Int. J. Advanced Networking and Applications
Volume: 17, Issue: 02, Pages : 6821-6833 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.202517204

A Review of Video Text Retrieval Research

Wu Tianzi
School of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China
Email: 19965390178@163.com
Wang Yudi[1],Wu JiaHui[1],Li Zhengang[1],Wang Ya[1],Jin Ran[1,2]Corresponding Author
1. School of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China
2. Ningbo University of Finance & Economics, Ningbo, China
Email: 1148908353@qq.com;838540257@qq.com;2023882023@zwu.cn;18854022702@163.com;ran.jin@163.com

----------ABSTRACT----------

Video text retrieval is a hot research topic in artificial intelligence, with the core challenge being the semantic gap between visual dynamic features and discrete linguistic symbols. In recent years, with the development of large-scale models, cross-modal modeling capabilities have significantly improved, driving continuous evolution in retrieval methods regarding granularity modeling strategies. This article provides a systematic review of research methods in video text retrieval, categorizing them into single-granularity retrieval and multi-granularity retrieval. Single-granularity retrieval focuses on modeling a single semantic layer. Coarse-grained methods achieve efficient retrieval through global feature matching using pre-trained models, but they suffer from incomplete semantic coverage. Fine-grained methods enhance semantic analysis accuracy through local alignment mechanisms, but they are constrained by inherent limitations. In contrast, multi-granularity retrieval combines global scene understanding and local detail perception through hierarchical feature fusion strategies, with typical technical approaches including dynamic fusion frameworks. Analysis results indicate that multi-granularity retrieval can more comprehensively capture cross-modal semantic associations, providing a more effective solution for video text retrieval.

Keywords - **Modal mechanisms; semantic gap; multi-granularity modeling;**

Date of Submission: July 02, 2025 Date of Acceptance: August 09, 2025

I.INTRODUCTION

With the rapid growth of multimodal data, achieving accurate and efficient video text retrieval in complex semantic scenarios has become a core challenge in cross-modal understanding. In recent years, researchers have proposed various semantic modeling strategies, with one significant trend being the classification and modeling of methods based on semantic granularity. Specifically, existing methods can be broadly categorized into single-granularity methods that focus on a single semantic level and multi-granularity methods that integrate multi-level semantic information. The latter demonstrates stronger expressive capabilities and adaptability through the collaborative modeling of global, local, and segment-level semantics, gradually becoming the mainstream direction. The rapid development of multimodal large models has significantly driven the evolution of granularity modeling strategies. Early language large models (LLMs), such as GPT[2] and T5 [58], established the foundation for semantic generation and reasoning capabilities through large-scale text pre-training. However, their unimodal architecture limited their ability to perceive complex video scenes. To address this, vision-language large models (VLMs), such as CLIP [21] and VideoCLIP[72], introduced image/video encoders into language modeling frameworks, achieving preliminary unification of cross-modal semantic spaces through joint training, thereby providing foundational support for video-text retrieval tasks. However, such architectures still rely on explicit granularity alignment and global representations, making it difficult to effectively capture dynamic semantic and detailed information in videos. The emergence of generative multimodal models has further driven the formation of granularity fusion strategies. Alayrac et al.[42] proposed Flamingo, which uses cross-modal cross-attention mechanisms to map multimodal inputs to a unified semantic space, achieving granular alignment under contextual awareness. Yang et al.[5] further introduced text-guided decoding processes in CogVideoX, enhancing the model's semantic expression capabilities in video generation and retrieval tasks. Cheng et al.[76] proposed VideoLLaMA2, which, while maintaining language alignment capabilities, introduces spatial-temporal modeling and audio understanding modules, further expanding the expressive dimensions of multimodal large models in complex video scenarios. A common feature of these architectures is that they reduce reliance on explicit granularity design, instead dynamically learning and adjusting the granularity of multimodal alignment through the model's contextual understanding capabilities and training strategies, thereby enhancing its modeling capabilities for complex video scenarios. Compared to earlier methods, it demonstrates stronger robustness and generalization capabilities in complex video contexts, representing the current trend toward more flexible, context-driven semantic granularity modeling strategies.

Although existing reviews have provided detailed analyses of model structures, embedding mechanisms, and loss functions [39], systematic analyses of how “ semantic granularity ” evolves with advancements in large-scale model technology

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 02, Pages : 6821-6833 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.202517204

remain scarce. Therefore, this paper takes semantic granularity as a thread, combining representative technological paths from language large models (LLMs) to vision-language large models (VLMs) and then to generative multimodal large models (MLLMs), to analyze the differences and trends in their granularity modeling capabilities and mechanism designs. It attempts to construct a comparative perspective on granularity modeling across modality types, providing methodological summaries and directional references for future research.

The specific contributions of this paper are as follows:

(1) Further analysis of the understanding of the cross-modal semantic gap from the perspective of granularity modeling, and analysis of the theoretical foundation and research motivations of " multi-granularity retrieval " as a core technical approach;

(2) Systematic of the technical principles, evolutionary trends, and key mechanisms of single-granularity and multi-granularity retrieval methods, with summaries and comparisons presented through diagrams and tables;

(3) Summarizes the challenges faced by current methods and outlines future research trends in data construction, model design, and application, providing references for subsequent research.

II.PARTICLE SIZE CLASSIFICATION AND TECHNICAL PATH

In video text retrieval tasks, how models select and organize semantic granularity has become one of the key factors influencing semantic expression accuracy and cross-modal alignment capabilities. Granularity typically refers to the semantic level of detail that models focus on when processing modal data, such as entire video segments (global granularity), local action segments (local granularity), or frame-level features (fine granularity). To address this, the paper categorizes existing methods into two main types: single-granularity methods and multi-granularity methods. The former focuses on a single semantic level (e.g., frame-word or sentence-video), including coarse-grained methods based on global semantics and fine-grained methods targeting local details; the latter introduces multi-level feature co-modeling strategies, leveraging frame-segment joint modeling, global-local collaboration, or cross-level attention mechanisms to enhance the model's ability to express complex semantic structures. Figure 1 illustrates this granularity modeling perspective-based classification framework and its representative pathways. This classification is not only an engineering-level categorization but also rooted in the hierarchical processing mechanisms of human cognition [74], the distributed representation theory of multimodal learning [75], and the demand for task-adaptive capabilities in large-scale model training. Driven by large models, retrieval methods have gradually shifted from static global alignment to dynamic granularity adjustment and task-aware alignment. This classification helps identify the granularity modeling strategies required for different retrieval tasks, such as event recognition tasks that favor fine-grained modeling, while semantic search relies more on global semantic consistency. By clearly defining method affiliations, this framework facilitates subsequent research in task-adaptive modeling design.

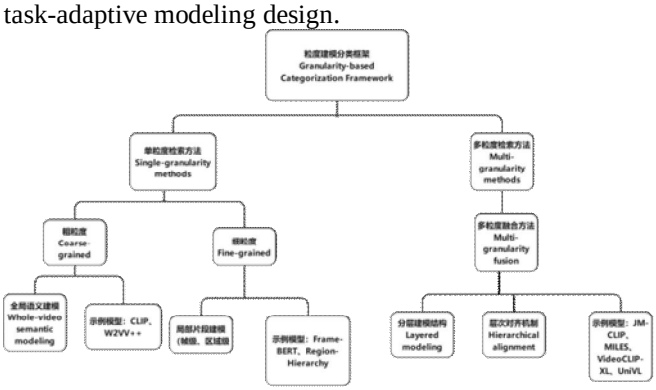


Figure 1. Granularity-based Categorization Framework for Video-Text Retrieval Methods

This figure categorizes and organizes existing video text retrieval methods from the perspective of semantic modeling granularity. The left side shows single-granularity methods, which are further divided into coarse-grained (global semantic modeling) and fine-grained (frame-level or region modeling) methods. The right side shows multi-granularity methods, which include strategies for integrating multi-layer semantic features through hierarchical structures and alignment mechanisms, and lists typical representative models. This diagram illustrates the current mainstream methods' evolution in terms of granularity modeling.

