. Αυτή η ιεραρχική δομή ανταποκρίνεται καλύτερα στη φυσική οργάνωση του video content και επιτρέπει την πιο αποτελεσματική μοντελοποίηση μακροπρόθεσμων χρονικών εξαρτήσεων [59].

3.5.2 Transformer-Based Caption Generation

Η επανάσταση του Transformer paradigm μετέβαλε ριζικά τη μοντελοποίηση ακολουθιών στη φυσική γλώσσα, και οι επιπτώσεις της στο video captioning ήταν εξίσου σημαντικές [19]. Τα Transformer-based caption generation μοντέλα αντικαθιστούν τα recurrent networks με στοίβες από self-attention και feed-forward layers, επιτρέποντας την παράλληλη επεξεργασία των ακολουθιών και την αποτελεσματικότερη αξιοποίηση των υπολογιστικών πόρων. Ο encoder-decoder Transformer για video captioning αποτελείται από έναν visual encoder που επεξεργάζεται την ακολουθία των frame features μέσω multi-head self-attention, και έναν language decoder που παράγει την περιγραφή αυτορεγρεσσιακά, συνδυάζοντας masked self-attention με cross-attention προς τις visual representations [19].

Η αξιοποίηση προ-εκπαιδευμένων language models ως decoders αποτελεί μία ισχυρή στρατηγική που επιτρέπει στα caption generation μοντέλα να επωφεληθούν από την πλούσια γλωσσική γνώση που έχει αποκτηθεί μέσω εκπαίδευσης σε τεράστια text corpora [60]. Μοντέλα όπως τα GPT, BART και T5 έχουν χρησιμοποιηθεί ως αφετηρία για video captioning tasks, με τα visual features να εισάγονται είτε ως πρόσθετες προθέματα στην είσοδο είτε μέσω cross-attention layers που εισάγονται στην αρχιτεκτονική [60]. Το fine-tuning αυτών των μεγάλων μοντέλων σε video captioning datasets οδηγεί συνήθως σε σημαντική βελτίωση της ποιότητας των παραγόμενων περιγραφών, ιδιαίτερα όσον αφορά τη γραμματική ορθότητα, τη φυσικότητα και την ποικιλία της γλώσσας [60].

Οι τεχνικές constrained decoding επιτρέπουν τον έλεγχο συγκεκριμένων χαρακτηριστικών των παραγόμενων captions, όπως το μήκος, το ύφος, ή η συμπερίληψη συγκεκριμένων λέξεων [61]. Αυτό επιτυγχάνεται μέσω της τροποποίησης των πιθανοτήτων που παράγει το μοντέλο κατά τη διάρκεια της decoding διαδικασίας, χρησιμοποιώντας penalty terms που αποθαρρύνουν ανεπιθύμητες συμπεριφορές ή bonus terms που ενθαρρύνουν επιθυμητά χαρακτηριστικά [61]. Για παράδειγμα, length penalties μπορούν να αποτρέψουν την παραγωγή υπερβολικά σύντομων ή μακροσκελών περιγραφών, ενώ coverage mechanisms μπορούν να διασφαλίσουν ότι διαφορετικές πτυχές του οπτικού περιεχομένου αναφέρονται στην περιγραφή [61].

3.5.3 Στρατηγικές Decoding και Παραγωγής

Η στρατηγική decoding που χρησιμοποιείται κατά την inference επηρεάζει σημαντικά την ποιότητα, την ποικιλία και την αποδοτικότητα της παραγωγής captions. Το greedy decoding αποτελεί την απλούστερη προσέγγιση, επιλέγοντας σε κάθε timestep το token με την υψηλότερη πιθανότητα [66]. Ενώ είναι υπολογιστικά αποδοτικό, το greedy decoding συχνά οδηγεί σε suboptimal αποτελέσματα καθώς δεν εξετάζει το overall quality της πλήρους ακολουθίας [66].

Το **beam search** αντιμετωπίζει αυτόν τον περιορισμό διατηρώντας ταυτόχρονα πολλαπλές υποψήφιες ακολουθίες (beams) και επιλέγοντας τελικά αυτή με την υψηλότερη συνολική πιθανότητα [66]. Το beam width ελέγχει τον αριθμό των υποψηφίων που διατηρούνται, με μεγαλύτερα widths να οδηγούν σε καλύτερη ποιότητα αλλά υψηλότερο

υπολογιστικό κόστος [66]. Παραλλαγές όπως το diverse beam search ενθαρρύνουν την ποικιλία μεταξύ των beams προσθέτοντας diversity penalty terms, αποφεύγοντας την παραγωγή πολλών σχεδόν πανομοιότυπων υποψηφίων [67]. Το constrained beam search επιτρέπει την επιβολή συγκεκριμένων περιορισμών όπως η υποχρεωτική συμπερίληψη ή αποκλεισμός συγκεκριμένων λέξεων [67].

Οι **sampling methods** προσφέρουν εναλλακτική προσέγγιση που εισάγει στοχαστικότητα στη διαδικασία παραγωγής, οδηγώντας σε μεγαλύτερη ποικιλία outputs [68]. Το temperature scaling τροποποιεί την κατανομή πιθανοτήτων πριν το sampling, με χαμηλές θερμοκρασίες να κάνουν την κατανομή πιο peaked (προτιμώντας τα highprobability tokens) και υψηλές θερμοκρασίες να την κάνουν πιο uniform (επιτρέποντας περισσότερη ποικιλία) [68]. Το top-k sampling περιορίζει το sampling στα k tokens με τις υψηλότερες πιθανότητες, ενώ το nucleus sampling (top-p) επιλέγει από το μικρότερο σύνολο tokens των οποίων οι συσσωρευτικές πιθανότητες ξεπερνούν ένα threshold p [68]. Αυτές οι μέθοδοι βρίσκουν εφαρμογή ιδιαίτερα στην παραγωγή πολλαπλών diverse captions για το ίδιο event [68].

Οι length penalties και repetition penalties αποτελούν σημαντικά εργαλεία για τον έλεγχο των χαρακτηριστικών των παραγόμενων captions [69]. Τα length penalties τροποποιούν τα scores των υποψηφίων ακολουθιών βασισμένα στο μήκος τους, αποθαρρύνοντας υπερβολικά σύντομες ή μακροσκελείς περιγραφές [69]. Τα repetition penalties μειώνουν τις πιθανότητες tokens που έχουν ήδη παραχθεί, αποφεύγοντας την ανεπιθύμητη επανάληψη λέξεων ή φράσεων [69]. Η κατάλληλη ρύθμιση αυτών των παραμέτρων είναι κρίσιμη για την εξισορρόπηση μεταξύ fluency, informativeness και diversity [69].

3.5.4 Πολυτροπική Ενσωμάτωση (Multimodal Fusion)

Παρά το γεγονός ότι το Κεφάλαιο 2 κάλυψε τις βασικές αρχιτεκτονικές fusion μεθόδους των Vision-Language Models, στο πλαίσιο της παραγωγής λεζάντων για Dense Video Captioning υπάρχουν εξειδικευμένες στρατηγικές που αξίζει να αναφερθούν. Πρόσφατες προσεγγίσεις ενσωματώνουν πολλαπλές modalities πέρα από το οπτικό σήμα, όπως audio features και textual information (π.χ. subtitles), για να εμπλουτίσουν την περιγραφική ικανότητα του μοντέλου.

Οι multi-stage fusion στρατηγικές προτείνουν να συγχωνεύονται διαφορετικές modalities σε διαφορετικά επίπεδα της αρχιτεκτονικής. Για παράδειγμα, το MS-FTN (Multi-Stage Fusion Transformer Network) ενσωματώνει πρώτα audio και visual features σε χαμηλότερο επίπεδο, και στη συνέχεια τις συνδυάζει με μια global-shared text representation σε υψηλότερο επίπεδο, παράγοντας πλούσιες multimodal context features που βελτιώνουν την ποιότητα των λεζάντων. Η χρήση text representation είναι ιδιαίτερα αποτελεσματική, καθώς το κείμενο (π.χ. subtitles) μοιράζεται παρόμοια δομή με τις παραγόμενες λεζάντες, διευκολύνοντας τη μεταφορά γνώσης και την alignment μεταξύ των modalities.

Το **MART** [64] χρησιμοποιεί memory-augmented recurrent transformers που προ-εκπαιδεύονται σε video paragraph captioning tasks. Η memory component επιτρέπει στο μοντέλο να διατηρεί πληροφορία για μεγάλα χρονικά διαστήματα, κάτι που είναι κρίσιμο για την κατανόηση μακρών video. Το **BMT (Bi-modal Transformer)** [65] επεκτείνει αυτήν την ιδέα ενσωματώνοντας τόσο οπτικές όσο και ακουστικές πληροφορίες, αναγνωρίζοντας ότι το audio περιέχει πολύτιμες πληροφορίες για την κατανόηση του video content (π.χ. dialogs, background sounds).

3.5.5 Evaluation-Guided Generation και Reinforcement Learning

Οι evaluation-guided generation τεχνικές χρησιμοποιούν τις evaluation metrics ως μέρος της εκπαίδευσης για να βελτιώσουν απευθείας τα χαρακτηριστικά που μετρώνται κατά την αξιολόγηση. Το **Self-Critical Sequence Training (SCST)** αποτελεί μία ευρέως χρησιμοποιούμενη μέθοδο που εφαρμόζει reinforcement learning για να βελτιστοποιήσει τα caption generation μοντέλα σχετικά με non-differentiable metrics όπως τα BLEU, METEOR και CIDEr [70]. Το SCST χρησιμοποιεί το ίδιο το μοντέλο για να παράγει baseline captions μέσω greedy decoding, και στη συνέχεια δειγματοληπτεί captions από την τρέχουσα πολιτική του μοντέλου. Το reward για κάθε sampled caption υπολογίζεται ως η διαφορά του score του από το baseline score, και το μοντέλο ενημερώνεται για να αυξήσει την πιθανότητα captions με θετικά rewards [70]. Αυτή η προσέγγιση έχει αποδειχθεί ιδιαίτερα αποτελεσματική για τη βελτίωση των CIDEr scores, που συσχετίζονται ισχυρά με την ανθρώπινη κρίση της caption quality [70].

Το **retrieval-augmented generation** για Dense Video Captioning αξιοποιεί εξωτερικές πηγές πληροφοριών για να ενισχύσει την παραγωγή περιγραφών [71]. Η βασική ιδέα είναι η ανάκτηση σχετικών captions από ένα μεγάλο corpus βασισμένη στην ομοιότητα των visual features, και η χρήση αυτών των ανακτηθέντων captions ως πρόσθετου context για το generation model. Αυτό μπορεί να βελτιώσει σημαντικά την ποιότητα των περιγραφών, ιδιαίτερα για uncommon events ή domain-specific content όπου το μοντέλο μπορεί να έχει περιορισμένη εμπειρία από το training set. Οι retrieval-augmented προσεγγίσεις απαιτούν προσεκτικό σχεδιασμό του retrieval mechanism και της fusion strategy για την αποτελεσματική ενσωμάτωση της ανακτηθείσας πληροφορίας [71].

3.6 Post-Processing και Βελτιστοποίηση Αποτελεσμάτων

Η διαδικασία post-processing αποτελεί κρίσιμο στάδιο στα συστήματα Dense Video Captioning, καθώς βελτιώνει την ποιότητα και τη συνέπεια των αποτελεσμάτων που παράγονται από τα βασικά μοντέλα. Οι τεχνικές post-processing εφαρμόζονται τόσο στα temporal proposals όσο και στις παραγόμενες περιγραφές, με στόχο τη μείωση της πλεονάζουσας πληροφορίας, τη βελτίωση της χρονικής συνοχής και την αύξηση της ποιότητας των κειμένων.

3.6.1 Non-Maximum Suppression για Temporal Proposals

Το Non-Maximum Suppression (NMS) αποτελεί τη βασική τεχνική για την απαλοιφή επικαλυπτόμενων temporal proposals που αφορούν το ίδιο γεγονός. Κατά τη διαδικασία localization, τα proposal-based συστήματα συχνά παράγουν πολλαπλά proposals με υψηλό βαθμό επικάλυψης. Το NMS διατηρεί το proposal με το υψηλότερο confidence score και απορρίπτει τα υπόλοιπα proposals που έχουν Intersection over Union (IoU) μεγαλύτερο από ένα καθορισμένο threshold [48], [49].

Η κλασική διαδικασία NMS ακολουθεί τα εξής βήματα. Αρχικά, τα proposals ταξινομούνται κατά φθίνουσα σειρά με βάση το confidence score τους. Στη συνέχεια, επιλέγεται το proposal με το υψηλότερο score και προστίθεται στη λίστα των τελικών proposals. Για κάθε ένα από τα υπόλοιπα proposals, υπολογίζεται το IoU με το επιλεγμένο proposal και αν το IoU υπερβαίνει το threshold (συνήθως 0.5 έως 0.7), το proposal απορρίπτεται. Η διαδικασία επαναλαμβάνεται μέχρι να εξεταστούν όλα τα proposals.

Παρά την απλότητά του, το κλασικό NMS παρουσιάζει περιορισμούς όταν πολλά γεγονότα συμβαίνουν ταυτόχρονα ή έχουν μερική χρονική επικάλυψη. Για αυτόν το λόγο, έχουν προταθεί βελτιωμένες εκδοχές όπως το Soft-NMS, το οποίο αντί να απορρίπτει πλήρως τα επικαλυπτόμενα proposals, μειώνει σταδιακά τα confidence scores τους ανάλογα με το IoU. Επιπλέον, προσεγγίσεις όπως το Boundary-Matching Network [49]

ενσωματώνουν πιο εξελιγμένους μηχανισμούς επιλογής proposals που λαμβάνουν υπόψη τόσο τα confidence scores όσο και την ποιότητα των ορίων των temporal segments.

3.6.2 Φιλτράρισμα Πλεονασμού και Temporal Consistency

Το φιλτράρισμα πλεονάζουσας πληροφορίας αποτελεί κρίσιμο βήμα για την αποφυγή επαναλαμβανόμενων ή σχεδόν ταυτόσημων περιγραφών σε συνεχόμενα χρονικά τμήματα. Συστήματα που δεν εφαρμόζουν τέτοιες τεχνίκες συχνά παράγουν πολλαπλές παρόμοιες περιγραφές για γεγονότα που εκτείνονται σε διαφορετικά temporal windows.

Μια προσέγγιση για το φιλτράρισμα πλεονασμού βασίζεται στον υπολογισμό της ομοιότητας μεταξύ διαδοχικών περιγραφών. Χρησιμοποιώντας μετρικές όπως η **cosine similarity σε sentence embeddings** (π.χ. μέσω **BERT** ή **Sentence-BERT**), το σύστημα μπορεί να εντοπίσει περιγραφές που είναι σημασιολογικά παρόμοιες και να κρατήσει μόνο αυτές που ξεπερνούν ένα threshold διαφορετικότητας. Εναλλακτικά, μπορούν να χρησιμοποιηθούν απλούστερες μετρικές όπως το n-gram overlap ή η edit distance.

Η **temporal consistency** αποτελεί επίσης σημαντική παράμετρο ποιότητας. Οι περιγραφές για συνεχόμενα χρονικά segments θα πρέπει να διατηρούν μια λογική συνοχή, αποφεύγοντας απότομες αλλαγές που δεν ανταποκρίνονται στο οπτικό περιεχόμενο. Προσεγγίσεις όπως το **MART** [64] χρησιμοποιούν memory-augmented transformers που διατηρούν πληροφορία από προηγούμενα segments, εξασφαλίζοντας έτσι μεγαλύτερη συνοχή στις παραγόμενες περιγραφές. Άλλες μέθοδοι εφαρμόζουν smoothing techniques που προσαρμόζουν τα confidence scores των proposals με βάση το ιστορικό των προηγούμενων predictions.

3.6.3 Linguistic Post-Processing

Η γλωσσική μετεπεξεργασία περιλαμβάνει τεχνικές που στοχεύουν στη βελτίωση της γραμματικής και συντακτικής ποιότητας των περιγραφών. Παρά την εντυπωσιακή πρόοδο των νευρωνικών μοντέλων, τα παραγόμενα κείμενα μπορεί να περιέχουν γραμματικά λάθη, επαναλήψεις λέξεων ή συντακτικές ανωμαλίες.

Βασικές τεχνικές linguistic post-processing περιλαμβάνουν το length normalization, όπου οι πιθανότητες των ακολουθιών κανονικοποιούνται ως προς το μήκος τους για να αποφευχθεί η τάση των μοντέλων να προτιμούν σύντομες περιγραφές. Επίσης, χρησιμοποιούνται coverage mechanisms που παρακολουθούν ποιες πληροφορίες έχουν ήδη αναφερθεί, αποτρέποντας την επανάληψη των ίδιων λέξεων ή φράσεων. Πιο σύνθετες προσεγγίσεις ενσωματώνουν grammar checking modules που ελέγχουν και διορθώνουν γραμματικά λάθη, ή χρησιμοποιούν language models για reranking των candidate captions με βάση τη γλωσσική τους ευφράδεια.

3.6.4 Ensemble Methods

Οι ensemble προσεγγίσεις αποτελούν μια ισχυρή τεχνική για τη βελτίωση των αποτελεσμάτων συνδυάζοντας τις προβλέψεις πολλαπλών μοντέλων. Στο Dense Video Captioning, το ensembling μπορεί να εφαρμοστεί τόσο στο επίπεδο των temporal proposals όσο και στο επίπεδο των caption generators.

Για το temporal localization, ένα ensemble μπορεί να συνδυάζει proposals από διαφορετικά μοντέλα (π.χ. BSN και BMN) μέσω weighted voting ή late fusion. Τα proposals που προτείνονται από πολλαπλά μοντέλα λαμβάνουν υψηλότερα confidence scores, ενώ proposals που προέρχονται από ένα μόνο μοντέλο υπόκεινται σε αυστηρότερα κριτήρια επιλογής. Για την παραγωγή περιγραφών, μπορούν να χρησιμοποιηθούν πολλαπλά caption generation models με διαφορετικές αρχιτεκτονικές ή εκπαιδευμένα σε διαφορετικά

δεδομένα. Οι τελικές περιγραφές επιλέγονται μέσω consensus voting ή με βάση μετρικές ποιότητας όπως το SPICE [61].

Τέλος, αξίζει να σημειωθεί ότι οι τεχνικές post-processing δεν αποτελούν απλώς προσθήκες, αλλά αναπόσπαστο μέρος των σύγχρονων Dense Video Captioning συστημάτων. Σύμφωνα με εμπειρικές μελέτες [40], [51], η σωστή εφαρμογή post-processing μπορεί να βελτιώσει την τελική απόδοση στις βασικές μετρικές αξιολόγησης, αποδεικνύοντας τη σημαντικότητά τους στην επίτευξη state-of-the-art αποτελεσμάτων.

3.6.5 Large Language Models για Narrative Coherence

Μια πρόσφατη και ιδιαίτερα υποσχόμενη προσέγγιση είναι η χρήση Large Language Models (LLMs) ως post-processing module για τη βελτίωση της narrative coherence σε επίπεδο ολόκληρου video. Σε αυτή τη στρατηγική, ένα standard Dense Video Captioning σύστημα παράγει αρχικά μεμονωμένες λεζάντες για κάθε event proposal. Στη συνέχεια, το σύνολο αυτών των λεζάντων, μαζί με τις αντίστοιχες χρονικές τους θέσεις, τροφοδοτείται σε ένα LLM (όπως το GPT-4 ή το Qwen2) με ένα κατάλληλο prompt που ζητά από το μοντέλο να παράγει μια ενοποιημένη, συνεκτική αφήγηση.

Το LLM, αξιοποιώντας τις ικανότητές του σε commonsense reasoning και discourse understanding, μπορεί να:

- Αναδιοργανώσει τις περιγραφές ώστε να ακολουθούν μια λογική χρονική ροή.
- Προσθέσει συνδετικούς όρους και cohesive devices (π.χ. "then", "meanwhile", "after that") που κάνουν την αφήγηση πιο φυσική.
- Εξαλείψει περιττές επαναλήψεις και συνοψίσει παρόμοια γεγονότα.
- Εισαγάγει αιτιακές σχέσεις και εξηγήσεις που δεν ήταν ρητές στις αρχικές λεζάντες.

Για παράδειγμα, αν οι αρχικές λεζάντες είναι:

i. [0:00-0:05]: "A man enters a kitchen"
ii. [0:05-0:10]: "A man opens the fridge"
iii. [0:10-0:15]: "A man takes out ingredients"

Το LLM μπορεί να τις συνδυάσει σε: "A man enters the kitchen, opens the fridge, and takes out ingredients, likely preparing to cook a meal." Αυτή η συνοψισμένη και εμπλουτισμένη αφήγηση είναι πιο κατανοητή και προσφέρει μεγαλύτερη αξία στον χρήστη.

Πειραματικά αποτελέσματα σε πολυδάστατα Dense Video Captioning benchmarks δείχνουν ότι η ενσωμάτωση LLMs ως narrative post-processing module μπορεί να βελτιώσει τη METEOR score κατά έως και 9%, ενώ παράλληλα αυξάνει την ανθρώπινη αξιολόγηση της συνοχής και της φυσικότητας των περιγραφών. Ωστόσο, η προσέγγιση αυτή εισάγει επιπλέον υπολογιστικό κόστος και εξάρτηση από το LLM, και απαιτεί προσεκτικό prompt engineering για να αποφευχθεί η παραγωγή hallucinations (εσφαλμένων ή μη υπαρκτών πληροφοριών που δεν υποστηρίζονται από το οπτικό περιεχόμενο).

3.7 Σύγχρονες Προσεγγίσεις με Vision-Language Models

Οι σύγχρονες προσεγγίσεις στο Dense Video Captioning αξιοποιούν τις δυνατότητες των Vision-Language Models (VLMs) που έχουν προ-εκπαιδευτεί σε τεράστια datasets εικόνων και κειμένου. Αυτά τα μοντέλα έχουν μάθει πλούσιες multimodal representations που επιτρέπουν την αποτελεσματική μεταφορά γνώσης (transfer learning) στο πρόβλημα του video captioning, συχνά με εντυπωσιακά αποτελέσματα ακόμα και με περιορισμένα video-specific training data.

3.7.1 Αξιοποίηση Pre-trained Vision-Language Models

Η άνοδος μοντέλων όπως το CLIP [27], BLIP, και άλλων multimodal transformers έχει μεταμορφώσει την προσέγγιση στο Dense Video Captioning. Αντί να εκπαιδεύονται μοντέλα από την αρχή χρησιμοποιώντας μόνο video captioning datasets, οι σύγχρονες μέθοδοι ξεκινούν από μοντέλα που έχουν ήδη μάθει να συσχετίζουν οπτικό περιεχόμενο με φυσική γλώσσα μέσω contrastive learning ή generative pretraining σε εκατομμύρια image-text pairs.

Για το Dense Video Captioning, τα video frames μπορούν να κωδικοποιηθούν χρησιμοποιώντας τον visual encoder του CLIP, παράγοντας representations που είναι ήδη aligned με γλωσσικές έννοιες. Αυτό διευκολύνει σημαντικά την εκπαίδευση caption generators, καθώς τα visual features έχουν ήδη σημασιολογικό πλούτο που συσχετίζεται άμεσα με λεκτικές περιγραφές. Το SimVLM [71] προσφέρει μια διαφορετική προσέγγιση μέσω prefix language modeling, όπου το μοντέλο μαθαίνει να προβλέπει tokens κειμένου conditioned σε visual inputs. Η προ-εκπαίδευση με weakly supervised data (εικόνες με noisy captions από το web) επιτρέπει στο SimVLM να μάθει robust multimodal representations χωρίς να απαιτούνται high-quality annotations. Όταν προσαρμόζεται για Dense Video Captioning, το SimVLM μπορεί να αξιοποιήσει αυτήν την προ-εκπαίδευση για να παράγει πιο λεπτομερείς και σημασιολογικά πλούσιες περιγραφές.

3.7.2 Prompt-Based Approaches

Οι prompt-based προσεγγίσεις αποτελούν μια πρόσφατη τάση που αξιοποιεί την ικανότητα των VLMs να ανταποκρίνονται σε φυσικές γλωσσικές οδηγίες. Αντί να fine-tuned ολόκληρο το μοντέλο για το συγκεκριμένο task, αυτές οι μέθοδοι χρησιμοποιούν carefully designed prompts για να καθοδηγήσουν το μοντέλο στην επιθυμητή συμπεριφορά [63].

Για παράδειγμα, ένα VLM μπορεί να δεχτεί ως είσοδο frames από ένα video segment μαζί με ένα prompt όπως "Describe what is happening in this video clip in detail:" και να παράγει την αντίστοιχη περιγραφή. Πιο εξελιγμένες προσεγγίσεις χρησιμοποιούν task-specific prompts που περιλαμβάνουν πληροφορίες για το domain, το στυλ της περιγραφής, ή συγκεκριμένες πτυχές που πρέπει να καλυφθούν (π.χ. "Describe the actions, objects, and environment in this sports video segment").

Το prompt tuning και το prefix tuning αποτελούν τεχνικές που βελτιστοποιούν learnable prompt embeddings αντί να fine-tuned όλες οι παράμετροι του μοντέλου. Αυτές οι μέθοδοι είναι ιδιαίτερα αποδοτικές όταν τα διαθέσιμα training data είναι περιορισμένα, καθώς απαιτούν την εκπαίδευση μόνο ενός μικρού αριθμού παραμέτρων ενώ διατηρούν τη γενική γνώση του προ-εκπαιδευμένου μοντέλου. Επιπλέον, διαφορετικά prompts μπορούν να χρησιμοποιηθούν για διαφορετικά domains ή styles χωρίς να απαιτείται πλήρης επανεκπαίδευση του μοντέλου.

3.7.3 Zero-Shot και Few-Shot Learning

Μια από τις πιο ενδιαφέρουσες δυνατότητες των σύγχρονων VLMs είναι η ικανότητά τους για zero-shot και few-shot learning. Λόγω της εκτενούς προ-εκπαίδευσής τους, αυτά τα μοντέλα μπορούν να εκτελέσουν Dense Video Captioning χωρίς καθόλου task-specific training (zero-shot) ή με ελάχιστα παραδείγματα (few-shot).

Στο πλαίσιο **zero-shot inference**, μοντέλα όπως το GPT-4 Vision ή το Gemini μπορούν να εφαρμοστούν χωρίς περαιτέρω εκπαίδευση, αξιοποιώντας τη γενικευμένη πολυτροπική γνώση που έχει αποκτηθεί κατά την προ-εκπαίδευσή τους. Η γενικής του κατανόηση του οπτικού περιεχομένου και της γλώσσας επιτρέπει την παραγωγή λογικών περιγραφών, αν και η ποιότητα μπορεί να υστερεί σε σύγκριση με task-specific μοντέλα, ιδιαίτερα όσον αφορά την temporal precision.

Το **few-shot learning** αποτελεί μια ενδιάμεση προσέγγιση μεταξύ zero-shot και πλήρους fine-tuning, παρέχοντας στο μοντέλο περιορισμένο αριθμό παραδειγμάτων (π.χ. 1–10) Dense Video Captioning outputs ως in-context examples. Μέσω αυτής της διαδικασίας, το μοντέλο προσαρμόζει τη συμπεριφορά του κατά το inference, χωρίς να τροποποιεί τις παραμέτρους του, αξιοποιώντας τα παρεχόμενα παραδείγματα για την καλύτερη ευθυγράμμιση του ύφους, της δομής και του επιπέδου λεπτομέρειας των παραγόμενων περιγραφών. Πρόσφατες μελέτες δείχνουν ότι με carefully selected examples και proper prompt engineering, few-shot VLMs μπορούν να πλησιάσουν ή ακόμα να ξεπεράσουν την απόδοση πλήρως supervised μεθόδων σε ορισμένα benchmarks.

3.8 Σύνοψη, Σύγκριση και Τάσεις

Οι μέθοδοι Dense Video Captioning που εξετάστηκαν στο κεφάλαιο διαφέρουν ουσιαστικά στον αρχιτεκτονικό σχεδιασμό, τις υπολογιστικές απαιτήσεις και την απόδοση. Η κατανόηση των trade-offs μεταξύ αυτών των προσεγγίσεων καθορίζει την επιλογή της κατάλληλης μεθόδου για κάθε εφαρμογή.

3.8.1 Σύγκριση Proposal-Based, End-to-End και Pre-Trained Προσεγγίσεων

Οι proposal-based προσεγγίσεις προσφέρουν interpretability και modularity, κάθε component βελτιστοποιείται ανεξάρτητα, διευκολύνοντας το debugging και την ανάλυση errors. Ωστόσο, το error propagation από το temporal localization στο caption generation περιορίζει την τελική απόδοση. Αντίθετα, οι end-to-end μέθοδοι με transformers μαθαίνουν από κοινού temporal structure και semantic content, επιτυγχάνοντας υψηλότερη ακρίβεια αλλά με αυξημένη πολυπλοκότητα εκπαίδευσης [43].

Τα παραδοσιακά task-specific μοντέλα εξειδικεύονται στην temporal συνέχεια και event detection, παραμένοντας υπολογιστικά αποδοτικά. Τα Vision-Language Models φέρουν πλούσια σημασιολογική κατανόηση από large-scale pretraining, παράγοντας πιο λεπτομερείς περιγραφές ακόμα και για uncommon objects. Η temporal precision τους όμως υστερεί, καθώς πολλά VLMs προ-εκπαιδεύονται σε static images και απαιτούν εκτεταμένο fine-tuning για ακριβή temporal grounding [27], [53].

3.8.2 Trade-offs: Ακρίβεια, Ταχύτητα και Πόροι

Η επιλογή μεθόδου Dense Video Captioning συχνά περιλαμβάνει trade-offs μεταξύ ακρίβειας, ταχύτητας inference και υπολογιστικών απαιτήσεων.

Οι proposal-based μέθοδοι με lightweight caption generators μπορούν να επεξεργαστούν videos σχετικά γρήγορα, κάνοντάς τις κατάλληλες για εφαρμογές που απαιτούν near-real-time processing. Για παράδειγμα, συστήματα που χρησιμοποιούν BSN

[48] για proposals και ένα μικρό LSTM για captions μπορούν να επιτύχουν αποδεκτή απόδοση με χαμηλό υπολογιστικό κόστος.

Οι end-to-end transformer-based μέθοδοι, όπως το PDVC [43], προσφέρουν καλύτερη ακρίβεια αλλά με αυξημένες υπολογιστικές απαιτήσεις. Η χρήση self-attention σε μεγάλα sets από video frames είναι computationally expensive, ιδιαίτερα για μακρά video. Τεχνικές όπως το sparse attention ή το temporal downsampling μπορούν να μειώσουν το κόστος, αλλά συχνά με μικρή απώλεια στην ακρίβεια.

Τα μεγάλης κλίμακας Vision–Language Models καταλαμβάνουν μια διαφορετική θέση στο σχεδιαστικό φάσμα των προσεγγίσεων για την περιγραφή video, προσφέροντας ιδιαίτερα υψηλή ποιότητα παραγόμενων λεζάντων με αντάλλαγμα αυξημένες υπολογιστικές απαιτήσεις. Ενδεικτικά, ένα μοντέλο όπως το Vid2Seq [53], με τάξη μεγέθους δισεκατομμυρίων παραμέτρων, ενδέχεται να απαιτεί τη χρήση πολλαπλών GPUs ακόμη και κατά το στάδιο του inference, ενώ η ταχύτητα επεξεργασίας περιορίζεται σε μικρό αριθμό video ανά δευτερόλεπτο. Ως αποτέλεσμα, η πρακτική αξιοποίησή τους είναι συχνά δυσχερής σε σενάρια πραγματικού χρόνου ή σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους. Παρά τους περιορισμούς αυτούς, το βασικό τους πλεονέκτημα έγκειται στην ισχυρή ικανότητα γενίκευσης, καθώς μπορούν να παράγουν υψηλής ποιότητας περιγραφές σε zero-shot ή few-shot σενάρια, ακόμη και χωρίς περαιτέρω fine-tuning.

