Birzeit University
Faculty of Engineering & Technology
Department of Electrical & Computer Engineering

CONTENT-BASED VIDEO RETRIEVAL (MULTIMODAL APPROACH)

Prepared By:
Diaeddin Tahboub - 1200136
Ahmad Zubaidia - 1200105
Hamza Najar – 1192605

Supervised By:

Dr. Aziz Qaroush

A graduation project submitted to the Department of Electrical and Computer Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Systems Engineering.

BIRZEIT

February 2025

Abstract

This research addresses the growing need for effective video retrieval systems that can handle user queries in natural language rather than relying on constrained keyword inputs. To bridge the gap between diverse user descriptions and complex video content, we explored various vision models for robust object and scene detection and integrated them with automated captioning to generate higher-level semantic representations. The resulting multi-modal framework fuses these visual and textual features, enabling fine-grained recognition of video content while capturing overarching contextual information. To optimize both efficiency and accuracy, different frame extraction strategies are investigated, including uniform down-sampling, scene-based segmentation, and full-frame processing. Additionally, large language models are used to augment the textual queries, mitigating mismatches that arise from varied user phrasings. Experiments on standard benchmarks, notably MSVD and MSR-VTT, show that query augmentation and the combined use of vision-based embeddings with caption-derived features offer significant improvements over single-modality approaches. Multiple ranking methods, such as mode-based, random, and medoid, are further evaluated to account for multiple captions associated with each video. The findings demonstrate that systematically refining textual queries while merging both low-level visual details and high-level linguistic information leads to a more comprehensive and precise video retrieval system, paving the way for user-friendly, context-sensitive solutions in large-scale video collections.

Arabic Abstract

يتناول هذا البحث الحاجة المتزايدة إلى أنظمة فعّالة لاسترجاع مقاطع الفيديو يمكنها التعامل مع استفسارات المستخدمين باللغة الطبيعية، عوضًا عن الاقتصار على كلمات مفتاحية محدودة. ولردم الفجوة بين توصيفات المستخدم المتنوعة والمحتوى البصري المعقّد للفيديو، تم استكشاف نماذج رؤية حاسوبية متقدمة للتعرّف على الأجسام والمشاهد ودمجها مع تقنيات العنونة التلقائية، بهدف توليد تمثيلات دلالية عالية المستوى. ينشئ الإطار المُقترح تمثيلًا ثنائيّ المكوّنات (مرئيًّا ولغويًّا) لمقاطع الفيديو، ما يتيح فهمًا أكثر تفصيلًا للمضمون البصري مع المحافظة على السياق العام. لتحقيق التوازن بين الكفاءة والدقة، جرت دراسة استراتيجيات متعددة لاستخراج الإطارات، شملت التنقيص المنتظم وتقسيم الفيديو إلى مشاهد فضلاً عن معالجة جميع الإطارات بالكامل. إضافةً إلى ذلك، اُستُخدِمت النماذج اللغوية الضخمة في تعزيز الاستعلامات النصية، بغرض الحدّ من التباين الناجم عن اختلاف صِيَغ التعبير لدى المستخدمين. وقد بيّنت التجارب على أنّ الجمع بين المعلومات البصرية المنبثقة من نماذج الرؤية والعنونة التلقائية، إلى MSR-VTTو MSVD معايير مرجعية مثل جانب تعزيز الاستعلامات عبر النماذج اللغوية، يُحسّن أداء الاسترجاع مقارنةً بالأساليب أحادية الجانب. علاوةً على ذلك، أظهرت نتائج الترتيب المعتمد على آليات مختلفة (النمط الأكثر تكرارًا، والعشوائي، والميدويد) أنّ مراعاة تعددية العناوين المصاحبة لكل فيديو يعزّز دقة البحث ومواءمته للاحتياجات المتنوّعة للمستخدمين. تؤكد هذه النتائج أن المزج المنهجي بين التفاصيل البصرية منخفضة المستوى والمعلومات اللغوية رفيعة المستوى، مضافًا إليه تحسين الاستعلامات النصية، يؤدّي إلى نظام أشمل وأكثر دقة لاسترجاع مقاطع الفيديو. وبذلك يمهّد هذا الإطار المقترح السبيل لبناء نظم بحث مرنة وواعية بالسياق، قادرة على التعامل مع مجموعات فيديو ضخمة بصورة أكثر فعالية وموثوقية.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

The rapid growth of online video collections across platforms such as YouTube, social media networks, and educational websites has turned videos into a principal medium for sharing knowledge and experiences. Traditional methods for video search often rely on keyword matching or manual annotations, yet these strategies cannot fully capture the richness of visual and temporal information within a video. Users increasingly express their search needs in everyday language, describing actions, contexts, or events rather than relying on narrow keywords. This shift in how queries are phrased introduces a fundamental challenge: bridging the gap between the multifaceted content of a video's frames and the varied linguistic forms used by viewers to describe their queries. In response, the present work aims to establish a more direct alignment between user intent and the low-level and high-level details embedded in video data, thereby enhancing the capabilities of existing retrieval systems.

1.2 Introduction

Content-based video retrieval aims to match text-based queries with videos that show what the user is asking for. Early solutions depended heavily on adding human-generated tags or other manual descriptions. These methods are labor-intensive and frequently miss out on important visual or contextual details. In recent years, progress in deep learning has led to more sophisticated ways to extract meaningful features from video frames. Meanwhile, improvements in natural language

processing have made it possible to create text captions automatically, translating visual events into words. Bringing these two methods visual and textual into a single system can help users quickly find the videos they need, even if the videos show complex actions or scenes.

Still, there are major challenges. First, videos are more than just a series of images; they show changes over time that can be important to understand. Second, users can phrase their queries in many different ways, which can cause mismatches between the words they use and the captions in the system. Third, combining image-based features and caption-based features is not always straightforward, since each type of feature represents different parts of the video's information. Addressing these challenges requires careful selection of frames or scenes, effective fusion of textual and visual features, and flexible handling of user queries.

1.3 Problem statement

Though deep learning models can extract powerful features from individual frames and generate coherent captions, several issues remain unresolved. One of the most prominent is the semantic gap between a user's natural language query often containing subtle or idiomatic expressions and the raw pixel data of a video. Additionally, exhaustively analyzing every frame of extensive video collections is computationally impractical, necessitating more strategic frame selection or scene detection methods. Moreover, the variability of user queries, both in structure and terminology, increases the risk of mismatches between textual descriptions and visual or caption-based embeddings. The aim of this research is to tackle these complexities by developing a retrieval pipeline that can flexibly adapt to different data scales, incorporate deep semantic cues from both vision and language models, and refine user queries to ensure higher retrieval precision.

1.4 Methodology

To tackle these problems, this research follows a step-by-step approach. First, it tests different strategies for choosing frames from a video, such as sampling frames evenly, detecting major scene changes, or using all frames when necessary. The next step involves turning visual content into numeric vectors using a vision transformer, while a separate model creates text captions for each video. These captions are then turned into their own numeric vectors through a text encoder. The system explores two ways to combine these representations. In one, they are fused into a single vector. In the other, each vector's similarity score is calculated separately, and then the scores are combined. Because users may phrase their queries in many different ways, a large language model (GPT-3.5 Turbo) generates variations of each text query, which can help reduce mismatches between what users type and the captions associated with the video. Once the system has computed

how closely each video matches a query, different ranking methods consolidate these results. For example, we can look at the most frequent rank a video receives or choose a “medoid” query one that best represents the set of possible captions. The system’s performance is then measured using standard metrics like Recall@k (how often the correct video is found in the top k matches), the average value of all ranks(ranks of all videos) and the median value of these ranks.

1.5 Contribution

This research provides several contributions to content-based video retrieval. First, it combines visual embeddings created by a powerful vision model with caption-based textual embeddings generated automatically, bridging the gap between raw images and language. Second, it examines multiple ways to extract frames uniform sampling, scene detection, and processing every frame to find the balance between speed and accuracy. Third, it offers an in-depth comparison of two methods of fusion: one that mixes features at the vector level and another that mixes their similarity scores. Fourth, it uses GPT-3.5 Turbo to enrich the textual queries, demonstrating that adding different versions of each query can noticeably improve search results. Lastly, it presents various ranking strategies (mode-based, random, median, and medoid) to see which best handles the multiple captions often linked to each video in large datasets.

1.6 Report outline

The report is organized to provide a logical flow from conceptual background to practical evaluation. It begins in Chapter 2 with a focused review of existing literature, discussing the historical context and state-of-the-art methods in visual feature extraction, automated captioning, and multi-modal fusion. Chapter 3 details the specific architecture and algorithms employed in this work, describing each component of the pipeline and illustrating how visual and textual embeddings are generated, fused, and integrated with user queries. Chapter 4 presents the experimental results from multiple benchmarks, offering quantitative metrics and qualitative insights into how different strategies ranging from frame extraction methods to fusion techniques affect retrieval performance. Finally, Chapter 5 concludes by summarizing the contributions and limitations of the proposed approach, suggesting directions for future research that could further refine and scale multi-modal video retrieval systems.

Chapter 2

Related Work

Contents

2.1 Introduction

Recent advances in multimedia retrieval have increasingly focused on leveraging multiple modalities—visual, textual, and auditory—to capture richer semantic information from video data. In response to the challenges of processing unstructured video content, researchers have proposed various architectures that integrate multi-modal cues, temporal dynamics, and cross-modal interactions. In this chapter, we review six influential studies that provide insights into multi-modal representation, fusion strategies, pre-training methods, and evaluation techniques.
These works collectively inform the design of our proposed retrieval system.

2.2 Multi-Modal Video Retrieval: Techniques and Trends

Gabeur et al. [1] introduce a multi-modal transformer architecture that jointly encodes video features across visual, audio, and text streams by leveraging self-attention mechanisms to capture temporal dependencies and cross-modal correlations, thereby establishing a robust basis for multi-modal fusion in video retrieval systems; in a complementary study, Shvetsova et al. [2] propose a modality-agnostic fusion transformer that processes heterogeneous data streams through a combinatorial loss, bypassing the need for explicit positional or modality encodings and informing our fusion and ranking methods with effective strategies for integrating diverse features.
Radford et al. [3] present their seminal Contrastive Language–Image Pre-training (CLIP) framework, which aligns image and text representations in a shared latent space using contrastive loss, and, by pre-training on large-scale image–caption pairs, produces transferable visual features

that justify incorporating CLIP's ViT-B/32 model into our visual feature extraction pipeline.