III.SINGLE-GRAIN RETRIEVAL

Single-granularity retrieval methods focus on establishing cross-modal connections between video and text at a unified semantic level. Existing methods primarily revolve around global matching, local alignment, or temporal structure modeling, emphasizing modal alignment at a specific level. While these methods offer clear structures and high computational efficiency, they remain limited when handling complex semantics or multi-scale fusion, posing key constraints for further development. Generally speaking, the design of single-granularity methods can be summarized into two core issues: first, how to model video or text independently to extract modality-specific features with semantic expressive capabilities; second, how to effectively align these two modalities at a unified granularity level. Based on this analysis, this paper reviews the representative approaches and evolutionary trends of current mainstream single-granularity retrieval methods from two dimensions: feature modeling mechanisms and cross-modal alignment strategies. Therefore, the methods are categorized into two dimensions: " feature modeling " and " alignment strategies, " focusing respectively on the quality of intra-modal representations and the mechanisms for inter-modal fusion — two critical factors influencing the performance of retrieval systems.

3.1FEATURE MODELING MECHANISML

3.1.1MULTI-SCALE TRANSFORMER ARCHITECTURE

Text modal modeling plays a foundational role in video text retrieval. Traditional methods such as Word2Vec[68] and FastText[70] can capture static word meanings but have limited capabilities in understanding context, making them unsuitable for modeling long texts and cross-sentence

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 02, Pages : 6821-6833 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.202517204

structures. The emergence of Transformer architectures, such as BERT [61], ActBert [12], RoBERTa [57], T5 [58-60], and other pre-trained models, has driven the evolution of text semantic modeling to a deeper level. Current research primarily focuses on optimizing three key issues: long text adaptation, fine-grained semantic modeling, and multi-scale information fusion. Li et al.[4] proposed DeepSeek-VL as a representative method. The model adopts a dual-path structure composed of a SigLIP encoder and a SAM-B local modeling module, responsible for global semantic extraction and detailed content characterization, respectively. To coordinate information fusion across different scales, DeepSeek-VL introduces the MLA (Multi-head Latent Attention) mechanism, dynamically activating semantic subpaths in the feature space, enabling the model to adaptively adjust its granularity focus based on task objectives. This structure not only enhances the model's ability to express complex syntactic structures and multi-layer semantics but also demonstrates strong robustness and scalability in cross-modal retrieval. In terms of fine-grained modeling, the hybrid expert model proposed by Miech et al.[52] utilizes a dynamic gating mechanism to achieve adaptive regulation between feature paths, enabling the collaborative modeling of local and global information. This mechanism significantly improves the model's ability to jointly identify text details and structural actions, making it particularly suitable for fragment-level retrieval tasks. The MDMMT [35] model further introduces a multi-layer self-attention mechanism to perform staged filtering and aggregation of multi-scale features, effectively enhancing the semantic relevance and discriminative power of feature representations. Additionally, Su et al.[1] proposed RoRE, which converts absolute positions into relative dependency structures to enhance the modeling of long-distance dependencies; Lei et al.[19] proposed ClipBERT, which employs a sparse sampling strategy to adapt to long-text scenarios, alleviating the positional information decay issue in long-distance modeling for Transformers; Clip4Clip[37] enhances local semantic capture capabilities through segment segmentation and sentence phrase matching, achieving good results in short video retrieval tasks; while the Flamingo framework proposed by Alayrac[42] combines cross-layer cross-attention mechanisms to enhance dynamic alignment capabilities between fine-grained semantics in generative modeling paths, expanding the boundaries of multi-scale expression.

Research on multi-scale Transformer architectures has demonstrated diverse approaches under different task requirements, including structural segmentation and sparse modeling, as well as cross-layer interaction and dynamic path mechanisms. However, from a theoretical perspective, current methods are largely empirically driven and lack a unified modeling framework for granularity selection mechanisms and task-driven adaptive expressions. Future research can further incorporate task-aware signals to drive granularity modeling from static definition toward dynamic regulation, thereby enhancing the model's expressive accuracy and adaptability across multiple semantic scales. From the perspective of large model adaptability, such multi-scale Transformer structures exhibit high structural compatibility with current mainstream multimodal generative models, particularly in terms of MLA mechanisms and cross-attention design, providing a transferable foundation for large models to adaptively model multi-semantic layers.

3.1.2 SPATIO-TEMPORAL ATTENTION MECHANISM

In video text retrieval tasks, effectively modeling the dynamic semantics of videos across temporal and spatial dimensions is key to enhancing cross-modal understanding capabilities. Early methods primarily relied on 3D convolutions for local temporal modeling, but these approaches suffer from high computational costs and weak cross-modal adaptability, limiting their large-scale application. With the development of Transformer architectures, researchers have increasingly shifted toward constructing unified encoding frameworks with spatio-temporal attention mechanisms to enhance modeling capabilities and expressive flexibility for continuous dynamics. Current research methods can be broadly categorized into three types: first, constructing unified spatio-temporal modeling structures, such as the ViViT[23] and Swin Transformer[20] models, which emphasize global and local dynamic relationship modeling, respectively, laying the foundation for structural frameworks; second, introducing local attention mechanisms to capture action keyframes or segments, with methods including ViCLIP [66], VideoCLIP-XL[63], and Long-CLIP[64], which enhance model robustness in dynamic scenes through action masking and temporal attention mechanisms; third, focusing on efficiency optimization and sparse modeling, represented by STAM [69], STAR[15], and Uni-AdaFocus[14], which use sparse activation strategies to reduce redundant computations while preserving key temporal features. Beyond temporal modeling, some studies further incorporate audio modalities as supplementary temporal information. Representative methods like ECLIPSE[73] employ audiovisual joint modeling mechanisms to enhance semantic coverage breadth, while TEFAL[50] uses modal decoupling structures to strengthen audio-visual synchronization accuracy, making it suitable for professional retrieval scenarios. The aforementioned methods differ in modeling strategies, structural design, and task adaptation, as detailed in Table 1, which compares and summarizes the core mechanisms of representative methods, aiding in understanding the evolution of current research and future optimization directions. Against the backdrop of the continuous evolution of multimodal large models, the fine-grained alignment and intermodal collaboration strategies achieved by spatio-temporal attention mechanisms provide important strategic support for semantic tracking and dynamic scene understanding in generative modeling tasks.