3.8.3 Σύγχρονες Τάσεις και Ερευνητικές Κατευθύνσεις

Οι πρόσφατες ερευνητικές εξελίξεις στο Dense Video Captioning συγκλίνουν προς την κατεύθυνση της κλιμάκωσης, της χρονικής ευελιξίας και της πολυτροπικής κατανόησης. Μία κυρίαρχη τάση αφορά την ανάπτυξη video foundation models, όπως τα Vid2Seq και InternVideo2, τα οποία προεκπαιδεύονται σε μεγάλης κλίμακας σύνολα δεδομένων με εκατομμύρια video. Η αξιοποίηση weak ή pseudo labels, προερχόμενων από speech transcripts ή αυτόματες περιγραφές, επιτρέπει την εκμάθηση πλούσιων σημασιολογικών αναπαραστάσεων χωρίς την ανάγκη εκτενούς χειροκίνητης επισημείωσης [53].

Παράλληλα, αυξανόμενο ενδιαφέρον παρουσιάζουν οι streaming αρχιτεκτονικές, οι οποίες στοχεύουν στην επεξεργασία arbitrarily long videos σε πραγματικό χρόνο. Τεχνικές όπως το deformable attention και οι memory-based μηχανισμοί επιτρέπουν την προοδευτική ενημέρωση της χρονικής πληροφορίας, καθιστώντας εφικτή την εφαρμογή Dense Video Captioning σε live περιβάλλοντα, όπως παρακολούθηση συμβάντων και online analytics [72].

Τέλος, η πολυτροπική ολοκλήρωση αποτελεί πλέον βασικό άξονα έρευνας. Η συνδυαστική αξιοποίηση οπτικής πληροφορίας, ήχου, ομιλίας και υποτίτλων έχει αποδειχθεί κρίσιμη για την παραγωγή συνεκτικών και σημασιολογικά πλούσιων λεζάντων. Συστήματα όπως τα InternVideo2 και MS-FTN καταδεικνύουν ότι η ενσωμάτωση πολλαπλών modalities οδηγεί σε ανώτερη κατανόηση σύνθετων γεγονότων και αφηγηματικών δομών [60].

3.8.4 Ανοιχτές Προκλήσεις και Περιορισμοί

Παρά τη σημαντική πρόοδο του πεδίου, το Dense Video Captioning εξακολουθεί να αντιμετωπίζει ουσιώδεις προκλήσεις. Ένα βασικό πρόβλημα αφορά την ασάφεια στον καθορισμό των χρονικών ορίων των γεγονότων, ιδιαίτερα σε περιπτώσεις επικαλυπτόμενων ή σταδιακά εξελισσόμενων ενεργειών. Η χρήση αυστηρών IoU thresholds στα benchmarks δυσχεραίνει την αξιόπιστη αξιολόγηση και ενδέχεται να υποτιμά ποιοτικά ορθές προβλέψεις.

Επιπλέον, η αφηγηματική ασυνέπεια των παραγόμενων λεζάντων παραμένει σημαντικός περιορισμός. Φαινόμενα όπως επαναλήψεις, αντιφατικές αναφορές ή έλλειψη

συνδετικών στοιχείων μειώνουν τη χρησιμότητα των αποτελεσμάτων, καθιστώντας συχνά απαραίτητη την εφαρμογή post-processing ή LLM-based refinement τεχνικών.

Η κατανόηση μακρών video αποτελεί επίσης ανοιχτό ερευνητικό ζήτημα. Τα περισσότερα υφιστάμενα benchmarks επικεντρώνονται σε σύντομα clips, ενώ η κλιμάκωση σε μεγάλης διάρκειας περιεχόμενο συνοδεύεται από εκθετική αύξηση των υπολογιστικών απαιτήσεων. Τέλος, τα pre-trained Vision–Language Models εμφανίζουν φαινόμενα hallucinations, όπου περιγράφονται αντικείμενα, χρώματα ή ενέργειες που δεν υφίστανται στο video, αναδεικνύοντας την ανάγκη για ισχυρότερους μηχανισμούς grounding και ελέγχου αξιοπιστίας.

ΚΕΦΑΛΑΙΟ 4 Προτεινόμενο μοντέλο

4.1 Αρχιτεκτονική Επισκόπηση και Σχεδιαστικές Αρχές

Το προτεινόμενο σύστημα Dense Video Captioning σχεδιάστηκε με στόχο την αξιόπιστη και συστηματική αξιοποίηση σύγχρονων προ-εκπαιδευμένων μοντέλων όρασης–γλώσσας για την παραγωγή περιγραφικών λεζάντων σε χρονικά εντοπισμένα τμήματα video. Σε αντίθεση με μεγάλο μέρος της σχετικής βιβλιογραφίας, όπου η έμφαση δίνεται στην εκπαίδευση εξειδικευμένων end-to-end μοντέλων σε μεγάλα επισημειωμένα σύνολα δεδομένων, η παρούσα εργασία υιοθετεί μια διαφορετική προσέγγιση. Κεντρική επιδίωξη δεν αποτελεί η ανάπτυξη ενός νέου μοντέλου, αλλά η σχεδίαση ενός ενιαίου και παραμετροποιήσιμου συστήματος επεξεργασίας, το οποίο επιτρέπει τη δίκαιη συγκριτική αξιολόγηση ετερογενών αρχιτεκτονικών υπό κοινές συνθήκες.

Η επιλογή αυτή υπαγορεύεται τόσο από πρακτικούς όσο και από επιστημονικούς λόγους. Η εκπαίδευση σύγχρονων Dense Video Captioning μοντέλων από την αρχή προϋποθέτει πρόσβαση σε μεγάλης κλίμακας υπολογιστικούς πόρους και εκτεταμένα επισημειωμένα δεδομένα. Παράλληλα, η ραγδαία εξέλιξη των προ-εκπαιδευμένων Vision-Language Models καθιστά ιδιαίτερα ενδιαφέρουσα τη διερεύνηση του κατά πόσο τέτοια μοντέλα μπορούν να αξιοποιηθούν αποτελεσματικά σε απαιτητικές εργασίες όπως το Dense Video Captioning, ακόμη και χωρίς περαιτέρω fine-tuning.

Βασική σχεδιαστική αρχή του συστήματος αποτελεί η υιοθέτηση μιας αρχιτεκτονικής διακριτών σταδίων, στην οποία η ανάλυση του video, ο χρονικός εντοπισμός σκηνών, η εξαγωγή οπτικών χαρακτηριστικών και η παραγωγή λεζάντων υλοποιούνται ως ανεξάρτητα αλλά αλληλένδετα υποσυστήματα. Η modular αυτή δομή επιτρέπει την απομόνωση και τη μελέτη της συνεισφοράς κάθε επιμέρους συνιστώσας, διευκολύνοντας τόσο την ανάλυση των αποτελεσμάτων όσο και την πραγματοποίηση ablation studies. Επιπλέον, εξασφαλίζει ότι διαφορετικά μοντέλα παραγωγής λεζάντων μπορούν να ενσωματωθούν στο ίδιο πλαίσιο χωρίς τροποποίηση της συνολικής ροής επεξεργασίας.

Η επεξεργασία ενός video διέρχεται από πέντε διακριτά στάδια που εκτελούνται διαδοχικά, σχηματίζοντας έναν σύγχρονο pipeline επεξεργασίας δεδομένων. Σε πρώτο στάδιο πραγματοποιείται η **ανίχνευση σκηνών (Scene Detection)**, κατά την οποία εντοπίζονται αυτόματα τα χρονικά όρια των επιμέρους σκηνών με βάση αλλαγές στο οπτικό περιεχόμενο. Στη συνέχεια ακολουθεί η **εξαγωγή keyframes (Keyframe Extraction)**, όπου επιλέγονται τα πλέον αντιπροσωπευτικά frames από κάθε σκηνή, σύμφωνα με τις απαιτήσεις της εκάστοτε αρχιτεκτονικής μοντέλου. Τα frames αυτά χρησιμοποιούνται στο στάδιο της **παραγωγής λεζάντων (Caption Generation)**, όπου παρέχεται η δυνατότητα επιλογής και ενσωμάτωσης διαφορετικών μοντέλων (όπως τα BLIP, GIT-VATEX, Qwen2-VL) για τη δημιουργία φυσικών γλωσσικών περιγραφών. Ακολούθως, εφαρμόζεται **σημασιολογική συγχώνευση διαδοχικών σκηνών (Semantic Merging)** με παρόμοιο περιεχόμενο, με στόχο τη μείωση της πλεοναστικότητας και τη βελτίωση της συνοχής των αποτελεσμάτων. Τέλος, τα **παραγόμενα αποτελέσματα εξάγονται σε δομημένες μορφές (Export)**, κατάλληλες τόσο για πειραματική αξιολόγηση όσο και για πρακτική αξιοποίηση, όπως αρχεία JSON, CSV και SRT.

Στο υψηλότερο επίπεδο, το σύστημα οργανώνεται γύρω από έναν κοινό πυρήνα επεξεργασίας που υλοποιείται στο module main.py για εκτέλεση μέσω γραμμής εντολών και επαναχρησιμοποιείται στο module app.py για διαδραστική χρήση μέσω περιβάλλοντος Streamlit. Και στις δύο περιπτώσεις ακολουθείται η ίδια ακολουθία επεξεργασίας, διασφαλίζοντας συνέπεια αποτελεσμάτων ανεξάρτητα από τον τρόπο κλήσης του συστήματος. Η κεντρική παραμετροποίηση υλοποιείται μέσω του module Config, το οποίο καθορίζει τις διαδρομές δεδομένων, την επιλογή μοντέλου, τις παραμέτρους ανίχνευσης, τα thresholds, καθώς και τις σημαίες για την επιλεκτική απενεργοποίηση συνιστωσών

(ablation studies). Με αυτόν τον τρόπο εξασφαλίζεται ότι όλες οι κρίσιμες ρυθμίσεις βρίσκονται σε ένα σημείο και υπόκεινται σε έλεγχο εγκυρότητας πριν από την εκτέλεση.

Ο σχεδιασμός του συστήματος βασίζεται σε τρεις θεμελιώδεις αρχές που διέπουν την όλη υλοποίηση. Η πρώτη αρχή είναι η **αρθρωτότητα (Modularity)**, όπου κάθε στάδιο της επεξεργασίας υλοποιείται ως ανεξάρτητη μονάδα με συγκεκριμένες ευθύνες. Αυτό επιτρέπει την εύκολη αντικατάσταση ή αναβάθμιση επιμέρους συστατικών χωρίς να επηρεάζεται η υπόλοιπη αρχιτεκτονική. Η δεύτερη αρχή είναι η **επεκτασιμότητα (Extensibility)**, η οποία υλοποιείται μέσω του factory design pattern που επιτρέπει την εύκολη ενσωμάτωση νέων μοντέλων λεζάντας ή εναλλακτικών αλγορίθμων επεξεργασίας. Η τρίτη αρχή είναι η **επαληθευσιμότητα (Verifiability)**, με ενσωματωμένη υποστήριξη για configuration-driven ablation studies που διευκολύνουν τη συστηματική αξιολόγηση της συνεισφοράς κάθε επιμέρους τεχνικής στην τελική απόδοση του συστήματος.

Συνοψίζοντας, το προτεινόμενο σύστημα λειτουργεί ως ένα ελεγχόμενο πειραματικό πλαίσιο για τη μελέτη του Dense Video Captioning με προ-εκπαιδευμένα μοντέλα. Η έμφαση στη modular σχεδίαση, στη δίκαιη συγκριτική αξιολόγηση και στην παραμετροποιήσιμη επεξεργασία καθιστά το σύστημα κατάλληλο τόσο για την ανάλυση των πλεονεκτημάτων και των περιορισμών διαφορετικών αρχιτεκτονικών όσο και για την εξαγωγή συμπερασμάτων που μπορούν να γενικευθούν σε πραγματικά σενάρια εφαρμογής.

4.2 Ανίχνευση Σκηνών και Χρονικός Κατακερματισμός video

Η ανίχνευση σκηνών αποτελεί το πρώτο κρίσιμο στάδιο της προτεινόμενης υλοποίησης Dense Video Captioning, καθώς καθορίζει τα χρονικά τμήματα πάνω στα οποία θα εξαχθούν frames και θα παραχθούν οι αντίστοιχες λεζάντες. Η επιλογή των χρονικών ορίων δεν επηρεάζει μόνο τον αριθμό των παραγόμενων περιγραφών, αλλά και τη σημασιολογική συνοχή κάθε τμήματος, άρα και την ικανότητα των μοντέλων να παράγουν συγκεκριμένες και ακριβείς λεζάντες. Για τον λόγο αυτό, ο χρονικός κατακερματισμός σχεδιάστηκε ως content-based διαδικασία και ενσωματώνει μηχανισμό προσαρμογής ευαισθησίας, ώστε να παραμένει σταθερά λειτουργικός σε διαφορετικού τύπου video.

4.2.1 Επιλογή Scene-Based Segmentation

Στην υλοποίηση αποφεύγεται ο κατακερματισμός σε σταθερά χρονικά παράθυρα, διότι τα γεγονότα σε πραγματικά video παρουσιάζουν μεταβλητή διάρκεια και δεν ευθυγραμμίζονται με αυθαίρετα χρονικά όρια. Η διαίρεση σε fixed windows μπορεί να οδηγήσει είτε σε τεμαχισμό ενός ενιαίου γεγονότος σε πολλαπλά τμήματα είτε σε συγχώνευση διαφορετικών γεγονότων στο ίδιο τμήμα. Και στις δύο περιπτώσεις, η οπτική είσοδος που θα χρησιμοποιηθεί στη συνέχεια γίνεται λιγότερο συνεκτική, αυξάνοντας την πιθανότητα γενικών ή ασύνδετων περιγραφών [4].

Αντίθετα, η scene-based προσέγγιση επιδιώκει να ορίσει χρονικά τμήματα που αντιστοιχούν σε διακριτές ενότητες περιεχομένου. Η πρακτική συνέπεια είναι ότι τα downstream στάδια, τόσο η εξαγωγή αντιπροσωπευτικών frames όσο και η παραγωγή λεζάντας, τροφοδοτούνται με δεδομένα που περιγράφουν πιο καθαρά ένα γεγονός, κάτι που είναι ιδιαίτερα σημαντικό για Dense Video Captioning [4].

4.2.2 Content-Based Ανίχνευση έναντι Shot Boundary Detection

Η ανίχνευση σκηνών βασίζεται σε μεταβολές του οπτικού περιεχομένου, και όχι αποκλειστικά σε shot boundaries που σχετίζονται με τεχνικές αλλαγές κάμερας. Ένα cut ή μια μεταβατική αλλαγή λήψης δεν συνεπάγεται απαραίτητα αλλαγή γεγονότος, ενώ αντίστροφα ένα γεγονός μπορεί να εκτείνεται σε περισσότερα από ένα shots. Για Dense

Video Captioning, ο στόχος είναι η περιγραφή γεγονότων και όχι η περιγραφή κινηματογραφικών μεταβάσεων [1].

Η content-based ανίχνευση είναι καταλληλότερη επειδή επιχειρεί να εντοπίσει ουσιαστικές μεταβολές στη σκηνή, δηλαδή αλλαγές που τείνουν να συνοδεύονται από διαφοροποίηση αντικειμένων, χώρου ή δραστηριότητας. Με αυτόν τον τρόπο, τα χρονικά τμήματα που προκύπτουν είναι πιο πιθανό να αντιστοιχούν σε εννοιολογικές ενότητες, κάτι που βελτιώνει τη συνολική συνέπεια του pipeline.

4.2.3 Υλοποίηση με PySceneDetect και ContentDetector

Για την ανίχνευση σκηνών χρησιμοποιείται η βιβλιοθήκη **PySceneDetect** και ειδικότερα ο αλγόριθμος **ContentDetector**. Η content-based ανίχνευση σκηνών βασίζεται στη μέθοδο **ContentDetector**, η οποία αναλύει τις χρωματικές και φωτομετρικές μεταβολές μεταξύ διαδοχικών πλαισίων του video. Σε αντίθεση με απλούστερες προσεγγίσεις που στηρίζονται αποκλειστικά στη μεταβολή της φωτεινότητας (luminance-based methods), η συγκεκριμένη μέθοδος αξιοποιεί το σύνολο της χρωματικής πληροφορίας κάθε πλαισίου. Πιο συγκεκριμένα, κάθε frame μετατρέπεται από τον χρωματικό χώρο RGB στον χώρο HSV (Hue, Saturation, Value), επιτρέποντας τον διαχωρισμό της πληροφορίας χρώματος από τη φωτεινότητα. Η επιλογή του χρωματικού χώρου HSV είναι ιδιαίτερα σημαντική, καθώς καθιστά τον αλγόριθμο πιο ανθεκτικό σε μεταβολές φωτισμού που δεν αντιστοιχούν σε πραγματική αλλαγή σκηνής. Μεταβολές στη φωτεινότητα, όπως σκιές ή παροδικές αλλαγές έκθεσης, μπορούν να επηρεάσουν έντονα luminance-based μετρικές, χωρίς όμως να σηματοδοτούν ουσιαστική αλλαγή περιεχομένου. Αντίθετα, η ανάλυση στον χώρο HSV επιτρέπει πιο αξιόπιστο εντοπισμό αλλαγών που σχετίζονται με διαφοροποίηση αντικειμένων, χώρου ή δραστηριότητας.

Για κάθε ζεύγος διαδοχικών πλαισίων f_i και f_{i+1}, το σύστημα υπολογίζει έναν δείκτη ομοιότητας S(i, i+1), ο οποίος βασίζεται στη μέση απόλυτη διαφορά των τιμών των τριών καναλιών του HSV χώρου. Πιο συγκεκριμένα, για κάθε κανάλι $c \in \{H, S, V\}$, υπολογίζεται η μέση τιμή της απόλυτης διαφοράς ανάμεσα σε όλα τα αντίστοιχα pixels των δύο πλαισίων. Το τελικό σκορ προκύπτει από τη σταθμισμένη μέση τιμή των τριών καναλιών, με μεγαλύτερο βάρος στο κανάλι της φωτεινότητας (Value) που είναι πιο ευαίσθητο σε δραματικές αλλαγές σκηνής όπως cuts και fades. Η σχέση που περιγράφει αυτόν τον υπολογισμό μπορεί να εκφραστεί ως:

$$S(i, i+1) = w_H \cdot \delta_H + w_S \cdot \delta_S + w_V \cdot \delta_V$$

όπου δH, δS, δV αντιστοιχούν στις μέσες απόλυτες διαφορές των καναλιών Hue, Saturation και Value αντίστοιχα, ενώ οι συντελεστές στάθμισης w_H, w_S, w_V ικανοποιούν τη σχέση $w_H + w_S + w_V = 1$. Μια αλλαγή σκηνής ανιχνεύεται όταν το S(i, i+1) υπερβαίνει ένα προκαθορισμένο κατώφλι θ, το οποίο ρυθμίζει την ευαισθησία της ανίχνευσης.

4.2.4 Ρόλος Παραμέτρων και Μηχανισμός Προσαρμοστικής Ευαισθησίας

Η συμπεριφορά της ανίχνευσης καθορίζεται κυρίως από δύο παραμέτρους. Η πρώτη είναι το **threshold**, το οποίο ελέγχει την ευαισθησία ανίχνευσης, δηλαδή το πόσο μεγάλη πρέπει να είναι η μεταβολή περιεχομένου ώστε να θεωρηθεί ότι ξεκινά νέα σκηνή. Η δεύτερη είναι το **min_scene_len**, το οποίο επιβάλλει ελάχιστο μήκος σκηνής και αποτρέπει τη δημιουργία πολύ μικρών τμημάτων που συνήθως είναι θόρυβος ή στιγμιαίες μεταβολές χωρίς σημασιολογική αξία [74].

Η παραμετροποίηση της ανίχνευσης σκηνών υλοποιείται μέσω ενός μηχανισμού δύο επιπέδων, ο οποίος επιτρέπει την προσαρμογή της ευαισθησίας ανάλογα με τα χαρακτηριστικά του video. Στο πρώτο επίπεδο, το οποίο χαρακτηρίζεται ως Standard Mode, χρησιμοποιούνται συντηρητικές παράμετροι που είναι κατάλληλες για την πλειονότητα των περιπτώσεων. Συγκεκριμένα, το content threshold ορίζεται στην τιμή $\theta_{std} = 27.0$, η

οποία έχει επιλεγεί εμπειρικά ώστε να ανιχνεύει μόνο σημαντικές αλλαγές περιεχομένου, αποφεύγοντας την υπερβολική κατάτμηση λόγω μικρών κινήσεων της κάμερας ή παροδικών αλλαγών φωτισμού. Παράλληλα, επιβάλλεται ελάχιστο μήκος σκηνής $\text{min\_scene\_length}_{\text{std}}$ = 25 frames, το οποίο λειτουργεί ως φίλτρο για την απόρριψη εξαιρετικά σύντομων τμημάτων που συχνά οφείλονται σε τεχνικά artifacts, όπως frame drops ή transition effects. Μετά την ολοκλήρωση της ανίχνευσης στο Standard Mode, το σύστημα αξιολογεί την ποιότητα της τμηματοποίησης με βάση τον αριθμό των ανιχνευθέντων σκηνών. Αν ο αριθμός αυτός είναι μικρότερος από δύο, υπάρχει ένδειξη ότι το video περιλαμβάνει λεπτές ή σταδιακές μεταβάσεις που δεν ανιχνεύθηκαν με τις συντηρητικές ρυθμίσεις.

Σε αυτήν την περίπτωση ενεργοποιείται το δεύτερο επίπεδο, το οποίο χαρακτηρίζεται ως Fallback ή Sensitive Mode. Στο επίπεδο αυτό, το content threshold μειώνεται στην τιμή θ_{fallback} = 17.0, επιτρέποντας την ανίχνευση αλλαγών με μικρότερη οπτική διαφορά. Ταυτόχρονα, το ελάχιστο μήκος σκηνής μειώνεται σε $\text{min\_scene\_length}_{\text{fallback}}$ = 15 frames, προσφέροντας μεγαλύτερη ευελιξία στην αποδοχή σύντομων αλλά δυνητικά σημαντικών τμημάτων. Η δυναμική αυτή προσαρμογή αυξάνει την ικανότητα ανίχνευσης λεπτών μεταβάσεων, βελτιώνοντας το recall σε απαιτητικά video, χωρίς να θυσιάζεται σημαντικά η ακρίβεια σε περιεχόμενο με απότομα cuts [74].

4.2.5 Επίδραση του Temporal Segmentation στην Περιγραφή

Η χρονική τμηματοποίηση (Temporal Segmentation) επηρεάζει άμεσα τη μορφή της οπτικής εισόδου που λαμβάνουν τα μοντέλα παραγωγής λεζάντας και, συνεπώς, την ποιότητα των παραγόμενων περιγραφών. Στην περίπτωση του **BLIP**, όπου η σκηνή αναπαρίσταται από περιορισμένο αριθμό frames, η υπερβολικά μεγάλη διάρκεια σκηνών αυξάνει την πιθανότητα να συμπεριληφθούν πολλαπλά γεγονότα στο ίδιο τμήμα, οδηγώντας σε γενικές λεζάντες. Αντίθετα, η υπερκατάτμηση μπορεί να παράγει πολλές, παρόμοιες σκηνές και κατ' επέκταση πλεονάζουσες περιγραφές.

Για τα **GIT-VATEX** και **Qwen2-VL**, τα οποία αξιοποιούν πολλαπλά frames ανά σκηνή, τα χρονικά όρια καθορίζουν το εύρος μέσα στο οποίο γίνεται το sampling frame. Όταν οι σκηνές είναι συνεκτικές, το σύνολο των frames τείνει να αποτυπώνει με σαφήνεια την ίδια δραστηριότητα, διευκολύνοντας τη δημιουργία περιγραφών με μεγαλύτερη πληρότητα και χρονική συνέπεια. Αντίθετα, όταν μια σκηνή είναι υπερβολικά εκτεταμένη ή περιλαμβάνει ετερογενές περιεχόμενο, η είσοδος γίνεται πιο "θορυβώδης" από σημασιολογικής πλευράς, δυσκολεύοντας τη συμπύκνωση σε μία ενιαία λεζάντα.

Συνολικά, η επιλογή content-based scene detection και η χρήση προσαρμοστικής ευαισθησίας μέσω εναλλακτικών ρυθμίσεων threshold και ελάχιστου μήκους σκηνής συνιστούν κρίσιμες σχεδιαστικές αποφάσεις της υλοποίησης. Οι αποφάσεις αυτές στοχεύουν στη δημιουργία χρονικών τμημάτων με επαρκή διάρκεια και υψηλή συνοχή, ώστε να υποστηρίζεται αποτελεσματικά η παραγωγή περιγραφών τόσο από frame-based όσο και από video-based και multimodal μοντέλα.

4.3 Εξαγωγή και Επιλογή Keyframes

Η εξαγωγή και επιλογή keyframes αποτελεί το επόμενο κρίσιμο στάδιο της προτεινόμενης υλοποίησης, καθώς λειτουργεί ως ο σύνδεσμος μεταξύ του χρονικού κατακερματισμού και της παραγωγής λεζάντων. Τα χρονικά όρια που προκύπτουν από την ανίχνευση σκηνών καθορίζουν το εύρος μέσα στο οποίο πραγματοποιείται η δειγματοληψία οπτικής πληροφορίας, ωστόσο η ποιότητα και ο τρόπος επιλογής των frame επηρεάζουν άμεσα την αποτελεσματικότητα των μοντέλων παραγωγής λεζάντας.

Το προτεινόμενο σύστημα υλοποιεί διαφορετικές στρατηγικές εξαγωγής ανάλογα με τον τύπο του μοντέλου λεζάντας που χρησιμοποιείται, αναγνωρίζοντας ότι frame-based και video-based μοντέλα έχουν διαφορετικές απαιτήσεις εισόδου και δυνατότητες επεξεργασίας. Η διαφοροποίηση αυτή είναι αναγκαία, καθώς τα μοντέλα που αξιολογούνται εμφανίζουν διαφορετικές απαιτήσεις ως προς την ποσότητα και τη φύση της οπτικής εισόδου, ιδιαίτερα ως προς τον τρόπο ενσωμάτωσης της χρονικής πληροφορίας.

4.3.1 Διαφοροποίηση Στρατηγικών Ανάλογα με το Μοντέλο

Στην υλοποίηση διακρίνονται δύο βασικές κατηγορίες στρατηγικών εξαγωγής frames. Η πρώτη αφορά frame-based μοντέλα, όπου η σκηνή αναπαρίσταται μέσω ενός περιορισμένου αριθμού στατικών εικόνων. Η δεύτερη αφορά video-based και multimodal μοντέλα, τα οποία αξιοποιούν πολλαπλά frames για την ενσωμάτωση χρονικής δυναμικής.

Αφού καθοριστούν τα χρονικά όρια κάθε σκηνής, το σύστημα προχωρά στην εξαγωγή των κατάλληλων οπτικών εισόδων. Η διαδικασία αυτή διαφοροποιείται ανάλογα με το μοντέλο και υλοποιείται μέσω εξειδικευμένων αρθρωμάτων (modules) που λειτουργούν ως "Engines". Αυτά τα modules ακολουθούν μια ενιαία προγραμματιστική διεπαφή (interface) που ορίστηκε για τις ανάγκες του συστήματος, εκθέτοντας κοινές μεθόδους όπως η extract_frames_batch για την προετοιμασία των frames και η generate_caption για την παραγωγή του κειμένου.

Για frame-based μοντέλα όπως το **BLIP**, η επιλογή ενός μοναδικού και αντιπροσωπευτικού frame ανά σκηνή είναι κρίσιμη, καθώς η λεζάντα θα βασιστεί αποκλειστικά στην πληροφορία που περιέχεται σε αυτό το στιγμιότυπο. Αντίθετα, για μοντέλα όπως τα **GIT-VATEX** και **Qwen2-VL**, η σκηνή αναπαρίσταται μέσω ενός συνόλου frame κατανεμημένων στο χρονικό της εύρος, επιτρέποντας στο μοντέλο να αντλήσει πληροφορία για την εξέλιξη της δραστηριότητας μέσα στη σκηνή. Η διάκριση αυτή ενσωματώνεται ρητά στο σύστημα και διασφαλίζει ότι κάθε μοντέλο τροφοδοτείται με οπτική είσοδο που είναι συμβατή με τον τρόπο λειτουργίας του.

4.3.2 Αλγόριθμος Ποιοτικής Επιλογής Keyframe για Frame-Based Μοντέλα

Για τα frame-based μοντέλα, και ειδικότερα για το BLIP που χρησιμοποιείται στην παρούσα υλοποίηση, η επιλογή ενός αντιπροσωπευτικού frame ανά σκηνή αποτελεί κρίσιμο παράγοντα για την ποιότητα της παραγόμενης λεζάντας. Δεδομένου ότι η περιγραφή βασίζεται αποκλειστικά σε στατική οπτική πληροφορία, η επιλογή ενός μη αντιπροσωπευτικού ή χαμηλής ποιότητας πλαισίου μπορεί να οδηγήσει σε ασαφείς ή λανθασμένες περιγραφές, ανεξάρτητα από τη γενική ικανότητα του μοντέλου.

Για τον λόγο αυτό, αναπτύχθηκε ένας αλγόριθμος έξυπνης επιλογής keyframe, ο οποίος αξιολογεί κάθε υποψήφιο πλαίσιο με βάση δύο συμπληρωματικές μετρικές ποιότητας. Η **πρώτη μετρική** αφορά την **ευκρίνεια του frame (sharpness)**, ενώ η **δεύτερη μετρική** αφορά την **πληροφοριακή εντροπία (information entropy)**. Ο συνδυασμός των δύο αυτών κριτηρίων επιτρέπει την επιλογή πλαισίων που είναι ταυτόχρονα τεχνικά άρτια και πληροφοριακά πλούσια.

Η **ευκρίνεια** υπολογίζεται μέσω της εφαρμογής του τελεστή Laplacian στο grayscale μετασχηματισμό του πλαισίου. Ο Laplacian είναι ένας δεύτερης τάξης διαφορικός τελεστής που εντοπίζει απότομες μεταβολές έντασης, οι οποίες αντιστοιχούν σε άκρες και υφές της εικόνας. Ένα καλά εστιασμένο πλαίσιο παρουσιάζει έντονη παρουσία τέτοιων μεταβολών, ενώ ένα πλαίσιο με motion blur ή εκτός εστίασης εμφανίζει σημαντικά χαμηλότερη απόκριση. Μαθηματικά, για ένα πλαίσιο f(x, y) ο Laplacian ορίζεται ως:

$$\nabla^2 f(x,y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Στην πράξη, εφαρμόζεται μια διακριτή προσέγγιση συνέλιξης με χρήση πυρήνα (kernel) 3×3, αξιοποιώντας την αντίστοιχη υλοποίηση της βιβλιοθήκης OpenCV. Η τελική μετρική ευκρίνειας ορίζεται ως η διακύμανση (variance) των τιμών του μετασχηματισμού Laplace για το σύνολο των εικονοστοιχείων του πλαισίου:

$$Sharpness(f) = Var(L) = \frac{1}{N}\sum_{i=1}^{N}(L_i - \mu)^2$$

όπου L_i είναι οι τιμές του Laplacian και μ η μέση τιμή τους. Υψηλότερες τιμές διασποράς υποδηλώνουν καλύτερη εστίαση και μικρότερη παρουσία blur.