Addressing the challenge of generating meaningful semantic captions for video content, Ye et al. [4] propose the "mPLUG-Owl" modular framework that synergizes visual processing with language model capabilities by decomposing the captioning task into separate yet collaborative modules, thus enhancing both caption generation and multi-modal understanding a strategy that aligns with our use of mPLUG-Owl for automatic video captioning to ensure high-level semantic cues complement low-level visual details. Dong et al. [5] introduce a dual-encoding strategy for video retrieval that employs separate encoders for visual and textual modalities before merging them into a common embedding space, and their analysis of trade-offs between independent and fused representations provides valuable insights that resonate with our design, which evaluates the merits of both independent modality assessment and integrated multi-modal fusion. Finally, recognizing that practical system performance extends beyond algorithmic accuracy, Alpay et al. [6] conduct a user study on CLIP-based video retrieval systems, demonstrating that considerations of retrieval precision and user satisfaction are essential for assessing the usability and robustness of multi-modal retrieval solutions a perspective that is central to our comprehensive evaluation framework.

Chapter 3

Proposed Work

Contents

3.1 System Overview

3.1.1 Overview of the Retrieval Framework

Content-based video retrieval (CBVR) is an essential field in multimedia research, driven by the increasing volume of video data and the necessity for efficient indexing and retrieval methods. Traditional retrieval techniques often rely exclusively on either visual features extracted from frames or textual metadata, limiting their effectiveness in capturing the full semantic richness of video content. This study proposes a multi-modal retrieval system that integrates both visual and caption-based representations, enabling more accurate and semantically meaningful video retrieval. The retrieval framework follows a multi-step pipeline, consisting of frame extraction, feature representation, multi-modal fusion, query expansion, and ranking. Each stage of this pipeline addresses key challenges in CBVR, including the selection of representative frames, robust feature extraction using deep learning models, and multi-modal fusion of visual and textual information. The system leverages Contrastive Language-Image Pretraining (CLIP) for visual representation and mPLUG-Owl for automatic video captioning, ensuring that both low-level visual details and high-level semantic descriptions are incorporated into the retrieval process.

This section outlines the overall workflow of the proposed system, detailing its major components and emphasizing its key contributions in the context of CBVR.

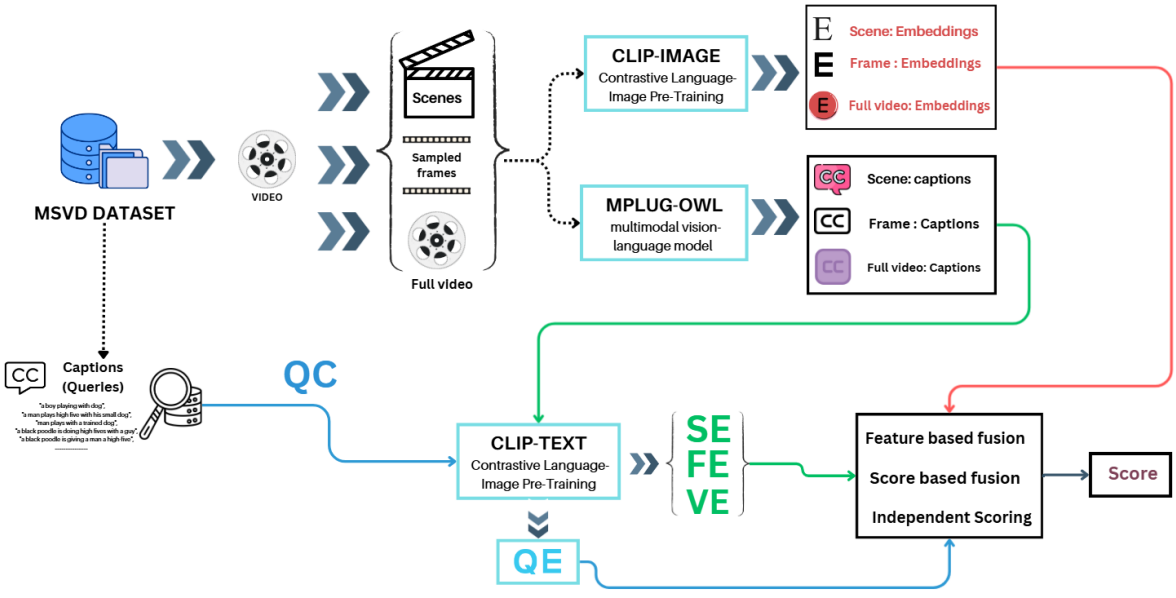

Figure 1:Overview of the Retrieval Framework

3.1.2 Processing Pipeline

The system follows a structured processing pipeline comprising five major components: frame extraction, feature representation, fusion strategies, query expansion, and ranking. Each of these components contributes to enhancing retrieval efficiency and accuracy.

3.1.2.1 Frame Extraction

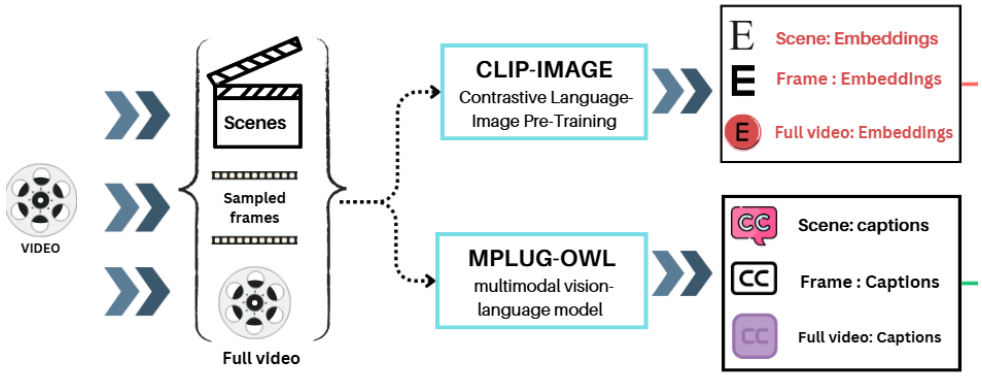

Figure 2: Frame Extraction

Frame extraction is a critical preprocessing step in content-based video retrieval, determining how video content is sampled and subsequently represented. The system explores three distinct frame extraction strategies to evaluate their impact on retrieval effectiveness:

1. Uniform Down-sampling: Frames are sampled at fixed intervals (N) to ensure a structured and computationally efficient representation of the video. While this method provides uniform coverage, it may fail to capture key moments that are crucial for retrieval.
2. Scene Detection-Based Extraction: A scene boundary detection algorithm identifies significant visual transitions within the video, ensuring that frames are selected at scene boundaries rather than at fixed intervals. This method preserves contextual integrity by extracting visually distinct moments, enhancing retrieval relevance.
3. All-Frames Processing: Every frame of the video is utilized, providing the most comprehensive representation of visual information. While this method captures all visual variations, it significantly increases computational cost and may introduce redundant information.

The selection of a frame extraction strategy impacts the balance between computational efficiency and retrieval performance. By evaluating all three approaches, this study aims to identify the optimal strategy for content-aware frame selection.

3.1.2.2 Feature representation

Once frames are extracted, the system generates multi-modal embeddings, encoding both visual and textual information. This step ensures that video content is meaningfully represented in a shared feature space, allowing for effective similarity computation during retrieval.

1. Visual Representation: Each extracted frame is processed using CLIP ViT-B/32, a vision transformer-based model that converts the image into a 512-dimensional feature embedding. To create a unified video-level representation, frame embeddings are aggregated using the average function producing a single compact vector that captures the overall visual semantics of the video.
2. Textual Representation: The video is also processed using mPLUG-Owl, a state-of-the-art video captioning model, which generates descriptive text summarizing the visual content. These captions are then encoded using CLIP's text encoder, producing a corresponding 512-dimensional textual embedding.

This dual-representation ensures that both low-level object details (captured via CLIP's vision model) and high-level semantic meaning (captured via captions) contribute to retrieval. The effectiveness of each modality is independently assessed, and their combined impact is evaluated through fusion strategies.

3.1.2.4 Fusion Strategies

To integrate multi-modal information, the system explores two fusion techniques that combine visual embeddings and textual embeddings to enhance retrieval accuracy:

1. Feature-Based Fusion: In this approach, visual and textual embeddings are merged at the feature level using weighted summation, resulting in a balanced 512-dimensional representation. This fusion method ensures that complementary information from both modalities is effectively preserved.

2. Score-Based Fusion: Instead of merging features, similarity scores are computed separately for visual and textual embeddings and subsequently combined using a weighted summation approach. This allows for a flexible fusion strategy where the contribution of each modality can be dynamically adjusted based on retrieval requirements.

By comparing feature-level and score-level fusion, the system evaluates the advantages and limitations of each approach, providing empirical insights into the most effective fusion strategy for CBVR.

3.1.2.5 Query Expansion and Augmentation

To enhance retrieval robustness, the system employs query augmentation using GPT-3.5 Turbo, a large language model (LLM) capable of generating semantically diverse query rephrasings. This process ensures that retrieval remains effective even when users phrase their queries differently.
Expanded queries are encoded using CLIP's text encoder, allowing the system to match multiple linguistic variations to the same video content. This augmentation technique improves retrieval by ensuring that the system remains resilient to different query formulations, particularly in cases where semantic equivalence is maintained across distinct linguistic structures.