Table 1. Comparison of Spatio-temporal Modeling Methods

method	Modeling strategy	Structural features	advantage	limitations
ViViT[23]	Spatio-temporal joint modeling	Pure Transformer architecture, unified spatio-temporal encoding	Capturing long-distance dependencies	Lack of local detail modeling capabilities
Swin-V[20]	Local sliding window modeling	Multi-scale sliding window, hierarchical spatio-temporal modeling	Improving local motion representation capabilities	The structure is relatively complex, and the inference speed is limited.
ViCLIP [66]	Patch mask + action focus mechanism	Introducing an attention suppression mechanism to suppress	Better suited for complex videos	Highly task-dependent, lacking in

		redundant areas		generalization
VideoCLIP-X[63]	Frame-level focusing mechanism	Reinforce key frame timing expressions and combine them with semantic attention control.	Improve sequential comprehension skills	The structure is relatively heavy, and the training cost is relatively high.
STAM[69]	Attention weighting filter fragment	Integration of multi-scale feature selection module	Improve timing focus capabilities	Limited processing performance for long videos
STAR[15]	Sparse Attention + Token Selection Mechanism	Introducing sparse activation strategies and selective modeling	Significantly reduce computing costs	Expression ability limited by token selection strategy
Uni-AdaFocus[14]	Dynamic focus mechanism	Modality-guided attention sparsification	Retain semantically significant fragments	Limited by the accuracy of the attention mechanism
ECLIPSE[73]	Audio-visual feature fusion	Dual video and audio branches, integrating ResNet and ASTstructure	Heterogeneous modal joint modeling	Susceptible to audio-visual synchronization issues
TEFAL[50]	Modal separation + semantic-guided attention	Decoupling audio-visual feature pathways, emphasizing audio-visual synchronization segments	Suitable for professional settings (such as surgical videos)	Limited versatility, complex structure

3.2 Cross-modal alignment path

The cross-modal matching mechanism in single-granularity retrieval methods is primarily designed around semantic alignment granularity. Based on different levels of modal information integration, these methods can be divided into three categories: global feature matching, local semantic alignment, and spatio-temporal structure alignment. They focus on overall semantic consistency, fine-grained alignment accuracy, and cross-frame dynamic modeling capabilities, respectively, and are core branches of current mainstream technical approaches.

3.2.1 Global feature matching

Global feature matching methods focus on constructing a unified semantic space, emphasizing the overall semantic consistency between text and video, and are suitable for efficient retrieval tasks in large-scale databases. These methods typically use a dual-tower structure or contrastive learning framework to encode text and video separately and map them to the same embedding space for feature-level matching. Although this approach has advantages in terms of retrieval efficiency and modeling cost, it often suffers from semantic loss when dealing with complex contexts and local dynamic information. Luo et al.[58] proposed the Video-RAG model, which introduces a generative path to break through the expressive bottleneck of purely discriminative structures. This method designs a lightweight language generation module on the basis of a shared semantic space to generate potential textual descriptions of videos, and enhances the coverage of ambiguous semantics through similarity assessment. This structure not only enhances the model's ability to perceive detailed semantics but also expands the robustness of global matching under ambiguous queries. Other studies have also optimized the global matching framework. Li et al.[67] proposed the W2VV++ model, which combines a GRU text encoder with a triplet loss function to improve modeling capabilities for complex sentence structures and semantic similarity. Hao et al.[6] proposed a bidirectional nearest neighbor matching strategy, effectively strengthening the semantic correspondence between video-text pairs; while the DGL model[10] introduced a dynamic prompt mechanism to guide the model to focus on semantically prominent local regions, improving expression accuracy without significantly increasing structural complexity. In addition, the CNN + Word2Vec framework proposed by Frome et al.[48] early on, despite its simple structure, laid the foundational path for shared semantic spaces, providing important inspiration for subsequent methods. Overall, global matching methods have significant advantages in terms of efficiency and structural simplicity, making them suitable for rapid screening and coarse-grained ranking scenarios. However, there is still room for optimization in terms of expression layers and semantic density. Future research could explore more efficient context-guided mechanisms or fusion strategies for lightweight generation modules to enhance the model's coverage of semantic details while maintaining inference speed. The dual-tower structure and shared semantic space mechanism align closely with the current mainstream large-model representation paradigm, offering good module decoupling and transfer potential, particularly for building scalable multi-stage retrieval systems.

3.2.2 Local semantic alignment

Local semantic alignment methods aim to achieve more precise modal correspondences at micro-levels such as frame-word and segment-phrase. These methods typically employ cross-attention mechanisms, dynamic token weighting, and contrastive learning to optimize interaction paths between cross-modal semantic units, significantly improving the alignment accuracy and semantic resolution capabilities of retrieval models. Local semantic alignment methods shift toward fine-grained modeling at the frame-word and segment-phrase levels. Fang et al.[3] proposed the PROTA model, which designs a token-level attention weighting mechanism that dynamically adjusts the matching strength between visual and text tokens based on semantic intensity, thereby effectively enhancing the ability to capture semantic core regions. During the training phase, PROTA simultaneously introduces alignment loss and token confidence adjustment mechanisms, enabling the model to maintain high sensitivity to key semantics in complex phrases or long sentences, effectively alleviating the performance degradation issues of traditional alignment strategies in redundant semantic or ambiguous contexts. Building on this, Jind et al.[8] proposed the HBI model, which introduces Banzhaf interaction values to model token influence, further enhancing the model's perception of micro-semantic structures; Wang et al.[63] proposed the VideoCLIP model, which significantly improves the local expression of action semantics through frame-phrase contrastive learning; while DIVA [53] introduced a diffusion model structure to optimize the modeling of action continuity and local semantic scenes. Additionally, pre-training models such as CLIP proposed by Radford et al.[21] and ALIGN proposed by Jia et al.[28] established the foundational framework for local alignment mechanisms in shared space encoding, demonstrating good transferability and scalability. Therefore, local semantic alignment methods demonstrate a clear advantage in improving fine-grained alignment accuracy, particularly for retrieval tasks with specific query descriptions or complex video scene structures. However, such methods are typically structurally complex, computationally intensive, and highly sensitive to data

distribution stability. Future research could combine task difficulty awareness mechanisms, structural compression strategies, or multi-modal token reordering mechanisms to further enhance their generalization capabilities and deployment efficiency.

3.2.3SPATIAL-TEMPORAL STRUCTURE ALIGNMENT

In dynamic scenes, the spatio-temporal structure alignment mechanism has emerged. Its primary objective is to address the challenge of preserving semantic consistency while enabling the localization and mapping of actions and events across frames. Yang et al.[7] proposed TubeDETR, which achieves direct prediction of action boundaries by explicitly modeling the video intervals and spatial locations corresponding to text. Luo et al.[37] introduced CLIP4Clip, which employs InfoNCE loss to enhance segment-level alignment performance. Researchers Wang et al.[14] proposed the Uni-AdaFocus architecture, which combines text content for key frame selection, thereby reducing computational complexity while maintaining expressive capability. Current spatio-temporal alignment methods have demonstrated certain action modeling capabilities, but challenges remain in handling blurry event boundaries, semantic drift, and compression of long videos with multiple events. Further breakthroughs are needed in structural adaptation and data mechanisms. Compared to static matching strategies, spatio-temporal structural alignment methods better align with the dynamic flow characteristics of video semantics, demonstrating unique advantages in modeling action sequences and spatial associations. However, current methods still face challenges in handling issues such as blurred action boundaries and complex event structures, which affect the model's stability and interpretability. Future research could explore introducing self-supervised temporal masking strategies or combining multi-scale event modeling mechanisms to enhance the model's ability to capture key dynamic segments and adapt to complex temporal semantics. Methods that enhance spatio-temporal structural modeling capabilities have the potential to address the current shortcomings of large models in understanding dynamic scenes, particularly in generative retrieval and long-video understanding tasks, where they could expand the contextual window and enhance event perception.

IV.MULTI-GRANULARITY SEARCH

Multi-granularity methods in video text retrieval (VTR) aim to overcome the limitations of traditional coarse-grained and fine-grained modeling in terms of semantic coverage and matching accuracy. By introducing information fusion mechanisms across multiple semantic levels, these methods enhance the model's ability to perceive and express complex semantic structures. In recent years, related research has emerged, with researchers exploring more expressive structural design paths through strategies such as hierarchical modeling, granularity coordination, and multimodal alignment. For example, the UniVL model proposed by Luo et al. [38] employs a hierarchical feature fusion mechanism to uniformly model global and local features of both video and text, achieving good results in multimodal retrieval tasks. To systematically summarize the evolutionary trends and key mechanisms of such methods, this section will analyze them from two dimensions: feature modeling mechanisms and cross-modal fusion strategies, with a focus on their structural design concepts and representative implementation methods.