Η **δεύτερη μετρική** αφορά την **πληροφοριακή εντροπία Shannon**, η οποία ποσοτικοποιεί τον πλούτο και την πολυπλοκότητα του οπτικού περιεχομένου. Η εντροπία υπολογίζεται από το κανονικοποιημένο ιστόγραμμα του grayscale πλαισίου και εκφράζεται ως:

$$Entropy(f) = -\sum_{i}^{p}(i)\log_2 p(i)$$

όπου p(i) είναι η πιθανότητα εμφάνισης της τιμής έντασης i. Πλαίσια με χαμηλή εντροπία αντιστοιχούν συνήθως σε ομοιόμορφα ή φτωχά οπτικά περιβάλλοντα, ενώ υψηλότερες τιμές εντροπίας συνδέονται με πλουσιότερο περιεχόμενο, περισσότερα αντικείμενα και λεπτομέρειες.

Στο προτεινόμενο σύστημα, η διαδικασία αυτή δεν εφαρμόζεται εξαντλητικά σε όλα τα frames, αλλά ακολουθεί μια **στρατηγική δειγματοληψίας** που ελέγχεται από την παράμετρο διαμόρφωσης NUM_SAMPLES. Συγκεκριμένα, για κάθε ανιχνευμένη σκηνή, το σύστημα απομονώνει ένα σύνολο υποψήφιων πλαισίων σε ομοιόμορφα κατανεμημένες χρονικές θέσεις. Η τιμή της παραμέτρου ορίστηκε σε NUM_SAMPLES = 5, μια επιλογή που προέκυψε ως η βέλτιστη ισορροπία (trade-off) μεταξύ της υπολογιστικής απόδοσης και της αντιπροσωπευτικότητας του δείγματος. Για κάθε υποψήφιο πλαίσιο υπολογίζονται οι τιμές ευκρίνειας και εντροπίας. Πριν από τη συγχώνευσή τους, εφαρμόζεται μια διαδικασία κανονικοποίησης εντός της σκηνής (intra-scene normalization), διαιρώντας κάθε τιμή με τη μέγιστη που παρατηρείται στο τρέχον σύνολο δειγμάτων. Αυτό το βήμα διασφαλίζει ότι η αξιολόγηση είναι σχετική με τα διαθέσιμα δεδομένα της συγκεκριμένης λήψης και όχι απόλυτη.

Το τελικό σκορ κατάταξης προκύπτει από το σταθμισμένο άθροισμα των κανονικοποιημένων μετρικών, σύμφωνα με τον τύπο:

$$Score(f) = w_{sharp} \cdot Sharpness_{norm}(f) + w_{entropy} \cdot Entropy_{norm}(f)$$

όπου τα βάρη w_{sharp} και $w_{entropy}$ αποτελούν κρίσιμες παραμέτρους του συστήματος. Βάσει εκτενών πειραματικών δοκιμών σε ποικίλα datasets, οι συντελεστές καθορίστηκαν σε w_{sharp} **= 0.70** και $w_{entropy}$ = **0.30** . Η βαρύτητα αυτή (70%) προς την ευκρίνεια αντικατοπτρίζει την εμπειρική παρατήρηση ότι τα Vision-Language μοντέλα είναι ιδιαίτερα ευαίσθητα σε θολά (blurred) στιγμιότυπα, τα οποία οδηγούν συχνά σε hallucinations. Παράλληλα, η διατήρηση της εντροπίας (30%) λειτουργεί ως δικλείδα ασφαλείας, αποτρέποντας την επιλογή τεχνικά άρτιων αλλά πληροφοριακά κενών εικόνων. Το frame με το υψηλότερο συνολικό σκορ

επιλέγεται ως keyframe της σκηνής και χρησιμοποιείται ως οπτική είσοδος για το frame-based μοντέλο.

4.3.3 Uniform Temporal Sampling για Video-Aware Μοντέλα

Σε αντίθεση με τις frame-based προσεγγίσεις που περιορίζονται στην ανάλυση μεμονωμένων στιγμιότυπων, τα **video-aware** μοντέλα (όπως τα **GIT-VATEX** και **Qwen2-VL**) έχουν σχεδιαστεί για την επεξεργασία ακολουθιών πολλαπλών frames, ενσωματώνοντας ουσιαστικά τη χρονική διάσταση. Αυτό καθιστά εφικτή την κατανόηση της εξέλιξης μιας δράσης, της κίνησης, καθώς και των αιτιακών σχέσεων μεταξύ των γεγονότων, στοιχεία κρίσιμα για την περιγραφή περίπλοκων σκηνών με χρονική εξάρτηση (π.χ. αθλητικά στιγμιότυπα).

Για την τροφοδοσία των μοντέλων αυτών, αναπτύχθηκε και ενσωματώθηκε στον πυρήνα του συστήματος η στρατηγική της **Ομοιόμορφης Χρονικής Δειγματοληψίας (Uniform Temporal Sampling)**. Η υλοποίηση αυτή εξάγει πλαίσια σε ισαπέχοντα χρονικά διαστήματα, διασφαλίζοντας προγραμματιστικά την αντιπροσωπευτική κάλυψη όλης της διάρκειας της σκηνής και εξαλείφοντας τη μεροληψία επιλογής (selection bias) προς συγκεκριμένα τμήματα.

Η λογική παραμετροποίησης σχεδιάστηκε ώστε να προσαρμόζεται δυναμικά στις ιδιαιτερότητες κάθε αρχιτεκτονικής. Συγκεκριμένα, για το **GIT-VATEX** διατηρήθηκε η επιλογή των 6 frames, ακολουθώντας τις προδιαγραφές του fine-tuning στο dataset VATEX. Αντιθέτως, για το **Qwen2-VL**, καθορίστηκε ρητά μέσω κώδικα το όριο των 5 frames, μια απόφαση που προέκυψε ως αναγκαίος περιορισμός για τη βέλτιστη διαχείριση της μνήμης GPU (VRAM) και την αποφυγή υπερφόρτωσης του συστήματος.

Τέλος, στον αλγόριθμο υπολογισμού των χρονικών σημείων προστέθηκε ένας μηχανισμός ασφαλείας που εισάγει περιθώρια (margins) στα άκρα. Η λειτουργία αυτή αποτρέπει τη λήψη frames από τα όρια της σκηνής, καθώς έχει παρατηρηθεί ότι τα σημεία αυτά περιέχουν συχνά τεχνουργήματα μετάβασης (transition artifacts) που υποβαθμίζουν την ποιότητα της περιγραφής.

4.4 Παραγωγή Λεζάντων με Προ-Εκπαιδευμένα Μοντέλα

Το στάδιο παραγωγής λεζάντων υλοποιείται ως modular υποσύστημα που υποστηρίζει πολλαπλές αρχιτεκτονικές όρασης–γλώσσας και επιτρέπει τη συγκριτική τους αξιολόγηση υπό κοινές συνθήκες. Το σύστημα ενσωματώνει τρία διαφορετικά μοντέλα, ένα **frame-based**, ένα **video-based** και ένα **γενικού σκοπού multimodal** μοντέλο, τα οποία επιλέγονται δυναμικά μέσω ενός **μηχανισμού factory (model_factory.py).** Με τον τρόπο αυτό, καθίσταται δυνατή η εναλλαγή μοντέλων χωρίς μεταβολή της υπόλοιπης ροής επεξεργασίας, διασφαλίζοντας ενιαία είσοδο, συγκρίσιμες εξόδους και ελεγχόμενο πειραματικό περιβάλλον.

Και στις τρεις περιπτώσεις, τα μοντέλα χρησιμοποιούνται σε zero-shot καθεστώς, χωρίς fine-tuning, έτσι ώστε οι παραγόμενες λεζάντες να αντικατοπτρίζουν τις εγγενείς δυνατότητες των αρχιτεκτονικών και όχι προσαρμογή σε συγκεκριμένο dataset. Οι υλοποιήσεις βασίζονται στη βιβλιοθήκη HuggingFace Transformers, ενώ οι παράμετροι παραγωγής ελέγχονται κεντρικά μέσω αρχείου ρυθμίσεων, επιτρέποντας συστηματική ανάλυση της συμπεριφοράς κάθε μοντέλου.

4.4.1 Δυναμική Επιλογή Μοντέλου και Ενοποιημένη Ροή Εκτέλεσης

Η επιλογή του μοντέλου πραγματοποιείται δυναμικά μέσω του model_factory, ο οποίος επιστρέφει την κατάλληλη μηχανή παραγωγής λεζάντας με βάση τη ρύθμιση του συστήματος. Κάθε μηχανή υλοποιεί κοινή λογική εισόδου και εξόδου, αλλά διαφοροποιείται ως προς τον τρόπο προετοιμασίας των οπτικών δεδομένων και τον μηχανισμό παραγωγής κειμένου.

Η σχεδίαση αυτή επιτρέπει όχι μόνο τη συγκριτική αξιολόγηση της ποιότητας των λεζάντων, αλλά και την αποτίμηση του υπολογιστικού κόστους και του χρόνου εκτέλεσης κάθε προσέγγισης. Παράλληλα, διευκολύνει τη μελλοντική επέκταση του συστήματος με νέα μοντέλα, χωρίς να απαιτούνται αλλαγές στα υπόλοιπα στάδια του pipeline.

4.4.2 Frame-Based Παραγωγή Λεζάντας

Για την frame-based προσέγγιση, το σύστημα αξιοποιεί τη βασική έκδοση του **BLIP** ως image captioning model. Κάθε σκηνή αναπαρίσταται από ένα μοναδικό keyframe, το οποίο έχει επιλεγεί μέσω του αλγορίθμου ποιοτικής αξιολόγησης που παρουσιάστηκε στο Κεφάλαιο 4.4.2. Η παραγωγή λεζάντας υλοποιείται μέσω της κλάσης **CaptionEngine**, η οποία αναλαμβάνει την προετοιμασία της εισόδου και την κλήση του μοντέλου.

Η εικόνα προεπεξεργάζεται μέσω του **BlipProcessor**, ενώ η διαδικασία text generation πραγματοποιείται με την κλάση **BlipForConditionalGeneration** της βιβλιοθήκης HuggingFace Transformers. Αντί για **greedy decoding**, εφαρμόζεται **beam search decoding με beam width ίσο με 5**, επιτρέποντας την εξερεύνηση πολλαπλών υποψήφιων ακολουθιών κατά τη γλωσσική παραγωγή και οδηγώντας σε πιο φυσικές και ποιοτικές περιγραφές.

Επιπλέον, ορίζεται **minimum generation length ίσο με 10 tokens**, ώστε να αποφεύγονται υπερβολικά σύντομες και μη πληροφοριακές λεζάντες, καθώς και **maximum generation length ίσο με 60 tokens**, περιορίζοντας την παραγωγή υπερβολικά μακροσκελών κειμένων. Τέλος, εφαρμόζεται **repetition penalty ίσο με 1.2**, το οποίο μειώνει την επανάληψη λέξεων και φράσεων, συμβάλλοντας στη δημιουργία πιο ποικίλων και φυσικών περιγραφών. Οι επιλογές αυτές στοχεύουν στην ισορροπία μεταξύ περιγραφικής πληρότητας και αναγνωσιμότητας.

4.4.3 Video-Based Παραγωγή Λεζάντας

Το GIT-VATEX αξιοποιεί πολλαπλά frames ανά σκηνή και επιχειρεί να ενσωματώσει βασική temporal information στη διαδικασία caption generation. Η υλοποίηση πραγματοποιείται μέσω της κλάσης **VideoCaptionEngine**, η οποία κατασκευάζει τα κατάλληλα input tensors από τα frames που έχουν εξαχθεί μέσω uniform temporal sampling.

Η προετοιμασία των δεδομένων γίνεται μέσω του **AutoProcessor**, ενώ η παραγωγή κειμένου βασίζεται στο **AutoModelForCausalLM**, αξιοποιώντας τη γενική διεπαφή της βιβλιοθήκης Transformers. Όπως και στην περίπτωση του BLIP, η διαδικασία generation ελέγχεται μέσω παραμέτρων **minimum και maximum generation length 10 και 60 αντίστοιχα**, εξασφαλίζοντας συγκρίσιμη έξοδο μεταξύ διαφορετικών μοντέλων.

Η αναμενόμενη συμπεριφορά του GIT-VATEX είναι η παραγωγή λεζάντων που αποδίδουν καλύτερα actions και temporal transitions, καθώς το μοντέλο δεν περιορίζεται σε ένα static snapshot αλλά επεξεργάζεται μια σύντομη sequence frames.

4.4.4 Multimodal Παραγωγή Λεζάντας με Vision–Language Models

Το Qwen2-VL ενσωματώνεται στο σύστημα ως γενικού σκοπού Vision-Language Model, το οποίο υποστηρίζει multimodal input και text generation μέσω δομής τύπου chat. Η υλοποίηση πραγματοποιείται μέσω της κλάσης **QwenVideoEngine**, η οποία διαμορφώνει ρητά ένα **multimodal prompt** που περιλαμβάνει τόσο την οπτική όσο και τη γλωσσική συνιστώσα.

Η οπτική είσοδος παρέχεται ως video component, το οποίο περιλαμβάνει τα επιλεγμένα frames και τον ρυθμό δειγματοληψίας, ενώ η γλωσσική καθοδήγηση του μοντέλου πραγματοποιείται μέσω του ακόλουθου prompt:

"Describe the scene in a single sentence. Focus on the main subject, the specific action taking place, and the visible environment. Mention visual details like colors or objects explicitly."

Το prompt αυτό έχει σχεδιαστεί ώστε να κατευθύνει το μοντέλο στην παραγωγή σύντομων αλλά περιγραφικών λεζάντων, δίνοντας έμφαση στο κύριο υποκείμενο, τη δράση και το περιβάλλον της σκηνής, ενώ παράλληλα ενθαρρύνει την αναφορά συγκεκριμένων οπτικών λεπτομερειών.

Για λόγους αποδοτικότητας, το μοντέλο φορτώνεται με **4-bit quantization μέσω του BitsAndBytesConfig**, επιτρέποντας την εκτέλεση σε GPU περιορισμένης μνήμης χωρίς σημαντική απώλεια ποιότητας. Παράμετροι όπως ο αριθμός των **new tokens**, η **temperature** και το **top-p sampling** ορίζονται κεντρικά στο αρχείο ρυθμίσεων, επιτρέποντας έλεγχο της στοχαστικότητας και της ποικιλίας των παραγόμενων λεζάντων.

4.5 Semantic Merging και Post-Processing

Η πυκνή χρονική τμηματοποίηση που εφαρμόζεται στο πλαίσιο του Dense Video Captioning επιτρέπει την ακριβή ανίχνευση και απομόνωση οπτικών γεγονότων, ωστόσο οδηγεί συχνά στην παραγωγή διαδοχικών σκηνών με υψηλό βαθμό σημασιολογικής επικάλυψης. Στην πράξη, συνεχόμενες σκηνές μπορεί να αντιστοιχούν στην ίδια δραστηριότητα και να διαφέρουν μόνο στη γλωσσική διατύπωση της περιγραφής τους, ιδιαίτερα όταν χρησιμοποιούνται σύγχρονα Vision-Language Models με έντονη παραφραστική ικανότητα.

Για την αντιμετώπιση αυτού του φαινομένου, το προτεινόμενο σύστημα ενσωματώνει ένα στάδιο σημασιολογικής συγχώνευσης και μετα-επεξεργασίας, το οποίο έχει σχεδιαστεί ως οργανικό μέρος του συνολικού pipeline. Στόχος του σταδίου αυτού δεν είναι απλώς η μείωση του αριθμού των παραγόμενων λεζάντων, αλλά η ενοποίηση διαδοχικών σκηνών όταν αυτές αντιστοιχούν στην ίδια νοηματική ενότητα, με τρόπο που διατηρεί τη χρονική ακρίβεια και βελτιώνει τη συνοχή της τελικής περιγραφής του video.

Η λειτουργικότητα αυτή υλοποιείται από την κλάση **SceneMerger**, η οποία εφαρμόζει σημασιολογική σύγκριση λεζάντων, διαχείριση προσωρινών ομάδων συγχώνευσης και επιλογή αντιπροσωπευτικής περιγραφής για κάθε συγχωνευμένη σκηνή.

4.5.1 Σημασιολογική Σύγκριση Λεζάντων και Μηχανισμός Συγχώνευσης

Με την ολοκλήρωση της παραγωγής των περιγραφών, το σύστημα προχωρά στο στάδιο της **Συγχώνευσης Σκηνών (Scene Merging)**. Η διαδικασία αυτή σχεδιάστηκε ώστε να βασίζεται στη σημασιολογική σύγκριση των λεζάντων και όχι σε απλή λεκτική ομοιότητα. Για την επίτευξη αυτού του στόχου, η κλάση **SceneMerger** ενσωματώνει το προ-εκπαιδευμένο μοντέλο **Sentence Transformer all-MiniLM-L6-v2**, το οποίο φορτώνεται κατά την αρχικοποίηση. Μέσω αυτού, κάθε λεζάντα μετατρέπεται σε **διανυσματική αναπαράσταση (embedding)**, επιτρέποντας τη χαρτογράφηση όλων των περιγραφών σε έναν κοινό σημασιολογικό χώρο.

Η ροή εκτέλεσης ακολουθεί αυστηρή χρονική σειρά, υπολογίζοντας την **ομοιότητα συνημιτόνου (cosine similarity) μεταξύ των embeddings διαδοχικών σκηνών**. Όταν η τιμή της ομοιότητας υπερβαίνει το καθορισμένο κατώφλι **MERGE_THRESHOLD**, οι σκηνές χαρακτηρίζονται ως σημασιολογικά συνεκτικές. Σε αυτό το σημείο, αντί για την άμεση οριστικοποίηση της νέας σκηνής, ο αλγόριθμος ενεργοποιεί έναν προσωρινό buffer συγχώνευσης, ο οποίος συγκεντρώνει δυναμικά τις διαδοχικές σκηνές που ανήκουν στην ίδια νοηματική ενότητα.

Ο buffer εμπλουτίζεται όσο οι επόμενες σκηνές διατηρούν υψηλή σημασιολογική ομοιότητα. Μόλις η ομοιότητα υποχωρήσει κάτω από το όριο του threshold, ο buffer θεωρείται ολοκληρωμένος και το σύστημα προχωρά στη δημιουργία μιας ενιαίας συγχωνευμένης σκηνής. Η προσέγγιση αυτή επιτρέπει γραμμική επεξεργασία των σκηνών, αποφεύγοντας την πολυπλοκότητα της σύγκρισης όλων των λεζάντων μεταξύ τους, διατηρώντας χαμηλό υπολογιστικό κόστος και αυστηρή χρονική συνέπεια.

4.5.2 Τελική Αναπαράσταση Γεγονότων

Το πιο κρίσιμο στάδιο της σημασιολογικής συγχώνευσης αφορά την επιλογή της λεζάντας που θα εκπροσωπεί τη συγχωνευμένη σκηνή. Η επιλογή αυτή δεν μπορεί να βασιστεί σε απλοϊκά κριτήρια, όπως η χρονικά πρώτη ή τελευταία περιγραφή, καθώς τέτοιες στρατηγικές συχνά οδηγούν σε λεζάντες που δεν αποτυπώνουν επαρκώς το κοινό νόημα της ομάδας σκηνών.

Για τον λόγο αυτό, το προτεινόμενο σύστημα εφαρμόζει συνδυαστική αξιολόγηση των υποψήφιων λεζάντων του buffer. Αρχικά, για κάθε λεζάντα υπολογίζεται ένας **δείκτης αντιπροσωπευτικότητας (centroid score)**, ο οποίος εκφράζει τον μέσο βαθμό σημασιολογικής ομοιότητας της συγκεκριμένης λεζάντας προς όλες τις υπόλοιπες λεζάντες της ίδιας ομάδας. Ο δείκτης αυτός αποτυπώνει το κατά πόσο μια λεζάντα βρίσκεται κοντά στο «νοηματικό κέντρο» της ομάδας και, συνεπώς, περιγράφει με τον πιο συνοπτικό και ακριβή τρόπο το κοινό περιεχόμενο των συγχωνευόμενων σκηνών.

Παράλληλα, όταν υπάρχει προηγούμενη σκηνή, υπολογίζεται και ένας **δείκτης συνέχειας συμφραζομένου (context score)**, ο οποίος μετρά τη σημασιολογική εγγύτητα της υποψήφιας λεζάντας με την αμέσως προηγούμενη τελική περιγραφή. Το κριτήριο αυτό επιτρέπει στο σύστημα να λαμβάνει υπόψη τη ροή της αφήγησης σε επίπεδο video και να αποφεύγει απότομες σημασιολογικές μεταβάσεις.

Η τελική κατάταξη προκύπτει από τον γραμμικό συνδυασμό των δύο δεικτών. Στο πλαίσιο της παρούσας υλοποίησης και κατόπιν πειραματικών δοκιμών, οι συντελεστές βαρύτητας καθορίστηκαν σε 0.70 για το **centroid score** και 0.30 για το **context score**. Η συγκεκριμένη ρύθμιση (70/30) επιλέχθηκε διότι διασφαλίζει ότι η λεζάντα θα είναι πρωτίστως πιστή στο περιεχόμενο της τρέχουσας ομάδας, διατηρώντας ωστόσο μια σημαντική εξάρτηση από τα συμφραζόμενα για ομαλή ροή. Η λεζάντα που μεγιστοποιεί αυτή τη σταθμισμένη συνάρτηση υιοθετείται ως η τελική περιγραφή, ενώ τα χρονικά όρια του buffer ενοποιούνται σε ένα ενιαίο διάστημα.

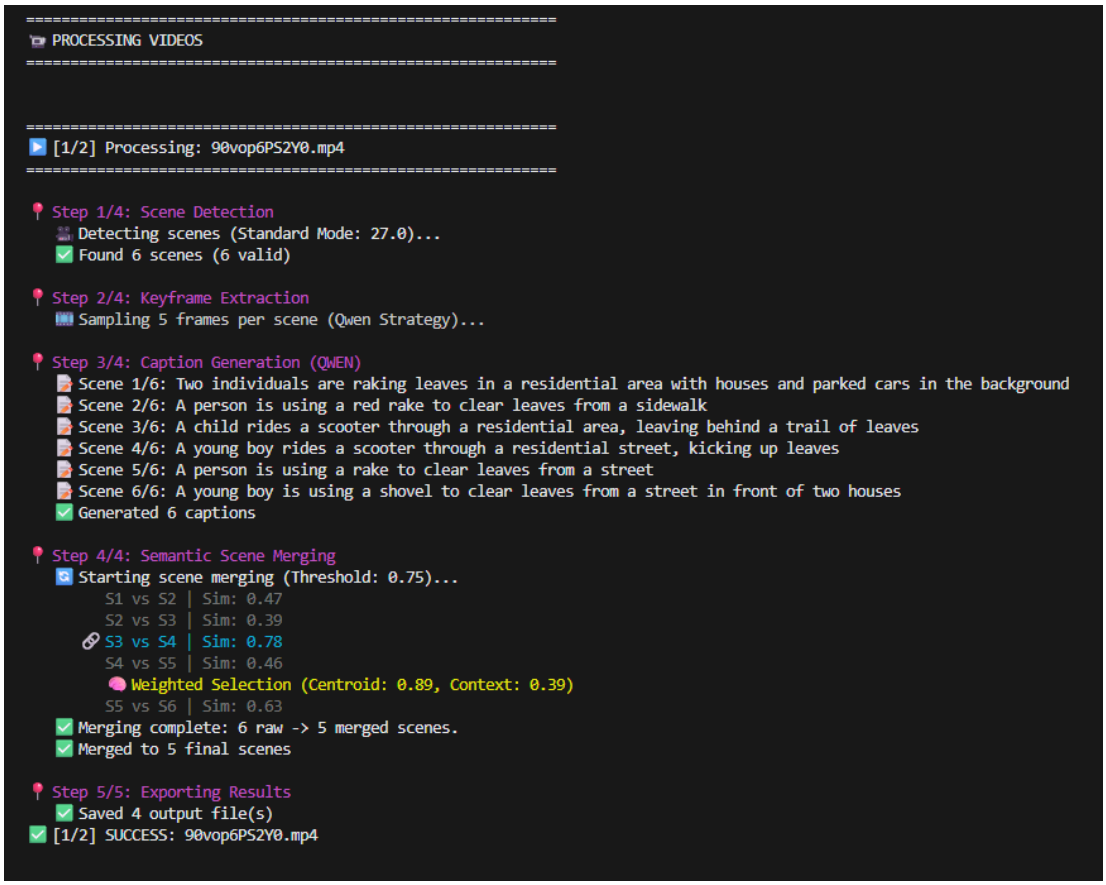
Με τον τρόπο αυτό, το σύστημα μεταβαίνει από μια λεπτομερή αλλά αποσπασματική αναπαράσταση του video σε μια πιο συμπαγή και αφηγηματικά συνεκτική περιγραφή γεγονότων. Η σημασιολογική συγχώνευση λειτουργεί ως ουσιαστικό στάδιο εξευγενισμού των αποτελεσμάτων και προετοιμάζει το έδαφος για την πειραματική αποτίμηση που ακολουθεί, όπου η ποιότητα, η συνοχή και η πληρότητα των τελικών λεζάντων μπορούν να αξιολογηθούν ποσοτικά και ποιοτικά.

4.6 Εξαγωγή Αποτελεσμάτων και Διεπαφή Χρήστη

Το σύστημα Dense Video Captioning υποστηρίζει εκτέλεση μέσω γραμμής εντολών για αυτοματοποιημένες διαδικασίες, καθώς και διαδικτυακή διεπαφή για διαδραστική χρήση. Τα αποτελέσματα εξάγονται σε πολλαπλές μορφές για την κάλυψη διαφορετικών περιπτώσεων χρήσης και την ενσωμάτωση με άλλα συστήματα.

4.6.1 Εκτέλεση Pipeline

Η εκτέλεση του συστήματος μέσω της γραμμής εντολών παρέχει λεπτομερή πληροφόρηση για κάθε στάδιο επεξεργασίας σε πραγματικό χρόνο. Η **Εικόνα 4.1** παρουσιάζει την εκτέλεση του πλήρους pipeline για ένα δείγμα video, όπου εμφανίζονται διαδοχικά τα πέντε στάδια επεξεργασίας που έχουν περιγραφεί στις προηγούμενες ενότητες.



Εικόνα 4.1: *Εκτέλεση του pipeline επεξεργασίας από τη γραμμή εντολών με λεπτομερή αναφορά για κάθε στάδιο και τις παραγόμενες σκηνές με τις αντίστοιχες περιγραφές.*

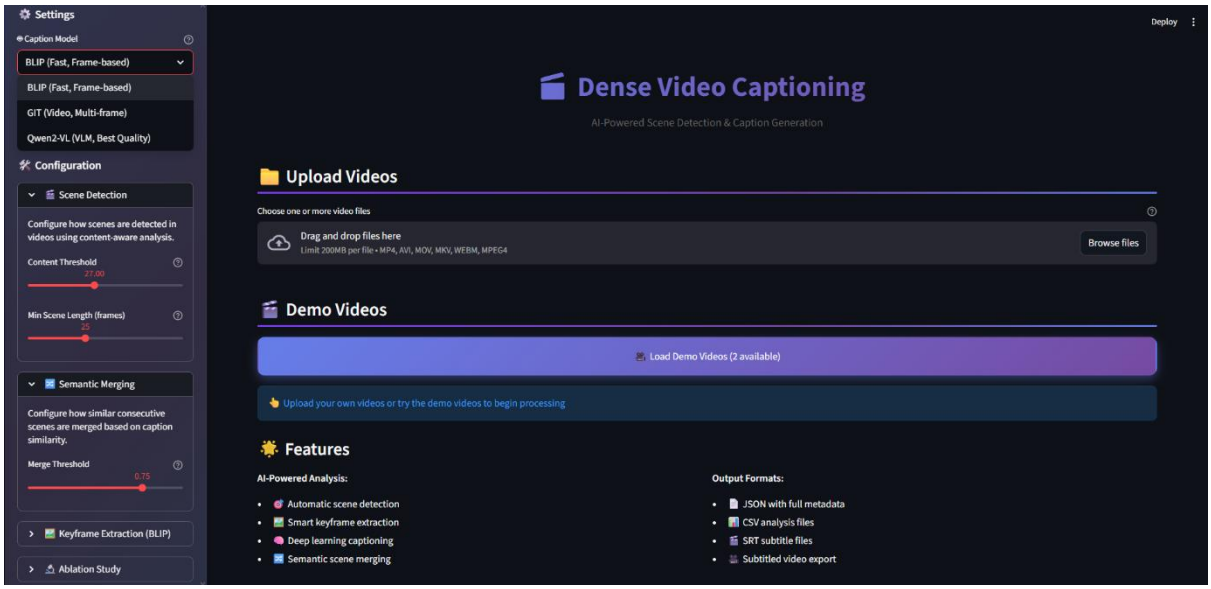
Η εικόνα αποτυπώνει τη διακριτή οπτικοποίηση των σταδίων επεξεργασίας, όπου η ροή εργασιών (pipeline) αναλύεται βήμα προς βήμα. Μέσα από τη δομημένη καταγραφή (structured logging), παρέχεται πλήρης διαφάνεια στην εσωτερική λογική του συστήματος, επιτρέποντας την άμεση εποπτεία των ενδιάμεσων υπολογισμών (όπως τα scores ομοιότητας) και των κριτηρίων συγχώνευσης πριν την τελική εξαγωγή.

Η εκτέλεση μέσω CLI επιτρέπει την ενσωμάτωση του συστήματος σε αυτοματοποιημένες διαδικασίες και την επεξεργασία μεγάλων datasets. Η structured έξοδος μπορεί να καταγραφεί σε log files για μεταγενέστερη ανάλυση και monitoring της απόδοσης του συστήματος.

4.6.2 Διαδικτυακή Διεπαφή (Streamlit)

Για την παροχή φιλικής προς τον χρήστη διεπαφής αναπτύχθηκε διαδικτυακή εφαρμογή με το framework Streamlit. Η διεπαφή καθιστά δυνατή τη φόρτωση video, την παραμετροποίηση του συστήματος, και την παρουσίαση των αποτελεσμάτων χωρίς να απαιτείται γνώση προγραμματισμού.

Στην **Εικόνα 4.2** παρουσιάζεται η αρχική οθόνη του συστήματος, η οποία είναι σχεδιασμένη με τρόπο που εξασφαλίζει τη λειτουργικότητα και την εύκολη χρήση. Η διεπαφή οργανώνει τις διαθέσιμες επιλογές σε δύο διακριτά τμήματα για την καλύτερη διαχείριση της διαδικασίας.



Εικόνα 4.2: *Η διαδικτυακή διεπαφή του συστήματος με το sidebar παραμέτρων και την περιοχή φόρτωσης video.*

Στην αριστερή πλευρά, η πλευρική στήλη (sidebar) περιλαμβάνει το σύνολο των ρυθμίσεων. Μέσω αυτού του πίνακα, ο χρήστης ορίζει το μοντέλο ανάλυσης και προσαρμόζει τις παραμέτρους για την ανίχνευση και τη συγχώνευση των σκηνών, ελέγχοντας τη λειτουργία του αλγορίθμου.

Το κεντρικό τμήμα της οθόνης αφορά την εισαγωγή δεδομένων. Στο επάνω μέρος βρίσκεται η περιοχή φόρτωσης αρχείων, η οποία υποστηρίζει τη μέθοδο drag-and-drop, ενώ ακριβώς από κάτω διατίθεται η ενότητα "Demo Videos". Η συγκεκριμένη επιλογή παρέχει πρόσβαση σε έτοιμα παραδείγματα video, επιτρέποντας την άμεση εκτέλεση και τον έλεγχο του συστήματος χωρίς να απαιτείται η μεταφόρτωση νέου υλικού.

