Rethinking Video-Language Model from the Language Input Perspective

Xiang Fang[1], Wanlong Fang[2], Changshuo Wang[3*], Xiaoye Qu[4], Daizong Liu[5]
[1]School of Software Engineering, Huazhong University of Science and Technology
[2]Nanyang Technological University, Singapore
[3]University College London
[4]Huazhong University of Science and Technology
[5]Wuhan University
xfang9508@gmail.com, wanlongfang@gmail.com, wangchangshuo1@gmail.com, xiaoye@hust.edu.cn, daizongliu@whu.edu.cn

Abstract

Driven by the wave of large language models, Video-Language Models (VLMs) have become a significant yet challenging technology to bridge the gap between videos and texts. Although previous VLM works have made significant progress, almost all of them implicitly assume that all the texts are predefined by the specific template. In real-world applications, such a strict assumption is impossible to satisfy since 1) predefining all the texts is extremely time-consuming and labor-intensive. 2) these predefined text inputs are too restrictive and user-unfriendly, limiting their applications. It is observed that given a video input, texts with similar semantics but different templates lead to various performances. To this end, in this paper, we propose a novel plug-and-play framework for various VLM-based methods to fully bridge videos and texts. Specifically, we first generate positive and negative texts from the original ones to target specific text components. Then, we propose an attribute-based text reasoning strategy to mine fine-grained textual semantics of generated texts. Finally, we utilize videos as guidance to conduct cross-modal bridging by designing a self-weighted loss. Extensive experiments show that the proposed method can serve as the plug-and-play module to effectively improve the performance of state-of-the-art VLMs.

Introduction

Due to remarkable success, Video-Language Models (VLMs) have attracted more and more attention (Rizve et al. 2024; Fang, Zhang, and Chan 2026; Fang et al. 2025a, 2023c, 2022; Jia et al. 2025b,a, 2024, 2020; Qiu et al. 2023, 2025b, 2024b, 2025a, 2024a; Liang et al. 2023a, 2025, 2023b). VLMs require cooperation from both computer vision and natural language processing for precise semantic alignment and have a wide range of applications such as video summarization (Abdar et al. 2024; Fang et al. 2023b; Fang, Fang, and Wang 2025; Fang, Easwaran, and Genest 2025; Fang and Fang 2026; Fang, Fang, and Wang 2026; Fang et al. 2026, 2025c, 2024b, 2025d,b, 2024a,c, 2023a, 2021b; Fang, Easwaran, and Genest 2025; Fang et al. 2020, 2021a; Fang, Easwaran, and Genest 2024; Fang and Hu 2020) and video question answering (Yu et al. 2024). Benefiting from the strong knowledge integration ability in large language models (LLMs) (Fang, Zhang, and Chan 2026; Zhang et al. 2025; Liu et al. 2023, 2024), VLMs show superior performances in solving complex image-language tasks by utilizing appropriate human-instructed prompts (Hakim et al. 2023; Wang et al. 2025a,b; Wang, Fang, and Tiwari 2025; Wang et al. 2026, 2025c). In VLMs, the sentence text is the most important input that accompanies the video due to its human-friendly and descriptive nature (Zhang 2018; Zhang et al. 2022).

Current VLMs contain three main popular yet challenging tasks in Figure 1: video question answering (VideoQA) (Gao et al. 2023), video sentence grounding (VSG) (Zhang et al. 2023) and video-text retrieval (VTR) (Zhu et al. 2023). VideoQA is a significant multi-modal task where a model is given a video along with a natural language question about the video content, and it must generate or select the correct answer. The task requires the model to understand the visual cues in the video, as well as the language of the question, to provide relevant and accurate responses. Given a language text and an untrimmed video, VSG aims at retrieving the start and end timestamps of the target video moment, semantically according to a sentence text. Given a language text, VTR targets to retrieve relevant videos from a large video database, which can be either text-based (text-to-video retrieval) or video-based (video-to-video retrieval). The significant goal of VTR is to find videos that best match the given input by analyzing visual content, actions, and sometimes audio cues. The performance in above three downstream tasks depends on their capability to extract video features and align them with text features. Some VLM-based methods only perform data augmentations on the video input to improve the model robustness during training in every epoch. Unfortunately, existing methods only utilize predefined texts without any augmentation. In real-world applications, these sentence texts with similar textual semantics might be inputted with different structure/vocabulary variations from various users. As shown in Figure 1(b), the text ("Person pours water into a glass") shares the same semantics as the text ("Water is poured into a glass by person"). However, previous methods yield dissimilar grounding re-

*Corresponding Author.