4.1MULTI-GRANULARITY FEATURE EXTRACTION MODELING

The introduction of multi-granularity modeling aims to address the issue of semantic hierarchy mismatch between video and text. Videos often contain overall scenes, local actions, and instantaneous details, while text descriptions may also involve multiple granularities ranging from word-level phrases to complete sentences. Therefore, how to effectively incorporate semantic hierarchy and modality collaboration during the modeling stage has become one of the core objectives in current method design. Some researchers have chosen to explicitly introduce granularity divisions in the structural design. For example, Jiang et al.[44] proposed the HCMI model, which constructs a nested structure of words, phrases, and sentences on the text side, and separately models frames, segments, and global semantics on the video side. They employ a triple-path alignment strategy to enhance granularity matching capabilities. The key mechanism lies in achieving hierarchical semantic alignment through multi-granularity comparison loss, and introducing a granularity gating module to dynamically adjust the fusion weights of features at different granularity levels based on context, thereby maintaining expressive fineness while mitigating redundant interference. Meanwhile, Ge et al.[49] proposed MILES, which further incorporates granularity nested structures into the pre-training stage, combined with an inter-layer parameter sharing mechanism, to enhance the consistency and generalization capabilities of multi-granularity representations in structural design.

Some researchers have also attempted to incorporate granularity differences into modality fusion. For example, Kim et al.[29] and Yu et al.[30] introduced bilinear pooling and matrix decomposition mechanisms, respectively, to construct joint representations during the visual-linguistic feature fusion stage. This approach captures the importance of semantically relevant regions in an implicit manner, rather than pre-defining a fixed granularity structure. These methods simplify the modeling process and exhibit a certain degree of semantic adaptability. However, when dealing with long texts or complex temporal actions, their granularity discrimination capabilities are relatively limited, and they are prone to feature dilution issues. Additionally, the introduction of the Transformer architecture has driven the development of multi-layer semantic extraction mechanisms. HiT[25] distinguishes semantic scales by introducing multi-layer contrastive losses, while Dzabraev et al.[36] propose MDMMT-2, which incorporates positional encoding and staged training strategies to enhance hierarchical learning of granular semantic information. Arnab et al.'s ViViT[22] uniformly divides videos into spatio-temporal patches, emphasizing the integrated modeling capability of temporal and spatial granularity. Additionally, Gabeur et al.[34] proposed MMT, which further introduces multi-scale collaborative mechanisms in multi-modal paths to ensure consistency

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 02, Pages : 6821-6833 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.202517204

between visual and linguistic representations across different semantic levels. Rahman et al.[17] expanded this into a multi-scale dynamic fusion strategy for multi-modal inputs. Existing methods have evolved from early feature concatenation to a complex mechanism system combining structural partitioning, modal interaction, and semantic fusion in granularity modeling. However, most methods still rely on static structural settings or fixed depths to encode granularity, lacking the ability to adaptively adjust granularity paths for specific tasks and content. This poses challenges in balancing granularity redundancy, modeling conflicts, and computational overhead. Future research can further explore structures with granularity-aware and dynamic control capabilities.

This will enhance the model's adaptability, expressive accuracy, and inference efficiency in complex semantic scenarios.

The problem of granularity alignment is a core challenge in video text retrieval. The diversity and dynamism of video modalities make it difficult for a single global feature to cover complex semantics. Radford et al.[21] proposed CLIP, which encodes videos and texts into global vectors and calculates similarity through dot product:

$$S(v,t) = \langle f_v(v), f_t(t) \rangle \quad \textbf{(1)}$$

$S(v,t)$ Indicates the similarity score between videos and text. $f_v(v)$ Indicates video feature vector, $f_t(t)$ Indicates text feature vector.This strategy is efficient and concise, but it cannot analyze local details and temporal changes, resulting in limited performance in complex scenarios. To compensate for the shortcomings of global matching, Luo et al.[37] proposed CLIP4Clip, which introduces a contrastive learning framework and uses InfoNCE loss to achieve fragment-phrase alignment:

$$\mathcal{L}\text{InfoNCE} = -\log \frac{\exp(s(v_i,t_i)/\tau)}{\sum j=1^{M} \exp(s(v_i,t_j)/\tau)} \quad \textbf{(2)}$$

v is the vector representation of the video clip, t is the vector representation v of the video clip, $sim(v,t)$ is the similarity between the video clip and the phrase, τ is the temperature hyperparameter used to control the smoothness of the distribution, N is the number of negative samples, and represents other phrases related to the video clip in the same batch.This local alignment approach improves the model's ability to perceive fine-grained semantics, but it relies too heavily on explicit granularity design, which increases computational complexity. Further work includes JM-CLIP[62], which proposes a multi-granularity hierarchical modeling framework by weighting and combining alignment losses of different granularities:

$$\mathcal{L}\text{JM-CLIP} = \lambda_1 \mathcal{L}\text{frame-phrase} + \lambda_2 \mathcal{L}\text{segment-sentence} + \lambda_3 \mathcal{L}\text{video-text} \quad \textbf{(3)}$$

Researchers explicitly introduced granularity hierarchy relationships, enhancing the model's ability to cover complex semantics, but at the cost of flexibility and increased modeling and training costs. With the emergence of large models, BLIP-2 [4] proposed the Query Attention mechanism, which guides visual features into the language space through query vectors, achieving a weakening of explicit granularity alignment paths:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad \textbf{(4)}$$

Q Represents the query vector, which is an input representation. K Represents the key vector, indicating the association degree between the query and each key. V Represents the value vector, which plays a role in transmitting information in attention calculation. QK^t Represents the dot product between the Q query vector and the K key vector K. Represents the dimension of the vector.This approach simplifies granularity design and enhances end-to-end capabilities, but granularity control capabilities are implicit in the context representation, lacking interpretability and adjustability. Overall, the formula evolution of granularity alignment methods reflects ongoing exploration in the field to balance efficiency, expressive power, and task adaptability. Future work should consider how to introduce task-aware granularity adjustment mechanisms within large-scale model frameworks to balance the completeness, interpretability, and computational feasibility of multi-granularity expressions.

4.2MULTI-GRANULARITY ALIGNMENT AND FUSION STRATEGY

To facilitate understanding of the differences in granularity modeling between different structural mechanisms, Figure 3 shows a comparison of the semantic alignment strategies of the three current mainstream methods, including explicit structure, unified structure, and generative structure.

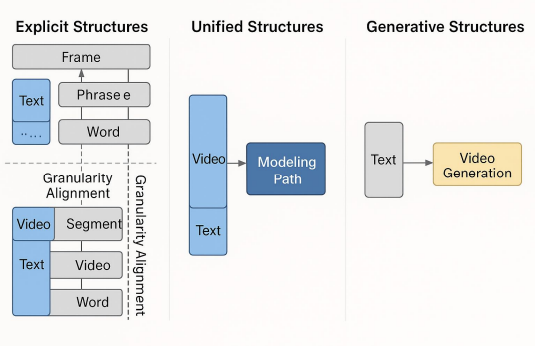


figure3.Comparison chart of multi-model granularity alignment mechanisms

Figure 3. Comparison of granularity alignment mechanisms among three types of video text retrieval methods. The left side represents explicit structural mechanisms, which typically align text words, phrases, and sentences with video frames or segments using multi-granularity features; the middle represents unified structural mechanisms, which model video and text in a unified representation space through shared encoding paths; the right side represents generative structural mechanisms, which directly generate representations or content consistent with video semantics from input text, bypassing explicit alignment steps.