Μετά την ολοκλήρωση της επεξεργασίας, το σύστημα παρουσιάζει τα αποτελέσματα σε οργανωμένη μορφή. Η **Εικόνα 4.3** εμφανίζει τη σελίδα αποτελεσμάτων με το video player, τα στατιστικά επεξεργασίας, τις αναλυτικές περιγραφές σκηνών, και τις επιλογές λήψης των αρχείων εξόδου.

Εικόνα 4.3: *Παρουσίαση αποτελεσμάτων με video player, στατιστικά επεξεργασίας, αναλυτικές περιγραφές σκηνών και κουμπιά λήψης αρχείων.*

Η σελίδα αποτελεσμάτων περιλαμβάνει video player με ενσωματωμένους υποτίτλους, panel με στατιστικά μετρικά της επεξεργασίας, λίστα των σκηνών με τα χρονικά τους όρια και τις αντίστοιχες περιγραφές, και κουμπιά λήψης για τα αρχεία JSON, SRT και το video με burned subtitles. Η διαδραστική παρουσίαση επιτρέπει την άμεση επαλήθευση της ποιότητας των παραγόμενων περιγραφών σε σχέση με το οπτικό περιεχόμενο.

4.6.3 Μορφές Εξόδου

Το σύστημα εξάγει τα παραγόμενα αποτελέσματα σε τέσσερις διαφορετικές μορφές, με στόχο την κάλυψη πολλαπλών σεναρίων χρήσης και την ευκολότερη αξιοποίησή τους σε διαφορετικά περιβάλλοντα. Η κύρια μορφή εξόδου είναι ένα δομημένο αρχείο JSON, το οποίο ακολουθεί προκαθορισμένο schema, διασφαλίζοντας συνέπεια στην αναπαράσταση των δεδομένων και διευκολύνοντας τη διαλειτουργικότητα με άλλα εργαλεία και εφαρμογές. Επιπλέον, παράγονται αρχεία υποτίτλων σε μορφή SRT, ένα βίντεο με ενσωματωμένες (burned-in) λεζάντες μέσω της βιβλιοθήκης FFmpeg, καθώς και αρχείο CSV για συνοπτική παρουσίαση και ανάλυση των αποτελεσμάτων σε tabular μορφή.

Οι **Εικόνες 4.4 και 4.5** παρουσιάζουν ενδεικτικά τη δομή του παραγόμενου JSON output. Το πρώτο τμήμα περιλαμβάνει μεταδεδομένα της επεξεργασίας, καθώς και τις παραμέτρους εκτέλεσης του pipeline, οι οποίες τεκμηριώνουν τις συνθήκες υπό τις οποίες παρήχθησαν τα αποτελέσματα και υποστηρίζουν την αναπαραγωγιμότητά τους. Το δεύτερο τμήμα περιέχει τη λίστα των ανιχνευμένων σκηνών, με τα αντίστοιχα χρονικά όρια και τις παραγόμενες περιγραφές σε φυσική γλώσσα.

```
"CAPTION_MODEL_TYPE": "qwen",
"DEVICE": "cuda",
"BLIP_MODEL_NAME": "Salesforce/blip-image-captioning-base",
"GIT_MODEL_NAME": "microsoft/git-large-vatex",
"QWEN_MODEL_NAME": "Qwen/Qwen2-VL-2B-Instruct",
"BLIP_NUM_BEAMS": 5,
"BLIP_MIN_LENGTH": 10,
"BLIP_MAX_LENGTH": 60,
"GIT_MIN_LENGTH": 10,
"GIT_MAX_LENGTH": 60,
"QWEN_MAX_LENGTH": 200,
"QWEN_DO_SAMPLE": true,
"QWEN_TEMPERATURE": 0.2,
"QWEN_TOP_P": 0.9,
"SIMILARITY_MODEL": "all-MiniLM-L6-v2",
"MERGE_THRESHOLD": 0.75,
"CONTENT_THRESHOLD": 27.0,
"MIN_SCENE_LEN": 25,
"CONTENT_THRESHOLD_FALLBACK": 17.0,
"MIN_SCENE_LEN_FALLBACK": 8,
"NUM_SAMPLES": 5,
"WEIGHT_SHARP": 0.7,
"WEIGHT_ENTROPY": 0.30000000000000004,
"EVAL_IOU_THRESHOLD": 0.3,
"ENABLE_SCENE_MERGING": true,
"ENABLE_KEYFRAME_SELECTION": true,
```


***Εικόνα 4.4**: Τμήμα metadata και παραμέτρων του JSON output.*


```
"scenes": [
    {
        "scene_id": 1,
        "start": 0.0,
        "end": 21.0,
        "duration": 21.0,
        "caption": "Two individuals are raking leaves in a residential area with houses and parked cars in the background",
        "meta": {}
    },
    {
        "scene_id": 2,
        "start": 21.0,
        "end": 24.933333333333334,
        "duration": 3.93,
        "caption": "A person is using a red rake to clear leaves from a sidewalk",
        "meta": {}
    },
    {
        "scene_id": 3,
        "start": 24.933333333333334,
        "end": 168.93333333333334,
        "duration": 144.0,
        "caption": "A child rides a scooter through a residential area, leaving behind a trail of leaves",
        "meta": {}
    },
    {
        "scene_id": 5,
        "start": 168.93333333333334,
        "end": 191.33333333333334,
        "duration": 22.4,
        "caption": "A person is using a rake to clear leaves from a street",
        "meta": {}
    },
    {
        "scene_id": 6,
        "start": 191.33333333333334,
        "end": 271.23333333333335,
        "duration": 79.9,
        "caption": "A young boy is using a shovel to clear leaves from a street in front of two houses",
        "meta": {}
    }
]
}
```


***Εικόνα 4.5:** Structured array σκηνών με χρονικά όρια, διάρκεια και περιγραφές σε φυσική γλώσσα.*

Το JSON αρχείο περιλαμβάνει schema version, timestamp επεξεργασίας, video identifier, επιλεγμένο μοντέλο και mode, καθώς και configuration object με όλες τις παραμέτρους που χρησιμοποιήθηκαν. Το scenes array περιέχει για κάθε σκηνή τον αύξοντα αριθμό, τα χρονικά όρια σε δευτερόλεπτα, τη διάρκεια, και την περιγραφή που παράγεται από το μοντέλο.

Εκτός από το JSON, το σύστημα παράγει τρεις επιπλέον μορφές εξόδου. **Η Εικόνα 4.6** παρουσιάζει το SRT αρχείο υποτίτλων που ακολουθεί το standard SubRip format.

```
1
00:00:00,000 --> 00:00:21,000
Two individuals are raking leaves in a residential area with houses and parked cars in the background

2
00:00:21,000 --> 00:00:24,933
A person is using a red rake to clear leaves from a sidewalk

3
00:00:24,933 --> 00:02:48,933
A child rides a scooter through a residential area, leaving behind a trail of leaves

5
00:02:48,933 --> 00:03:11,333
A person is using a rake to clear leaves from a street

6
00:03:11,333 --> 00:04:31,233
A young boy is using a shovel to clear leaves from a street in front of two houses
```

Εικόνα 4.6: *Αρχείο υποτίτλων σε μορφή SRT με timecodes και περιγραφές σκηνών.*

Το SRT format περιλαμβάνει για κάθε σκηνή έναν αύξοντα αριθμό, το χρονικό εύρος σε μορφή timecode, και το κείμενο της περιγραφής. Αυτή η μορφή υποστηρίζεται από όλα τα video players και επιτρέπει την εύκολη ενσωμάτωση των υποτίτλων κατά την αναπαραγωγή. Επιπλέον, το σύστημα παράγει CSV αρχείο σε tabular μορφή για ανάλυση σε spreadsheet applications.

Προαιρετικά, το σύστημα μπορεί να παράγει video με burned subtitles μέσω FFmpeg. Η **Εικόνα 4.7** παρουσιάζει δείγμα frame από το subtitled video όπου οι υπότιτλοι είναι μόνιμα ενσωματωμένοι στο video stream.

Εικόνα 4.7: *Frame από video με burned subtitles.*

Οι burned subtitles είναι ιδανικοί για sharing σε platforms που δεν υποστηρίζουν external subtitles ή για περιπτώσεις όπου επιθυμείται η μόνιμη εμφάνιση των περιγραφών. Η παραγωγή του subtitled video μπορεί να απενεργοποιηθεί για εξοικονόμηση χρόνου και αποθηκευτικού χώρου όταν δεν είναι απαραίτητη.

ΚΕΦΑΛΑΙΟ 5 Πειραματική αποτίμηση

5.1 Πειραματικό Πλαίσιο και Μεθοδολογία Αξιολόγησης

Η αξιολόγηση του προτεινόμενου συστήματος Dense Video Captioning διεξήχθη με στόχο την ολοκληρωμένη μέτρηση της απόδοσής του, τόσο ως προς την **ακρίβεια του χρονικού εντοπισμού των γεγονότων**, όσο και ως προς την **ποιότητα των παραγόμενων γλωσσικών περιγραφών**. Το πειραματικό πλαίσιο σχεδιάστηκε με τρόπο που να επιτρέπει τη συστηματική σύγκριση των τριών υποστηριζόμενων μοντέλων παραγωγής λεζάντας, συγκεκριμένα των **BLIP**, **GIT-VATEX** και **Qwen2-VL**, καθώς και την **αποτίμηση της συνεισφοράς των επιμέρους υποσυστημάτων**. Για την επίτευξη αυτού του στόχου, υιοθετήθηκε μια διμερής στρατηγική αξιολόγησης που περιλαμβάνει, αφενός, την **end-to-end αξιολόγηση του πλήρους συστήματος** και, αφετέρου, την **αξιολόγηση υπό συνθήκες oracle temporal segmentation**. Στην περίπτωση της oracle temporal segmentation αξιολόγησης χρησιμοποιούνται τα χρονικά όρια των ground truth timestamps, επιτρέποντας τη μέτρηση της ικανότητας παραγωγής λεζάντων ανεξάρτητα από τα σφάλματα της διαδικασίας χρονικού εντοπισμού (temporal localization).

Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι το **ActivityNet Captions**, ένα από τα πλέον καθιερωμένα και αναγνωρισμένα benchmarks στον τομέα του Dense Video Captioning. Το εν λόγω dataset περιλαμβάνει video που καλύπτουν ευρύ φάσμα δραστηριοτήτων, από αθλητικά γεγονότα και οικιακές εργασίες έως κοινωνικές αλληλεπιδράσεις και επαγγελματικές δεξιότητες. Κάθε video συνοδεύεται από πολλαπλές, χρονικά προσδιορισμένες περιγραφές που έχουν σχολιαστεί από ανθρώπινους αξιολογητές, καθορίζοντας ακριβώς την αρχή και το τέλος κάθε σημαντικού γεγονότος καθώς και μια φυσική γλωσσική περιγραφή του περιεχομένου. Για τις ανάγκες της παρούσας μελέτης, χρησιμοποιήθηκε ένα υποσύνολο 100 video από το validation set. Ο περιορισμός του δείγματος κρίθηκε αναγκαίος λόγω της μη διαθεσιμότητας σημαντικού μέρους του αρχικού υλικού (broken URLs/διαγραμμένα video), αλλά και λόγω των πεπερασμένων υπολογιστικών πόρων που καθιστούσαν δύσκολη την επεξεργασία του πλήρους όγκου δεδομένων σε εύλογο χρονικό διάστημα. Παρά ταύτα, η επιλογή του validation set διασφαλίζει την εγκυρότητα της αξιολόγησης, ενώ το συγκεκριμένο υποσύνολο προσφέρει επαρκή ποικιλία και πολυπλοκότητα για την αξιόπιστη εξαγωγή συμπερασμάτων.

Η μεθοδολογία αξιολόγησης βασίζεται σε ένα σύνολο ποσοτικών μετρικών που αξιολογούν διαφορετικές διαστάσεις της απόδοσης του συστήματος. Για την αξιολόγηση της χρονικής ακρίβειας του εντοπισμού γεγονότων, χρησιμοποιούνται οι μετρικές **Precision** και **Recall** που υπολογίζονται με βάση το **Intersection over Union** των προβλεπόμενων χρονικών διαστημάτων με τα ground truth annotations. Η μετρική **Intersection over Union**, συχνά αναφερόμενη ως IoU, ορίζεται ως ο λόγος της χρονικής τομής δύο διαστημάτων προς την ένωσή τους. Ένα προβλεπόμενο γεγονός θεωρείται επιτυχώς εντοπισμένο (True Positive) όταν το IoU του με κάποιο ground truth γεγονός υπερβαίνει ένα καθορισμένο κατώφλι, το οποίο στην παρούσα αξιολόγηση ορίστηκε τυπικά στο 30% (tIoU > 0.3). Η **Precision** μετρά το ποσοστό των προβλέψεων του συστήματος που αντιστοιχούν σωστά σε πραγματικά γεγονότα, ενώ το **Recall** μετρά το ποσοστό των πραγματικών γεγονότων που εντοπίστηκαν επιτυχώς από το σύστημα. Επιπλέον, για την εκτίμηση της ευαισθησίας του συστήματος σε διαφορετικά επίπεδα χρονικής ακρίβειας, υπολογίζονται **καμπύλες χρονικού εντοπισμού (temporal localization curves)** που παρουσιάζουν την Precision και το Recall για διαφορετικές τιμές του IoU threshold, από ελαστικότερα κριτήρια όπως 0.1 έως αυστηρότερα κριτήρια όπως 0.9.

Όσον αφορά την ποιότητα των παραγόμενων γλωσσικών περιγραφών, χρησιμοποιείται ένα σύνολο καθιερωμένων μετρικών από τον τομέα της επεξεργασίας φυσικής γλώσσας και της μηχανικής μετάφρασης. Η μετρική **BLEU**, η οποία είναι ευρέως διαδεδομένη στην αξιολόγηση συστημάτων παραγωγής κειμένου, μετρά την ομοιότητα μεταξύ του

παραγόμενου κειμένου και των reference περιγραφών βασιζόμενη στην επικάλυψη n-grams. Στην παρούσα αξιολόγηση υπολογίζονται τόσο το **BLEU-3** όσο και το **BLEU-4**, τα οποία αντιστοιχούν σε τριάδες και τετράδες διαδοχικών λέξεων αντίστοιχα. Το **BLEU-4** θεωρείται ιδιαίτερα απαιτητικό καθώς απαιτεί ακριβή αντιστοίχιση τεσσάρων διαδοχικών λέξεων, ενώ το **BLEU-3** είναι ελαφρώς πιο επιεικές. Η μετρική **METEOR** παρέχει μια πιο εξελιγμένη αξιολόγηση λαμβάνοντας υπόψη όχι μόνο την ακριβή αντιστοίχιση λέξεων αλλά και συνώνυμα, παραφράσεις και γραμματικές παραλλαγές, καθιστώντας την πιο ανθεκτική σε διαφορετικούς τρόπους έκφρασης της ίδιας σημασίας. Τέλος, η μετρική **ROUGE-L** εστιάζει στη μακρύτερη κοινή υπακολουθία (Longest Common Subsequence) μεταξύ του παραγόμενου και του reference κειμένου, αξιολογώντας τη δομική ομοιότητα και τη συνοχή των περιγραφών.

Ένα κρίσιμο τεχνικό ζήτημα στην αξιολόγηση είναι η **αντιστοίχιση των προβλεπόμενων γεγονότων με τα ground truth annotations**, ειδικά όταν το πλήθος τους διαφέρει. Για τη βέλτιστη επίλυση αυτού του προβλήματος ανάθεσης, υιοθετήθηκε ο **αλγόριθμος Hungarian Matching**. Ο αλγόριθμος αυτός κατασκευάζει έναν πίνακα κόστους βασισμένο στο αρνητικό IoU και εντοπίζει τη διμερή αντιστοίχιση που μεγιστοποιεί τη συνολική επικάλυψη, διασφαλίζοντας ότι κάθε πρόβλεψη αντιστοιχίζεται το πολύ σε ένα πραγματικό γεγονός. Προβλέψεις που δεν αντιστοιχίζονται ή έχουν IoU κάτω από το κατώφλι χαρακτηρίζονται ως False Positives, ενώ τα μη αντιστοιχισμένα ground truth γεγονότα ως False Negatives.

Η πειραματική διαδικασία εκτελέστηκε σε **υπολογιστικό σύστημα με κάρτα γραφικών NVIDIA GeForce GTX 1650 (4 GB VRAM)**. Το **περιβάλλον υλοποίησης βασίστηκε στο PyTorch 2.6 με υποστήριξη CUDA 12.4**, εξασφαλίζοντας την αποδοτική επιτάχυνση των υπολογιστικά απαιτητικών σταδίων του pipeline, όπως η εξαγωγή οπτικών αναπαραστάσεων και η διαδικασία caption generation. Η χρήση GPU ήταν ιδιαίτερα κρίσιμη για το μοντέλο Qwen2-VL, το οποίο, παρά την εφαρμογή 4-bit quantization για τη μείωση του μνημονικού αποτυπώματος, διατηρεί υψηλές υπολογιστικές απαιτήσεις λόγω του μεγέθους και της πολυτροπικής αρχιτεκτονικής του. Για τη διασφάλιση της αναπαραγωγιμότητας των αποτελεσμάτων, όλες οι στοχαστικές διαδικασίες ελέγχθηκαν μέσω καθορισμένων random seeds, ενώ οι παράμετροι εκτέλεσης και τα μεταδεδομένα κάθε πειράματος καταγράφηκαν συστηματικά.

5.2 Στρατηγικές Αξιολόγησης

Η αξιολόγηση ενός συστήματος Dense Video Captioning αποτελεί μια σύνθετη διαδικασία, καθώς η συνολική απόδοση είναι συνάρτηση της συνεργασίας δύο διακριτών υποσυστημάτων: του μηχανισμού χρονικού εντοπισμού γεγονότων (temporal event localization) και του μοντέλου παραγωγής περιγραφών (caption generation). Προκειμένου να απομονωθούν οι πηγές σφάλματος και να κατανοηθεί σε βάθος η συμπεριφορά των εξεταζόμενων αρχιτεκτονικών (BLIP, GIT-VATEX, Qwen), η πειραματική διαδικασία διαρθρώνεται σε δύο αλληλοσυμπληρούμενους άξονες. Η πρώτη προσέγγιση **End-to-End Evaluation αφορά την ολιστική αξιολόγηση του πλήρους συστήματος**, ελέγχοντας την ικανότητα του συστήματος να εντοπίζει και να περιγράφει γεγονότα αυτόνομα. Η δεύτερη προσέγγιση **oracle temporal segmentation αποσυνδέει τη διαδικασία εντοπισμού**, **εστιάζοντας αποκλειστικά στη γλωσσική ικανότητα των μοντέλων υπό ιδανικές συνθήκες χρονισμού**. Στις ενότητες που ακολουθούν αναλύονται οι τεχνικές λεπτομέρειες και η σκοπιμότητα της κάθε στρατηγικής.

5.2.1 End-to-End Αξιολόγηση Πλήρους Pipeline

Η end-to-end αξιολόγηση αντιπροσωπεύει το πιο ρεαλιστικό σενάριο χρήσης του συστήματος, όπου ένα ακατέργαστο video τροφοδοτείται ως είσοδος και το σύστημα πρέπει να παράγει αυτόνομα τόσο τα χρονικά όρια των γεγονότων όσο και τις αντίστοιχες γλωσσικές περιγραφές. Σε αυτή την προσέγγιση, το σύστημα εκτελεί διαδοχικά όλα τα στάδια του pipeline που περιγράφηκαν στο Κεφάλαιο 4, ξεκινώντας από την ανίχνευση σκηνών μέσω content-aware detection, συνεχίζοντας με την εξαγωγή αντιπροσωπευτικών keyframes, προχωρώντας στην παραγωγή λεζάντων από τα επιλεγμένα πλαίσια και ολοκληρώνοντας με τη σημασιολογική συγχώνευση διαδοχικών σκηνών με παρόμοιο σημασιολογικό περιεχόμενο.

Η αξιολόγηση της end-to-end απόδοσης πραγματοποιείται **συγκρίνοντας τα παραγόμενα χρονικά τμήματα και τις αντίστοιχες περιγραφές με τα ground truth annotations του ActivityNet Captions dataset**. Για κάθε video στο σύνολο αξιολόγησης, το σύστημα παράγει ένα σύνολο προβλέψεων που αποτελείται από ζεύγη χρονικών διαστημάτων και περιγραφών. Αυτές οι προβλέψεις αντιστοιχίζονται στα ground truth γεγονότα χρησιμοποιώντας τον **αλγόριθμο Hungarian matching** που περιγράφηκε προηγουμένως, με κριτήριο αποδοχής το **Intersection over Union threshold με τιμή 0.3**. Μια πρόβλεψη θεωρείται επιτυχημένη μόνο όταν ικανοποιούνται συγχρόνως δύο προϋποθέσεις. Πρώτον, το προβλεπόμενο χρονικό διάστημα πρέπει να έχει επαρκή χρονική επικάλυψη με κάποιο ground truth γεγονός, όπως μετριέται από το IoU. Δεύτερον, η παραγόμενη λεζάντα αξιολογείται ως προς την ποιότητά της χρησιμοποιώντας τις μετρικές BLEU, METEOR και ROUGE-L σε σύγκριση με τη reference περιγραφή του αντίστοιχου ground truth γεγονότος.

Αυτή η προσέγγιση αξιολόγησης παρέχει μια ρεαλιστική εικόνα της πρακτικής χρησιμότητας του συστήματος, καθώς οποιοδήποτε σφάλμα σε οποιοδήποτε στάδιο του pipeline επηρεάζει το τελικό αποτέλεσμα. Για παράδειγμα, ένα εξαιρετικό μοντέλο παραγωγής λεζάντων δεν μπορεί να αποδώσει καλά αν το στάδιο της scene detection αποτύχει να εντοπίσει σωστά τα όρια των γεγονότων ή αν η επιλογή keyframes δεν εξάγει αντιπροσωπευτικά πλαίσια. Αντίστοιχα, τέλεια χρονική τμηματοποίηση δεν οδηγεί σε υψηλές συνολικές επιδόσεις αν οι παραγόμενες περιγραφές είναι ανακριβείς ή γενικευμένες. Επομένως, η end-to-end αξιολόγηση μετρά την ολιστική απόδοση του συστήματος ως ενιαίου μηχανισμού και αποκαλύπτει πώς τα επιμέρους συστατικά συνεργάζονται για την επίτευξη του τελικού στόχου.

5.2.2 Αξιολόγηση υπό συνθήκες oracle temporal segmentation

Για την απομόνωση και την αποκλειστική μέτρηση της ικανότητας των μοντέλων να παράγουν ποιοτικές γλωσσικές περιγραφές, εφαρμόζεται μια δεύτερη στρατηγική αξιολόγησης που παρακάμπτει πλήρως τα στάδια του temporal localization και χρησιμοποιεί τα ground truth χρονικά όρια που παρέχονται από το ActivityNet Captions dataset. Σε αυτή την προσέγγιση, τα μοντέλα caption generation εφαρμόζονται απευθείας στα χρονικά τμήματα που έχουν σχολιαστεί από ανθρώπινους αξιολογητές, εξαλείφοντας έτσι την επίδραση τυχόν σφαλμάτων που προκύπτουν από ανακριβή ανίχνευση σκηνών ή υποβέλτιστη επιλογή keyframes.

Η διαδικασία αυτής της αξιολόγησης ξεκινά με την ανάγνωση των ground truth annotations για κάθε video, τα οποία καθορίζουν ακριβώς τα χρονικά σημεία έναρξης και λήξης κάθε σημαντικού γεγονότος. Για κάθε τέτοιο χρονικό τμήμα, το σύστημα εξάγει τα κατάλληλα frames από το video χρησιμοποιώντας την ίδια στρατηγική που εφαρμόζεται και στην end-to-end αξιολόγηση. Για τα frame-based μοντέλα όπως το BLIP, εφαρμόζεται ο αλγόριθμος smart keyframe selection που επιλέγει το βέλτιστο πλαίσιο με βάση τη σταθμισμένη αξιολόγηση sharpness και entropy. Για τα video-aware μοντέλα όπως το GIT-VATEX και το Qwen2-VL, εξάγονται πολλαπλά πλαίσια σε ομοιόμορφα κατανεμημένα

χρονικά σημεία εντός του τμήματος. Τα εξαγόμενα πλαίσια τροφοδοτούνται στο αντίστοιχο μοντέλο παραγωγής λεζάντων, το οποίο παράγει μια φυσική γλωσσική περιγραφή για το γεγονός. Η παραγόμενη περιγραφή συγκρίνεται στη συνέχεια με την reference περιγραφή του ground truth χρησιμοποιώντας τις καθιερωμένες μετρικές BLEU, METEOR και ROUGE-L.

Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι ότι **παρέχει μια καθαρή μέτρηση της ποιότητας των μοντέλων παραγωγής λεζάντων ανεξάρτητα από την απόδοση των υπόλοιπων συστατικών του συστήματος.** Με αυτόν τον τρόπο, μπορεί να προσδιοριστεί το ανώτερο όριο απόδοσης που μπορεί να επιτευχθεί αν η scene detection και η keyframe extraction ήταν τέλειες. Η διαφορά μεταξύ των αποτελεσμάτων της oracle αξιολόγησης και της end-to-end αξιολόγησης αποκαλύπτει την επίδραση που έχουν τα σφάλματα temporal localization στη συνολική απόδοση του συστήματος. Επιπλέον, η oracle αξιολόγηση επιτρέπει την αντικειμενική σύγκριση των διαφορετικών μοντέλων caption generation σε ίσους όρους, καθώς όλα λαμβάνουν ακριβώς την ίδια χρονική πληροφορία ως είσοδο. Αυτό είναι ιδιαίτερα σημαντικό για την αξιολόγηση του trade-off μεταξύ της πολυπλοκότητας των μοντέλων και της ποιότητας των αποτελεσμάτων τους.

5.3 Αποτελέσματα Συνολικής Απόδοσης

Σε αυτή την ενότητα παρουσιάζονται συγκεντρωτικά τα ποσοτικά ευρήματα της πειραματικής διαδικασίας για τα τρία εξεταζόμενα μοντέλα BLIP, GIT-VATEX, Qwen2-VL, καλύπτοντας τόσο το σενάριο End-to-End (E2E) όσο και το σενάριο Oracle βάσει ground truth χρονικών διαστημάτων. Η συνολική αποτίμηση εξετάζει τρεις συμπληρωματικές διαστάσεις, **την ποιότητα των παραγόμενων λεζάντων**, **την ακρίβεια του χρονικού εντοπισμού στο E2E σενάριο**, **καθώς και την χρονική αποδοτικότητα.**

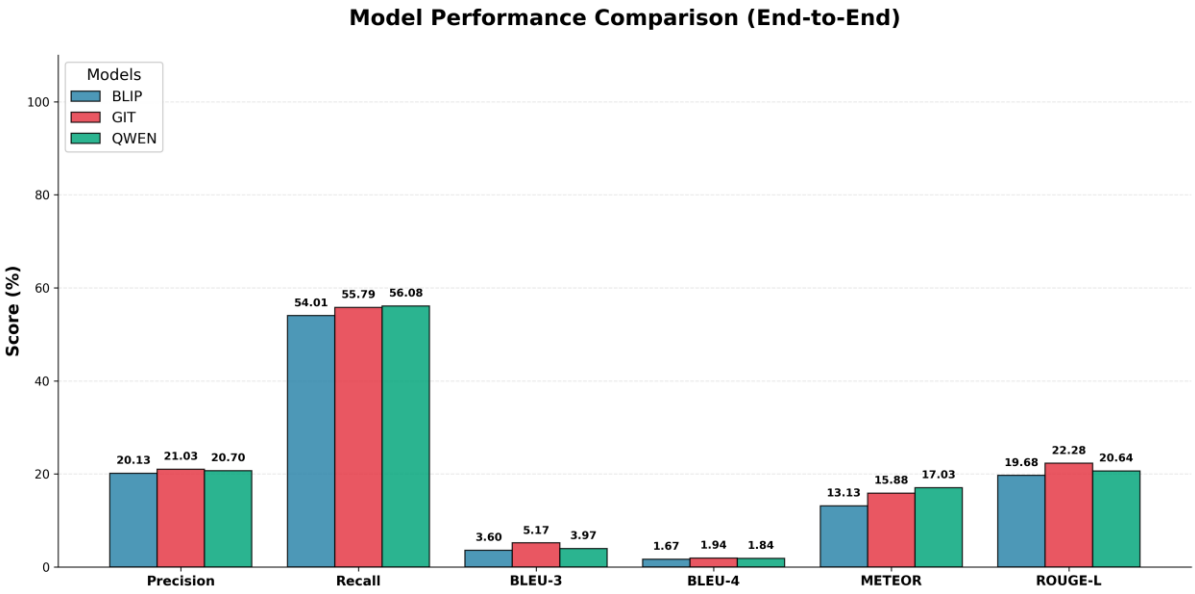
Ωστόσο, **η ερμηνεία των δεικτών ποιότητας κειμένου απαιτεί ιδιαίτερη προσοχή**. Οι συμβατικές μετρικές, όπως το BLEU και το ROUGE, βασίζονται πρωτίστως στη λεξιλογική επικάλυψη (lexical overlap) με τα ground truths, γεγονός που ενδέχεται να **υποτιμά την επίδοση μοντέλων που παράγουν πιο ελεύθερες ή σημασιολογικά πλούσιες περιγραφές**. Το φαινόμενο αυτό καθίσταται εντονότερο στα prompt-driven μοντέλα, όπως το Qwen2-VL, όπου η δομή της εντολής εισόδου (prompt) δύναται να διαφοροποιήσει αισθητά το ύφος, την έκταση και το επίπεδο λεπτομέρειας της λεζάντας. Αυτό οδηγεί συχνά σε περιγραφές που, αν και ποιοτικά ορθές, αποκλίνουν λεξιλογικά από τα ground truths. Υπό αυτό το πρίσμα, τα ποσοτικά δεδομένα πλαισιώνονται και ολοκληρώνονται μέσω της ποιοτικής ανάλυσης που παρατίθεται στην Ενότητα 5.7.

5.3.1 Σύγκριση Μοντέλων σε End-to-End Αξιολόγηση

Τα αποτελέσματα της end-to-end αξιολόγησης παρουσιάζονται στον **Πίνακα 5.1** και οπτικοποιούνται στο **Σχήμα 5.1**. Η αξιολόγηση διεξήχθη σε υποσύνολο 100 video από το validation set του ActivityNet Captions dataset, εφαρμόζοντας το πλήρες pipeline επεξεργασίας που περιλαμβάνει scene detection, keyframe extraction, caption generation και semantic merging. Οι μετρικές temporal localization υπολογίζονται **με IoU threshold 0.3**, το οποίο αποτελεί το τυποποιημένο κριτήριο στη βιβλιογραφία.

Πίνακας 5.1: Συγκριτικά Αποτελέσματα End-to-End Αξιολόγησης.