3.1.2.6 Matching and Ranking

For each video, the system computes similarity scores between the query embedding and the video representations derived from both visual and textual features. To evaluate the effectiveness of different ranking strategies, the system employs three ranking methods:

1. Mode-Based Ranking: The most frequently occurring rank across multiple queries is selected as the final rank for a given video.
2. Random Query-Based Ranking: A single random query is used to rank the video, providing a query-specific ranking.
3. Medoid Query-Based Ranking: Queries are clustered based on semantic similarity, and the most representative query (medoid) is used to determine the final rank.
4. Median Based Ranking: Each video receives multiple ranks from different queries, and the median rank is used as the final rank to ensure robustness against outliers and extreme values.

By systematically comparing these ranking approaches, the study identifies the most effective ranking method for ensuring consistency and accuracy in retrieval results.

3.2 System Modules

This section details the various modules that constitute the proposed content-based video retrieval (CBVR) system. It outlines how video frames are extracted, processed, and represented, and describes the techniques used to incorporate both visual and textual information. Additionally, it explains how scene detection and representation enhance retrieval performance by segmenting videos into smaller, semantically coherent units. Finally, it addresses the fusion of multi-modal features, the augmentation of queries, and the ranking methods that ensure the system retrieves the most relevant content.

3.2.1 CLIP-Based Representation

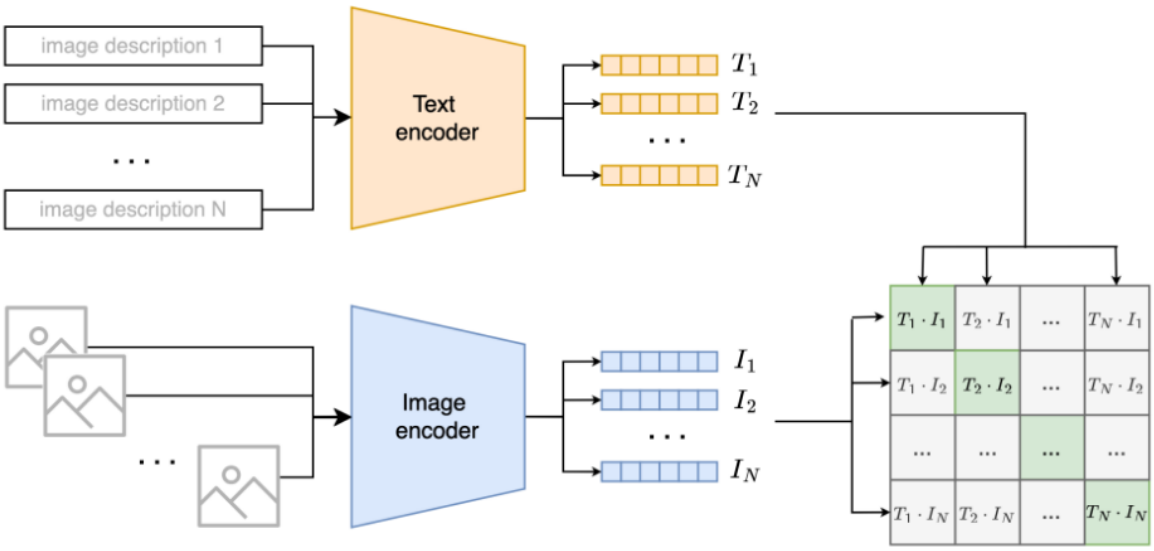

Figure 3: Clip model including image encoder and text encoder [6].

CLIP (Contrastive Language–Image Pretraining) introduces two encoders—one for images and another for text that map data into a shared embedding space. The image encoder processes visual inputs into 512-dimensional embeddings, capturing color, shape, texture, and conceptual cues, while the text encoder transforms textual descriptions into similarly sized vectors that reflect meaning and context. Placing these representations in the same embedding space allows for direct comparisons between text and image data, enabling seamless cross-modal retrieval [6].

3.2.2 mPLUG-Owl for Video Captioning

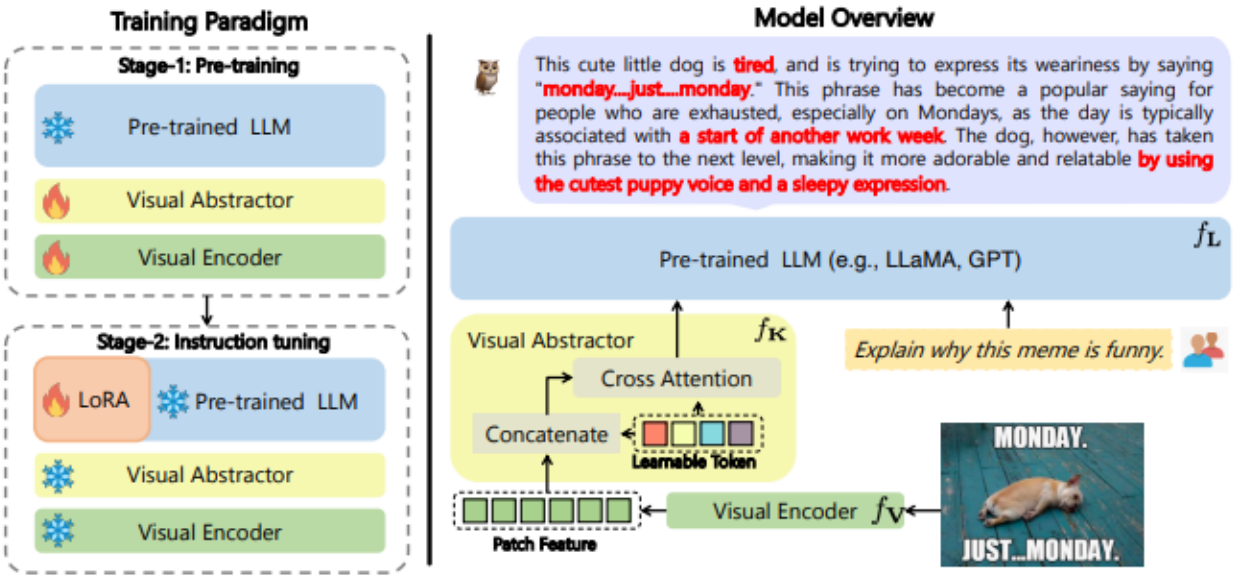

Figure 4: MPLUG-OWL overview [4]

mPLUG-Owl is a state-of-the-art video captioning model designed to generate contextually rich and semantically accurate textual descriptions for video content. Built on a large-scale transformer architecture, the model is pre-trained on diverse video datasets, enabling it to capture complex object interactions, scene transitions, and temporal dynamics. By leveraging both spatial and temporal cues, mPLUG-Owl produces captions that not only identify key objects and actions but also convey higher-level contextual relationships such as causality or intent. In the proposed retrieval framework, these captions serve as crucial textual representations that complement low-level visual features extracted via CLIP. This dual-modality strategy allows the system to bridge the semantic gap between raw pixel data and user queries, enhancing retrieval performance for various query formulations, including those with nuanced linguistic structures or context-heavy descriptions. Ultimately, mPLUG-Owl's robust captioning capabilities provide a powerful means of translating visual information into text, reinforcing the system's ability to retrieve videos based on both appearance and semantics [4].

3.2.3 Down-sampling and Representation in Frame Extraction

Down-sampling in frame extraction reduces the number of processed frames by selecting a fixed set of evenly spaced frames, ensuring a structured representation of the video while managing computational efficiency. These selected frames undergo preprocessing steps such as color normalization and resizing to maintain consistency across samples.

For representation, each extracted frame is passed through the CLIP ViT-B/32 visual encoder, which transforms visual features into 512-dimensional embeddings, numerically encoding the content of each frame. To generate a unified video representation, the individual frame embeddings are aggregated by averaging them, resulting in a (1, 512) video embedding. This vector serves as a compact descriptor of the video's overall visual content, capturing essential patterns across the extracted frames.

The resulting video embedding is used in two ways. First, it allows direct comparison with query embeddings, enabling retrieval based on visual similarity. In parallel, the video can be processed through a captioning model, such as mPLUG-Owl, to generate descriptive text. These captions are encoded using CLIP's text encoder, producing a corresponding 512-dimensional textual embedding. By combining visual and textual embeddings, the system enables a comparison of different retrieval methods, evaluating their effectiveness in capturing video content.

3.2.4 All-Frames Representation

In contrast to down-sampling, the all-frames approach processes every frame in the video. Two methods are employed to represent the resulting visual data:

1. CLIP-Based Frame Embedding Aggregation: Each frame is encoded into a 512-dimensional vector using CLIP ViT-B/32, after which the embeddings are aggregated (averaged) to form a single 512-dimensional video representation. This method preserves comprehensive detail but increases the computational burden.
2. mPLUG-Owl for Full-Video Captioning: Instead of encoding frames individually, the entire video is fed into mPLUG-Owl, which generates a textual summary. This summary is then encoded with CLIP's text encoder, yielding a 512-dimensional textual embedding. Because it relies on a holistic textual description, this method highlights semantic content and can streamline queries that rely on higher-level contextual cues.