4.2.1EXPLICIT STRUCTURAL MECHANISM

Explicit structural mechanisms explicitly establish semantic hierarchical correspondences between video and text by embedding multi-granularity feature segmentation and alignment paths within the model. Such methods typically decompose video into hierarchical representations such as frames, segments, and the entire video, and utilize cross-attention or feature fusion modules to achieve

multi-modal alignment, thereby enhancing the accuracy of understanding complex semantic content. Representative methods include JM-CLIP[62], VideoCLIP-XL[63], TEFAL[50], CLIP4Clip[37], TubeDETR[7], and TIB[50], among others. For example, Ge et al.'s [62] JM-CLIP is based on multi-level granular video structure segmentation, encoding frames, segments, and global information separately, and combining multi-level text alignment paths of words, phrases, and sentences. It introduces a temporal cross-attention mechanism to strengthen the collaborative modeling capabilities between local details and global semantics. This method addresses the issues of semantic sparsity and insufficient action representation in long videos, achieving a Text-Video R@1 of 62.5% on the MSR-VTT[18] dataset, significantly outperforming TEFAL [50]'s 48.1%. The specific structure is shown in Figure 4, which illustrates the fusion process of cross-layer granularity features and multi-modal alignment paths. Wang et al.[63] proposed VideoCLIP-XL, which focuses on the issue of expression efficiency in long video retrieval. It employs keyframe selection and a multi-scale interaction structure, combined with a distributed embedding strategy, to enhance the model's ability to capture temporal information. On the DiDeMo[24] dataset, it achieved a R@1 of 62.3%, outperforming VideoCLIP [72]'s 47.8%. Building on this, Ibrahimi et al.[50] introduced TEFAL, which incorporates a modal branch structure and semantic constraint strategy, independently modeling video-text and audio-text pairs to enhance the consistency of multimodal alignment. CLIP4Clip[37] is based on a cross-encoder design, enhancing dynamic association modeling between local segments and phrases, making it suitable for short video scenarios. TubeDETR[7] explicitly models the correspondence between text descriptions and video segments, enabling event-level semantic localization and expanding the boundaries of retrieval models in action recognition and semantic alignment. As demonstrated by the above methods, explicit structural mechanisms combine granularity division and modular design to effectively enhance the model's ability to express and match complex temporal semantics, particularly suitable for tasks such as long video understanding and structured event retrieval.

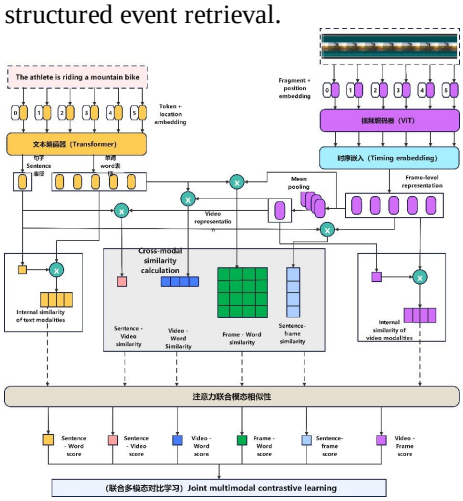


Fig4.JM-CLIP Framework diagram

The model is based on a video encoding path at the frame, clip, and global granularity levels, combined with text alignment channels at the word, phrase, and sentence levels, and achieves collaborative modeling and semantic fusion at the granularity level through a multimodal cross-attention mechanism.

4.2.2 Unified Modeling Structure

To alleviate the issues of redundant computation and information fragmentation between different granularities caused by explicit structural schemes, some methods have shifted toward integrating multi-granularity semantic alignment objectives within a unified structure. Among these, the work by Luo et al.[38] is one of the earlier efforts to introduce the encoder-decoder structure into video-text retrieval tasks. This method simultaneously processes global and local features of both video and text within a unified modeling pipeline, thereby avoiding the redundant information propagation issues between modules of different granularities in explicit structures. Its core design includes a shared encoder for intra-modal feature extraction and a collaborative decoder for cross-modal alignment, guided by a shared attention mechanism to facilitate semantic fusion. Additionally, UniVL[38] incorporates a multi-task pre-training strategy that balances description generation and retrieval, enhancing the model's generalization capabilities across different tasks. On datasets such as MSR-VTT, UniVL achieves superior retrieval performance compared to traditional modular methods while maintaining structural consistency. Other works include MMT[34] proposed by Gabeur and HiT proposed by Liu et al.[25], which construct a unified representation space through a multi-layer Transformer structure and combine multi-scale contrast mechanisms to achieve cross-granularity consistency constraints. Building on this, Madasu et al.[55] extended the approach to multilingual scenarios, proposing the MuMUR framework to align multilingual text and visual content in a unified semantic space, demonstrating the potential of unified structures for granularity generalization. Meanwhile, Rahman et al.[17] proposed a multi-modal, multi-scale mechanism to enhance multi-modal expression quality from a structural fusion perspective. Although unified modeling structures improve multi-granularity coupling capabilities and overall consistency, they still face issues such as high structural complexity and the absence of granularity control mechanisms. Future research could explore task-aware dynamic granularity control strategies to enhance adaptability and interpretability.

4.2.3 Generative Semantic Alignment

Unlike the traditional “ representation matching ” paradigm, generative semantic alignment methods attempt to bypass the explicit feature alignment step and directly generate semantic representations from the input text that match the target video, thereby constructing an end-to-end retrieval path. Yang et al.[5] proposed CogVideoX, one of the representative works in this direction, which combines diffusion models with a three-dimensional spatio-temporal attention mechanism to model the deep integration of semantic, spatial, and temporal information during the generation process. The model achieves frame-by-frame video synthesis through a dual-branch generation structure and introduces a time-aware diffusion generation process to simulate the dynamic transition of actions in long sequences. Additionally, CogVideoX employs a multi-layer semantic control mechanism to guide the generation results in

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 02, Pages : 6821-6833 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.202517204

structurally expressing composite actions, sequences, and relationships within complex language instructions. This method demonstrates strong semantic generalization capabilities in zero-shot video-text retrieval tasks, particularly in complex video scenarios involving multi-scene and multi-action mixed expressions. In contrast, Li et al.[43] proposed the T2VIndexer, which constructs a more streamlined generation path by directly mapping input text to video index vectors via a sequence-to-sequence structure, thereby omitting traditional feature matching and alignment processes and achieving higher inference efficiency. Other researchers, such as Yuan et al.[31], proposed the GDR framework to fuse generative and discriminative structures, while Madasu et al.[55] proposed MuMUR to achieve multilingual generative alignment, both of which expanded the expressive space of generative retrieval structures. Although generative paths have expanded the scope of understanding of semantic alignment mechanisms and broken the structural division limitations of traditional methods, there are still challenges in terms of retrieval accuracy, semantic control capabilities, and interpretability. Future research could consider introducing semantic granularity as a generative condition, using prompt guidance and structural constraints to enhance the consistency and controllability of generated semantics. For example, Zhang et al.[40] constructed a branch-based multimodal fusion framework, combining Video Q-Former and ImageBind to achieve cross-granularity embedding alignment;

4.2.4 Redesigning granularity fusion mechanisms driven by multimodal large models

Unlike traditional methods that rely on explicit granularity divisions such as frames, segments, and events, recent multimodal large models achieve higher expressive freedom by implicitly integrating information across different levels in the semantic space through a unified cross-modal encoding structure. Among these, Li et al.[4] proposed BLIP-2 as a typical representative method. The model uses the Q-Former module to select key information from image or video features using a set of learnable query vectors and maps it to an embedding space acceptable to the language model. This structure simplifies the cross-modal modeling process while also achieving implicit granularity filtering of visual information to some extent. Compared to models like CLIP[21] and VideoCLIP[72], which rely on global representations or fixed-resolution features, BLIP-2 demonstrates stronger expressive power when handling fine-grained semantics, complex scenes, and contextually ambiguous regions. However, this granularity selection is entirely dependent on the spontaneous distribution of attention weights, lacking explicit control and interpretability. When processing videos with continuous actions or strong temporal dependencies, the ambiguity of granularity focusing may still lead to alignment errors or semantic misjudgments.