Μοντέλο	Precision	Recall	BLEU-3	BLEU-4	METEOR	ROUGE-L
BLIP	20.13	54.01	3.60	1.67	13.13	19.68
GIT-VATEX	**21.03**	55.79	**5.17**	**1.94**	15.88	**22.28**
Qwen2-VL	20.70	**56.08**	3.97	1.84	**17.03**	20.64



Σχήμα 5.1: *Συγκριτική Απόδοση Μοντέλων σε End-to-End Αξιολόγηση.*

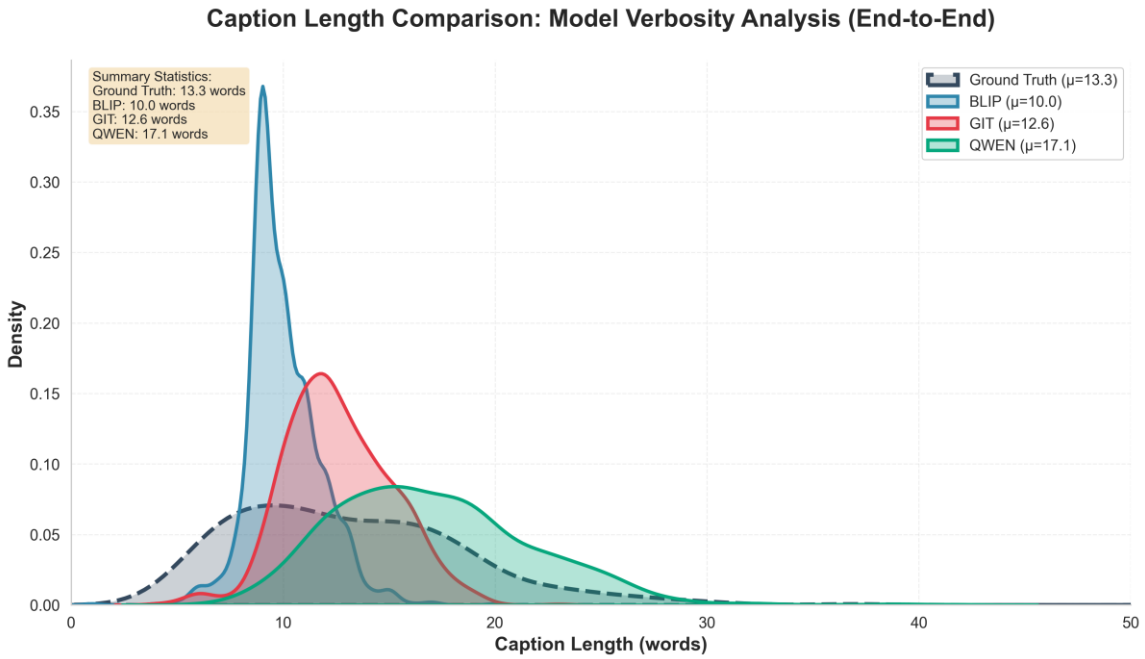
Η συνδυαστική εξέταση των αριθμητικών δεδομένων του **Πίνακα 5.1** και της οπτικής τους απεικόνισης στο **Σχήμα 5.1**, αναδεικνύει ορισμένα καίρια ευρήματα. **Πρώτον**, η απόδοση των τριών μοντέλων στο **temporal localization είναι εντυπωσιακά ομοιόμορφη**, με διαφορές μικρότερες του **1%** στην Precision και **2%** στο Recall. Αυτή η ομοιομορφία υποδηλώνει ότι η απόδοση του temporal localization module είναι σχετικά ανεξάρτητη από την επιλογή του μοντέλου caption generation, καθώς το στάδιο της scene detection εφαρμόζεται πριν την παραγωγή λεξάντων και χρησιμοποιεί τους ίδιους αλγόριθμους για όλα τα μοντέλα. Η μικρή διαφορά που παρατηρείται οφείλεται κυρίως στο στάδιο του semantic merging, το οποίο ενοποιεί διαδοχικές σκηνές με βάση τη σημασιολογική ομοιότητα των παραγόμενων λεξάντων. Επομένως, διαφορετικά μοντέλα που παράγουν διαφορετικού στυλ περιγραφές μπορεί να οδηγήσουν σε ελαφρώς διαφορετικά patterns συγχώνευσης.

Δεύτερον, παρατηρείται σημαντική ασυμμετρία μεταξύ Precision και Recall, με την Recall να υπερβαίνει κατά πολύ την Precision και στα τρία μοντέλα. Συγκεκριμένα, η Recall είναι κατά μέσο όρο περίπου 2.7 φορές υψηλότερη από την Precision. Το φαινόμενο αυτό υποδηλώνει ότι το σύστημα **τείνει να υπερ-τμηματοποιεί (over-segmentation) τα video**, παράγοντας περισσότερα προβλεπόμενα γεγονότα συγκριτικά με τον αριθμό των ground truth annotations. Η συμπεριφορά αυτή κρίνεται αναμενόμενη, δεδομένου ότι η content-based ανίχνευση σκηνών εντοπίζει οπτικές μεταβολές που δεν αντιστοιχούν απαραίτητα σε σημασιολογικά διακριτά γεγονότα, όπως αυτά που ορίζονται από τους ανθρώπινους σχολιαστές. Παρά τη διαδικασία σημασιολογικής συγχώνευσης (semantic merging) που εφαρμόζεται στο τελικό στάδιο, το σύστημα δεν επιτυγχάνει επαρκή μείωση του αριθμού των προβλέψεων ώστε να προσεγγίσει τη λεπτομέρεια (granularity) των ανθρώπινων επισημειώσεων.

Τρίτον, όσον αφορά την ποιότητα των παραγόμενων λεξάντων, το **GIT-VATEX υπερέχει στις μετρικές BLEU και ROUGE-L**, επιτυγχάνοντας **43%** υψηλότερο BLEU-3 έναντι του BLIP και **30%** υψηλότερο έναντι του Qwen2-VL. Η υπεροχή αυτή αποδίδεται στο fine-tuning του GIT στο σύνολο δεδομένων VaTeX και στην αποτελεσματική αξιοποίηση της χρονικής πληροφορίας. Ωστόσο, το **Qwen2-VL σημειώνει την υψηλότερη επίδοση στη μετρική METEOR με 17.03%**, υπερβαίνοντας το GIT-VATEX κατά 7.25%. Το αποτέλεσμα αυτό παρουσιάζει ιδιαίτερο ενδιαφέρον, δεδομένης της σημαντικής υστέρησης

του Qwen2-VL στις μετρικές BLEU. Η παρατηρούμενη απόκλιση ερμηνεύεται από τη διαφορετική φύση των τύπων μετρικών. Το BLEU βασίζεται σε ακριβή αντιστοίχιση n-grams και τιμωρεί αυστηρά τις παραφράσεις ή τις εναλλακτικές διατυπώσεις, ενώ το METEOR λειτουργεί πιο ευέλικτα, συνυπολογίζοντας συνώνυμα, λεξιλογικές παραλλαγές και τη σημασιολογική εγγύτητα. Το Qwen2-VL, ως ένα μεγάλο instruction-tuned multimodal μοντέλο, **τείνει να παράγει εκτενέστερες και πιο περιγραφικές λεζάντες, οι οποίες, αν και ενδέχεται να αποκλίνουν λεξιλογικά από τα ground truths, αποδίδουν με ακρίβεια το σημασιολογικό περιεχόμενο.**

Για την κατανόηση αυτής της συμπεριφοράς, το **Σχήμα 5.2** παρουσιάζει την ανάλυση του μήκους των παραγόμενων περιγραφών για κάθε μοντέλο. Όπως διαφαίνεται στο διάγραμμα, το Qwen2-VL παράγει αισθητά εκτενέστερες περιγραφές, με μέσο μήκος 17.1 λέξεων, συγκριτικά με το μέσο μήκος των ground truths που ανέρχεται στις 13.3 λέξεις. Αντίθετα, το BLIP παράγει συνοπτικότερες περιγραφές με μέσο μήκος 10.0 λέξεων, ενώ το GIT-VATEX προσεγγίζει περισσότερο τα δεδομένα αναφοράς με 12.6 λέξεις. Η κατανομή του μήκους των λεζάντων, όπως αποτυπώνεται στις καμπύλες πυκνότητας του **Σχήματος 5.2**, αποκαλύπτει ότι το BLIP εμφανίζει την πλέον συγκεντρωμένη κατανομή με κορύφωση περί τις 10 λέξεις, το GIT-VATEX παρουσιάζει ευρύτερη κατανομή με μέγιστο στις 12 λέξεις, ενώ το Qwen2-VL χαρακτηρίζεται από αυξημένη διασπορά, εκτεινόμενο προς μεγαλύτερα μήκη κειμένου.



Σχήμα 5.2: *Σύγκριση μέσου μήκους και πυκνότητας κατανομής των λεζάντων ανά μοντέλο.*

Η τάση του Qwen2-VL να παράγει μακροσκελείς περιγραφές οφείλεται στο instruction-tuning του μοντέλου και στο συγκεκριμένο prompt που χρησιμοποιήθηκε κατά την αξιολόγηση, το οποίο ενθαρρύνει λεπτομερείς περιγραφές. Αυτό αποτελεί μια σημαντική παρατήρηση καθώς υποδηλώνει ότι η απόδοση του Qwen2-VL είναι ιδιαίτερα ευαίσθητη στην επιλογή του prompt, και διαφορετικές διατυπώσεις του prompt θα μπορούσαν δυνητικά να οδηγήσουν σε σημαντικά διαφορετικά αποτελέσματα.

Εξίσου σημαντική παρατήρηση αποτελούν οι χαμηλές απόλυτες τιμές σε όλες τις μετρικές αξιολόγησης. Ακόμα και το μοντέλο **GIT-VATEX**, που επέδειξε τη μεγαλύτερη σταθερότητα, περιορίζεται σε **5.17% BLEU-3** και **22.28% ROUGE-L**. Αντίστοιχα, το **Qwen2-VL**, παρά την επικράτησή του στη σημασιολογική ακρίβεια, δεν ξεπερνά το **17.03%** στη μετρική **METEOR**. Τα αποτελέσματα αυτά δεν οφείλονται αποκλειστικά στη γλωσσική ποικιλομορφία, αλλά αντικατοπτρίζουν και τους περιορισμούς της υιοθετημένης αρχιτεκτονικής. Η προτεινόμενη προσέγγιση ακολουθεί μια πολυ-σταδιακή διαδικασία

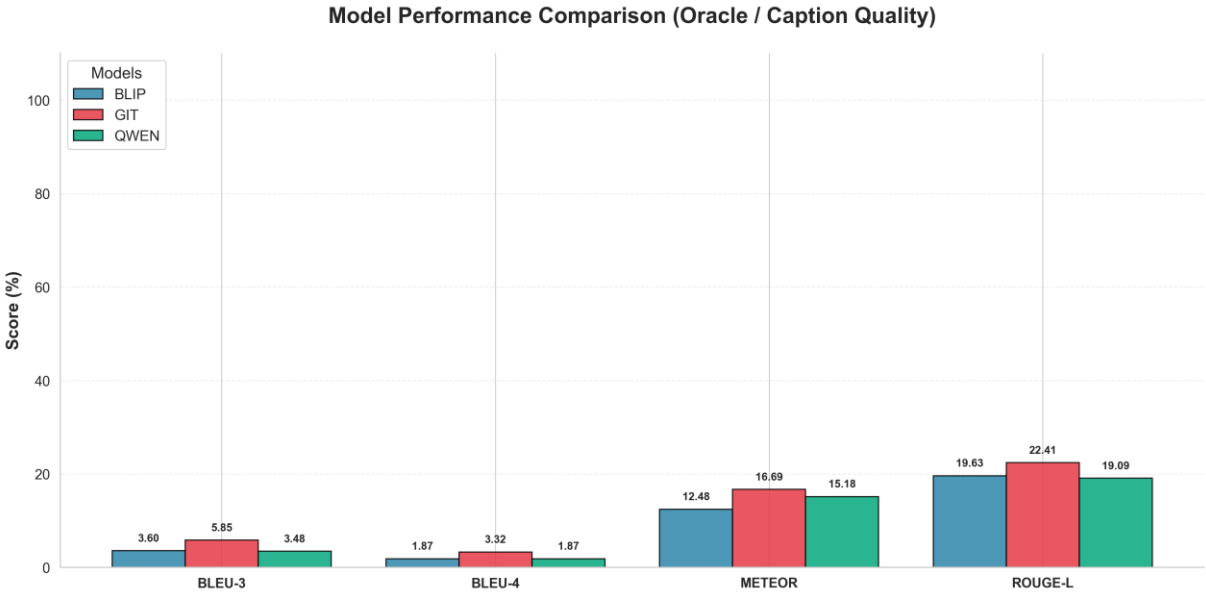
(multi-stage pipeline), όπου ο χρονικός εντοπισμός και η παραγωγή περιγραφής είναι διακριτά βήματα. Αυτό οδηγεί σε αναπόφευκτη διάδοση σφάλματος (error propagation), εάν ο αλγόριθμος ανίχνευσης σκηνών (visual scene detection) δεν ταυτιστεί απόλυτα με τα σημασιολογικά όρια των γεγονότων (semantic events), το μοντέλο λεζάντας καλείται να περιγράψει ένα ελλιπές ή θορυβώδες χρονικό τμήμα, μειώνοντας αναγκαστικά την ακρίβεια της πρόβλεψης έναντι των ground truths. Συνεπώς, τα scores αυτά είναι ενδεικτικά μιας μεθόδου που βασίζεται στη σύνθεση προ-εκπαιδευμένων μοντέλων και όχι σε μια end-to-end εκπαίδευση βελτιστοποιημένη αποκλειστικά για το dataset ActivityNet.

5.3.2 Αξιολόγηση Ποιότητας Λεζάντων βάσει Ground truth Χρονικών Ορίων

Για την απομόνωση της καθαρής ικανότητας των μοντέλων να παράγουν ποιοτικές γλωσσικές περιγραφές, διεξήχθη αξιολόγηση χρησιμοποιώντας τα ground truth χρονικά όρια από το dataset, παρακάμπτοντας πλήρως το temporal localization module. Τα αποτελέσματα παρουσιάζονται στον **Πίνακα 5.2** και οπτικοποιούνται στο **Σχήμα 5.3**.

Πίνακας 5.2: Αποτελέσματα Αξιολόγησης με Ground truth Χρονικά Όρια.

Μοντέλο	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**
BLIP	3.60	1.87	12.48	19.63
GIT-VATEX	**5.85**	**3.32**	**16.69**	**22.41**
Qwen2-VL	3.48	1.87	15.18	19.09



Σχήμα 5.3: *Συγκριτική Απόδοση Μοντέλων σε Oracle Αξιολόγηση (Ground truth Χρονικά Όρια)*

Τα αποτελέσματα της αξιολόγησης υπό συνθήκες Oracle αναδεικνύουν ενδιαφέροντα πρότυπα συμπεριφοράς, τα οποία διαφοροποιούνται από την End-to-End αξιολόγηση.

Πρώτον, στο πλαίσιο του παρόντος πειραματικού σχεδιασμού, το μοντέλο GIT-VATEX σημειώνει τις υψηλότερες επιδόσεις σε όλες τις εξεταζόμενες μετρικές, **BLEU-3 5.85%, BLEU-4 3.32%, METEOR 16.69% και ROUGE-L 22.41%**. Η συνεπής αυτή επικράτηση δεν συνεπάγεται απαραίτητα την καθολική ανωτερότητα της αρχιτεκτονικής, αλλά υποδεικνύει την υψηλή προσαρμοστικότητα του GIT-VATEX στα ειδικά χαρακτηριστικά του ActivityNet Captions dataset και τη συγκεκριμένη κατανομή των δεδομένων αναφοράς. Συνεπώς, για τη συγκεκριμένη εργασία (task) και υπό την προϋπόθεση ακριβών χρονικών

ορίων, το GIT-VATEX αποδεικνύεται η πλέον αποδοτική επιλογή μεταξύ των συγκρινόμενων μοντέλων.

Δεύτερον, εντοπίζεται ένα αξιοσημείωτο παράδοξο στα αποτελέσματα του Qwen2-VL. Συγκεκριμένα, ενώ κατά την end-to-end αξιολόγηση το μοντέλο επέτυχε την υψηλότερη τιμή στη μετρική **METEOR 17.03%**, στο σενάριο Oracle η επίδοση αυτή υποχωρεί στο **15.18%**, χαμηλότερο από το αντίστοιχο score του end-to-end evaluation. Παρόμοια τάση παρατηρείται και στο **BLEU-3**, όπου το score της Oracle αξιολόγησης **3.48%** υπολείπεται του αντίστοιχου end-to-end **3.97%**. Το συγκεκριμένο αποτέλεσμα, αν και εκ πρώτης όψεως αντι-διαισθητικό, ερμηνεύεται σε μεγάλο βαθμό από τη φύση της εντολής εισόδου (prompt) που χρησιμοποιήθηκε. Η επιλογή ενός prompt που στοχεύει ρητά σε «λεπτομερή περιγραφή» (detailed description), ωθεί το μοντέλο στην παραγωγή εκτενών και πιο αναλυτικών περιγραφών σε αντίθεση με τα ground truths.

Τρίτον, το μοντέλο **BLIP** επιδεικνύει αξιοσημείωτη σταθερότητα, με τις επιδόσεις στο σενάριο Oracle να κυμαίνονται σε επίπεδα παραπλήσια με εκείνα της End-to-End αξιολόγησης. Συγκεκριμένα, καταγράφεται ελαφρώς υψηλότερο **BLEU-4 (1.87% έναντι 1.67%)** και οριακά χαμηλότερο **METEOR (12.48% έναντι 13.13%)**. Οι μικρές αυτές αποκλίσεις εμπίπτουν στα όρια της στατιστικής διακύμανσης, υποδηλώνοντας ότι, στην περίπτωση του BLIP, η συνολική ποιότητα των παραγόμενων λεζάντων παρουσιάζει σχετική ανθεκτικότητα και δεν επηρεάζεται δραστικά από την ακρίβεια του χρονικού εντοπισμού.

Τέλος, το μοντέλο GIT-VATEX ακολουθεί το πλέον αναμενόμενο θεωρητικό πρότυπο, όπου όλες οι μετρικές υπό συνθήκες Oracle υπερέχουν των αντίστοιχων End-to-End. Η διαφορά, αν και μικρή, παραμένει σταθερή σε όλες τις διαστάσεις, με το **BLEU-3** να αυξάνεται κατά **0.68** ποσοστιαίες μονάδες, το **BLEU-4** κατά **1.38**, το **METEOR** κατά **0.81** και το **ROUGE-L** κατά **0.13**. Η παρατήρηση αυτή καταδεικνύει ότι, για το GIT-VATEX, το υποσύστημα χρονικού εντοπισμού (temporal localization module) εισάγει ένα διακριτό σφάλμα, το οποίο επιδρά μετρήσιμα, αν και όχι καθοριστικά στην τελική απόδοση του συστήματος.

Πίνακας 5.3: Σύγκριση Oracle vs End-to-End Gap (ποσοστιαίες μονάδες)

Μοντέλο	**Δ BLEU-3**	**Δ BLEU-4**	**Δ METEOR**	**Δ ROUGE-L**
BLIP	0.00	+0.20	-0.65	-0.05
GIT-VATEX	**+0.68**	**+1.38**	+0.81	+0.13
Qwen2-VL	-0.49	+0.03	**-1.85**	**-1.55**

Σημείωση: Θετικές τιμές υποδηλώνουν ότι το Oracle evaluation έχει υψηλότερο score (αναμενόμενο), ενώ αρνητικές τιμές υποδηλώνουν ότι το E2E evaluation έχει υψηλότερο score (απροσδόκητο).

Όπως αποτυπώνεται στον Πίνακα 5.3, το μοντέλο Qwen2-VL εμφανίζει αρνητική απόκλιση (negative gap) σε τρεις από τις τέσσερις μετρικές, επιβεβαιώνοντας την ασυνήθιστη συμπεριφορά του έναντι των υπολοίπων αρχιτεκτονικών. Το εύρημα αυτό κρίνεται βαρύνουσας σημασίας για την ορθή ερμηνεία των πειραματικών δεδομένων, καθώς υπογραμμίζει την αναγκαιότητα της διεξοδικής εξέτασης της αλληλεπίδρασης μεταξύ των επιμέρους συστατικών του συστήματος (pipeline). Σε αυτό το πλαίσιο, καθοριστικό ρόλο φαίνεται να διαδραματίζει η φύση των οδηγιών (instructions) και του prompt που δόθηκαν στο μοντέλο, καθώς η ρητή ενθάρρυνση για λεπτομερή περιγραφή ενδέχεται να οδηγεί σε σημασιολογικά πλούσιες αλλά λεκτικά αποκλίνουσες διατυπώσεις, οι οποίες "τιμωρούνται" από τις μετρικές ακόμη και υπό ιδανικές συνθήκες χρονισμού.

5.3.3 Χρονική Αποδοτικότητα

Πέρα από την ποιότητα των παραγόμενων αποτελεσμάτων, η υπολογιστική απόδοση αποτελεί κρίσιμο παράγοντα για την πρακτική εφαρμογή ενός dense video captioning συστήματος. Ο **Πίνακας 5.4** παρουσιάζει τους χρόνους επεξεργασίας που καταγράφηκαν κατά την end-to-end αξιολόγηση των εκατό video του dataset, παρέχοντας μια ρεαλιστική εικόνα των υπολογιστικών απαιτήσεων κάθε μοντέλου.

Πίνακας 5.4: Χρόνοι Επεξεργασίας End-to-End Pipeline

Μοντέλο	**Συνολικός Χρόνος (s)**	**Μέσος Χρόνος/Video (s)**	**Σχετική Ταχύτητα**
BLIP	758.3	7.6	1.0× (βάση)
GIT-VATEX	18,334.3	183.3	24.2×
Qwen2-VL	24,588.8	245.9	32.4×

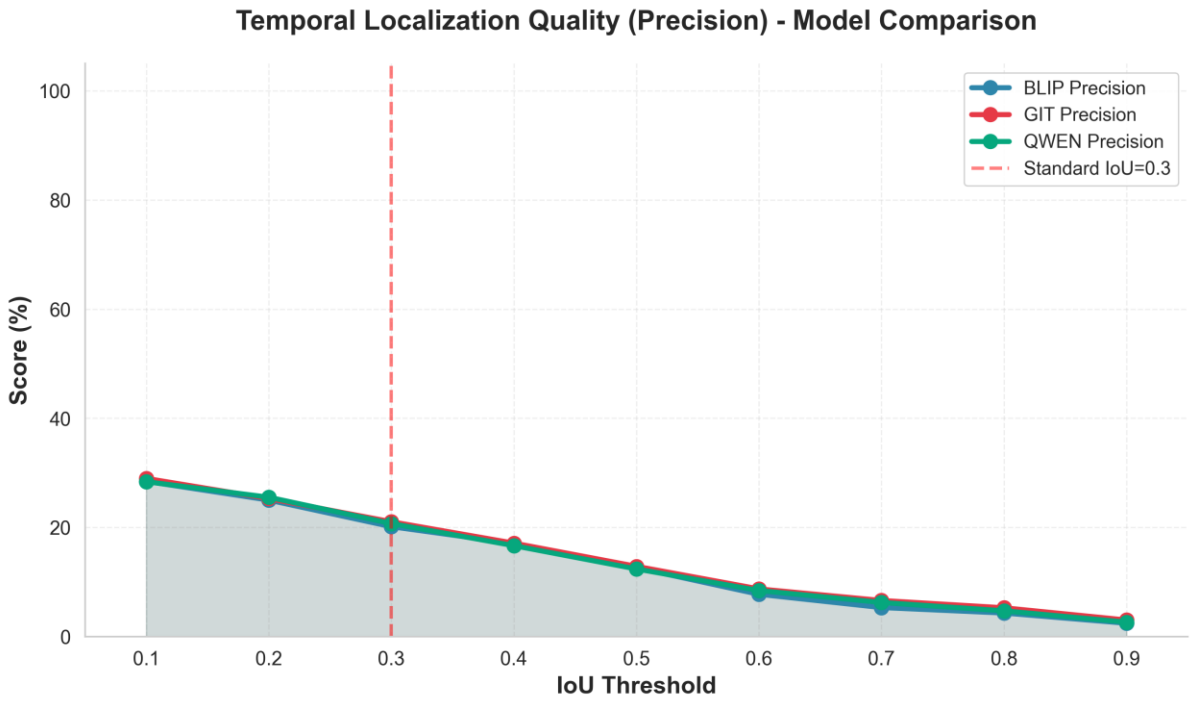
Το μοντέλο **BLIP** παρουσιάζει τη μέγιστη ταχύτητα με μέσο χρόνο εκτέλεσης 7.6 δευτερόλεπτα ανά video. Αντιθέτως, το **GIT-VATEX** απαιτεί 183.3 δευτερόλεπτα (~24x περισσότερο), ενώ το **Qwen2-VL** αναδεικνύεται ως το πλέον απαιτητικό με 245.9 δευτερόλεπτα (~32x περισσότερο από το BLIP). Οι αποκλίσεις αυτές αποδίδονται πρωτίστως σε αρχιτεκτονικές επιλογές: το BLIP επεξεργάζεται ένα μοναδικό keyframe, ενώ τα GIT-VATEX και Qwen2-VL αναλύουν πολλαπλά πλαίσια (6 και 5 αντίστοιχα) για την εξαγωγή χρονικών χαρακτηριστικών, πολλαπλασιάζοντας τον υπολογιστικό φόρτο. Επιπλέον, το Qwen2-VL επιβαρύνεται από την πολυπλοκότητα των 2 δισεκατομμυρίων παραμέτρων του, παρά τη χρήση 4-bit κβαντοποίησης.

5.4 Ανάλυση Temporal Localization

Η ανάλυση της χρονικής ακρίβειας του εντοπισμού γεγονότων πραγματοποιείται μέσω των temporal localization curves, οι οποίες αποτυπώνουν την εξέλιξη της Precision και του Recall σε διαφορετικά επίπεδα IoU threshold. Οι καμπύλες αυτές παρέχουν μια λεπτομερή εικόνα της απόδοσης του συστήματος ως προς το κατά πόσο τα προβλεπόμενα χρονικά όρια των γεγονότων ευθυγραμμίζονται με τα αντίστοιχα ground truth annotations. Όσο αυξάνεται το IoU threshold, η αξιολόγηση γίνεται αυστηρότερη, απαιτώντας μεγαλύτερη χρονική επικάλυψη μεταξύ πρόβλεψης και αναφοράς για την αποδοχή ενός match, γεγονός που επηρεάζει άμεσα τις τιμές Precision και Recall.

Σημειώνεται ότι στην παρούσα εργασία τα χρονικά segments των γεγονότων προκύπτουν από ένα κοινό στάδιο ανίχνευσης σκηνών μέσω του PySceneDetect (ContentDetector) και παραμένουν σταθερά ανεξαρτήτως του captioning μοντέλου. Ως αποτέλεσμα, οι temporal localization curves εμφανίζουν σε μεγάλο βαθμό παρόμοια συμπεριφορά μεταξύ των διαφορετικών αρχιτεκτονικών. Οι μικρές αποκλίσεις που παρατηρούνται οφείλονται κυρίως στο στάδιο της σημασιολογικής συγχώνευσης (semantic merging), όπου χρονικά γειτονικά segments με παρόμοιες περιγραφές ενδέχεται να ενοποιηθούν. Η διαδικασία αυτή μπορεί να μειώσει τον αριθμό των τελικών σκηνών, επηρεάζοντας έμμεσα την αντιστοίχιση με τα ground truth γεγονότα και οδηγώντας σε οριακές μεταβολές στις τιμές Precision και Recall.

Όπως παρουσιάζεται στο **Σχήμα 5.4**, η συμπεριφορά των καμπυλών σε όλο το εύρος των IoU thresholds επιτρέπει την εκτίμηση της σταθερότητας και της αξιοπιστίας του συστήματος ως προς τη χρονική ακρίβεια του εντοπισμού γεγονότων.



Σχήμα 5.4: *Σύγκριση Temporal Localization Precision ανά Μοντέλο*

Το **Σχήμα 5.4** παρουσιάζει τη συγκριτική απόδοση των τριών μοντέλων όσον αφορά την Precision σε διαφορετικά IoU thresholds. Το πιο χαρακτηριστικό εύρημα είναι η σχεδόν ταυτόσημη συμπεριφορά και των τριών μοντέλων σε όλο το φάσμα των IoU values, με τις καμπύλες να επικαλύπτονται σχεδόν πλήρως. Αυτή η ομοιομορφία επιβεβαιώνει οριστικά ότι η απόδοση του temporal localization module είναι ανεξάρτητη από την επιλογή του caption generation μοντέλου, καθώς το στάδιο της scene detection εκτελείται πριν την παραγωγή λεζάντων χρησιμοποιώντας αποκλειστικά οπτική πληροφορία. Οι ελάχιστες διαφορές που παρατηρούνται οφείλονται στο στάδιο του semantic merging, το οποίο χρησιμοποιεί τις παραγόμενες λεζάντες για την ενοποίηση παρόμοιων σκηνών, επηρεάζοντας έμμεσα τον τελικό αριθμό και τα χρονικά όρια των προβλεπόμενων γεγονότων.

Ο **Πίνακας 5.5** συγκεντρώνει τα κρίσιμα σημεία των temporal localization curves, παρέχοντας τις ακριβείς αριθμητικές τιμές για Precision και Recall σε επιλεγμένα IoU thresholds.

Πίνακας 5.5: Temporal Localization Performance σε Διαφορετικά IoU Thresholds

IoU	BLIP P/R (%)	GIT-VATEX P/R (%)	Qwen2-VL P/R (%)
0.1	28.43 / 76.26	28.97 / 76.85	28.37 / 76.85
0.3	20.13 / 54.01	21.03 / 55.79	20.70 / 56.08
0.5	12.61 / 33.83	12.86 / 34.12	12.38 / 33.53
0.7	5.31 / 14.24	6.60 / 17.51	6.24 / 16.91
0.9	2.43 / 6.53	3.02 / 8.01	2.63 / 7.12

Όπως αποτυπώνεται στον πίνακα, με ένα ελαστικό **threshold IoU** της τάξης του **10%**, τα μοντέλα επιτυγχάνουν **Recall** που κυμαίνεται μεταξύ **76-77%**. Το ποσοστό αυτό υποδηλώνει ότι το σύστημα εντοπίζει επιτυχώς περίπου 3 στα 4 ground truth γεγονότα με τουλάχιστον ελάχιστη χρονική επικάλυψη. Ωστόσο, η **Precision** παραμένει σε χαμηλά επίπεδα, της τάξης του **28-29%**, γεγονός που αντικατοπτρίζει το ζήτημα της **υπερ-τμηματοποίησης (over-segmentation)** που αναλύθηκε στην Ενότητα 5.3.1. Το σύστημα

παράγει σημαντικά μεγαλύτερο αριθμό προβλεπόμενων γεγονότων συγκριτικά με τις επισημειώσεις αναφοράς (ground truth annotations), οδηγώντας αναπόφευκτα σε χαμηλή ακρίβεια ακόμη και υπό χαλαρά κριτήρια επικάλυψης.

Η μετάβαση στο τυποποιημένο **threshold IoU** του **30%**, το οποίο αποτελεί το standard κριτήριο στη βιβλιογραφία του Dense Video Captioning και απεικονίζεται με την κόκκινη διακεκομμένη γραμμή στο **Σχήμα 5.4**, επιφέρει σημαντική πτώση στην απόδοση του συστήματος. Συγκεκριμένα, η μετρική **Recall** μειώνεται στο εύρος **54-56%**, ενώ **η Precision** υποχωρεί στο **20-21%**. Η απότομη αυτή κάμψη υποδηλώνει ότι ένας σημαντικός αριθμός των ανιχνευμένων σκηνών παρουσιάζει μόνον μερική επικάλυψη με τα ground truth γεγονότα, με τα χρονικά όρια να εμφανίζουν μετατοπίσεις ή ατελή ταύτιση σε σχέση με τις επισημειώσεις των ανθρώπινων αξιολογητών.