3.2.5 Scene Detection

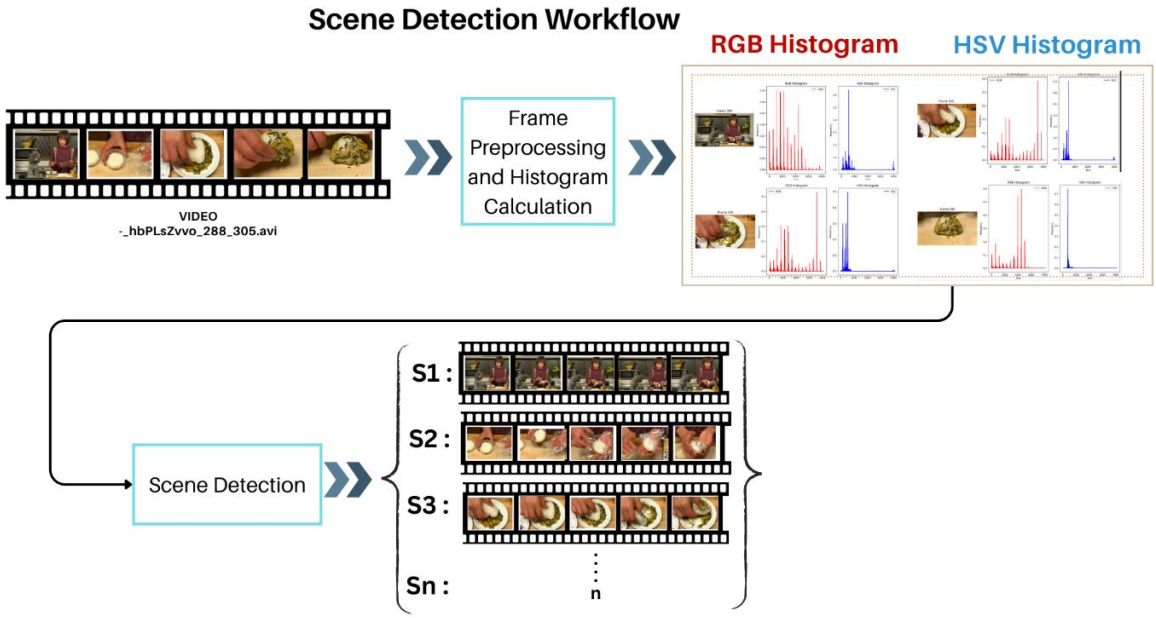

Figure 5: Scene detection workflow

Scene detection aims to divide lengthy videos into smaller segments based on major visual transitions, thereby focusing the retrieval process on well-defined scenes. The procedure starts by reading the video frame by frame and applying Gaussian blurring to reduce noise that might introduce spurious scene boundaries. Frame color properties are then analyzed in RGB and HSV color spaces: RGB captures primary color details, whereas HSV emphasizes hue, saturation, and brightness.

Using histogram comparison and the Chi-Square distance metric, the system identifies points where the visual difference between consecutive frames exceeds a predefined threshold. These points signify scene boundaries. Short scenes, which may arise from transient changes or noise, are merged with adjacent scenes to ensure contextual continuity. By segmenting the video into coherent scenes, the system reduces search space and improves retrieval precision.

3.2.6 Scene Representation

Once scenes are detected, each segment is characterized by two parallel representations:

1. Visual Representation: All frames within a scene (e.g., frames 80–120) are encoded via CLIP ViT-B/32. The 512-dimensional embeddings of these frames are averaged, forming a single scene-level embedding that reflects the collective visual content.
2. Textual Representation: Each detected scene is converted into a short video clip and processed by mPLUG-Owl, generating natural language captions describing the scene.

These captions are then encoded with CLIP's text encoder, producing 512-dimensional textual embeddings that encapsulate the semantic and contextual aspects of the scene.

By combining visual and textual representations, the system leverages multi-modal information to enhance retrieval accuracy, ensuring that both low-level visual details and high-level semantic cues are captured.

3.2.7 Caption-based representation

A caption-based representation utilizes descriptive text to effectively express the semantic meaning and contextual information of video content. Using mPLUG-Owl, the system produces captions that detail objects, actions, and contextual elements, translating visual scenes into textual form. The textual captions are then encoded with CLIP's text encoder, generating 512-dimensional embeddings that accurately capture the caption's semantic information. This approach is particularly valuable when users formulate queries in natural language, aligning high-level textual queries with corresponding semantic content in the video.

3.2.8 Fusion

Fusion is a critical stage in the video retrieval framework, enabling the integration of visual and textual features to enhance retrieval accuracy. By merging low-level details from visual embeddings with the semantic richness of textual representations, the system effectively captures both appearance-based and contextual information.

3.2.8.1 Feature-Based Fusion

Feature-based fusion combines textual and visual embeddings by computing a weighted sum, ensuring that the dimensionality of the resulting vector remains consistent. Let $\mathbf{t} \in \mathbb{R}^{512}$ represent the textual embedding and $\mathbf{v} \in \mathbb{R}^{512}$ represent the corresponding visual embedding. Because both embeddings lie in $\mathbb{R}^{512}$, their fusion is also a 512 -dimensional vector:

$$\mathbf{f} = \alpha\mathbf{t} + (1 - \alpha)\mathbf{v}, \alpha \in [0,1]$$

Here, α controls the relative emphasis of textual versus visual information. When $\alpha = 0.5$, both modalities contribute equally to $\mathbf{f}$. For values of α larger than 0.5, the textual embedding influences the final representation more heavily, whereas smaller values of α highlight the visual modality. Regardless of the chosen α, the dimensionality of the fused embedding $\mathbf{f}$ remains 512, ensuring direct comparability with other embeddings within the same feature space.

3.2.8.2 Score-Based Fusion

Score-based fusion aggregates modality-specific similarity scores rather than merging the underlying embeddings. Suppose $\mathbf{q} \in \mathbb{R}^{512}$ represents a query embedding. One may compute a visual similarity score, $\text{sim}_{\text{visual}}(\mathbf{q}, \mathbf{v}_i)$, and a textual similarity score, $\text{sim}_{\text{textual}}(\mathbf{q}, \mathbf{t}_i)$. These separate scores are then combined via a weighted sum:

$$\text{score}_{\text{fusion}} = \alpha \text{sim}_{\text{visual}} + (1 - \alpha)\text{sim}_{\text{textual}}$$

where $\alpha \in [0,1]$. In scenarios where multiple textual embeddings exist (such as scene-level captions), the maximum textual similarity across all available captions can be used, ensuring that the best-matching textual segment is selected.

In our implementation, we set $\alpha = 0.5$, ensuring a balanced contribution from both visual and textual modalities by default. Nevertheless, adjusting α\alphaα in other contexts allows the system to prioritize semantic context (textual) or visual appearance details, depending on the nature of the query. This approach provides greater flexibility in heterogeneous retrieval scenarios, where some queries hinge on nuanced visual attributes while others rely primarily on higher-level textual semantics.

3.2.9 Augmentation

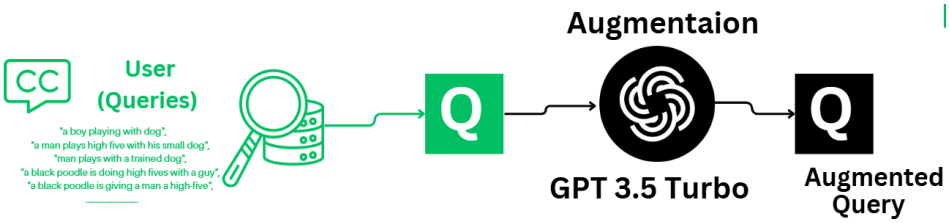

Figure 6:Query Augmentation

Augmentation extends query capabilities by expanding or rephrasing user queries, thereby increasing the system's robustness. Queries derived from ground truth annotations are enriched using GPT-3.5 Turbo, which generates semantically relevant variants. These augmented queries are encoded via CLIP's text encoder, producing additional 512-dimensional embeddings that capture linguistic variations. By integrating these varied embeddings, the system adapts more effectively to diverse user query formulations, improving overall retrieval performance. Matching and Ranking.

3.2.10 Matching and Ranking

Matching and ranking are important steps to evaluate how well the system retrieves relevant videos based on their associated queries. For each video, multiple ground truth queries are applied to the model, and the rank of the video is determined for each query. The rank indicates where the video appears in the list of results generated for that query. Since each video is linked to multiple queries, there will be several ranks calculated for the same video. To assign a single rank to the video, the most common rank (the mode) across all queries is chosen. This mode gives a consistent measure of how well the video performs overall in the retrieval process.

We also explored other approaches to ranking. One method involves selecting a random query from the queries associated with a video. The rank is then calculated for this single query, providing a simple, query-specific evaluation for the video.

Another approach uses the medoid query, which is the most representative query among all the queries linked to a video. To find the medoid query, all the queries are grouped into clusters based on their similarity, and the query closest to the center of the cluster is chosen. The rank of the video is then calculated using the results from this medoid query. These methods—mode-based, random query-based, and medoid query-based—offer different ways to assess the system's ability to match videos with their associated queries. Together, they provide a clear and comprehensive understanding of the system's retrieval performance.

3.3 Experimental Setup & Testing

Before presenting the tests and results, it is important to clarify key aspects of the evaluation process. In the MSVD and MSR-VTT test sets, each video includes multiple ground truth captions, which were converted into embeddings using the same CLIP model described in Chapter 3. In this context, "Textual features" refer to the embeddings of these captions (at the frame, scene, or video level), while "Visual features" refer to embeddings of the visual content. For video-level representations, visual embeddings are computed by averaging the embeddings of all frames in a video; for scene-level representations, they are obtained by averaging the embeddings of frames within a scene; and for frame-level representations, the embedding of each individual frame serves as the visual feature.

Ground truth captions act as queries for retrieval. The resulting query-response pairs are grouped according to two primary criteria. First, a *query-based* distinction is drawn between using a random caption from a different video (and then determining its rank) and using the *medoid* caption from a video's set of captions (i.e., the caption whose embedding has the minimal overall distance to all

other captions within that video). Second, a *rank-based* distinction involves determining the *mode* of the ranks observed in the dataset (i.e., the most frequently occurring rank), and the median of these ranks (for each video). These categories ensure methodological clarity prior to interpreting the results.