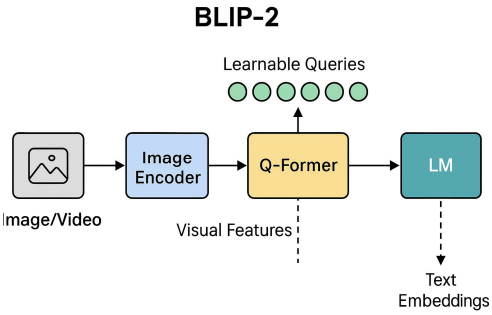


Figure 4. Schematic diagram of the BLIP-2 architecture

The model consists of three parts: an image/video encoder, a Q-Former module, and a language model. Q-Former uses learnable query vectors to extract key information from visual features and maps it to the input space of the language model, achieving cross-modal granular expression and alignment.

The overall modeling process of BLIP-2 is shown in Figure 4. Through a three-stage structure of visual feature extraction, semantic filtering, and language projection, it achieves multi-level semantic expression of the input image/video content. In contrast, Alayrac et al.[42]'s Flamingo feeds image tokens and text tokens into the language model together, demonstrating strong multi-turn dialogue capabilities, but it also faces similar bottlenecks in semantic alignment and temporal modeling. Subsequent researchers Yang et al.[5] proposed the CogVideoX model, which attempts to combine diffusion-based generation structures with expert path control mechanisms to enhance temporal and spatial granularity expression flexibility. Lu et al. proposed the DeepSeek-VL[51] model, which activates different expert paths via the MLA module to enhance adaptability to complex tasks. Although these designs have to some extent expanded the granularity control capabilities of large models, they still primarily rely on implicit mechanisms and lack task-aware dynamic adjustment strategies.

V.DATASET AND RESULTS

The performance of cross-modal retrieval methods for video text requires a reasonable evaluation system to measure. A good evaluation system can guide the future direction of cross-modal retrieval. This chapter will compare and review existing cross-modal retrieval methods for video text from the perspectives of evaluation datasets and evaluation metrics. Table 1 summarizes granularity-based retrieval methods and provides a comparative review of experiments conducted on some of these methods.

5.1DATASET

The following sections provide detailed descriptions of several commonly used datasets, which consist of numerous video clips and their corresponding annotated descriptions. A general overview of these datasets is presented in Table 3.

MSR-VTT[18]: Over 7,000 videos were collected using 257 popular queries from YouTube, and these videos were edited into 10,000 clips ranging from 10 to 30 seconds in length. Contains 100,000 GIF images and 120,000 descriptions, with an average duration of approximately 3.6 seconds per GIF. Of these, 80,000 GIFs are used for training, 10,000 for

validation, and the test set includes 11,000 GIFs, each accompanied by three descriptions.

ActivityNet[65]: Includes 20,000 video clips and 100,000 descriptions, with each clip averaging approximately 180 seconds, totaling 849 hours. Each clip averages 3.65 sentence descriptions, with time localization and annotation conducted via Amazon Mechanical Turk.

LSMDC[41]: Includes 118,081 video clips and 118,114 sentences, with videos sourced from 202 movies, each clip lasting approximately 45 seconds. The dataset is divided into 101,046 training videos, 7,408 validation videos, and a test set of 1,000 videos.

DiDeMo[24]: This dataset comprises 10,000 unedited 25–30-second videos, each accompanied by 4 descriptions, totaling 40,000 descriptions. Each video is divided into 5-second segments, with the dataset split into 8,395 training videos, 1,065 validation videos, and 1,004 test videos.

HowTo100M[33]: This dataset contains 136 million video clips from 1.22 million instructional videos on YouTube, categorized into 12 categories. Each clip lasts approximately 6.5 minutes, with annotations generated via YouTube ASR automatic transcription or API translation.

Vatex[32]: Includes 41,250 videos and 825,000 Chinese and English subtitles. Among these, 206,000 are parallel translations between English and Chinese. The dataset is divided into 25,991 training videos, 1,500 validation videos, and 1,500 test videos.

Table3. Overview of Common Datasets

database	Database type	Number of videos	Number of fragments	Text quantity	Source	Duration
MSR-VTT	Video/Text	7180	10000	200000	YouTube	41.2h
LSMDC	Video/Text	200	128118	128118	Movies	158h
Vatex	Video/Text	41250	41250	825000	YouTube	-
Activity Net	Video/Text	20000	100000	100000	YouTube	849h
HowTo100M	Video/Text	12210	13600	13600	YouTube	13447h

5.2EVALUATION CRITERIA

This section summarizes commonly used evaluation metrics for analyzing the performance of cross-modal retrieval methods, including recall, precision, and average precision.

(1) Precision (Prec) and recall (Rec) are defined as the ratio of true positives (TP) to TP + false positives (FP), where TP is the number of correctly retrieved samples and FP is the number of positive samples predicted as negative. Prec can be used to measure the success probability of an information retrieval system. Recall (Rec) is defined as the ratio of TP to TP + FN, where FN is the number of negative samples predicted as positive samples. Unlike precision, TP + FN represents the total number of matching data in the dataset rather than the total number of retrieved samples. Precision and recall can be expressed as follows:

$$Prec = \frac{TP}{TP + FP} \quad (5)$$

$$Rec = \frac{TP}{TP + FN} \quad (6)$$

(2) Recall@k (R@k)

In recall rate, the commonly used evaluation standard for cross-modal retrieval of video text is R@k, i.e., Recall@k. R@k calculates the percentage of times at least one correct result is found among the first k search results. For video-text cross-modal retrieval, common query ratios are R@1, R@5, and R@10, representing the recall rates for the first 1, 5, and 10 results, respectively. Generally, higher recall rates result in lower precision rates.

(3) MedR

In video retrieval, MedR (Median Rank) is a commonly used experimental evaluation metric, meaning the median rank. It represents the median position of the correct option in the retrieval result sequence, reflecting the situation where the correct result is in the middle of the retrieval list across all retrieval operations. The lower the MedR value, the closer the correct result is to the front of the retrieval list, indicating better performance of the retrieval model or algorithm.

5.3EVALUATION RESULTS

Table4. Comprehensive performance comparison of single-granularity retrieval methods on mainstream datasets

method	dataset	Text-Video				Video-Text			
		R@1	R@5	R@10	Medr	R@1	R@5	R@10	MEDR
CLIP[21]	MSR-VTT	31.2	53.7	64.2	4	27.2	51.7	62.6	5
CLIP4Clip[37]	ActivityNet	46.4	43.8	-	-	48.1	49.5	-	-
Clip4clip[37]	MSR-VTT	44.5	73.2	81.6	-	28.3	56.7	67.9	2
T2VLAD[26]	LSMDC	14.3	32.4	42.2	16	14.2	33.5	41.7	17
Clipbert[19]	ActivityNet	21.3	49	63.5	6	-	-	-	-
W2VV++[67]	VATEX	12.7	34.8	47.1	12	20.7	48.9	62.1	6
ActBERT[26]	MSR-VTT	8.6	23.4	33.1	36	-	-	-	-
HiT[25]	ActivityNet	29.6	60.7	95.6	3	-	-	-	-
TEFAL[50]	MSR-VTT	48.1	73.8	82.8	2	-	-	-	-
HCGC[45]	VATEX	43.6	72.4	80.1	-	29.3	57.8	68.5	39.7
TEFAL[50]	LSMDC	26.8	46.1	56.5	7	-	-	-	-
TEFAL[50]	VATEX	61	90.4	95.3	1	-	-	-	-