Η υιοθέτηση αυστηρότερων κριτηρίων οδηγεί σε ραγδαία υποβάθμιση της απόδοσης. Συγκεκριμένα, με **IoU** ίσο με **50%**, το **Recall** περιορίζεται στο εύρος **33-34%** και η **Precision** στο **12-13%**, εύρημα που υποδηλώνει ότι μόλις ένα στα τρία ground truth γεγονότα εντοπίζεται με χρονική επικάλυψη άνω του ημίσεος. Στο εξαιρετικά αυστηρό **threshold IoU 90%**, το οποίο προϋποθέτει σχεδόν τέλεια χρονική ταύτιση, οι επιδόσεις συρρικνώνονται περαιτέρω, με το **Recall** να κυμαίνεται στο **6-8%** και την **Precision** στο **2-3%**. Τα αποτελέσματα αυτά καταδεικνύουν ότι μόνο ένα ελάχιστο ποσοστό των προβλέψεων επιτυγχάνει ακριβή χρονική ευθυγράμμιση (alignment) με τις επισημειώσεις αναφοράς.

Το pattern του **Recall ακολουθεί ανάλογη πορεία με την Precision**, με συστηματική μείωση καθώς το **IoU threshold** αυξάνεται. Η σταθερή αναλογία μεταξύ Precision και Recall σε όλα τα επίπεδα IoU υποδηλώνει ότι το σύστημα διατηρεί την ίδια σχετική συμπεριφορά ανεξάρτητα από την αυστηρότητα του κριτηρίου αποδοχής. Η απόσταση μεταξύ των δύο μετρικών παραμένει σχετικά σταθερή, με το Recall να υπερβαίνει την Precision κατά περίπου δύο κόμμα πέντε με δύο κόμμα επτά φορές σε όλο το φάσμα των IoU values.

Η παρατηρούμενη **ομοιομορφία της απόδοσης μεταξύ των τριών μοντέλων σε όλα τα επίπεδα IoU** αποτελεί σημαντικό εύρημα, το οποίο απορρέει από την αρχιτεκτονική δομή του συστήματος και συγκεκριμένα από τον λειτουργικό διαχωρισμό του χρονικού εντοπισμού (temporal localization) από τη διαδικασία παραγωγής λεζάντας (caption generation). Το γεγονός ότι οι διαφορές μεταξύ των μοντέλων είναι μικρότερες του ένα τοις εκατό σε όλα τα thresholds δείχνει ότι οποιαδήποτε βελτίωση στην χρονική ακρίβεια του συστήματος πρέπει να επιτευχθεί μέσω της βελτίωσης του scene detection module και όχι μέσω της αλλαγής του caption generation μοντέλου.

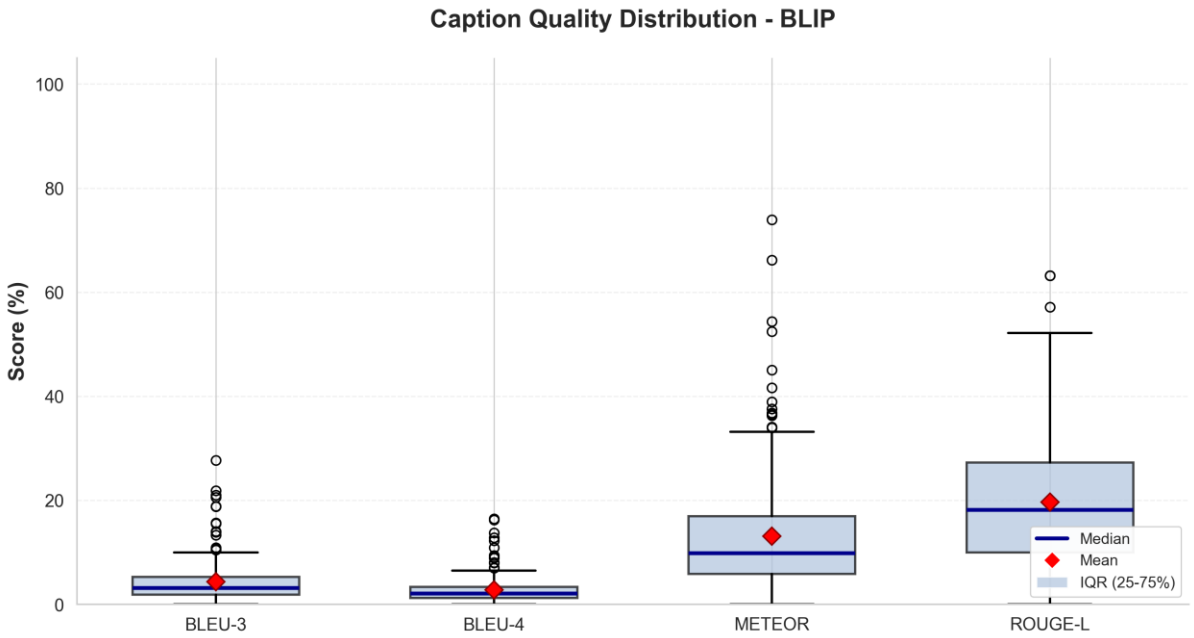
Η **σταθερή πτώση της απόδοσης** σε όλα τα μοντέλα υποδηλώνει ότι το θεμελιώδες πρόβλημα βρίσκεται στην ίδια την προσέγγιση της **content-based detection**. Η μέθοδος ανιχνεύει οπτικές αλλαγές που δεν αντιστοιχούν απαραίτητα στα σημασιολογικά όρια που θέτουν οι ανθρώπινοι αξιολογητές. Μια αλλαγή στη γωνία της κάμερας εντός της ίδιας δράσης θα ανιχνευθεί ως νέα σκηνή, ενώ οι annotators θεωρούν ολόκληρη τη δράση ως ενιαίο γεγονός. Αντίστροφα, μια λεπτή αλλαγή περιεχομένου που σηματοδοτεί νέα δράση μπορεί να μην ανιχνευθεί αν δεν συνοδεύεται από σημαντική οπτική μεταβολή. Η βελτίωση της χρονικής ακρίβειας απαιτεί θεμελιώδεις αλλαγές στο scene detection module, όπως την ενσωμάτωση σημασιολογικής πληροφορίας, την χρήση temporal context ή την εκπαίδευση ενός learned temporal proposal network που θα μπορεί να μαθαίνει τα σημασιολογικά όρια των γεγονότων από annotated δεδομένα.

5.5 Ανάλυση Κατανομής Ποιότητας Λεξάντων

Πέρα από τους μέσους όρους των μετρικών που παρουσιάστηκαν στην ενότητα 5.3, η ανάλυση της κατανομής των scores παρέχει σημαντική πληροφορία για τη σταθερότητα και την αξιοπιστία κάθε μοντέλου. Τα box plots απεικονίζουν την κατανομή των 4 βασικών μετρικών ποιότητας λεξάντων για κάθε μοντέλο στην end-to-end αξιολόγηση, αποκαλύπτοντας όχι μόνο την κεντρική τάση αλλά και τη διασπορά, την ύπαρξη outliers και τη συνολική συμπεριφορά σε διαφορετικά video.

5.5.1 Κατανομή Απόδοσης BLIP

Πέρα από τους συγκεντρωτικούς μέσους όρους, κρίνεται απαραίτητη η διερεύνηση της διακύμανσης των επιδόσεων ανά video, προκειμένου να αξιολογηθεί η σταθερότητα του μοντέλου σε διαφορετικού τύπου περιεχόμενο. Στο πλαίσιο αυτό, το **Σχήμα 5.5** που ακολουθεί αποτυπώνει την κατανομή των βαθμολογιών για το μοντέλο BLIP σε όλες τις εξεταζόμενες μετρικές, μέσω διαγραμμάτων θηκογράμματος (boxplots). Η γραφική αυτή απεικόνιση επιτρέπει την εποπτική εξέταση του εύρους τιμών, των διαμεσολαβητικών ορίων (quartiles), καθώς και τον εντοπισμό ακραίων τιμών (outliers), προσφέροντας μια σαφέστερη εικόνα για την ομοιογένεια της συμπεριφοράς του μοντέλου.



Σχήμα 5.5: *Κατανομή Ποιότητας Λεξάντων - BLIP*

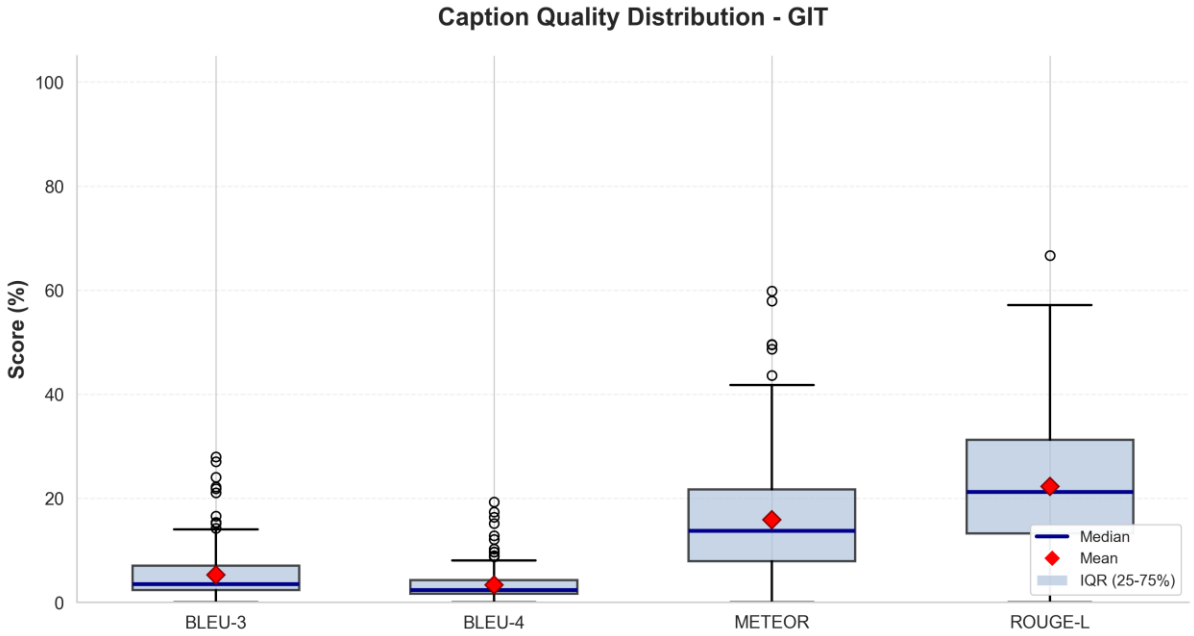
Το μοντέλο **BLIP** παρουσιάζει εξαιρετικά χαμηλές και συμπυκνωμένες κατανομές στις **BLEU** μετρικές, με **median** τιμές κοντά στο μηδέν και πολύ μικρό **interquartile range**. Το **50%** των video επιτυγχάνει **BLEU-3** κάτω από **3%** και **BLEU-4** κάτω από **2%**. Η κατανομή εμφανίζει σημαντικό αριθμό outliers στην άνω περιοχή, δηλαδή video όπου το BLIP επιτυγχάνει σχετικά καλές επιδόσεις, φτάνοντας μέχρι **15-20% BLEU-3**. Αυτό υποδηλώνει ότι η απόδοση του BLIP είναι εξαιρετικά ετερογενής, με το μοντέλο να λειτουργεί καλά σε συγκεκριμένα είδη περιεχομένου αλλά να αποτυγχάνει στην πλειονότητα των περιπτώσεων.

Στις μετρικές **METEOR** και **ROUGE-L**, το **BLIP** εμφανίζει ευρύτερη κατανομή με **median** γύρω στο **12%** για το **METEOR** και **20%** για το **ROUGE-L**. Το interquartile range είναι μεγαλύτερο, υποδηλώνοντας μεγαλύτερη διασπορά στην απόδοση. Η ύπαρξη πολλών outliers τόσο στην άνω όσο και στην κάτω περιοχή δείχνει ότι το BLIP έχει ασταθή

συμπεριφορά, με την ποιότητα των λεζάντων να εξαρτάται σημαντικά από το συγκεκριμένο video και το περιεχόμενό του. Η μεγάλη διασπορά καθιστά το BLIP λιγότερο αξιόπιστο για εφαρμογές που απαιτούν σταθερή ποιότητα εξόδου.

5.5.2 Κατανομή Απόδοσης GIT-VATEX

Ακολουθώντας την ίδια μεθοδολογία ανάλυσης, η προσοχή στρέφεται ακολούθως στο μοντέλο GIT-VATEX, η συμπεριφορά του οποίου παρουσιάζει σαφείς διαφοροποιήσεις συγκριτικά με το BLIP. Το **Σχήμα 5.6** που παρατίθεται, απεικονίζει την κατανομή των τιμών για τις τέσσερις μετρικές αξιολόγησης, επιτρέποντας τη συγκριτική διερεύνηση της σταθερότητας του μοντέλου, καθώς και τον προσδιορισμό του εύρους διακύμανσης μεταξύ της τυπικής και της βέλτιστης απόδοσης.



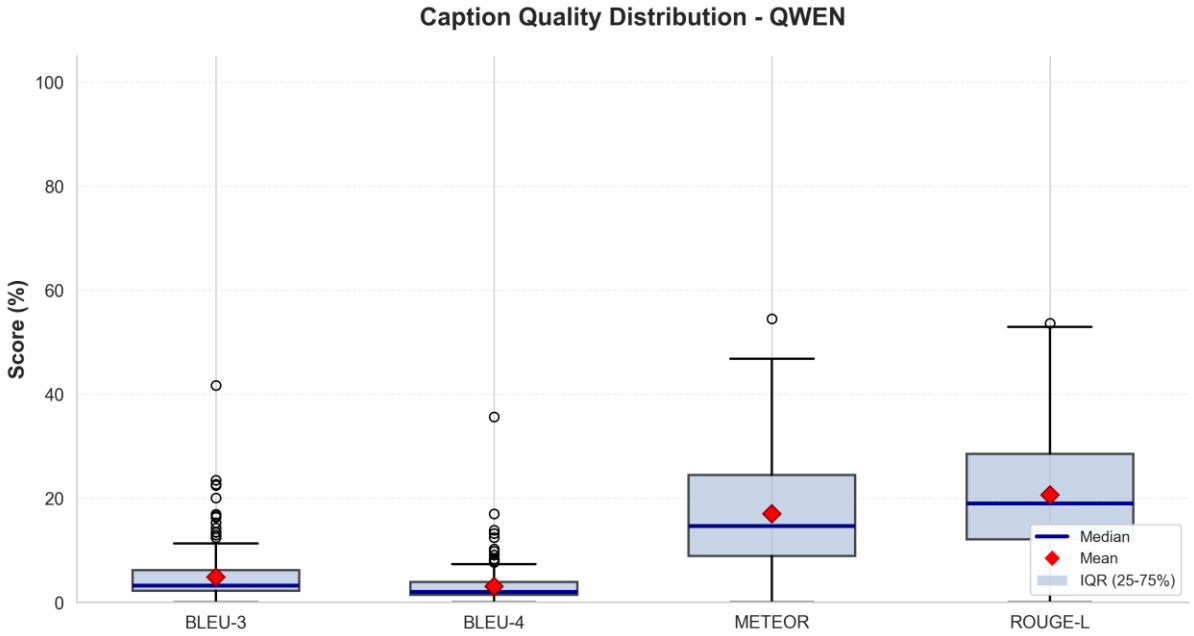
Σχήμα 5.6: *Κατανομή Ποιότητας Λεζάντων – GIT-VATEX*

Το μοντέλο **GIT-VATEX** παρουσιάζει σημαντικά διαφορετική κατανομή. Στις **BLEU** μετρικές, το **median** είναι υψηλότερο από του BLIP, γύρω στο **4-5%** για το **BLEU-3** και **2-3%** για το **BLEU-4**. Το πιο σημαντικό είναι ότι το **interquartile range** είναι ευρύτερο, υποδηλώνοντας ότι ένα μεγαλύτερο ποσοστό video επιτυγχάνει αξιοπρεπείς επιδόσεις. Η κατανομή του GIT-VATEX εμφανίζει λιγότερους αλλά πιο ακραίους outliers, με ορισμένα video να επιτυγχάνουν **BLEU-3** μέχρι **25-30%**. Οι συγκεκριμένοι outliers αντιστοιχούν πιθανότατα σε video με περιεχόμενο που παρουσιάζει υψηλή συνάφεια με το domain του VaTeX dataset, στο οποίο έχει πραγματοποιηθεί το fine-tuning του μοντέλου.

Στο **METEOR**, το **GIT-VATEX** εμφανίζει **median** γύρω στο **15-16%**, με μια πιο συμμετρική κατανομή και λιγότερους outliers σε σχέση με το BLIP. Αυτό υποδηλώνει μεγαλύτερη σταθερότητα και προβλεψιμότητα στην απόδοση του μοντέλου. Το **ROUGE-L** παρουσιάζει παρόμοιο pattern με **median** στο **21-22%** και ευρύ **interquartile range** που εκτείνεται από το **13% έως το 30%** περίπου. Η ύπαρξη outliers που φτάνουν μέχρι 50-60% δείχνει ότι το GIT-VATEX μπορεί να επιτύχει εξαιρετική απόδοση σε συγκεκριμένα video, κάτι που το καθιστά το πιο ισχυρό μοντέλο για εφαρμογές όπου η ποιότητα είναι κρίσιμη. Η συνολική εικόνα της κατανομής του GIT-VATEX δείχνει ένα μοντέλο με υψηλότερη baseline απόδοση και μεγαλύτερη δυνατότητα για excellence σε κατάλληλο περιεχόμενο.

5.5.3 Κατανομή Απόδοσης Qwen2-VL

Η συγκριτική ανάλυση των κατανομών ολοκληρώνεται με την εξέταση του Qwen2-VL, η συμπεριφορά του οποίου παρουσιάζει ιδιαίτερο ενδιαφέρον λόγω της διαφοροποιημένης αρχιτεκτονικής του δομής. Το **Σχήμα 5.7** αποτυπώνει τη διασπορά των τιμών για τις επιλεγμένες μετρικές, προσφέροντας μια εποπτική εικόνα της απόδοσης του μοντέλου και επιτρέποντας τη συσχέτιση των ποσοτικών δεδομένων με τα ποιοτικά χαρακτηριστικά που αναδείχθηκαν στην Ενότητα 5.3.



Σχήμα 5.7: *Κατανομή Ποιότητας Λεζάντων - Qwen2-VL*

Το **Qwen2-VL** εμφανίζει μια ενδιαφέρουσα κατανομή που αντανακλά τα χαρακτηριστικά του που αναλύθηκαν στην ενότητα 5.3. Στις **BLEU** μετρικές, η κατανομή είναι παρόμοια με του BLIP, με χαμηλό **median** και συμπυκνωμένες τιμές. Ωστόσο, στο **METEOR**, το **Qwen2-VL εμφανίζει την υψηλότερη median τιμή από τα 3 μοντέλα**, γύρω στο **16-17%**, με ευρύ **interquartile range** που εκτείνεται από το **10% έως το 23%**. Αυτή η ευρεία κατανομή αντανακλά την ευαισθησία του μοντέλου στο prompt και στο συγκεκριμένο περιεχόμενο, με το **METEOR** να επιβραβεύει τις λεπτομερείς και περιγραφικές λεζάντες που παράγει το Qwen2-VL ακόμα και όταν δεν ταιριάζουν λεκτικά με τα ground truth.

5.5.4 Συγκριτική Αξιολόγηση Κατανομών

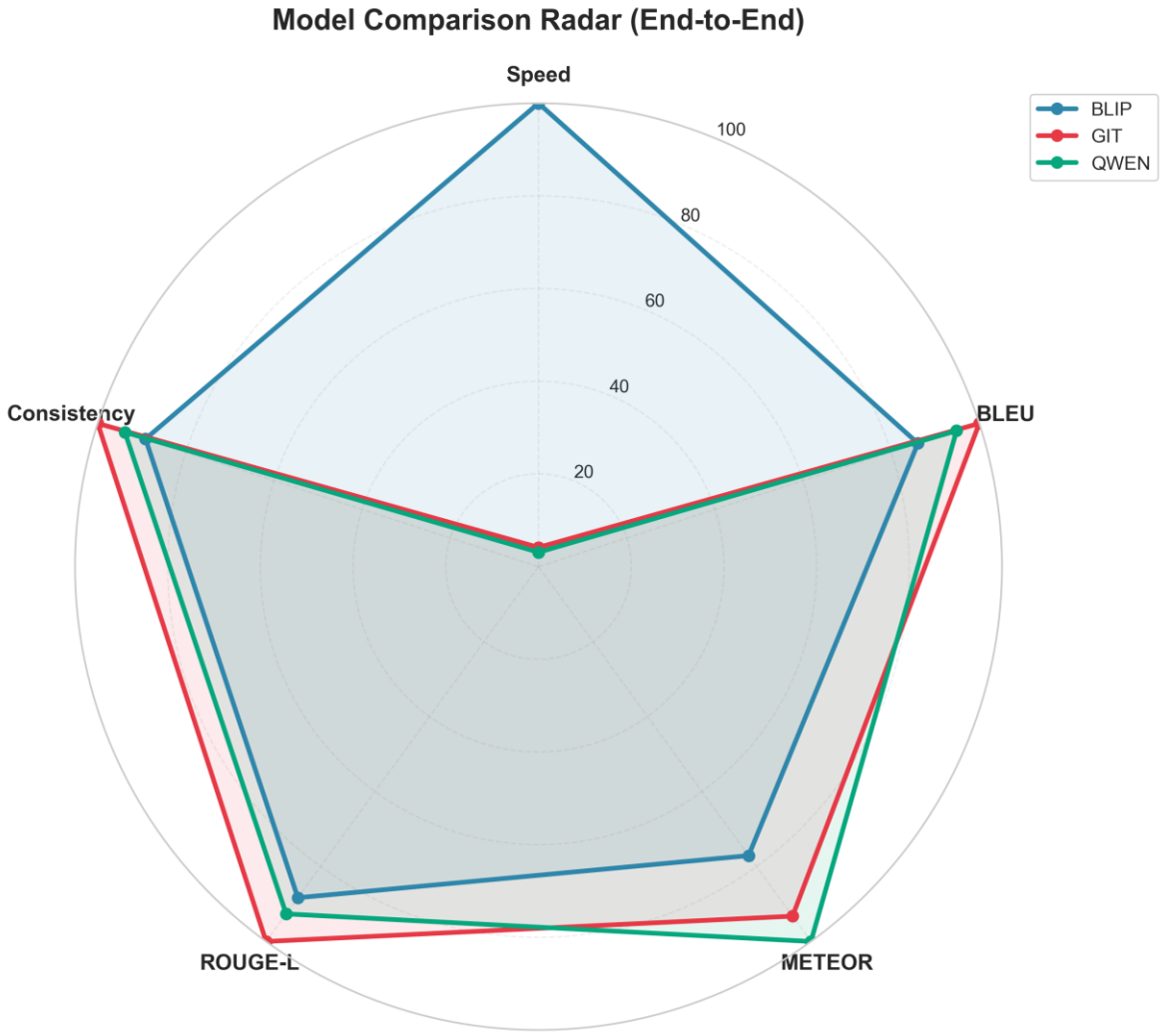
Η συγκριτική ανάλυση των κατανομών καταδεικνύει ότι το **GIT-VATEX** δεν περιορίζεται απλώς σε υψηλότερους μέσους όρους, αλλά παρουσιάζει και βελτιωμένα χαρακτηριστικά κατανομής, με υψηλότερη **διάμεση τιμή (median)** και **ευρύτερο ενδοτεταρτημοριακό εύρος (interquartile range)** στο σύνολο των μετρικών. Το εύρημα αυτό υποδηλώνει ότι το **GIT-VATEX** επιτυγχάνει ανώτερη απόδοση όχι μόνο ως μέση τιμή, αλλά και στην πλειονότητα των εξεταζόμενων video, επιδεικνύοντας τη μεγαλύτερη συνέπεια μεταξύ των συγκρινόμενων μοντέλων. Αντιθέτως, το BLIP, παρά το πλεονέκτημα της ταχύτητας, καταγράφει χαμηλή διάμεση απόδοση με αξιοσημείωτη διασπορά, στοιχείο που υποδεικνύει αδυναμία διατήρησης σταθερών ποιοτικών προτύπων σε όλο το εύρος του συνόλου δεδομένων.

Το **Qwen2-VL** καταλαμβάνει ενδιάμεση θέση, με την απόδοσή του να φαίνεται πως συσχετίζεται ισχυρά με τον βαθμό alignment μεταξύ του prompt και των χαρακτηριστικών του dataset. Η υψηλή επίδοση στο **METEOR**, σε αντιδιαστολή με τις χαμηλές τιμές στις μετρικές **BLEU**, επιβεβαιώνει την τάση του μοντέλου να παράγει σημασιολογικά πλούσιες περιγραφές, οι οποίες ωστόσο δεν ευθυγραμμίζονται λεξιλογικά με τις επισημειώσεις αναφοράς (ground truths).

Τέλος, η παρουσία **ακραίων τιμών (outliers)** και στα τρία μοντέλα υπογραμμίζει την πολυπλοκότητα και ετερογένεια του Dense Video Captioning. Συγκεκριμένα, video με σαφείς δράσεις και λεξιλόγιο που προσιδιάζει στα δεδομένα εκπαίδευσης οδηγούν σε υψηλές βαθμολογίες, ενώ περιπτώσεις με περίπλοκες αλληλεπιδράσεις ή εξειδικευμένο λεξιλόγιο συνεπάγονται πτώση της απόδοσης. Η μεταβλητότητα αυτή αναδεικνύει την ανάγκη για περαιτέρω έρευνα προς την κατεύθυνση μοντέλων ικανών να προσαρμόζονται δυναμικά στο ειδικό domain και το υφολογικό στυλ του εκάστοτε video.

5.6 Συγκριτική Αξιολόγηση και Συζήτηση

Η συνολική αξιολόγηση των 3 μοντέλων αποκαλύπτει διακριτά profiles απόδοσης που καθιστούν το καθένα κατάλληλο για διαφορετικά σενάρια εφαρμογής. Το **Σχήμα 5.8** συνοψίζει τη συγκριτική απόδοση παρουσιάζοντας 5 κρίσιμες διαστάσεις αξιολόγησης σε μορφή radar chart.



Σχήμα 5.8: *Συγκριτικό Radar Chart Απόδοσης Μοντέλων*

Το radar chart του **Σχήματος 5.8** απεικονίζει 5 διαστάσεις που έχουν κανονικοποιηθεί στην κλίμακα 0-100. Η διάσταση Speed υπολογίζεται ως inverse ratio του χρόνου επεξεργασίας (ταχύτερο = υψηλότερο score), οι διαστάσεις **BLEU, METEOR και ROUGE-L** αντιπροσωπεύουν τις αντίστοιχες caption quality μετρικές, ενώ η διάσταση Consistency υπολογίζεται από τον αντίστροφο μέσο Coefficient of Variation (CV) των 4 caption quality μετρικών, εξετάζοντας μόνο τα επιτυχημένα matches με IoU≥0.3. Χαμηλότερο CV υποδηλώνει υψηλότερη σταθερότητα στην απόδοση σε διαφορετικά video. Τα συγκεντρωτικά στατιστικά του **Radar Chart** παρουσιάζονται αναλυτικά στον **Πίνακα 5.6**.

Πίνακας 5.6: Συγκεντρωτικά Στατιστικά για Radar Chart

Μοντέλο	Speed	BLEU	METEOR	ROUGE-L	Consistency	CV avg
BLIP	**100.0**	86.1	77.1	88.3	89.2	0.907
GIT-VATEX	4.1	**100.0**	93.2	**100.0**	**100.0**	**0.809**
QWEN2 -VL	3.1	94.8	**100.0**	92.6	93.8	0.862

Σημείωση: Όλες οι τιμές εκτός από CV είναι κανονικοποιημένες στο εύρος 0-100. Το CV υπολογίζεται μόνο σε HITs με IoU≥0.3.

5.6.1 Profiling Μοντέλων

Το GIT-VATEX αναδεικνύεται ως το πιο ισορροπημένο και αποδοτικό μοντέλο για dense video captioning. Όπως φαίνεται στο **Σχήμα 5.8**, το κόκκινο polygon του GIT-VATEX καλύπτει σχεδόν ολόκληρη την επιφάνεια του radar chart εκτός από τη διάσταση Speed. Επιτυγχάνει την υψηλότερη απόδοση σε 4 από τις 6 βασικές μετρικές του Πίνακα 5.1 (**Precision, BLEU-3, BLEU-4, ROUGE-L**). Η video-aware προσέγγιση του και η εξειδίκευση στο VaTeX dataset του επιτρέπουν να συλλαμβάνει τη δυναμική των γεγονότων με ακρίβεια. Ο χρόνος επεξεργασίας των 183.3s ανά video (Speed score 4.1) είναι αποδεκτός για εφαρμογές όπου η ποιότητα προηγείται της ταχύτητας. Το trade-off μεταξύ ταχύτητας και ποιότητας είναι ευνοϊκό, με τη βελτίωση **43%** στο **BLEU-3** έναντι του BLIP να δικαιολογεί την **24×** αύξηση στον χρόνο επεξεργασίας. Το χαμηλό **CV (0.809)** επιβεβαιώνει την προβλέψιμη και σταθερή συμπεριφορά του σε διαφορετικά video.

Το **BLIP** κατακτά την κατηγορία της ταχύτητας με χρόνο επεξεργασίας μόλις 7.6s ανά video. Το μπλε polygon του στο **Σχήμα 5.8** εμφανίζει ένα δραματικό spike στη διάσταση Speed ενώ όλες οι άλλες διαστάσεις είναι στο εύρος **77-89**, δημιουργώντας ένα ασύμμετρο σχήμα που αντικατοπτρίζει το extreme trade-off του. Η απόδοση του στις caption quality μετρικές είναι η χαμηλότερη, με **BLEU-3 μόλις 3.6%** και normalized **BLEU score 86.1**. Το **Consistency score 89.2** είναι το χαμηλότερο από τα 3 μοντέλα, με **CV=0.907**, υποδηλώνοντας μεγαλύτερη διασπορά στην απόδοση μεταξύ διαφορετικών video. Η frame-based προσέγγιση του, που αναλύει μόνο ένα keyframe ανά σκηνή, περιορίζει την ικανότητά του να κατανοήσει temporal context και δυναμικές δράσεις. Είναι κατάλληλο για εφαρμογές που απαιτούν γρήγορη επισκόπηση μεγάλων βιβλιοθηκών video όπου η ακρίβεια μπορεί να θυσιαστεί για την ταχύτητα.

Το **Qwen2-VL** παρουσιάζει το πιο ενδιαφέρον αλλά και αντιφατικό profile. Το πράσινο polygon του στο **Σχήμα 5.8** είναι παρόμοιο με του GIT-VATEX αλλά ελαφρώς μικρότερο σε όλες τις διαστάσεις εκτός από το **METEOR** όπου επιτυγχάνει το μεγαλύτερο score. Με **Speed score μόλις 3.1 (245.9s ανά video)**, είναι το πιο αργό μοντέλο, **32× πιο αργό από το BLIP** και **34% πιο αργό ακόμα και από το GIT-VATEX**. Επιτυγχάνει υψηλό **METEOR (17.03%)** και **Recall (56.08%)**, αλλά υστερεί στις **BLEU** μετρικές με **normalized score 94.8**, χαμηλότερο από το GIT-VATEX. Το **Consistency score 93.8** με **CV=0.862** το τοποθετεί ενδιάμεσα μεταξύ GIT-VATEX και BLIP. Η απόδοσή του

επηρεάζεται έντονα από την επιλογή prompt, με την τάση προς verbosity (μέσος μήκος 17.1 λέξεων) να την καθιστά ασύμβατη με τα ground truth annotations που έχουν μέσο μήκος 13.3 λέξεων. Το αντι-διαισθητικό φαινόμενο όπου τα E2E scores υπερβαίνουν τα oracle scores υποδηλώνει ότι το semantic merging module λειτουργεί ως implicit quality filter.