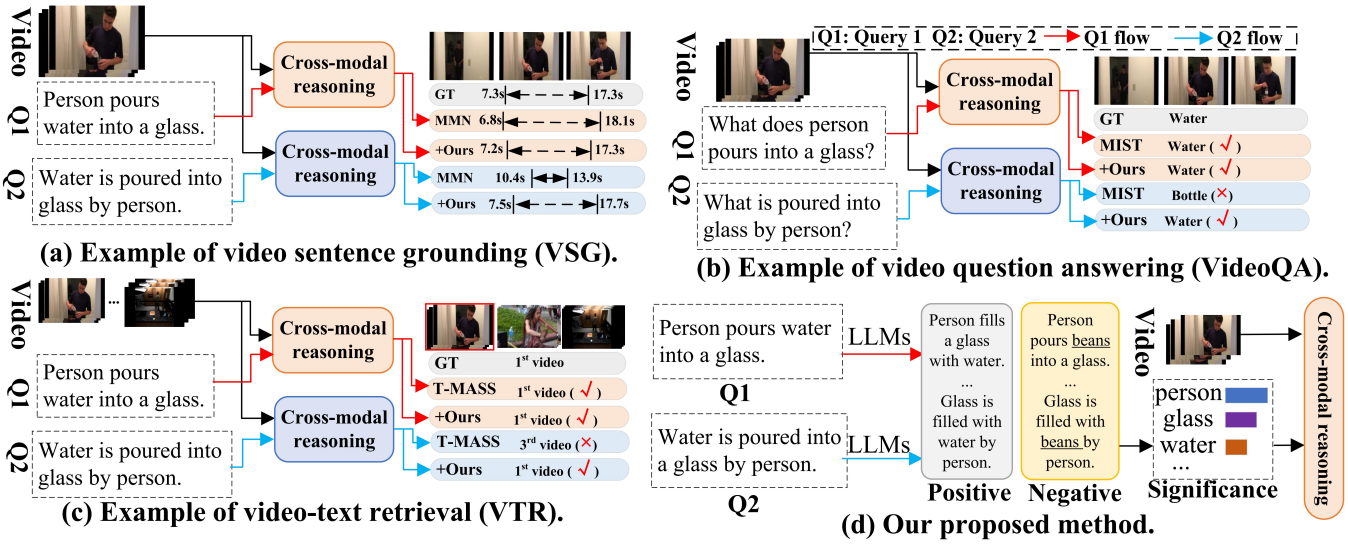

Figure 1: (a-c) Example of the VLM tasks (VSG, VideoQA and VTR), where our proposed method can serve as a plug-and-play module for previous VLM models to enhance their efficiency. (d) Pipeline of our method.

sults in Figure 1. The main reason is that these methods cannot utilize their weak text encoder to learn discriminative textual representations, which illustrates the significance of handling the text variations. Therefore, it is important to ensure that the designed VLM-based method is robust enough to deal with various texts with different templates. However, existing language augmentation approaches are not sufficiently effective to integrate the multi-modal inputs. Some methods target to replace or mask some words in a sentence, which only brings limited influence in diversifying the text structure/vocabulary. It is comparatively weaker than video augmentations. The target language augmentation approach should effectively rewrite sentence texts while reserving the core textual semantics. The approach is urgently required for model training to achieve the best results.

In this paper, we propose a simple yet highly effective framework to improve the robustness and performance of VLMs. Specifically, we generate multiple variants of each text in the video-text pairs. Based on the variation-origin pairs, we utilize them as examples to diversify all the texts in video-text datasets. Different from previous sentence augmentation works that only change some words to preserve sentence structures, we generate rich variations for diverse text inputs due to their extensive training datasets and emergent properties. Based on the above sentence augmentation, each video corresponds to diverse texts. Moreover, we utilize the original as the anchor to generate various hard negative texts by changing different sentence parts. In particular, we utilize precise prompt engineering to modify specific parts of the sentence with the rest parts unchanged. Also, we generate positive samples that lie relatively far from the anchor in the embedding space. To further understand the latent textual semantics, we design an attribute-based text reasoning strategy for fine-grained text mining. To bridge videos and texts, we incorporate these generated text samples with the video as guidance by a self-weighted cross-fusion loss. With these diverse texts, we target to train VLM models with augmentation from the text perspective.