HGR[47]	MSR-VTT	16.4	38.3	49.8	11	23	42.2	53.4	-
STAMttn[69]	DIDEMO	51.9	77.6	85	2	-	-	-	-
X-clip[56]	DiDeMo	47.8	79.3	-	-	47.8	76.8	-	-
X-clip[56]	LSMDC	26.1	48.4	-	-	26.9	46.2	-	-
T2VLAD[26]	MSRVTT	29.5	59	70.1	4	31.8	60	71.1	3

Table5. Comprehensive performance comparison of multi-granularity retrieval methods on mainstream datasets

method	dataset	Text-Video				Video-Text			
		R@1	R@5	R@10	Medr	R@1	R@5	R@10	MEDR
VideoCLIP-XL[63]	DiDeMo	62.3	-	-	-	62.7	-	-	-
UniVL[38]	MSR-VTT	21.2	49.6	63.1	6	-	-	-	-
VideoCLIP-XL[63]	ActivityNet	46.4	-	-	-	48.1	-	-	-
JM-Clip[62]	MSR-VTT	62.5	85.3	93.7	1	58.9	88	96	1
MILES[65]	LSMDC	17.8	35.6	44.1	15.5	-	-	-	-
JM-Clip[62]	ActivityNet	72	94.6	96.8	1	-	-	-	-
MuMUR[55]	MSR-VTT	44.8	72	82.5	2	-	-	-	-
MILES[49]	LSMDC	17.8	35.6	44.1	15.5	-	-	-	-
MILES[49]	MSR-VTT	37.7	63.6	73.8	3	-	-	-	-
MDMMT[35]	LSMDC	18.8	38.5	47.9	12.3	-	-	-	-

As shown in the table data, multi-granularity methods significantly outperform single-granularity methods in video text retrieval. On the MSR-VTT dataset, JM-Clip achieves a Text-Video R@1达62.5 and Video-Text R@1达58.9, far surpassing the best single-granularity method (TEFAL 48.1; CLIP4Clip 28.3). Multi-granularity methods enhance cross-modal alignment accuracy by fusing global and local features, with Medr generally reaching 1, while the best single-granularity method achieves 2. Performance varies significantly across different datasets: JM-Clip achieves R@1 of 72 on ActivityNet, while the highest score on LSMDC is only 34.2 (VideoCLIP-XL), indicating that long-video understanding is more challenging. Although the CLIP series has strong generalization capabilities, single-granularity methods exhibit significant gaps in cross-modal retrieval. This suggests that multi-granularity methods, by fusing features across multiple levels, can better bridge the semantic gap and improve retrieval performance.

VI. FUTURE PROSPECTS FOR VIDEO TEXT RETRIEVAL

As scholars continue to optimize video text retrieval methods, accuracy and efficiency have significantly improved. However, the increasing diversity and scale of data modalities have created new demands. In the future, existing model methods will still require continuous improvement. The following are some thoughts and outlooks:

(1) While existing multi-granularity retrieval methods have alleviated the semantic gap, they rely too heavily on pre-trained models. In the future, it will be necessary to explore joint learning mechanisms based on raw data to achieve autonomous modality feature extraction and reduce dependence on expert models.

(2) Video website-sourced datasets generally suffer from high noise levels, and existing methods such as hierarchical contrastive learning have not fundamentally resolved this issue. Enhancing video denoising techniques will become a key breakthrough for improving multi-granularity retrieval accuracy.

(3) Social media datasets suffer from insufficient category diversity. While current multi-modal datasets encompass multi-dimensional information such as video and audio, they still require the construction of high-quality multi-modal datasets to support model evolution in the face of complex retrieval demands.

ensure a high-quality product, diagrams and lettering MUST be either computer-drafted or drawn using India ink. Figure captions appear below the figure, are flush left, and are in lower case letters. When referring to a figure in the body of the text, the abbreviation "Fig." is used. Figures should be numbered in the order they appear in the text.

Table captions appear centered above the table in upper and lower case letters. When referring to a table in the text, no abbreviation is used and "Table" is capitalized.

VII. CONCLUSION

Video text retrieval, as a key task in cross-modal semantic understanding, has seen the development of a rich array of methods in recent years, particularly in feature extraction, semantic alignment, and granularity modeling. This paper takes “ semantic granularity ” as its starting point, constructing a classification framework that combines single-granularity and multi-granularity approaches, and systematically analyzes the modeling mechanisms,

alignment strategies, and structural evolution paths of different methods. Recently, related research has also been conducted in the fields of machine learning optimization[77] and cloud computing security technology[78]. These advances, combined with the findings of this paper, indicate that multi-granularity retrieval has advantages over single-granularity retrieval in terms of semantic coverage.Overall, single-granularity methods still hold an advantage in modeling efficiency but lack in detail analysis and adaptation to complex semantics. Multi-granularity methods, on the other hand, enhance semantic coverage and cross-modal reasoning capabilities by introducing hierarchical structures and alignment mechanisms. In recent years, certain large-scale model architectures that integrate unified modeling paths with context-guided mechanisms have demonstrated potential in granularity expression and modality fusion, expanding the design space for retrieval models. Future research could further focus on task-oriented granularity control strategies, generative mechanisms, and structural transparency optimization to enhance model stability and generalization capabilities in practical application scenarios.

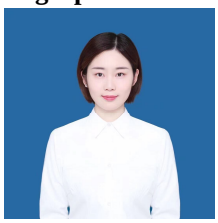
REFERENCES

[1] SU J, AHMED M, LU Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.
[2] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.
[3] FANG H, ZANG X, BAN C, et al. ProTA: Probabilistic token aggregation for text-video retrieval[C]//Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2024: 1-6.
[4] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2023: 19730-19742.
[5] YANG Z, TENG J, ZHENG W, et al. Cogvideox: Text-to-video diffusion models with an expert transformer[J]. arXiv preprint arXiv:2408.06072, 2024.
[6] HAO X, ZHANG W, WU D, et al. Dual alignment unsupervised domain adaptation for video-text retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18962-18972.
[7] YANG A, MIECH A, SIVIC J, et al. Tubedetr: Spatio-temporal video grounding with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16442-16453.
[8] JIN P, HUANG J, XIONG P, et al. Video-text as game players: Hierarchical Banzhaf interaction for cross-modal representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2472-2482.
[9] JIN P, LI H, CHENG Z, et al. Diffusionret: Generative text-video retrieval with diffusion model[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 2470-2481.
[10] YANG X, ZHU L, WANG X, et al. DGL: Dynamic global-local prompt tuning for text-video retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(7): 6540-6548.
[11] WANG R, CHEN D, WU Z, et al. Bevt: Bert pretraining of video transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14733-14743.
[12] GE Y, GE Y, LIU X, et al. Bridging video-text retrieval with multiple choice questions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16167-16176.
[13] ZHU L, YANG Y. Actbert: Learning global-local video-text representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8746-8755.
[14] WANG Y, ZHANG H, YUE Y, et al. Uni-AdaFocus: Spatial-temporal dynamic computation for video recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[15] SHI F, LEE C, QIU L, et al. Star: Sparse transformer-based action recognition[J]. arXiv preprint arXiv:2107.07089, 2021.
[16] LI P, XIE C W, ZHAO L, et al. Progressive spatio-temporal prototype matching for text-video retrieval[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4100-4110.
[17] RAHMAN M M, MARCULESCU R. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation[C]//Medical Imaging with Deep Learning. PMLR, 2024: 1526-1544.
[18] XU J, MEI T, YAO T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5288-5296.
[19] LEI J, LI L, ZHOU L, et al. Less is more: Clipbert for video-and-language learning via sparse sampling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7331-7341.
[20]LIU Z, NING J, CAO Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3202-3211.
[21] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.
[22] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.
[23] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.