5.6.2 Θεμελιώδεις Περιορισμοί και Σημεία Βελτίωσης

Η χαμηλή απόδοση των εξεταζόμενων μοντέλων αναδεικνύει τις εγγενείς προκλήσεις του Dense Video Captioning, με πρωτεύον εμπόδιο την ανεπάρκεια του υποσυστήματος temporal localization, η οποία αποτελεί το κύριο bottleneck της αρχιτεκτονικής. Η αδυναμία της content-based scene detection να διακρίνει επαρκώς τις οπτικές αλλαγές από τα σημασιολογικά γεγονότα επιβεβαιώνεται από την ταυτόσημη συμπεριφορά όλων των μοντέλων, γεγονός που υποδεικνύει την ανάγκη στροφής προς μεθόδους που ενσωματώνουν learned temporal proposal networks ή αξιοποιούν multimodal signals. Παράλληλα, παρατηρείται σημαντικό semantic gap μεταξύ των ground truth annotations και των παραγόμενων περιγραφών, καθώς η αυστηρότητα των n-gram μετρικών τείνει να υποτιμά έγκυρες αλλά στυλιστικά διαφοροποιημένες διατυπώσεις. Το φαινόμενο αυτό καθίσταται εντονότερο στα instruction-tuned μοντέλα όπως το Qwen2-VL, όπου η ευαισθησία στη δομή των prompts εισάγει επιπλέον πολυπλοκότητα και απαιτεί εκτενή πειραματισμό, καθιστώντας τα λιγότερο plug-and-play συγκριτικά με task-specific λύσεις όπως το GIT-VATEX.

Συμπερασματικά, προκύπτει ότι περίπου το 80% της συνολικής απόκλισης στην απόδοση αποδίδεται σε σφάλματα του temporal localization, ενώ μόλις το 20% αφορά την ποιότητα του caption generation. Η ποσοτική αυτή διαπίστωση οδηγεί στο συμπέρασμα ότι η βελτιστοποίηση του συστήματος προϋποθέτει κατά κύριο λόγο την αναβάθμιση του temporal localization module και δευτερευόντως την αλλαγή του γλωσσικού μοντέλου.

5.7 Ποιοτική Ανάλυση και Παραδείγματα

Η ποιοτική ανάλυση συμπληρώνει τις αριθμητικές μετρικές, επιτρέποντας την ερμηνεία των σφαλμάτων ως προς τη χρονική εντόπιση, τη γλωσσική ακρίβεια και τη σημασιολογική πληρότητα των παραγόμενων λεζάντων. Παρότι δείκτες όπως **BLEU, METEOR** και **ROUGE** παρέχουν μια συνοπτική ποσοτική εικόνα, δεν αποτυπώνουν πλήρως το είδος των αποκλίσεων που εμφανίζονται σε πραγματικά σενάρια Dense Video Captioning, όπως semantic drift, υπερ-γενίκευση ή υπερβολική εξειδίκευση.

Για τον λόγο αυτό, στην παρούσα ενότητα εξετάζονται αναλυτικά δύο αντιπροσωπευτικά video, τα οποία χρησιμοποιούνται σε όλες τις **υποενότητες (oracle, end-to-end και semantic merging)**, ώστε να είναι δυνατή η άμεση σύγκριση των μοντέλων υπό διαφορετικές συνθήκες αξιολόγησης.

Το **Video A (Video_ID: 2q_4I3ae0J4)** αντιστοιχεί σε αθλητική δραστηριότητα hurling, όπου το dataset διαχωρίζει το γεγονός σε διαδοχικά υπο-γεγονότα (serve, throw, hit). Το παράδειγμα αυτό είναι απαιτητικό ως προς τη χρονική εντόπιση, καθώς η πλήρης κάλυψη όλων των annotated segments προϋποθέτει λεπτομερή temporal segmentation.

Το **Video B (Video ID: 90vop6PS2Y0)** αφορά επαναλαμβανόμενη δραστηριότητα καθαρισμού φύλλων σε εξωτερικό χώρο, όπου μικρές μεταβολές στο οπτικό περιεχόμενο μπορούν να οδηγήσουν σε υπερ-κατατμήσεις και παραγωγή false positives. Το συγκεκριμένο παράδειγμα είναι κατάλληλο για την ανάλυση της συμπεριφοράς των μοντέλων σε παρατεταμένα γεγονότα και επαναλαμβανόμενες δράσεις.

Στις επόμενες υποενότητες, τα δύο αυτά παραδείγματα θα αναφέρονται ως **Video A και Video B**, και θα χρησιμοποιούνται για την παρουσίαση των **oracle captions (Πίνακες**

5.7A–5.7B), των **end-to-end αποτελεσμάτων του pipeline (Πίνακες 5.7C–5.7H)**, καθώς και της επίδρασης του **semantic merging (Πίνακες 5.9A–5.9B)**.

5.7.1 Ποιοτική Σύγκριση Oracle στα Ground truth Timestamps

Στο oracle σενάριο, τα μοντέλα παράγουν λεζάντες απευθείας πάνω στα ground truth χρονικά τμήματα του dataset, χωρίς την επίδραση των σταδίων scene detection και semantic merging. Με αυτόν τον τρόπο απομονώνεται η καθαρή ικανότητα caption generation, επιτρέποντας την αξιολόγηση της γλωσσικής και σημασιολογικής ποιότητας των μοντέλων υπό ιδανικές συνθήκες χρονικής εντόπισης.

Στους **Πίνακες 5.7A–5.7B** παρουσιάζονται πλήρως όλα τα annotated segments δύο αντιπροσωπευτικών video, μαζί με τις αντίστοιχες ground truth περιγραφές και τις παραγόμενες λεζάντες των **BLIP, GIT-VATEX και Qwen2-VL**. Η ανάλυση αυτή αναδεικνύει τις διαφορές των μοντέλων ως προς την περιγραφική ακρίβεια, τη χρονική συνέπεια και την τάση για semantic drift ή υπερβολική εξειδίκευση.

Στον **Πίνακα 5.7A παρατίθενται όλα τα ground truth segments του Video A**, καθώς και οι αντίστοιχες oracle λεζάντες των τριών μοντέλων, επιτρέποντας άμεση σύγκριση της σημασιολογικής τους ακρίβειας.

Πίνακας 5.7A: Λεζάντες Oracle για όλα τα GT segments του Video A

Segment (s)	Captions
0.00–37.51	**GT:** A man plays hurling and serves a ball with a stick. **BLIP:** A group of men standing on top of a soccer field **GIT-VATEX:** A man is practicing his sport with a ball and a ball **Qwen2-VL:** A man in a green and white uniform is standing on a field, holding a stick and throwing it
38.12–80.55	**GT:** The man throws the ball in the air and immediately hit the ball with the stick. **BLIP:** A man in a green shirt playing a game of soccer **GIT-VATEX:** A man is holding a bat and then throws a ball at it **Qwen2-VL:** A person in a green and white uniform is holding a stick and ball, preparing to hit the ball
81.78–120.51	**GT:** The man throws the ball in the air, then swing his body to hit the ball with the stick. **BLIP:** A man kicking a soccer ball on a field **GIT-VATEX:** A man is practicing his sport with a stick and a ball **Qwen2-VL:** A person in a green and white uniform is playing hurling on a field, swinging a stick

Στο συγκεκριμένο video, το ground truth περιγράφει με σαφήνεια τη δραστηριότητα hurling και τη χρονική ακολουθία “throw then hit”. Το **BLIP** μετατοπίζεται προς γενικότερα αθλητικά συμφραζόμενα (“soccer”), γεγονός που είναι συμβατό με τη frame-based λειτουργία του και την πιθανή σύγχυση αντικειμένων δράσης σε μεμονωμένα στιγμιότυπα. Το **GIT-VATEX** αποδίδει καλύτερα τη δράση μέσω ρημάτων (“throws a ball”), ενώ το **Qwen2-VL** εμφανίζει πιο συνεπή θεματική αντιστοίχιση και πλουσιότερη σημασιολογική περιγραφή.

Στον **Πίνακα 5.7B παρουσιάζονται όλα τα annotated segments του Video B** και οι παραγόμενες oracle λεζάντες των BLIP, GIT-VATEX και Qwen2-VL, αναδεικνύοντας διαφορές μεταξύ πιο γενικών frame-based περιγραφών και πιο action-centric video-based outputs.

Πίνακας 5.7Β: Λεζάντες Oracle για όλα τα GT segments (Video ID: 90vop6PS2Y0)

Segment (s)	Captions
0.00–35.26	**GT:** A boy pushes a rake while riding a scooter down the driveway. **BLIP:** A man is cleaning up leaves in the yard **GIT-VATEX:** A man is using a broom to sweep leaves on the sidewalk **Qwen2-VL:** A person is using a rake to clear leaves from a sidewalk in a residential area
36.62–82.73	**GT:** It pushes the leaves in the driveway out to the road. **BLIP:** A man is standing in front of a garage **GIT-VATEX:** A boy is riding a scooter in the street and then he pushes a pile of leaves **Qwen2-VL:** A person is using a rake to clear leaves from a sidewalk
86.79–271.23	**GT:** He continues the pattern until the driveway is clean of debris. **BLIP:** A young boy sweeping leaves in front of a house **GIT-VATEX:** A boy is sweeping leaves on the ground in front of a house **Qwen2-VL:** A person is using a rake to clear leaves from a driveway

Η oracle αξιολόγηση δείχνει ότι τα μοντέλα συλλαμβάνουν το γενικό γεγονός “cleaning leaves”, αλλά συχνά χάνεται η λεπτομέρεια “rake while riding a scooter”. Το **GIT-VATEX** διατηρεί πιο action-centric περιγραφές, ενώ το **BLIP** σε ορισμένα segments εστιάζει σε στατικό περιβάλλον (“standing in front of a garage”).

5.7.2 End-to-End Pipeline Outputs και Αντιστοίχιση με Ground Truth

Στο end-to-end σενάριο, τα χρονικά τμήματα δεν αντιστοιχούν στα έτοιμα timestamps του dataset, αλλά παράγονται δυναμικά από το προτεινόμενο pipeline μέσω scene detection και post-processing. Κατά συνέπεια, η τελική απόδοση των μοντέλων δεν εξαρτάται μόνο από την ικανότητα παραγωγής λεζάντας, αλλά και από τη σωστή χρονική τμηματοποίηση των γεγονότων.

Η πλήρης παράθεση όλων των predicted segments επιτρέπει να αναδειχθούν με σαφήνεια οι περιπτώσεις όπου εμφανίζονται false positives (δηλαδή προβλεπόμενα γεγονότα χωρίς αντιστοίχιση σε ground truth) και false negatives (δηλαδή ground truth γεγονότα που δεν καλύφθηκαν από κάποιο predicted segment). Στην παρούσα ενότητα παρουσιάζονται αναλυτικά όλα τα outputs του pipeline για δύο αντιπροσωπευτικά video.

Στους **Πίνακες 5.7C–5.7E** παρουσιάζονται για το **Video A** αναλυτικά όλα τα χρονικά τμήματα που παρήγαγε το end-to-end pipeline για τα τρία μοντέλα **BLIP, GIT-VATEX και Qwen2-VL**, μαζί με τις παραγόμενες λεζάντες, την αντιστοίχιση με το ground truth και τις περιπτώσεις **HIT, false positives και false negatives**. Η σύγκριση αυτή επιτρέπει να αναδειχθεί κατά πόσο η temporal τμηματοποίηση του συστήματος καλύπτει πλήρως τα annotated γεγονότα του dataset.

BLIP End-to-End Output Video A

Πίνακας 5.7C: Όλα τα pipeline segments για BLIP στο Video A

Segment (s)	Captions
0.00–25.88	**Pred:** A group of people playing soccer on a field **Match:** HIT (IoU=0.69) **GT Ref:** 0–37.51 "A man plays hurling and serves a ball with a stick."
25.88–117.52	**Pred:** A man holding a baseball bat on top of a field **Match:** HIT (IoU=0.46) **GT Ref:** 38.12–80.55 "The man throws the ball in the air and immediately hit the ball…"
117.52–122.88	**Pred:** A woman kicking a soccer ball on a field **Match:** FALSE POSITIVE
—	**FALSE NEGATIVE:** GT 81.78–120.51 "The man swings his body to hit the ball with the stick."

Το **BLIP** επιτυγχάνει δύο αντιστοιχίσεις, όμως οι λεζάντες του παραμένουν σχετικά γενικές και εμφανίζουν semantic drift προς "soccer". Επιπλέον, το τρίτο ground truth γεγονός δεν καλύπτεται επαρκώς, οδηγώντας σε false negative, ενώ εμφανίζεται και ένα σύντομο false positive segment. Το μοτίβο αυτό είναι ενδεικτικό temporal under-coverage, όπου το pipeline δεν παράγει ξεχωριστό segment για κάθε annotated υπο-γεγονός.

GIT-VATEX End-to-End Output Video A

Πίνακας 5.7D: Όλα τα pipeline segments για GIT-VATEX στο Video A

Segment (s)	Captions
0.00–25.88	**Pred:** A man is holding a stick and throws a ball **Match:** HIT (IoU=0.69) **GT Ref:** 0–37.51 "A man plays hurling…"
25.88–117.52	**Pred:** A man is holding a bat and then throws a ball **Match:** HIT (IoU=0.46) **GT Ref:** 38.12–80.55 "The man throws the ball…"
117.52–122.88	**Pred:** A woman is kicking a soccer ball on a field **Match:** FALSE POSITIVE
—	**FALSE NEGATIVE:** GT 81.78–120.51 "The man swings his body…"

Το **GIT-VATEX** παρουσιάζει παρόμοια χρονική συμπεριφορά με το BLIP, ωστόσο οι περιγραφές του είναι πιο action-centric, αποδίδοντας με μεγαλύτερη σαφήνεια την ακολουθία "throws a ball". Παρ' όλα αυτά, το pipeline εξακολουθεί να μην καλύπτει το τρίτο υπο-γεγονός του ground truth, κάτι που υποδεικνύει ότι το bottleneck εδώ δεν είναι η caption generation αλλά η temporal segmentation.

Qwen2-VL End-to-End Output Video A

Πίνακας 5.7E: Όλα τα pipeline segments για Qwen2-VL στο Video A

Segment (s)	Captions
0.00–25.88	**Pred:** A person in a green and white uniform is playing hurling on a field **Match:** HIT (IoU=0.69) **GT Ref:** 0–37.51 "A man plays hurling…"
117.52–122.88	**Pred:** A person is kicking a soccer ball on a field **Match:** FALSE POSITIVE
—	**FALSE NEGATIVE:** GT 38.12–80.55 "The man throws the ball…"
—	**FALSE NEGATIVE:** GT 81.78–120.51 "The man swings his body…"

Το **Qwen2-VL** αποδίδει πολύ σωστά το θεματικό πλαίσιο ("hurling"), όμως εμφανίζει μειωμένη temporal κάλυψη των επόμενων ground truth γεγονότων. Έτσι, παρότι στο oracle σενάριο παράγει πιο πλούσιες περιγραφές, στο end-to-end pipeline η απόδοσή του περιορίζεται από την έλλειψη επαρκών predicted segments, οδηγώντας σε περισσότερα false negatives.

Στους **Πίνακες 5.7F–5.7H** παρουσιάζονται για το **Video B** όλα τα **predicted segments του pipeline για τα BLIP, GIT-VATEX και Qwen2-VL** αντίστοιχα. Η πλήρης αυτή παράθεση επιτρέπει να φανεί με σαφήνεια η διαφορετική συμπεριφορά των μοντέλων ως προς την παραγωγή πλεονάζοντων χρονικών τμημάτων, καθώς και η επίδραση της επαναληπτικής φύσης της δράσης στην precision του συστήματος.

BLIP End-to-End Output Video B

Πίνακας 5.7F: Όλα τα pipeline segments για BLIP στο Video B

Segment (s)	Captions
0.00–21.00	**Pred:** A man and his dog are playing in the yard **Match:** HIT (IoU=0.60) **GT Ref:** 0–35.26 "A boy pushes a rake…"
21.00–24.93	**Pred:** A man is sweeping leaves on the street **Match:** FALSE POSITIVE
24.93–165.03	**Pred:** A man in a white shirt is sweeping leaves on the sidewalk **Match:** HIT (IoU=0.33) **GT Ref:** 36.62–82.73 "It pushes the leaves…"
165.03–191.33	**Pred:** A young boy playing with leaves **Match:** FALSE POSITIVE
191.33–271.23	**Pred:** A person is walking down the street in front of a house **Match:** HIT (IoU=0.43) **GT Ref:** 86.79–271.23 "He continues the pattern…"

Το **BLIP** εμφανίζει αρκετά false positives, γεγονός που υποδηλώνει ότι μικρές μεταβολές σκηνής ή ενδιάμεσα πλάνα οδηγούν σε παραγωγή νέων captions. Επιπλέον, οι περιγραφές του μετατοπίζονται σε δευτερεύουσες δραστηριότητες ("playing with dog"), μειώνοντας τη σημασιολογική ακρίβεια.

GIT-VATEX End-to-End Output Video B

Πίνακας 5.7G: Όλα τα pipeline segments για GIT-VATEX στο Video B

Segment (s)	Captions
0.00–24.93	**Pred:** A man is using a shovel to sweep leaves off the sidewalk **Match:** HIT (IoU=0.71) **GT Ref:** 0–35.26 "A boy pushes a rake…"
24.93–165.03	**Pred:** A boy is using a shovel to sweep the street **Match:** HIT (IoU=0.33) **GT Ref:** 36.62–82.73 "It pushes the leaves…"
165.03–191.33	**Pred:** A boy is pushing leaves with a shovel **Match:** FALSE POSITIVE
191.33–271.23	**Pred:** A boy is sweeping leaves on the ground in front of a house **Match:** HIT (IoU=0.43) **GT Ref:** 86.79–271.23 "He continues the pattern…"

Το **GIT-VATEX** παράγει πιο συγκρατημένο αριθμό segments, με λιγότερα false positives σε σχέση με το BLIP. Ωστόσο, εμφανίζει συστηματική αντικατάσταση του "rake" με "shovel", κάτι που δείχνει semantic approximation: η δράση αποδίδεται σωστά, αλλά το εργαλείο παρερμηνεύεται.

Qwen2-VL End-to-End Output Video B

Πίνακας 5.7H: Όλα τα pipeline segments για Qwen2-VL στο Video B

Segment (s)	Captions
0.00–21.00	**Pred:** Two individuals are raking leaves in a residential area **Match:** HIT (IoU=0.60) **GT Ref:** 0–35.26 "A boy pushes a rake…"
21.00–24.93	**Pred:** A person is using a red rake to clear leaves **Match:** FALSE POSITIVE
24.93–165.03	**Pred:** A person is raking leaves in front of a house **Match:** HIT (IoU=0.33) **GT Ref:** 36.62–82.73 "It pushes the leaves…"
165.03–191.33	**Pred:** A child is playing with leaves **Match:** FALSE POSITIVE
191.33–271.23	**Pred:** A person is using a rake to clear leaves from a driveway **Match:** HIT (IoU=0.43) **GT Ref:** 86.79–271.23 "He continues the pattern…"

Το **Qwen2-VL** διατηρεί υψηλή σημασιολογική πληρότητα και πιο περιγραφικό ύφος, αλλά παράγει επίσης false positives σε σύντομα ενδιάμεσα segments. Το μοτίβο αυτό υποδηλώνει ότι σε end-to-end ροές, ακόμη και μικρές αλλαγές κάδρου μπορούν να οδηγήσουν σε παραγωγή νέας λεξάντας, χωρίς να υπάρχει αντίστοιχο annotated γεγονός.

Τα παραπάνω παραδείγματα καταδεικνύουν ότι η τελική απόδοση στο end-to-end Dense Video Captioning καθορίζεται σε μεγάλο βαθμό από το temporal segmentation. Στο 2q_4I3ae0J4, το κύριο πρόβλημα είναι η ατελής κάλυψη όλων των ground truth υπο-γεγονότων, οδηγώντας σε false negatives. Στο 90vop6PS2Y0, το κυρίαρχο πρόβλημα είναι η παραγωγή επιπλέον segments που περιγράφουν δευτερεύουσες μικρο-μεταβάσεις, αυξάνοντας τα false positives.

Η σύγκριση των τριών μοντέλων δείχνει ότι οι video-based προσεγγίσεις (GIT-VATEX, Qwen2-VL) παράγουν πιο action-centric και σημασιολογικά πλούσιες περιγραφές, αλλά δεν

εξαλείφουν πλήρως τα localization errors που προκύπτουν από το segmentation στάδιο. Αυτό εξηγεί και το παρατηρούμενο gap μεταξύ oracle και end-to-end performance.

5.7.3 Ανάλυση Επιτυχημένων και Αποτυχημένων Περιγραφών

Η δεύτερη διάσταση της ποιοτικής ανάλυσης εξετάζει συγκεκριμένα παραδείγματα όπου τα μοντέλα επιτυγχάνουν ή αποτυγχάνουν να συλλάβουν το νόημα του γεγονότος, ανεξάρτητα από τη χρονική αντιστοίχιση. Παρότι οι περιπτώσεις HIT αποτυπώνουν την επιτυχία ως προς το IoU matching, η ποιότητα της παραγόμενης λεξάντας μπορεί να διαφέρει σημαντικά, από πλήρη περιγραφική ακρίβεια έως semantic drift ή πλήρη αποτυχία.

Ο **Πίνακας 5.8** κατηγοριοποιεί αντιπροσωπευτικά παραδείγματα βάσει του τύπου επιτυχίας ή αποτυχίας, αναδεικνύοντας χαρακτηριστικά μοτίβα των μοντέλων, όπως υπερ-γενίκευση στις frame-based προσεγγίσεις, action-centric πλεονεκτήματα στα video-based μοντέλα, αλλά και δυσκολίες σε σύνθετες κοινωνικές σκηνές με clutter και γρήγορες μεταβάσεις.

Πίνακας 5.8: Κατηγοριοποίηση Ποιότητας Περιγραφών

Category	Description
✓ Πλήρης Επιτυχία	**GT:** A race starts and people are running. **BLIP (✗):** A large crowd of people walking down the street **GIT-VATEX (✓):** A group of people are running in a marathon and cheering for the camera **Qwen2-VL (✓):** A woman in a fairy costume holds a sign that says SeaWheeze and waves to the camera **Analysis:** Το GIT-VATEX συλλαμβάνει την κύρια δράση (running, marathon). Το Qwen υπερ-εξειδικεύει με λεπτομέρειες που δεν υπάρχουν στο GT αλλά είναι σημασιολογικά σχετικές, καθώς στο μεγαλύτερο μέρος του video οι συμμετέχοντες εμφανίζονται να χορεύουν στο πλαίσιο του μαραθωνίου.
≈ Μερική Επιτυχία	**GT:** The coach helps the guy in red with the proper body placement and lifting technique. **BLIP (✗):** A man squats on a bench in a gym **GIT-VATEX (≈):** A man is demonstrating how to do leg exercises with a trainer **Qwen2-VL (≈):** A man in a red shirt is performing a leg exercise on a machine while a trainer assists him **Analysis:** Και τα δύο video-aware models αναγνωρίζουν τη σχέση trainer–trainee αλλά χάνουν το instructional focus του coaching.
✗ Semantic Drift	**GT:** Several shots of boats are shown riding around. **BLIP (✗):** A view of the water from a boat on a sunny day **GIT-VATEX (✗):** A person is swimming in the ocean and there are fish swimming around **Qwen2-VL (≈):** A person wearing scuba gear is exploring a coral reef

	Analysis: Το BLIP εστιάζει στο περιβάλλον. Το GIT-VATEX παρερμηνεύει τη δραστηριότητα (boating→swimming). Το Qwen συλλαμβάνει σχετική αλλά όχι ακριβή δραστηριότητα (underwater exploration).
✗ Πλήρης Αποτυχία	**GT:** People are dancing having fun at a party. **BLIP (✗):** A man in a striped shirt is holding a knife **GIT-VATEX (✗):** A group of people are performing a trick with a bow tie **Qwen2-VL (✗):** A man in a striped shirt performs with a baton in a city park, while a crowd watches **Analysis:** Όλα τα μοντέλα αποτυγχάνουν σε πολύπλοκες κοινωνικές σκηνές με γρήγορες μεταβάσεις και clutter.

Τα παραπάνω παραδείγματα επιβεβαιώνουν ότι η επιτυχία στο Dense Video Captioning δεν εξαρτάται αποκλειστικά από τη χρονική αντιστοίχιση των segments, αλλά και από τη σημασιολογική ακρίβεια και πληρότητα της παραγόμενης περιγραφής. Ακόμη και όταν ένα predicted segment θεωρείται επιτυχές ως προς το IoU matching, η γλωσσική απόδοση μπορεί να αποκλίνει σημαντικά από το ground truth, οδηγώντας σε μερική επιτυχία ή semantic drift.

Το **BLIP**, ως frame-based προσέγγιση, εμφανίζει συχνότερα τάση υπερ-γενίκευσης ή εστίασης στο οπτικό περιβάλλον αντί στη βασική δράση, γεγονός που εξηγεί περιπτώσεις όπου η περιγραφή παραμένει επιφανειακή ή άσχετη με το κεντρικό γεγονός. Αντίθετα, το **GIT-VATEX**, αξιοποιώντας χρονική πληροφορία από πολλαπλά frames, παράγει πιο action-centric λεζάντες και συλλαμβάνει καλύτερα τη βασική δυναμική των γεγονότων, ιδιαίτερα σε σκηνές με σαφή κινητική δραστηριότητα (π.χ. running, training).

Το **Qwen2-VL** παρουσιάζει γενικά υψηλότερη σημασιολογική πληρότητα και πιο πλούσιο περιγραφικό ύφος, αποδίδοντας συχνά λεπτομέρειες που δεν αναφέρονται ρητά στο ground truth αλλά παραμένουν συναφείς με το οπτικό περιεχόμενο. Ωστόσο, αυτή η ικανότητα μπορεί να οδηγήσει σε υπερ-εξειδίκευση ή σε παραγωγή περιγραφών που αποκλίνουν από την επισημείωση του dataset, ιδιαίτερα σε περιπτώσεις όπου το annotation είναι συνοπτικό ή αφαιρετικό.

Τέλος, οι περιπτώσεις πλήρους αποτυχίας αναδεικνύουν ότι ακόμη και ισχυρά Vision-Language Models δυσκολεύονται σε πολύπλοκες κοινωνικές σκηνές με έντονο clutter, πολλαπλές ταυτόχρονες αλληλεπιδράσεις και γρήγορες μεταβάσεις. Συνεπώς, η ποιοτική ανάλυση υποδεικνύει ότι η βελτίωση της απόδοσης απαιτεί όχι μόνο καλύτερο temporal localization, αλλά και πιο robust σημασιολογική κατανόηση σύνθετων γεγονότων.

5.7.4 Επίδραση Semantic Merging στην Ποιότητα των Περιγραφών

Ένα κρίσιμο στάδιο του προτεινόμενου end-to-end pipeline είναι το semantic merging, το οποίο εφαρμόζεται ως post-processing βήμα μετά το scene detection και την παραγωγή λεζάντων. Στόχος του είναι η μείωση της πλεονασματικότητας των παραγόμενων captions, συγχωνεύοντας διαδοχικά χρονικά τμήματα όταν οι περιγραφές τους παρουσιάζουν υψηλή σημασιολογική ομοιότητα.

Η ανάγκη για semantic merging προκύπτει από το γεγονός ότι τα scene detection modules συχνά οδηγούν σε υπερ-κατατμήσεις, ειδικά σε video με σταδιακές μεταβολές ή επαναλαμβανόμενες δράσεις. Σε τέτοιες περιπτώσεις, το pipeline μπορεί να δημιουργήσει πολλαπλά μικρά segments που περιγράφουν ουσιαστικά το ίδιο γεγονός, οδηγώντας σε αυξημένα false positives και μειωμένη συνοχή στην τελική έξοδο.

Για τον καθορισμό των συγχωνεύσεων εφαρμόζεται ένα σύνθετο μετρικό σύστημα που αξιολογεί κατά **70% την ομοιότητα των centroids** και κατά **30% την ομοιότητα με την προηγούμενη περιγραφή (context similarity)**. Η επιλογή ενός αυστηρού ορίου σε αυτή τη διαδικασία διασφαλίζει την αποφυγή υπερβολικών συγχωνεύσεων, συγκεκριμένα το όριο που χρησιμοποιήθηκε είναι **merging_threshold = 0.75**. Η αξία αυτής της στρατηγικής αποτυπώνεται στα παραδείγματα που ακολουθούν, όπου το semantic merging επιτυγχάνει αισθητή μείωση της πλεονασματικότητας και ενίσχυση της συνοχής.

Στην περίπτωση του Video A, το οποίο καταγράφει στιγμιότυπα αθλητικής δραστηριότητας (hurling), η δράση χαρακτηρίζεται από την παρουσία διαδοχικών segments που περιγράφουν παρεμφερείς κινήσεις με ελάχιστες χρονικές αποκλίσεις. Η υψηλή σημασιολογική επικάλυψη των λεζαντών, που προέκυψε από την αρχική ανίχνευση σκηνών, ενεργοποίησε τον μηχανισμό συγχώνευσης. Τα τμήματα που ενοποιήθηκαν παρατίθενται στον **Πίνακα 5.9A**, συνοδευόμενα από τις τιμές ομοιότητας (similarity scores) που τεκμηριώνουν την εγκυρότητα της απόφασης.

Πίνακας 5.9A: Semantic merging και similarity scores του Video A

Segment (s)	**Captions**
S1: 0.00–25.88	**Caption:** A man is holding a stick and throws a ball
S2: 25.88–117.52	**Caption:** A man is holding a bat and then throws a ball
S1 vs S2	**Pairwise Similarities:** Scentroid (S1, S2) = 0.89 Scontext (S1, S2) = 0.85 Total Score: $S=0.7 \cdot 0.89 + 0.3 \cdot 0.85 = 0.878$ $0.878 \geq 0.75 \rightarrow$ **Merge ✓**
Merged Output	**Final Caption:** A man is holding a bat and then throws a ball

Η συγχώνευση στον **Πίνακα 5.9A** ενεργοποιήθηκε επειδή τα δύο διαδοχικά captions περιγράφουν ουσιαστικά την ίδια αθλητική ενέργεια (“throws a ball”), με πολύ υψηλή σημασιολογική επικάλυψη. Οι τιμές **centroid και context similarity** οδηγούν σε συνολικό **weighted score $S = 0.7 \cdot 0.89 + 0.3 \cdot 0.85 = 0.878$**, το οποίο υπερβαίνει το αυστηρό **threshold 0.75**. Έτσι, το pipeline συγχωνεύει τα δύο segments σε μία ενιαία περιγραφή, μειώνοντας την πλεονασματικότητα και ενισχύοντας τη συνοχή της τελικής εξόδου χωρίς να χάνεται ουσιαστική πληροφορία.

Στην περίπτωση του Video B, το οπτικό περιεχόμενο αφορά διεργασίες καθαρισμού φύλλων, με τη ροή της δράσης να είναι συνεχής και να στερείται ευδιάκριτων ορίων μετάβασης (shot boundaries). Η συγκεκριμένη δομή οδήγησε το σύστημα scene detection σε κατακερματισμό του video και στην παραγωγή πλεονασματικών segments με σημασιολογικά ισοδύναμο σχολιασμό. Η διόρθωση αυτής της πλεονασματικότητας τεκμηριώνεται στον **Πίνακα 5.9B**, όπου παρουσιάζεται η σύμπτυξη δύο λεζαντών υψηλής συσχέτισης σε μία ενιαία αφηγηματική μονάδα.