Our main contributions are summarized as follows: 1) We make the first attempt to explore the effect of the template-free text for the robust VLM task, where we localize the target activity by a user-friendly text with any form instead of a predefined text. Also, we propose a novel plug-and-play framework for VLM-based methods to fully understand the text inputs. 2) We first generate diverse positive and negative texts by augment the original text from word and structure levels. To selectively integrate these generated texts, we design two modules (generating multi-level texts module and attribute-based text reasoning module) to understand the text input from different granularity. Besides, a video-guided self-weighted cross-modal bridging loss is proposed to integrate these sentence components by assigning adaptive weights to these components. 3) For three representative downstream tasks (VSG, VideoQA and VTR), we conduct experiments on many popular yet challenging datasets. To obtain different text types, we leverage LLMs to construct a small set of variation-origin by two strategies: original datasets and augmented datasets. 4) Extensive experimental results on both original datasets and augmented datasets show that our proposed model can serve as a plug-and-play module for state-of-the-art VLM-based methods.

Related Works

Video-language models (VLMs). The breakthrough of LLMs in language-oriented tasks (Ma et al. 2024; Zhang et al. 2025) and the emergence of GPT-4 have prompted researchers to explore the potential of LLMs in assisting with a range of tasks across multi-modal scenarios (Carolan, Fennelly, and Smeaton 2024). This has led to the development of a new field, namely VLMs. A variety of strategies and models have been proposed to address the discrep-

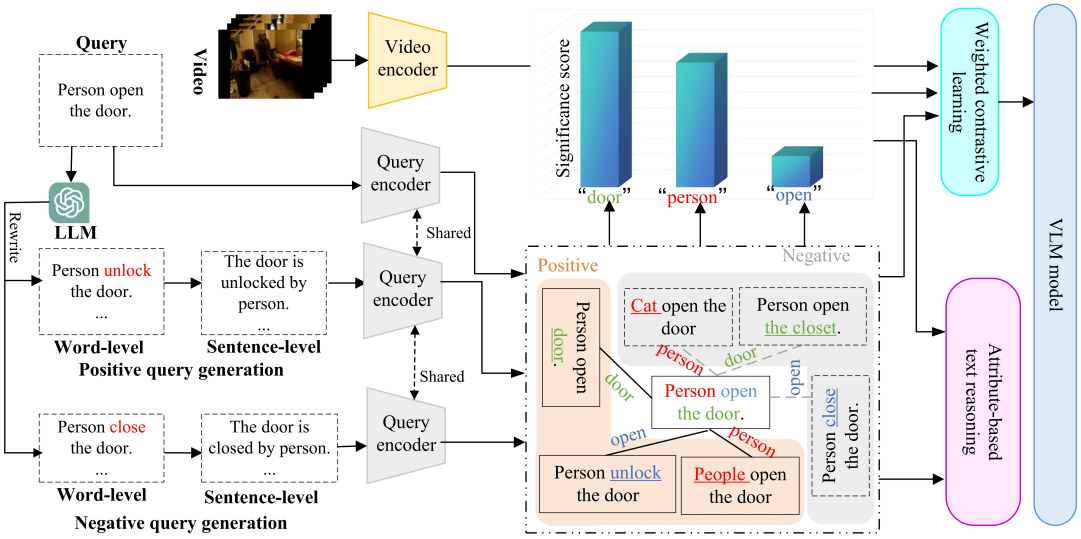

Figure 2: Illustration of our proposed framework.

ancy between text and other modalities. Some works employ learnable texts to extract visual information and generate language using LLMs conditioned on the visual features. Models including GPT-4o, MiniGPT-4 and LLaVA learn simple projection layers to align the visual features from visual encoders with text embeddings for LLMs. Additionally, parameter-efficient fine-tuning is adopted by introducing lightweight trainable adapters into models. Several benchmarks have verified that VLMs demonstrate satisfactory performance on visual perception and comprehension.