[24] JIN W, ZHAO Z, ZHANG P, et al. Hierarchical cross-modal graph consistency learning for video-text retrieval[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1114-1124.
[25] LIU S, FAN H, QIAN S, et al. Hit: Hierarchical transformer with momentum contrast for video-text retrieval[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11915-11925.
[26] WANG X, ZHU L, YANG Y. T2vlad: Global-local sequence alignment for text-video retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5079-5088.
[27] TONG Z, SONG Y, WANG J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[J]. Advances in Neural Information Processing Systems, 2022, 35: 10078-10093.
[28] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International Conference on Machine Learning. PMLR, 2021: 4904-4916.
[29] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[J]. arXiv preprint arXiv:1610.04325, 2016.
[30] YU Z, YU J, FAN J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1821-1830.
[31] YUAN P, WANG X, FENG S, et al. Generative dense retrieval: Memory can be a burden[J]. arXiv preprint arXiv:2401.10487, 2024.
[32] WANG X, WU J, CHEN J, et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4581-4591.
[33] MIECH A, ALAYRAC J B, SMAIRA L, et al. End-to-end learning of visual representations from uncurated instructional videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9879-9889.
[34] GABEUR V, SUN C, ALAHARI K, et al. Multi-modal transformer for video retrieval[C]//Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part IV 16. Cham: Springer International Publishing, 2020: 214-229.
[35] MIN S, KONG W, TU R C, et al. Hunyuan_tvr for text-video retrivial[J]. arXiv preprint arXiv:2204.03382, 2022.
[36] DZABRAEV M, KALASHNIKOV M, KOMKOV S, et al. Mdmmt: Multidomain multimodal transformer for video retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3354 – 3363.
[37] LUO H, JI L, ZHONG M, et al. Clip4clip: An empirical study of clip for end-to-end video clip retrieval[J]. arXiv preprint arXiv:2104.08860, 2021.
[38] LUO H, JI L, SHI B, et al. Univl: A unified video and language pre-training model for multimodal understanding and generation[J]. arXiv preprint arXiv:2002.06353, 2020.
[39] ZHU C, JIA Q, CHEN W, et al. Deep learning for video-text retrieval: A review[J]. International Journal of Multimedia Information Retrieval, 2023, 12(1): 3.
[40] ZHANG H, LI X, BING L. Video-llama: An instruction-tuned audio-visual language model for video understanding[J]. arXiv preprint arXiv:2306.02858, 2023.
[41] ROHRBACH A, TORABI A, ROHRBACH M, et al. Movie description[J]. International Journal of Computer Vision, 2017, 123: 94-120.
[42] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: A visual language model for few-shot learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.
[43] LI Y, YU J, GAI K, et al. T2VIndexer: A generative video indexer for efficient text-video retrieval[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 3955-3963.
[44] JIANG J, MIN S, KONG W, et al. Tencent text-video retrieval: Hierarchical cross-modal interactions with multi-level representations[J]. IEEE Access, 2022.
[45] JIN W, ZHAO Z, ZHANG P, et al. Hierarchical cross-modal graph consistency learning for video-text retrieval[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1114-1124.
[46] HUANG J, LI Y, FENG J, et al. Clover: Towards a unified video-language alignment and fusion model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14856-14866.
[47] CHEN S, ZHAO Y, JIN Q, et al. Fine-grained video-text retrieval with hierarchical graph reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10638-10647.
[48] FROME A, CORRADO G S, SHLENS J, et al. Devise: A deep visual-semantic embedding model[J]. Advances in Neural Information Processing Systems, 2013, 26.
[49] GE Y, GE Y, LIU X, et al. Miles: Visual BERT pre-training with injected language semantics for video-text retrieval[J]. arXiv preprint arXiv:2204.12408, 2022.
[50] IBRAHIMI S, SUN X, WANG P, et al. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 12054-12064.
[51] LU H, LIU W, ZHANG B, et al. DeepSeek-VL: Towards real-world vision-language understanding[J]. arXiv preprint arXiv:2403.05525, 2024.
[52] MIECH I, LAPTEV I, SIVIC J. Learning a text-video embedding from incomplete and heterogeneous data[J]. arXiv preprint arXiv:1804.02516, 2018.
[53] WANG W, SUN Q, ZHANG F, et al. Diffusion feedback helps CLIP see better[J]. arXiv preprint arXiv:2407.20171, 2024.
[54] TIAN K, CHENG Y, LIU Y, et al. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(6): 5207-5214.

[55] MADASU A, AFLALO E, STAN G B M, et al. Mumur: Multilingual multimodal universal retrieval[J]. Information Retrieval Journal, 2023, 26(1): 5.
[56] MA Y, XU G, SUN X, et al. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 638-647.
[57] LIU Y. RoBERTa: A robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019, 364.
[58] LUO Y, ZHENG X, YANG X, et al. Video-RAG: Visually-aligned retrieval-augmented long video comprehension[J]. arXiv preprint arXiv:2411.13093, 2024.
[59] XIA Y, ZHAO Q, LONG Y, et al. SensoryT5: Infusing sensorimotor norms into T5 for enhanced fine-grained emotion classification[J]. arXiv preprint arXiv:2403.15574, 2024.
[60] LI Y, SU Z, LI Y, et al. T5-SR: A unified seq-to-seq decoding strategy for semantic parsing[C]//ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.
[61] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171 – 4186.
[62] GE M, LI Y, WU H, et al. JM-CLIP: A joint modal similarity contrastive learning model for video-text retrieval[C]//ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 3010 – 3014.
[63] WANG J, WANG C, HUANG K, et al. VideoCLIP-XL: Advancing long description understanding for video CLIP models[J]. arXiv preprint arXiv:2410.00741, 2024.
[64] ZHANG B, ZHANG P, DONG X, et al. Long-CLIP: Unlocking the long-text capability of CLIP[C]//European Conference on Computer Vision. Cham: Springer, 2025: 310 – 325.
[65] KRISHNA R, HATA K, REN F, et al. Dense-captioning events in videos[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 706 – 715.
[66] WANG Y, HE Y, LI Y, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation[J]. arXiv preprint arXiv:2307.06942, 2023.
[67] LI X, XU C, YANG G, et al. W2VV++ fully deep learning for ad-hoc video search[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1786 – 1794.
[68] MIKOLOV T. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013, 3781.
[69] SHEN L, ZHAO S, ZHANG Y, et al. Spatio-temporal attention for text-video retrieval[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025.
[70] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135 – 146.
[71] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives[J]. arXiv preprint arXiv:1707.05612, 2017.
[72] XU H, GHOSH G, HUANG P Y, et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding[J]. arXiv preprint arXiv:2109.14084, 2021.
[73] LIN Y B, LEI J, BANSAL M, et al. Eclipse: Efficient long-range video retrieval using sight and sound[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 413 – 430.
[74] MARR D. Vision: A computational investigation into the human representation and processing of visual information[M]. Cambridge: MIT Press, 2010.
[75] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423 – 443.
[76] CHENG Z, LENG S, ZHANG H, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs[J]. arXiv preprint arXiv:2406.07476, 2024
[77]Tiwari D V K, Singh P. Classification of Motor Imaginary in EEG using feature Optimization and Machine Learning[J]. International Journal of Advanced Networking and Applications (IJANA), India, 2023, 15(02): 5887-5891.
[78]Abed A H. The Techniques of authentication in the Context of Cloud Computing[J]. Int. J. Advanced Networking and Applications, 2025, 16(04): 6515-6522.

Biographies and Photographs



Name:Wu Tianzi
Education:2024–Present, Zhejiang Wanli University, China
Research Interests: Multimodal, Computer Vision