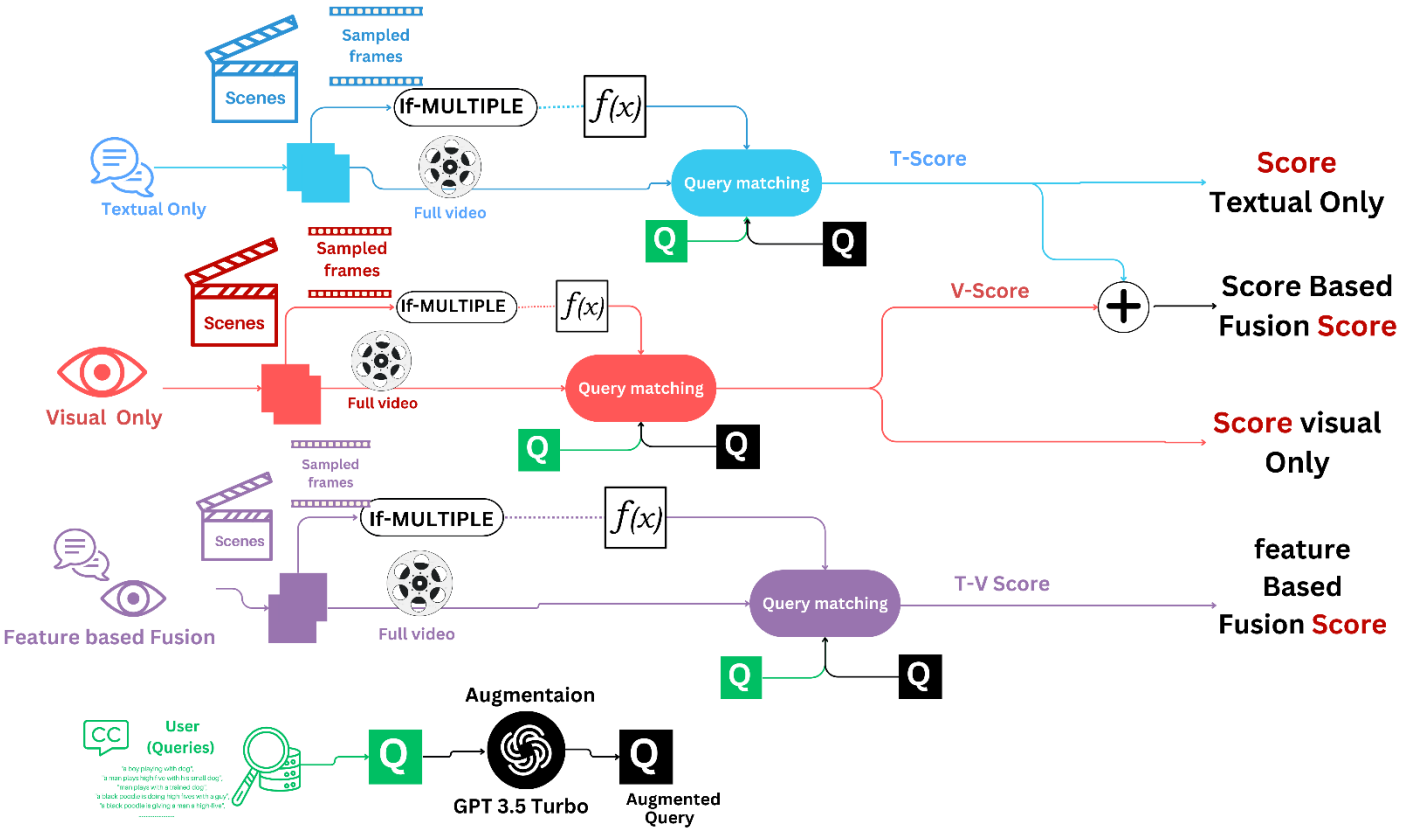

Figure 7: Test categories pipeline

The roadmap in the figure above shows the different test types for the MSVD dataset as an example. Some tests use only text, some only visuals, and some combine both. In all tests the ground truth captions are treated as queries. All tests are organized according to two main approaches. The first approach uses only textual or only visual information, while the second fuses textual and visual embeddings. Within these two approaches, there are additional distinctions based on whether a single feature or multiple features represent each video.

3.3.1 Textual or Visual Features Only

When a single modality is used (textual or visual), the experiments differ depending on whether each video has multiple embeddings or a single embedding.

3.3.1.1 Textual or Visual Features Only with (no averaging)

In this setting, each video is segmented into multiple frames or scenes, producing several feature

embeddings per video. Two main tests are conducted.

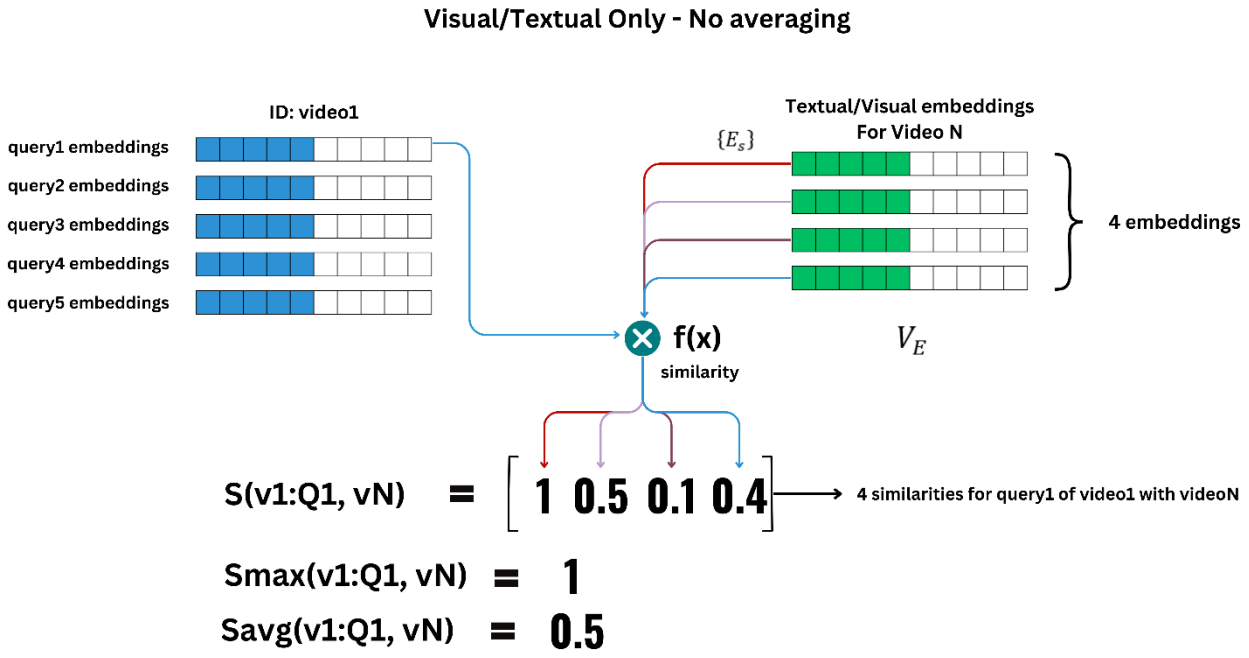

Figure 8:Visual/Textual Only - No averaging test

The figure above shows how the tests are performed on representations that has multiple embeddings per video like scene-based segmentation representation and down-sampled frames representation.

Test 1: Per-Feature Similarity Computation. All textual or visual feature embeddings, denoted $\{E_s\}$, are retrieved for each video V_E. The embedding of the query, Q, and its associated video identifier, V_{QID}, are also obtained. A similarity value is computed between Q and each embedding E in V_E, producing a set of similarity scores. These scores are then summarized in two ways: selecting the maximum similarity ("Max Similarity") or averaging them ("Average Similarity"). Once these summaries are obtained for every video, the videos are sorted in descending order of similarity, and the rank of the correct video V_Q is noted for that query. After repeating this for all queries, a random rank is chosen to simulate random-query retrieval, and the median, medoid, and mode of all ranks are computed.

3.3.1.2 Textual or Visual Features Only with (averaging applied)

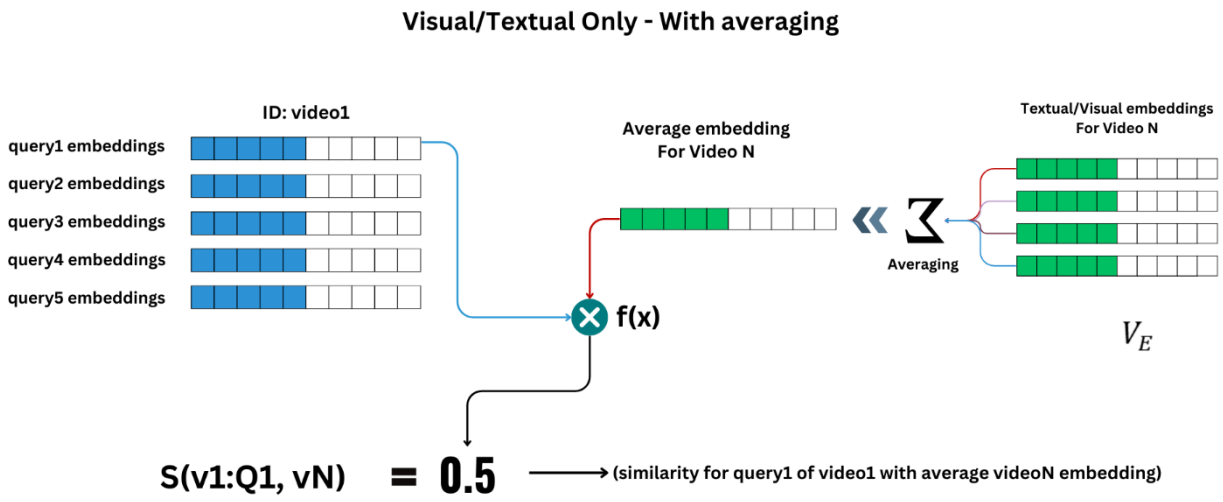

Figure 9: Visual/Textual Only - With averaging test

The figure above shows how the tests are performed on the representations that has multiple embeddings per video like scene-based segmentation representation and down-sampled frames representation but instead of comparing it independantly the representation is combined into single representation.

All feature embeddings $\{E_s\}$ for a given video V_E are averaged into a single vector E. Formally,

$$E = \frac{1}{n}\sum_{i=1}^{n} E_{s_i}$$

where n is the total number of feature embeddings for V_E. The query embedding Q is then compared to E for every video. The videos are ranked by descending similarity to Q, and the rank of the correct video V_Q is determined. After gathering these ranks for all queries, a random rank is sampled for random-query simulation. As in the previous test, the median, medoid, and mode of the rank distribution are also calculated.