Πίνακας 5.9B: Semantic merging και similarity scores του Video B

Segment (s)	**Captions**
S2: 24.93–165.03	**Caption:** A person is raking leaves in front of a house
S3: 191.33–271.23	**Caption:** A person is using a rake to clear leaves from a driveway
S2 vs S3	**Pairwise Similarities:** Scentroid (S2, S3) = 0.91 Scontext (S2, S3) = 0.88 Total Score: $S = 0.7 \cdot 0.91 + 0.3 \cdot 0.88 = 0.901$ $0.901 \geq 0.75 \rightarrow$ **Merge ✓**

Merged Output	**Final Caption:** A person is using a rake to clear leaves from a driveway

Στον **Πίνακα 5.9Β**, η διαδικασία merging εφαρμόστηκε σε δύο segments που αφορούν επαναλαμβανόμενη δράση καθαρισμού φύλλων. Τα captions παρουσιάζουν πολύ υψηλή ομοιότητα τόσο σε centroid όσο και σε context επίπεδο, με **συνολικό score S = 0.7 · 0.91 + 0.3 · 0.88 = 0.901 μεγαλύτερο από το threshold ενεργοποίησης 0.75**. Η συγχώνευση οδηγεί σε μία πιο συμπαγή και συνεκτική περιγραφή ενός παρατεταμένου γεγονότος, περιορίζοντας την παραγωγή πολλαπλών σχεδόν ταυτόσημων captions και συμβάλλοντας στη μείωση false positives που οφείλονται σε υπερ-κατατμήσεις.

Τα παραπάνω παραδείγματα καταδεικνύουν ότι το **semantic merging αποτελεί αποτελεσματικό εργαλείο post-processing**, ιδιαίτερα σε περιπτώσεις όπου το scene detection παράγει **επαναλαμβανόμενα ή υπερβολικά λεπτά segments**. Η μείωση redundancy οδηγεί σε πιο συμπαγή και συνεκτική τελική έξοδο, συμβάλλοντας σε καλύτερη ποιότητα παρουσίασης των captions και σε περιορισμό false positives.

Επιπλέον, η χρήση αυστηρού similarity threshold καθιστά τη διαδικασία merging συντηρητική, διασφαλίζοντας ότι οι συγχωνεύσεις πραγματοποιούνται μόνο όταν τα captions παρουσιάζουν πολύ υψηλή σημασιολογική επικάλυψη. Συνεπώς, το semantic merging ενισχύει τη συνοχή του pipeline χωρίς να υποβαθμίζει σημαντικά τη χρονική διάκριση των γεγονότων.

5.8 Συνοπτική Παρουσίαση Πειραματικών Ευρημάτων

Η ολοκληρωμένη πειραματική αποτίμηση του προτεινόμενου συστήματος Dense Video Captioning στο σύνολο δεδομένων ActivityNet Captions ανέδειξε τόσο τις δυνατότητες όσο και τους περιορισμούς μιας pipeline-based προσέγγισης που αξιοποιεί σύγχρονα pretrained vision-language μοντέλα χωρίς πρόσθετο fine-tuning. Η παρούσα ενότητα συνοψίζει τα κύρια ευρήματα της αξιολόγησης, παρέχοντας μια συγκεντρωτική εικόνα της απόδοσης των τριών μοντέλων BLIP, GIT-VATEX και Qwen2-VL υπό oracle και end-to-end συνθήκες.

Σε **επίπεδο συγκριτικής απόδοσης**, το **GIT-VATEX αναδείχθηκε ως το πλέον ισορροπημένο μοντέλο για το εξεταζόμενο task**, επιτυγχάνοντας τις καλύτερες επιδόσεις στις περισσότερες λεξικοκεντρικές μετρικές αξιολόγησης. Συγκεκριμένα, παρουσίασε **Precision 21.03%**, **Recall 55.79%**, **BLEU-3 5.17%** και **ROUGE-L 22.28%**, γεγονός που υποδηλώνει σταθερότητα και συνέπεια σε διαφορετικά είδη video. Η video-aware αρχιτεκτονική του, η οποία επεξεργάζεται πολλαπλά frames ανά σκηνή, επιτρέπει αποτελεσματικότερη κατανόηση της χρονικής δυναμικής των γεγονότων, ενώ το fine-tuning στο VATEX dataset ενισχύει την ικανότητά του να αποδίδει δράσεις και δραστηριότητες.

Το **BLIP**, από την άλλη πλευρά, **προσέφερε τη μεγαλύτερη υπολογιστική αποδοτικότητα, με μέσο χρόνο επεξεργασίας 7.6s ανά video, δηλαδή περίπου 24 φορές ταχύτερο από το GIT-VATEX.** Το χαρακτηριστικό αυτό καθιστά το BLIP ιδιαίτερα ελκυστικό για εφαρμογές μεγάλης κλίμακας ή real-time indexing. Ωστόσο, η frame-based φύση του περιορίζει την ικανότητά του να ενσωματώνει temporal context, οδηγώντας σε χαμηλότερες επιδόσεις στις μετρικές ποιότητας **BLEU-3 3.60%**, **METEOR 13.13%** και σε πιο γενικές, object-centric περιγραφές.

Το **Qwen2-VL παρουσίασε ένα πιο σύνθετο προφίλ απόδοσης**. Πέτυχε το υψηλότερο **METEOR score 17.03%**, αναδεικνύοντας ισχυρές δυνατότητες σημασιολογικής κατανόησης. Παράλληλα όμως, υστέρησε στις **BLEU μετρικές λόγω της τάσης του να παράγει εκτενείς και περιγραφικά πλούσιες λεζάντες,** οι οποίες αποκλίνουν λεξιλογικά από τα συνοπτικά ground truth captions. Επιπλέον, **αποτέλεσε το πιο αργό μοντέλο με μέσο χρόνο 245.9s ανά video**, περιορίζοντας την πρακτική του αξιοποίηση σε περιβάλλοντα περιορισμένων πόρων.

Ένα από τα σημαντικότερα συμπεράσματα της αξιολόγησης αφορά τον ρόλο της χρονικής εντόπισης γεγονότων ως κεντρικού bottleneck. Η ανάλυση των temporal localization curves έδειξε ότι η συνολική απόκλιση από state-of-the-art συστήματα οφείλεται σε μεγάλο βαθμό στην ανεπαρκή οριοθέτηση των γεγονότων. Με **Precision@50 (tIoU=0.5) μόλις 12.86% και εμφανές over-segmentation pattern, το pipeline παράγει κατά μέσο όρο 2.7 φορές περισσότερες προβλέψεις από τα αντίστοιχα ground truth events**. Το content-based scene detection, παρότι απλό και αποδοτικό, εντοπίζει κυρίως οπτικές μεταβολές χωρίς ρητή σημασιολογική κατανόηση, γεγονός που περιορίζει την ακρίβεια temporal localization.

Η σύγκριση oracle και end-to-end αξιολόγησης επέτρεψε την ποσοτικοποίηση της επίδρασης των επιμέρους components. Στο oracle setting, όπου η χρονική εντόπιση θεωρείται ιδανική, απομονώνεται η καθαρή ικανότητα caption generation. Αντίθετα, στο end-to-end σενάριο, η απόδοση υποβαθμίζεται από σφάλματα temporal segmentation. Το **GIT-VATEX** ακολουθεί το αναμενόμενο μοτίβο Oracle > E2E, επιβεβαιώνοντας ότι τα localization errors μειώνουν την τελική ποιότητα. Αντιθέτως, για τα **BLIP και Qwen2-VL** παρατηρήθηκαν περιπτώσεις όπου η end-to-end διαδικασία, σε συνδυασμό με το semantic merging, λειτουργεί ως μηχανισμός φιλτραρίσματος πλεονασματικών captions, οδηγώντας σε οριακή βελτίωση της μέσης ποιότητας.

Η ποιοτική ανάλυση του Κεφαλαίου 5.7 ανέδειξε επίσης διακριτά μοντελο-ειδικά patterns. Το **BLIP παράγει συχνότερα στατικές και περιβαλλοντικές περιγραφές**, το **GIT-VATEX αποδίδει πιο action-oriented captions με καλύτερη χρονική συνέπεια**, ενώ το **Qwen2-VL προσφέρει πλουσιότερες αλλά συχνά verbose περιγραφές**. Τα ευρήματα αυτά εξηγούν τις διαφοροποιήσεις μεταξύ λεξικοκεντρικών και σημασιολογικών μετρικών και υπογραμμίζουν την ανάγκη επιλογής evaluation framework ανάλογα με τον στόχο της εφαρμογής.

Συνολικά, η πειραματική αποτίμηση επιβεβαίωσε ότι τα σύγχρονα pretrained vision-language μοντέλα διαθέτουν επαρκείς γλωσσικές και σημασιολογικές ικανότητες για Dense Video Captioning ακόμη και χωρίς task-specific εκπαίδευση. Παράλληλα όμως, ανέδειξε ότι η κύρια πρόκληση μετατοπίζεται στο temporal localization, υποδεικνύοντας πως μελλοντικές προσεγγίσεις που συνδυάζουν learned temporal proposal modules με ισχυρά pretrained captioning models αποτελούν μία ιδιαίτερα υποσχόμενη κατεύθυνση για περαιτέρω έρευνα.

ΚΕΦΑΛΑΙΟ 6 Συμπεράσματα και μελλοντικές προεκτάσεις

6.1 Συμπεράσματα

Στο πλαίσιο της παρούσας εργασίας αναπτύχθηκε και αξιολογήθηκε ένα ολοκληρωμένο modular σύστημα Dense Video Captioning, το οποίο βασίζεται στη χρήση σύγχρονων pretrained Vision-Language Models σε zero-shot setting, χωρίς την ανάγκη περαιτέρω εκπαίδευσης στο ActivityNet Captions dataset. Κεντρικός άξονας της έρευνας αποτέλεσε η συγκριτική μελέτη τριών διακριτών προσεγγίσεων παραγωγής περιγραφών, οι οποίες διαφοροποιούνται ως προς την αρχιτεκτονική και τη διαχείριση της οπτικής πληροφορίας. Συγκεκριμένα, εξετάστηκε η frame-based προσέγγιση μέσω του μοντέλου BLIP που παράγει κείμενο από μεμονωμένα keyframes, η video-aware Transformer αρχιτεκτονική του GIT-VATEX που ενσωματώνει χρονική πληροφορία επεξεργαζόμενη πολλαπλά frames, καθώς και η αξιοποίηση ενός large-scale Vision-Language Model μέσω του Qwen2-VL, το οποίο χαρακτηρίζεται από αυξημένη σημασιολογική πληρότητα. Μέσω της συγκεκριμένης συγκριτικής ανάλυσης κατέστη δυνατή η διερεύνηση της ισορροπίας μεταξύ υπολογιστικής αποδοτικότητας, χρονικής συνοχής και γλωσσικής ακρίβειας, αναδεικνύοντας αφενός τη δυνατότητα αποτελεσματικής εφαρμογής μοντέλων γενικού σκοπού σε προβλήματα περιγραφής video και αφετέρου τα πλεονεκτήματα αλλά και τους περιορισμούς μιας στρατηγικής τύπου pipeline.

Η πειραματική αποτίμηση κατέδειξε ότι τα pretrained captioning models διαθέτουν την απαιτούμενη γλωσσική και σημασιολογική επάρκεια για την παραγωγή περιγραφών που είναι συγκρίσιμες με εκείνες εξειδικευμένων supervised συστημάτων. Ειδικότερα, το GIT-VATEX επέδειξε την πλέον σταθερή συμπεριφορά στις καθιερωμένες μετρικές αξιολόγησης, παράγοντας action-oriented captions με συνέπεια σε ποικίλα είδη γεγονότων. Από την άλλη πλευρά, το BLIP επιβεβαίωσε τη σημασία της υπολογιστικής αποδοτικότητας προσφέροντας ταχύτερη επεξεργασία εις βάρος όμως της βαθύτερης χρονικής κατανόησης, ενώ το Qwen2-VL ξεχώρισε για το πλούσιο περιγραφικό ύφος και τη σημασιολογική πληρότητα, αν και με αυξημένο υπολογιστικό κόστος και μια τάση για παραγωγή verbose captions. Επιπροσθέτως, η έρευνα ανέδειξε ότι το κυριότερο bottleneck του συστήματος εντοπίζεται στη διαδικασία του temporal localization. Η απόκλιση που παρατηρήθηκε μεταξύ της oracle και της end-to-end αξιολόγησης επιβεβαίωσε ότι τα σφάλματα στον χρονικό εντοπισμό επηρεάζουν καθοριστικά την τελική απόδοση ανεξαρτήτως της ποιότητας του captioning model, μετατοπίζοντας έτσι το ερευνητικό ενδιαφέρον στην ανάγκη αποτελεσματικότερης κατανόησης της χρονικής δομής. Τέλος, η διαδικασία του semantic merging αποδείχθηκε καθοριστική για τη βελτίωση της συνοχής της εξόδου, υπογραμμίζοντας ότι οι προσεκτικά σχεδιασμένες post-processing στρατηγικές δύνανται να αντισταθμίσουν αποτελεσματικά τις αδυναμίες της αρχικής temporal segmentation.

6.2 Περιορισμοί της Προσέγγισης

Παρά τα ενθαρρυντικά αποτελέσματα, η παρούσα έρευνα υπόκειται σε συγκεκριμένους περιορισμούς που απορρέουν τόσο από την υιοθέτηση ενός modular pipeline σχεδιασμού όσο και από την εγγενή πολυπλοκότητα του προβλήματος.

Αρχικά, η διαδικασία του temporal localization βασίστηκε στη μέθοδο του content-based scene detection, η οποία ανιχνεύει μεταβολές στο οπτικό περιεχόμενο χωρίς να διαθέτει ρητή σημασιολογική κατανόηση γεγονότων. Ως αποτέλεσμα, παρατηρήθηκε over-segmentation, με το σύστημα να παράγει περισσότερα predicted segments από τα αντίστοιχα ground truth events. Το γεγονός αυτό συνέβαλε στην αύξηση των false positives

και τον επακόλουθο περιορισμό του precision του pipeline, ανεξαρτήτως της γλωσσικής ορθότητας των παραγόμενων λεξάντων.

Επιπροσθέτως, η εφαρμογή των μοντέλων σε zero-shot setting χωρίς fine-tuning στο ActivityNet Captions, αν και καταδεικνύει την ισχύ γενίκευσης των pretrained Vision-Language Models, περιόρισε την προσαρμογή τους στα ειδικά annotation conventions του συνόλου δεδομένων όπου τα ground truth captions διακρίνονται συνήθως για τη συντομία και την αφαιρετικότητά τους. Η συγκεκριμένη συνθήκη αιτιολογεί τις περιπτώσεις όπου μοντέλα υψηλής σημασιολογικής αντίληψης, όπως το Qwen2-VL, παράγουν πλούσιες περιγραφές που αποκλίνουν λεξικολογικά από τις πρότυπες επισημειώσεις επηρεάζοντας αρνητικά την επίδοση στις n-gram based μετρικές αξιολόγησης.

Παράλληλα, η αποτίμηση των παραγόμενων περιγραφών στηρίχθηκε κατά κύριο λόγο σε αυτοματοποιημένες μετρικές όπως οι BLEU, METEOR και ROUGE-L. Παρότι χρήσιμες για συγκριτικούς σκοπούς, οι μετρικές αυτές αδυνατούν να αποτυπώσουν πλήρως την ανθρώπινη αντίληψη ποιότητας ειδικότερα σε multimodal generative tasks όπου η ποικιλομορφία στην έκφραση είναι αναμενόμενη.

Τέλος, οι υπολογιστικοί περιορισμοί επηρέασαν την έκταση των πειραμάτων, καθώς τα μεγαλύτερα Vision-Language Models απαιτούν σημαντικούς πόρους και χρόνο επεξεργασίας, γεγονός που περιορίζει την πρακτική αξιοποίηση τους σε large-scale Dense Video Captioning εφαρμογές.

6.3 Μελλοντικές Προεκτάσεις

Τα ευρήματα της παρούσας εργασίας αναδεικνύουν πολλαπλές προοπτικές για τη μελλοντική εξέλιξη και βελτιστοποίηση του συστήματος. Μία ιδιαίτερα υποσχόμενη κατεύθυνση αφορά την υιοθέτηση υβριδικών αρχιτεκτονικών temporal localization. Η ενσωμάτωση ενός lightweight pretrained μοντέλου εντοπισμού δράσεων, όπως τα **ActionFormer** ή **BMN**, ως αρχικό στάδιο επεξεργασίας, θα μπορούσε να εξασφαλίσει ακριβέστερο προσδιορισμό των γεγονότων, επιτρέποντας στο semantic merging να λειτουργήσει ως μηχανισμός refinement που συνδυάζει τη σημασιολογική εμβέλεια των learned μοντέλων με τη post-hoc διόρθωση των πλεονασματικών segments. Παράλληλα, η εφαρμογή στρατηγικών multi-model ensemble δύναται να αξιοποιήσει τα συμπληρωματικά πλεονεκτήματα διαφορετικών αρχιτεκτονικών, όπου για παράδειγμα το BLIP θα αναλάμβανε ρόλο ταχείας προεπισκόπησης, το GIT-VATEX θα εστίαζε σε action-centric captions και το Qwen2-VL θα παρείχε λεπτομερέστερες περιγραφές, οδηγώντας σε ένα πιο εύρωστο τελικό αποτέλεσμα.

Ειδικότερα για μοντέλα υψηλής περιγραφικότητας όπως το **Qwen2-VL**, η συστηματική εφαρμογή τεχνικών prompt engineering, μέσω της χρήσης few-shot παραδειγμάτων και της επιβολής ρητών περιορισμών ως προς το μήκος και την κατηγορία του video, κρίνεται απαραίτητη για τον περιορισμό της υπερβολικής πολυπλοκότητας και τη βελτιστοποίηση της ευθυγράμμισης με τα ground truth annotations. Επιπλέον, ο μηχανισμός του semantic merging επιδέχεται περαιτέρω αναβάθμιση μέσω προηγμένων αλγοριθμικών στρατηγικών, όπως η εισαγωγή temporal coherence penalties και η υιοθέτηση hierarchical clustering αντί της απλής σειριακής συγχώνευσης, στοχεύοντας σε πιο robust ενοποιήσεις και στην καλύτερη διατήρηση της χρονικής δομής. Μια εξίσου σημαντική επέκταση του pipeline αφορά την ενσωμάτωση multimodal fusion διαδικασιών πέρα από την οπτική πληροφορία, καθώς η συνεκτίμηση audio features, η χρήση OCR για on-screen text και η εφαρμογή speech recognition θα εμπλούτιζαν καθοριστικά το context προσφέροντας πληρέστερες περιγραφές.

Τέλος, το πλαίσιο αξιολόγησης μπορεί να εμπλουτιστεί περαιτέρω, ώστε να αποτυπώνει με μεγαλύτερη πληρότητα την ποιότητα των παραγόμενων αποτελεσμάτων. Η συνδυαστική αξιοποίηση των καθιερωμένων n-gram μετρικών με learned semantic metrics, όπως το BERTScore και το CLIPScore, θα μπορούσε να συμπληρώσει ουσιαστικά τις υφιστάμενες

μετρικές temporal localization, επιτρέποντας μια πιο ολοκληρωμένη αποτίμηση τόσο της χρονικής ακρίβειας όσο και της σημασιολογικής επάρκειας των παραγόμενων λεζάντων.

Συμπερασματικά, η παρούσα εργασία αναδεικνύει τη δυναμική των pretrained Vision-Language Models ως βασικών δομικών στοιχείων για το Dense Video Captioning, υποδεικνύοντας παράλληλα ότι η αποτελεσματική αξιοποίησή τους εξαρτάται σε μεγάλο βαθμό από την περαιτέρω ενίσχυση της χρονικής εντόπισης γεγονότων και τη στενότερη σύνδεσή της με τη διαδικασία γλωσσικής περιγραφής, μέσω πιο εκλεπτυσμένων και σημασιολογικά συνεπών στρατηγικών ενσωμάτωσης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] R. Krishna, K. Hata, F. Ren, L. Fei-Fei and J. C. Niebles, "Dense-Captioning Events in Videos," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 706-715, doi: 10.1109/ICCV.2017.83.

[2] M. Abdar et al., "A Review of Deep Learning for Video Captioning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2024.3522295.

[3] Y. Li, T. Yao, Y. Pan, H. Chao and T. Mei, "Jointly Localizing and Describing Events for Dense Video Captioning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7492-7500, doi: 10.1109/CVPR.2018.00782.

[4] Iqra Qasim, Alexander Horsch, and Dilip Prasad. 2025. Dense Video Captioning: A Survey of Techniques, Datasets and Evaluation Protocols. ACM Comput. Surv. 57, 6, Article 154 (June 2025), 36 pages. https://doi.org/10.1145/3712059

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 162, Baltimore, MD, USA, July 2022, pp. 12888-12900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[6] X. Wang et al., "VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research," arXiv preprint arXiv:1904.03493, June 2019.

[7] P. Wang et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv preprint arXiv:2409.12191, Sept. 2024.

[8] I. Bieda, A. Kisil and T. Panchenko, "An Approach to Scene Change Detection," 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Cracow, Poland, 2021, pp. 489-493, doi: 10.1109/IDAACS53288.2021.9660887.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. Conf. Empirical Methods Nat. Lang. Process. Syst. Demonstr. (EMNLP), Hong Kong, China, Nov. 2019, pp. 3982-3992, doi: 10.18653/v1/D19-1410.

[10] S. A. Mahmud et al., "Enhancing Video Understanding: Deep Neural Networks for Spatiotemporal Analysis," arXiv preprint arXiv:2502.07277, Feb. 2025.

[11] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International

Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4489-4497, doi: 10.1109/ICCV.2015.510.

[13] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," in IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994, doi: 10.1109/72.279181.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence -- Video to Text. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15). IEEE Computer Society, USA, 4534–4542, doi: 10.1109/ICCV.2015.515

[16] L. Yao et al., "Describing Videos by Exploiting Temporal Structure," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4507-4515, doi: 10.1109/ICCV.2015.512.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[18] Bahdanau, D., Cho, K. and Bengio, Y. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. arXiv: 1409.0473.

[19] A. Vaswani et al., “Attention Is All You Need,” arXiv: Computation and Language, Jun. 2017, [Online]. Available: https://arxiv.org/abs/1706.03762

[20] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv: Computer Vision and Pattern Recognition, Oct. 2020, [Online]. Available: https://arxiv.org/pdf/2010.11929.

[21] Shaw, Peter & Uszkoreit, Jakob & Vaswani, Ashish. (2018). Self-Attention with Relative Position Representations. 10.48550/arXiv.1803.02155.

[22] He, Kaiming & Chen, Xinlei & Xie, Saining & Li, Yanghao & Dollar, Piotr & Girshick, Ross. (2022). Masked Autoencoders Are Scalable Vision Learners. 15979-15988. 10.1109/CVPR52688.2022.01553.

[23] Caron, Mathilde & Touvron, Hugo & Misra, Ishan & Jégou, Hervé & Mairal, Julien & Bojanowski, Piotr & Joulin, Armand. (2021). Emerging Properties in Self-Supervised Vision Transformers. 9630-9640. 10.1109/ICCV48922.2021.00951.

[24] Arnab, Anurag & dehghani, Mostafa & Heigold, Georg & Sun, Chen & Lucic, Mario & Schmid, Cordelia. (2021). ViViT: A Video Vision Transformer. 6816-6826. 10.1109/ICCV48922.2021.00676.

[25] Bertasius, Gedas & Wang, Heng & Torresani, Lorenzo. (2021). Is Space-Time Attention All You Need for Video Understanding?. 10.48550/arXiv.2102.05095.

[26] Liu, Ze & Ning, Jia & Cao, Yue & Wei, Yixuan & Zhang, Zheng & Lin, Stephen & Hu, Han. (2021). Video Swin Transformer. 10.48550/arXiv.2106.13230.

[27] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," International Conference on Machine Learning, pp. 8748–8763, Jul. 2021, [Online]. Available: http://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[28] Bao, Liping & Wei, Longhui & Qiu, Xiaoyu & Zhou, Wengang & Li, Houqiang & Tian, Qi. (2023). Learning Transferable Pedestrian Representation from Multimodal Information Supervision. 10.48550/arXiv.2304.05554.

[29] Li, Junnan & Rs, Ramprasaath & Gotmare, Akhilesh & Joty, Shafiq & Xiong, Caiming & Hoi, Steven. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. 10.48550/arXiv.2107.07651.

[30] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

[31] Oord, Aaron & Li, Yazhe & Vinyals, Oriol. (2018). Representation Learning with Contrastive Predictive Coding. 10.48550/arXiv.1807.03748.

[32] Lu, Jiasen & Batra, Dhruv & Parikh, Devi & Lee, Stefan. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 10.48550/arXiv.1908.02265.

[33] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in Proc. 40th Int. Conf. Mach. Learn. (ICML), Honolulu, HI, USA, July 2023, pp. 19730-19742. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[34] J.-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning," in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), vol. 35, New Orleans, LA, USA, Nov. 2022, pp. 23716-23736. [Online]. Available: https://arxiv.org/abs/2204.14198

[35] H. Liu et al., "Visual Instruction Tuning," in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), vol. 36, New Orleans, LA, USA, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[36] W. Dai et al., "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), vol. 36, New Orleans, LA, USA, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2305.06500

[37] H. Zhang et al., "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," arXiv preprint arXiv:2306.02858, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2306.02858

[38] M. Maaz et al., "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," arXiv preprint arXiv:2306.05424, June 2023. [Online]. Available: https://arxiv.org/abs/2306.05424

[39] J. Wang, Z. Yang, X. Hu, L. Li, K. Q. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A Generative Image-to-text Transformer for Vision and Language," arXiv preprint arXiv:2205.14100, May 2022. [Online]. Available: https://arxiv.org/abs/2205.14100

[40] Liu, Z.; Song, R. Survey of Dense Video Captioning: Techniques, Resources, and Future Perspectives. Appl. Sci. 2025, 15, 4990. https://doi.org/10.3390/app15094990

[41] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, June 2018, pp. 7492-7500.

[42] D. Yang and C. Yuan, "Hierarchical Context Encoding for Events Captioning in Videos," in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Athens, Greece, Oct. 2018, pp. 1288-1292, doi: 10.1109/ICIP.2018.8451692.

[43] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-End Dense Video Captioning with Parallel Decoding," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 8739-8748, doi: 10.1109/ICCV48922.2021.00863.

[44] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-End Dense Video Captioning with Masked Transformer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, June 2018, pp. 8739-8748, doi: 10.1109/CVPR.2018.00911.

[45] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action

Localization in Untrimmed Videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 5734-5743, doi: 10.1109/CVPR.2017.155.

[46] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "SST: SingleStream Temporal Action Proposals," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 6373-6382, doi: 10.1109/CVPR.2017.675.

[47] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal Action Detection with Structured Segment Networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 2914-2923, doi: 10.1109/ICCV.2017.317.

[48] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary Sensitive Network for Temporal Action Proposal Generation," in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sept. 2018, pp. 3-19, doi: 10.1007/978-3-030-01225-0_1.

[49] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-Matching Network for Temporal Action Proposal Generation," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 3889-3898, doi: 10.1109/ICCV.2019.00399.

[50] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Learning Salient Boundary Feature for Anchor-free Temporal Action Localization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, June 2021, pp. 3320-3329, doi: 10.1109/CVPR46437.2021.00333.

[51] A. Wajid, G. Saleem, M. U. Farooq, M. U. G. Khan, and F. Shafait, "Deep Learning and Knowledge Graph for Image/Video Captioning: A Review of Datasets, Evaluation Metrics, and Methods," Engineering Reports, vol. 6, no. 10, e12785, Oct. 2024, doi: 10.1002/eng2.12785.

[52] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, "Grounded Video Description," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, June 2019, pp. 6578-6587, doi: 10.1109/CVPR.2019.00674.

[53] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, June 2023, pp. 10714-10726, doi: 10.1109/CVPR52729.2023.01032.

[54] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu, "Event-Centric Hierarchical Representation for Dense Video Captioning," IEEE Trans. Circuits Syst. Video

Technol., vol. 31, no. 5, pp. 1890-1900, May 2021, doi: 10.1109/TCSVT.2020.3014606.

[55] X. Long, C. Gan, and G. de Melo, "Video Captioning with Multi-Faceted Attention," Trans. Assoc. Comput. Linguistics, vol. 6, pp. 173-184, 2018, doi: 10.1162/tacl_a_00013.

[56] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), vol. 27, Montreal, QC, Canada, Dec. 2014, pp. 3104-3112. [Online]. Available: https://arxiv.org/abs/1409.3215

[57] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, June 2015, pp. 2625-2634, doi: 10.1109/CVPR.2015.7298878.

[58] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Lisbon, Portugal, Sept. 2015, pp. 1412-1421, doi: 10.18653/v1/D15-1166.

[59] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, June 2016, pp. 1029-1038, doi: 10.1109/CVPR.2016.117.

[60] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI Blog, vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://cdn.openai.com/better-languagemodels/language_models_are_unsupervised_multitask_learners.pdf

[61] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 382-398, doi: 10.1007/978-3-319-46454-1_24.

[62] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, June 2016, pp. 2361-2369, doi: 10.1109/CVPR.2016.259.

[63] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," Int. J. Comput. Vis., vol. 130, no. 9, pp. 2337-2348, Sept. 2022, doi: 10.1007/s11263-022-01653-1.

[64] J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal, "MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Online, July 2020, pp. 2603-2614, doi: 10.18653/v1/2020.acl-main.233.

[65] V. Iashin and E. Rahtu, "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer," in Proc. British Mach. Vis. Conf. (BMVC), Virtual, Sept. 2020. [Online]. Available: https://www.bmvc2020-conference.com/assets/papers/0488.pdf

[66] M. Freitag and Y. Al-Onaizan, "Beam Search Strategies for Neural Machine Translation," in Proc. 1st Workshop Neural Mach. Transl., Vancouver, BC, Canada, Aug. 2017, pp. 56-60, doi: 10.18653/v1/W17-3207.

[67] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra, "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models," in Proc. AAAI Conf. Artif. Intell., New Orleans, LA, USA, Feb. 2018, pp. 8151-8159, doi: 10.1609/aaai.v32i1.12340.

[68] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The Curious Case of Neural Text Degeneration," in Proc. Int. Conf. Learn. Represent. (ICLR), Addis Ababa, Ethiopia, Apr. 2020. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH

[69] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," in Proc. ACL 2017, Syst. Demonstr., Vancouver, BC, Canada, July 2017, pp. 67-72, doi: 10.18653/v1/P17-4012

[70] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-Critical Sequence Training for Image Captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 7008-7024, doi: 10.1109/CVPR.2017.131.

[71] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022. [Online]. Available: https://openreview.net/forum?id=GUrhfTuf_3

[72] Zhou, Xingyi & Arnab, Anurag & Buch, Shyamal & Yan, Shen & Myers, Austin & Xiong, Xuehan & Nagrani, Arsha & Schmid, Cordelia. (2024). Streaming Dense Video Captioning. 18243-18252. 10.1109/CVPR52733.2024.01727.

[73] H. Xu, B. Li, V. Ramanishka, L. Sigal and K. Saenko, "Joint Event Detection and Description in Continuous Video Streams," 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2019, pp. 25-26, doi: 10.1109/WACVW.2019.00011.

[74] S. Park et al., " Zero-Shot Scene Change Detection," arXiv preprint arXiv:2406.11210, 2025. [Online]. Available: https://arxiv.org/pdf/2406.11210v3.pdf

[75] A. -C. Jitaru and B. Ionescu, "High Density Crowd Scene Detection in Untrimmed Streaming Videos for Surveillance Purpose," 2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, Romania, 2023, pp. 1-6, doi: 10.1109/ECAI58194.2023.10194094.