3.3.1.3 Single feature per video

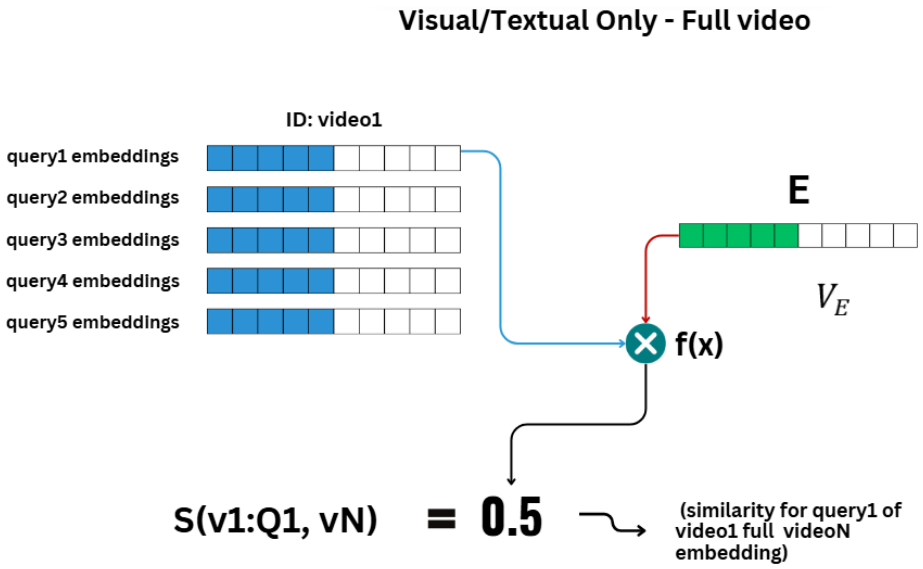

Figure 10: Single Feature per Video test

The figure above shows how the tests are performed on the representations that has single embedding per video in this case the full video. When each video is represented by only one textual, visual embedding, or both summed together (like in Feature-based fusion which we will discuss later), the procedure is simpler. The embedding E for every video V_E and the embedding of the query Q are used to compute a single similarity score per video-query pair. These scores are sorted in descending order, and the rank of the correct video V_Q is recorded. To simulate random-query retrieval, a random rank is selected, and the median, medoid, and mode of the overall rank list are then calculated.

3.3.2 Fusion of Textual and Visual Features

To determine whether combining textual and visual embeddings improves retrieval, two fusion methods are examined: feature-based fusion and score-based fusion.

3.3.2.1 Feature-based fusion

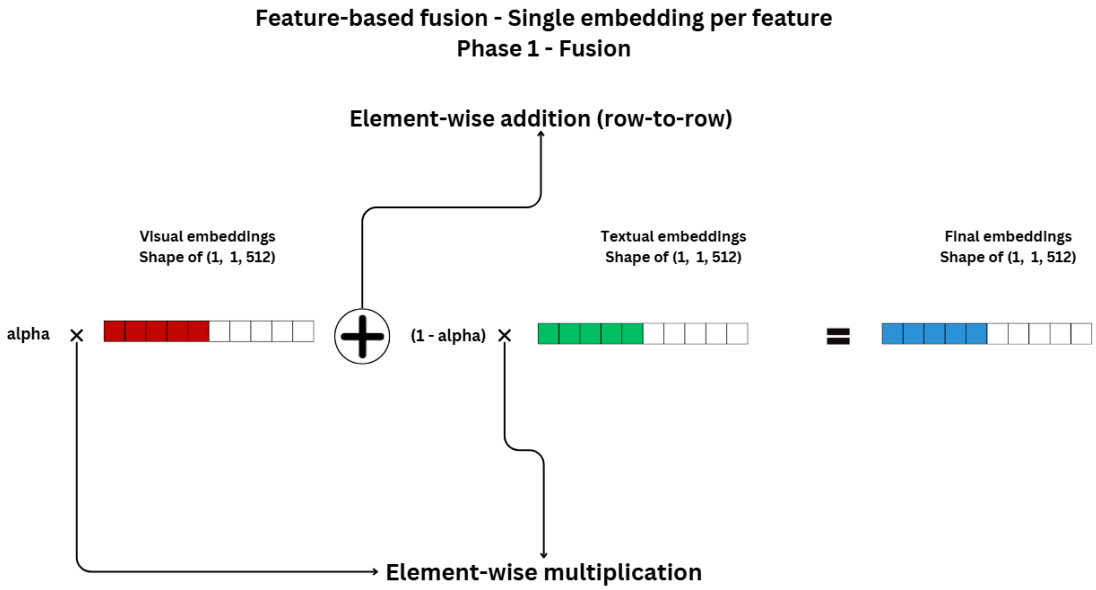

Figure 11: Feature-based fusion - Single embedding per feature pipeline

The figure above shows how the textual and visual embedding are fused to create a final embedding as a result of weighted sum. A single fused embedding E is created by computing a weighted sum of the visual embedding E_V and the textual embedding E_T. The formula is

$$E = \alpha E_V + (1 - \alpha) E_T \text{ with } 0 \leq \alpha \leq 1$$

Here, α specifies the contribution of each modality. In frame-level or scene-level settings, each video may yield multiple fused embeddings, which are processed according to the per-feature or averaged-embedding strategy. For video-level representations, there is only one fused embedding per video, and the single-feature procedure applies. After computing the similarity scores for each query, a random rank is drawn for random-query retrieval simulation, and the median, medoid, and mode ranks are computed to characterize the distribution.

3.3.2.2 Score-based fusion

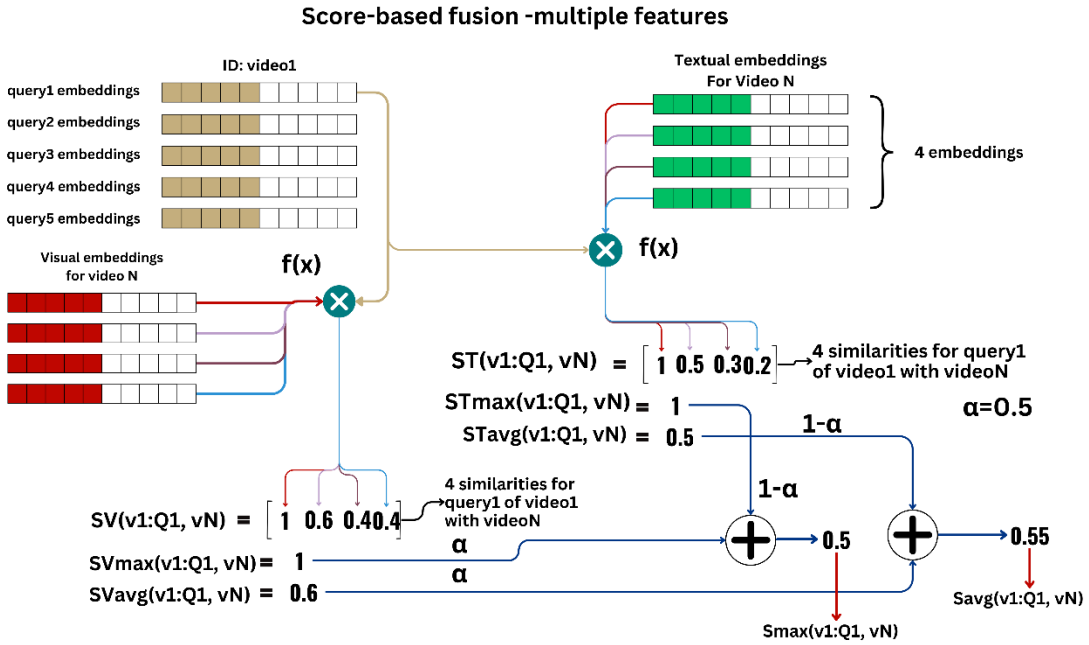

Figure 12: Score-based fusion -multiple features

The figure above shows how the system calculates separate scores for multiple textual and visual embeddings representations then fuses these scores to find the best match.

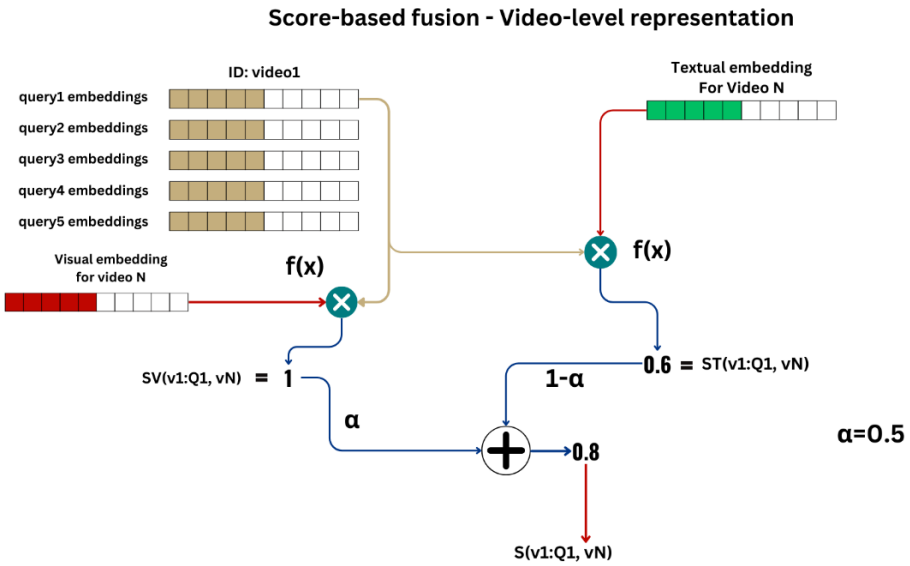

Figure 13: Score-based fusion - Video-level representation

The figure above shows how the single embeddings (visualand texual) representation's score is calculated and fused together to find the best match. In score-based fusion, separate similarity scores are computed for each modality before they are combined. If S_V is the similarity between Q and E_V, and S_T is the similarity between Q and E_T, the final score S is defined as

$$S = \alpha S_V + (1 - \alpha) S_T \text{ with } 0 \leq \alpha \leq 1$$

For multiple frame- or scene-level embeddings, a score is computed for each embedding pair, and these scores are then summarized by taking either the maximum or the average. The final scores for all videos are sorted to find the correct video V_Q and determine its rank. A random rank is chosen to simulate random-query retrieval, and the median, medoid, and mode of the resulting rank distribution are calculated. In a video-level representation, where each video has a single textual and a single visual embedding, only one fused similarity score per video is generated. The same procedure of sorting, determining V_Q 's rank, and analyzing rank statistics is followed.

These tests, covering textual-only, visual-only, feature-based fusion, and score-based fusion methods at multiple representation levels (frame, scene, and video), provide a comprehensive overview of how each approach influences retrieval performance. By examining the rank distributions and simulating random-query selection, these evaluations reveal which configurations most effectively retrieve the correct video under varying conditions.

Chapter 4

Experiments and results

Contents

This chapter presents the datasets, evaluation metrics, and experimental procedures used to assess the performance of the proposed retrieval and ranking system. The experiments are organized into separate sections, first detailing the datasets, then describing the measures employed, and finally outlining the tests conducted under various configurations of textual and visual features.

4.1 Datasets

4.1.1 Microsoft Video Description (MSVD)

The Microsoft Video Description (MSVD) dataset is commonly used in video captioning and multimodal learning research. It comprises 1,970 short video clips collected from YouTube, depicting a diverse range of activities such as cooking, exercising, playing musical instruments, and interacting with animals. Each video is annotated with multiple textual descriptions, typically ranging from 40 to 55 captions per video, written in various languages. These characteristics enable both monolingual and multilingual investigations. Most videos last from a few seconds to under one minute and capture real-world scenarios with dynamic actions, thereby providing rich and varied visual content. MSVD is frequently used for video captioning, video retrieval, and the evaluation of vision-language models. The dataset poses challenges involving noisy real-world data, complex activity understanding, and the generation of coherent, semantically accurate captions. Its diverse content and extensive annotations make it an influential resource for advancing video understanding and multimodal artificial intelligence.

4.1.2 Microsoft Research Video-to-Text (MSR-VTT)

The Microsoft Research Video-to-Text (MSR-VTT) dataset is one of the largest and most extensively used datasets for video captioning and video understanding tasks. It consists of 10,000 video clips, each approximately 10 to 30 seconds in duration. These clips span a broad range of 20 categories, including sports, cooking, news, and gaming, among others. Each clip is paired with around 20 human-annotated captions, contributing to a total of 200,000 textual descriptions. The dataset is sourced from YouTube and varies considerably in complexity, visual content, and narrative structure. MSR-VTT is widely utilized for tasks such as video captioning, video retrieval, and vision-language model evaluation. It presents challenges related to complex scene understanding and multimodal information processing, and it demands the generation of accurate, contextually relevant captions. Due to its scale and diversity, MSR-VTT has become a foundational resource for research in video-language understanding and cross-modal learning.

4.1.3 Importance of Multiple Ground Truth Captions

In this project, each video in both datasets is associated with multiple ground truth captions, which serve as critical inputs for retrieval and ranking experiments. The presence of multiple captions allows for flexible testing procedures that involve different ranking strategies, including

choosing the mode (the most frequently occurring caption), the medoid (the representative caption that minimizes the cumulative distance to all other captions), or a random rank. These diverse evaluation approaches facilitate a thorough analysis of the robustness and reliability of the retrieval and ranking system.

4.2 Measures

In this project, the focus is on evaluating the effectiveness of the retrieval and ranking process in identifying relevant videos for a given query. To assess the performance of the system, a set of well-established evaluation metrics has been selected, namely Recall@k (R@k), Median Rank (MdR), and Mean Rank (MnR). These metrics are closely associated with ranking performance and provide meaningful insights into the quality of the retrieval system.

1. Recall@k (R@k) measures the proportion of queries for which at least one correct video is retrieved within the top-k results. For instance, Recall@10 evaluates the frequency with which the correct video appears within the top ten ranked results. This metric is particularly important in practical applications, as users typically expect accurate and relevant content to appear among the highest-ranked results.
2. Median Rank (MdR) determines the middle rank of the first correctly retrieved video across all queries. A lower MdR signifies that, on average, relevant videos are positioned closer to the top of the ranked list, thereby minimizing the need for users to navigate through lower-ranked results. This metric is particularly useful in reflecting the retrieval system's consistency in ranking relevant content prominently.

3. Mean Rank (MnR) calculates the average rank of the first correctly retrieved video for all queries. While similar to MdR, MnR is more sensitive to outliers, thereby offering a broader perspective on retrieval performance. It provides an indication of the system's ability to rank relevant videos optimally across different queries, highlighting potential inconsistencies in ranking performance.

By employing these metrics, a comprehensive evaluation of the system's ability to retrieve and rank relevant videos is conducted. This approach ensures that the retrieval process aligns with the project's objective of delivering high-quality and user-friendly search results.

4.3 Results

Insights from Data: Medoid & Random Query Rankings

In our evaluation, we focused on medoid and random query rankings because these metrics offer a more realistic assessment of retrieval performance; mode rankings tend to be unrealistically high, while median rankings are consistently very low, making them less representative of real-world scenarios. This approach provides a balanced and reliable measure of our system's ability to retrieve relevant content across frames, scenes, and full video representations. To see the complete results, please refer to the full study available in our references [7] .

MSRVTT and MSVD Results: Downsampling – Before vs. After Augmentation

MSRVTT Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.100	0.193	0.256	76.0	206.78	0.115	0.199	0.255	71.5	192.81
Visual features only	0.228	0.460	0.551	**7.0**	55.226	0.263	0.516	0.619	5.0	**35.654**
Score based fusion	**0.238**	**0.465**	**0.557**	**7.0**	**52.811**	**0.277**	**0.538**	**0.630**	**4.0**	35.960
Feature-based fusion	**0.238**	**0.465**	**0.557**	**7.0**	**52.811**	**0.277**	**0.538**	**0.630**	**4.0**	35.960

Table 1: Medoid Query Ranking, before and after augmentation for Average-similarity summarization method - MSRVTT – Down-sampled frames representation

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.103	0.210	0.263	78.0	204.91	0.110	0.218	0.273	60.0	190.96
Visual features only	0.230	0.436	0.526	9.0	65.405	0.238	0.458	0.568	7.0	50.515
Score based fusion	0.244	0.453	0.536	8.0	65.245	**0.258**	**0.497**	**0.591**	**6.0**	50.459
Feature-based fusion	**0.254**	**0.467**	**0.568**	**7.0**	**63.619**	0.242	0.465	0.562	7.0	**50.153**

Table 2: Random Query Ranking, before and after augmentation for Average-similarity summarization method - MSRVTT – Down-sampled frames representation

MSVD Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.176	0.378	0.472	13.0	77.794	0.207	0.399	0.494	11.0	68.613
Visual features only	**0.378**	**0.663**	**0.772**	**2.0**	**12.201**	**0.400**	**0.696**	**0.796**	**2.0**	**10.197**
Score based fusion	0.299	0.524	0.622	4.0	39.621	0.318	0.551	0.652	4.0	33.348
Feature-based fusion	0.299	0.524	0.622	4.0	39.621	0.318	0.551	0.652	4.0	33.348

Table 3: Medoid Query Ranking for MSVD Frames (Before vs. After Augmentation, Average)

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.236	0.440	0.512	10.0	77.282	0.234	0.458	0.546	7.0	70.160
Visual features only	**0.330**	**0.622**	**0.710**	**3.0**	**21.140**	**0.364**	**0.636**	**0.739**	**3.0**	**17.333**
Score based fusion	0.300	0.545	0.634	4.0	45.452	0.335	0.581	0.678	4.0	31.749
Feature-based fusion	0.313	0.546	0.639	4.0	52.915	0.334	0.606	0.681	**3.0**	34.343

Table 4: Random Query Ranking for MSVD Frames (Before vs. After Augmentation, Max)

In both MSRVTT and MSVD experiments using downsampled frames, fusion methods that combine visual and text features consistently outperform unimodal approaches. Notably, our analysis indicates that average aggregation yields better performance than max aggregation for most metrics. For example, in MSRVTT (Table 1), fusion methods with average aggregation improve the R@1 score from about 0.238 to 0.277 and reduce the median rank from 7.0 to 4.0 after query augmentation, whereas max aggregation results are slightly less favorable. Similarly, in MSVD (Table 3), average aggregation with fusion and visual features achieves R@1 scores around 0.378–0.400 with median ranks as low as 2.0, outperforming the corresponding max aggregation outcomes presented in Table 4. Overall, these findings underscore that integrating both modalities with query augmentation and using average aggregation leads to more robust and precise video retrieval performance.

MSRVTT and MSVD Results: Scenes–Before vs. After Augmentation

MSRVTT Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.084	0.190	0.250	96.0	218.81	0.101	0.211	0.259	75.0	203.582
Visual features only	0.224	0.462	0.559	**7.0**	49.031	0.230	0.481	0.612	**6.0**	**33.826**
Score based fusion	0.244	0.453	0.536	8.0	65.245	**0.258**	**0.497**	**0.591**	**6.0**	50.459
Feature-based fusion	**0.254**	**0.467**	**0.568**	**7.0**	**63.619**	0.242	0.465	0.562	7.0	**50.153**

Table 5: Medoid Query Ranking for MSRVTT Scenes (Before vs. After Augmentation, Average)

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.082	0.180	0.252	85.0	214.83	0.104	0.205	0.265	76.5	206.724
Visual features only	0.208	0.437	0.534	**8.0**	64.097	0.218	0.449	0.561	8.0	52.463
Score based fusion	0.227	**0.444**	**0.544**	**8.0**	**61.49**	0.236	0.463	0.569	7.0	52.422
Feature-based fusion	**0.228**	0.442	0.519	9.0	63.759	**0.231**	**0.456**	**0.566**	**7.0**	**56.110**

Table 6: Random Query Ranking for MSRVTT Scenes (Before vs. After Augmentation, Average)

MSVD Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.164	0.337	0.415	19.0	87.773	0.190	0.345	0.443	16.0	76.573
Visual features only	0.312	0.597	0.704	**3.0**	20.490	0.319	0.643	0.737	**3.0**	15.784
Score based fusion	0.310	0.597	0.706	**3.0**	20.324	0.328	0.642	0.739	**3.0**	15.654
Feature-based fusion	**0.346**	**0.615**	**0.709**	**3.0**	**18.254**	**0.360**	**0.666**	**0.773**	**3.0**	**14.242**

Table 7: Medoid Query Ranking for MSVD Scenes (Before vs. After Augmentation, Average)

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.149	0.282	0.370	31.0	98.493	0.175	0.330	0.394	23.5	91.239
Visual features only	0.284	0.563	**0.669**	**4.0**	28.769	0.293	0.567	0.669	4.0	25.548
Score based fusion	0.284	**0.564**	**0.669**	**4.0**	32.033	0.278	0.557	0.676	4.0	21.658
Feature-based fusion	**0.303**	0.549	0.655	**4.0**	**28.442**	**0.316**	**0.610**	**0.707**	**3.0**	**20.033**

Table 8: Random Query Ranking for MSVD Scenes (Before vs. After Augmentation, Average)

In both MSRVTT and MSVD experiments using scene-level representations, fusion methods, as observed in the frame-based experiments, consistently outperform unimodal approaches when query augmentation is applied. For example, in MSRVTT (Table 6), fusion methods show a modest improvement in R@1 and a noticeable drop in median rank after augmentation, while textual-only features remain very low and text-only features are only slightly better. Similarly, in MSVD (Table 8), visual-only methods exhibit only a modest enhancement, whereas visual and fusion approaches achieve substantially higher R@1 scores with very low median ranks. The random query ranking results in Tables 7 and 9 further confirm that integrating both modalities with query augmentation leads to more robust and precise scene-level video retrieval performance.

MSRVTT and MSVD Results: Full Video–Before vs. After Augmentation

MSRVTT Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.106	0.171	0.216	134.0	239.43	0.106	0.201	0.250	117.5	224.267
Visual features only	0.238	**0.470**	**0.568**	**7.0**	53.978	0.271	0.526	0.629	5.0	**35.294**
Score based fusion	0.168	0.299	0.361	39.5	153.71	0.197	0.339	0.405	27.0	134.622
Feature-based fusion	**0.240**	0.457	0.562	**7.0**	**53.845**	**0.295**	**0.545**	**0.633**	**4.0**	38.374

Table 9: Medoid Query Ranking: before and after augmentation - MSRVTT-All frames representations

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.101	0.199	0.259	126.0	234.94	0.117	0.205	0.269	111.0	220.705
Visual features only	0.238	0.436	0.537	8.0	**60.516**	0.251	0.469	0.577	7.0	55.557
Score based fusion	0.176	0.320	0.385	34.0	145.013	0.180	0.335	0.403	25.5	136.177
Feature-based fusion	**0.258**	**0.459**	**0.556**	**7.0**	62.921	**0.276**	**0.496**	**0.594**	**6.0**	**52.094**

Table 10: Random Query Ranking: before and after augmentation MSRVTT-All frames representation

MSVD Results

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.185	0.355	0.440	19.0	96.382	0.204	0.399	0.488	12.0	85.951
Visual features only	0.328	0.606	0.699	3.0	20.078	0.354	0.646	0.752	3.0	13.975
Score based fusion	0.278	0.485	0.579	6.0	53.997	0.309	0.536	0.642	4.0	45.085
Feature-based fusion	**0.361**	**0.642**	**0.746**	**3.0**	**16.379**	**0.387**	**0.694**	**0.793**	**2.0**	**12.448**

Table 11: Medoid Query Ranking: before and after augmentation - MSVD-All frames representation

Method	Before Augmentation					After Augmentation				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
Text features only	0.170	0.340	0.409	28.0	110.25	0.175	0.342	0.439	18.5	97.706
Visual features only	0.267	0.552	0.661	4.0	33.204	0.330	0.569	0.691	**3.0**	25.046
Score based fusion	**0.351**	**0.604**	**0.690**	**2.0**	**25.769**	**0.345**	**0.600**	**0.693**	**3.0**	**21.430**
Feature-based fusion	0.282	0.467	0.537	2.0	34.860	0.260	0.485	0.585	6.0	57.801

Table 12: Random Query Ranking: before and after augmentation - MSVD-All frames representation

In both MSRVTT and MSVD experiments using full video representations, fusion methods with query augmentation outperform unimodal approaches—mirroring the trends seen in frame and scene experiments. For instance, in MSRVTT (Table 10), feature-based fusion increases R@1 from roughly 0.258 to 0.276 and lowers the median rank from 7.0 to 6.0, while in MSVD, fusion and visual methods achieve R@1 scores around 0.330–0.387 with median ranks as low as 2.0–3.0. Overall, integrating visual and text features with query augmentation leads to more robust video retrieval performance.

Chapter 5

Conclusion and Future Work

Contents

5.1 Conclusion

The research presented in this work set out to develop and analyze a multi-modal video retrieval framework that combines visual and textual features, applies query augmentation, and uses diverse ranking strategies. By experimenting on both the MSVD and MSR-VTT datasets, the study aimed to identify how different elements such as fusion methods, frame extraction levels, and augmented queries affect retrieval accuracy.A key finding was that multi-modal approaches consistently outperform single-modality methods, with visual embeddings providing a strong baseline but further improved by integrating caption-based textual embeddings. The study also showed that query augmentation using GPT-3.5 Turbo is a practical and impactful technique for closing the gap between varied user phrasing and the system's textual representations. This approach was particularly beneficial for text-only retrieval scenarios, but it also helped fusion-based strategies achieve more robust results.

The results show that in the MSVD dataset, which contains shorter clips generally centered on a single action, processing uniformly down-sampled frames can slightly outperform both a single full-video embedding and scene-based segmentation. Even though each clip is relatively short and appears to require fewer frames for comprehensive coverage, taking multiple snapshots (frames) still helps capture subtle details that may be missed by a single global representation or a scene-level approach. These improvements, however, remain moderate because MSVD videos rarely have complex substructures or abrupt transitions.By comparison, MSR-VTT with its longer, more varied

clips benefits significantly when videos are divided into multiple frames or scenes. This method captures a wider range of content within each clip, offering clear performance gains over using one full-video embedding.

The overall findings thus highlight that the scale and complexity of a video dataset should guide the decision to use uniform frame-down-sampling or scene-based segmentation.In terms of fusion, feature-based methods that merge textual and visual embeddings into a single vector consistently matched or exceeded the performance of score-based fusion. This was true for both MSVD and MSR-VTT, indicating that a coherent multi-modal embedding space is typically more effective than fusing similarity scores at a later stage. Finally, the system's ranking strategies demonstrated that mode-based ranking, which selects the most frequent rank among all ground-truth queries, tends to yield stable and reliable performance. Although random query selection and medoid ranking have value in specific use cases, they are less consistent overall.

5.2 Future Work

Although the current multi-modal retrieval framework has shown promising performance by integrating visual embeddings, caption-based textual features, and query augmentation, there remain multiple avenues to extend and refine its capabilities. The following sections outline the main directions for advancing this work, ranging from deeper temporal modeling to adaptive fusion strategies and user-centric enhancements.

5.2.1 Enhanced Visual Representations

A key step toward improving retrieval quality lies in exploring video-centric models such as VideoCLIP or other transformers that capture the temporal progression of actions and transitions. By going beyond static, frame-level embeddings, these architectures can align entire video sequences with textual descriptions, potentially uncovering nuanced object interactions, state changes, and scene progressions. For specialized domains such as medical imaging or sports analysis, domain-specific vision encoders can be trained on curated datasets to improve recognition of context-dependent elements, thereby enhancing retrieval precision.

5.2.2 Advanced Fusion Strategies

While the current system compares feature-based and score-based fusion, future work can investigate hierarchical and multi-stage fusion approaches. A hierarchical strategy might involve an initial coarse alignment of global video semantics, followed by a refined step that fuses finer-grained embeddings from individual frames, scenes, or object regions. Another promising direction is to replace the static weighting factor in fusion with dynamic or learnable parameters, thereby enabling the system to place greater emphasis on visual details for appearance-focused queries or rely more on textual data for context-oriented questions.

5.2.3 Query Processing and User Interaction

Although GPT-3.5 Turbo has been used to create alternative query formulations, upcoming research could integrate richer context-aware expansions that incorporate user history or session data. This would allow the system to generate query variations that align more closely with individual search patterns. Another consideration is to support multilingual or cross-lingual queries through large language models specifically designed for translation and multilingual understanding, thereby broadening accessibility and applicability to global user bases.

5.2.4 Temporal and Sequential Reasoning

Future research should examine fine-grained action recognition techniques to better capture the sequential nature of events within longer video segments. Enhanced scene transition modeling, which encodes both abrupt and smooth scene shifts, can further refine the system’s grasp of narrative flow. By treating transitional frames as distinct entities, retrieval results can better reflect the temporal context of changing actions or scenes, a vital aspect of more complex video narratives.

Bibliography

[1] "Gabeur, V., Sun, C., Alahari, K., & Schmid, C. (2020). Multi-modal Transformer for Video Retrieval.," [Online]. Available: https://arxiv.org/pdf/2007.10639

[2] "Shvetsova, N., et al. (2022). Everything at Once – Multi-modal Fusion Transformer for Video Retrieval.," [Online]. Available: https://arxiv.org/pdf/2112.04446

[3] "Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision (CLIP).," [Online]. Available: https://arxiv.org/pdf/2103.00020

[4] "Ye et al. (2023). mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality.," [Online]. Available: https://arxiv.org/pdf/2304.14178

[5] "Dong, J., et al. (2021). Dual Encoding for Video Retrieval by Text.," [Online]. Available: https://arxiv.org/pdf/2009.05381

[6] "Alpay, T., et al. (2023). Multimodal Video Retrieval with CLIP: A User Study.," [Online]. Available: https://link.springer.com/article/10.1007/s10791-023-09425-2

[7] "Video Retrieval System Dataset," [Online]. Available: https://tinyurl.com/2v3989jm

[8] "pareto _CLIP," [Online]. Available: https://www.pareto.si/blog/computer-vision-with-clip/