Although these methods have achieved promising results, all of them heavily rely on correctly aligned multi-modal datasets. Therefore, it is highly expected to develop a VLM model that is robust to different texts with similar semantics, which has not been studied as far as we know. Thus, *we make the first attempt to reveal the text understanding problem in VLM task and propose to eliminate the negative impact of the different texts with any template.*

Methodology

We elaborate on the proposed method, which strengthens the text encoder to obtain consistent representations for various semantically similar texts in real-world multi-modal datasets (*e.g.*, Charades-STA (Sigurdsson et al. 2016)), multiple semantics-similar texts often share a video moment with the target activity. For example, "Person opens the door" and "The door is opened by person" have similar semantics. Since the text template is fixed, it is still challenging to diversify the text input. Thus, we design a text augmentation module to generate semantically similar texts. The overall framework is shown in Figure 2.

Problem statement. Due to the strong language processing ability of of LLMs, we utilize LLMs to generate various texts by replacing different components for simulating the practical labeling process in the format-free setting. Previous VLM methods (Yu et al. 2023; Wang et al. 2022, 2024) cannot well handle these texts with similar semantics since they do not fully understand the textual information in the sentence. To address the existing models' limitations in correlating major sentence parts with suitable video representations, we present a novel plug-and-play framework to fully understand the language input by generating negative and positive samples targeting specific sentence parts. These samples facilitate improved perception of specific parts of the sentence, eventually enhancing the understanding of video-language correlation. We use the generated samples as auxiliary samples alongside the original training samples by employing a novel video-guided self-weighted cross-modal bridging loss. The proposed approach is application-agnostic and can be adopted successfully in various downstream multi-modal tasks.

Multi-level Text Augmentation

By treating original texts as anchors, we leverage LLMs for generating positive and negative texts to fully understand the texts, where we regard the generated text sharing similar semantics as a positive text; otherwise, the generated text is negative. Real-world datasets include multiple video-text pairs (V, Q), where V denotes the video and Q denotes one of corresponding texts. We denote $\{Q_1, \ldots, Q_M\}$ as the textual input set in the VLM task, where M denotes the total number of sentences. For the m-th text, we denote the generated positive text as P_m and the negative text as N_m . To obtain diverse generated texts, we adopt three text augmentation approaches: human-based , chat robot-based and open-source LLM-based. For convenience, we term the pair of generated text and original text as variation-origin text pairs. These texts differ in two levels: word level and sentence structure level. Therefore, we have two types of text augmentation by a two-step process: word-level augmentation and structure-level augmentation. For convenience, we take the positive text augmentation as an example.

Word-level augmentation. In the first step, we directly rewrite the original text by changing some words. With pre-

trained LLMs, we can rewrite all the texts by the following prompt: $P'_{i,m} \leftarrow \texttt{LLM}(Q_m$, "Rewrite the text '**text**' concisely by changing the i-th word while keeping the meaning"), where **text** is substituted with the given text, and we utilize the underlined text to prompt our model for producing morphologically diverse text expressions.

To evaluate the significance of q_i on the semantics of Q_m, we can evaluate the semantic change before and after removing this word: $S_1(q_i, Q_m, c) = 1 - \cos(c \cup Q_m, c \cup Q_m \setminus \{q_i\})$, where c denotes the prompt and $\cos(\cdot,\cdot)$ is the cosine similarity function. In real-world applications, larger $S_1(q_i, Q_m, c)$ denotes that removing q_i will lead to significant semantic changing, indicating that q_i is more relevant.

Structure-level augmentation. Since different users tend to utilize various text structures for video grounding, we need to augment the text structure for more diverse texts. Similarly, we utilize the following prompt for structure-level rewriting: $P_{i,m} \leftarrow \texttt{LLM}(P'_{i,m}$, "Rewrite the text '**text**' concisely by changing the text structure while keeping the meaning"). Given a sentence Q_m, we define the sentence-level relevance of Q^i_m as the probability-weighted semantic similarity with other sentences: $S_2(Q^i_m, Q^j_m, c) = \sum_{j=1,j\neq i} \cos(Q^i_m, Q^j_m) p(Q^j_m|c)$, where $p(Q^j_m, c)$ denotes the generative probability that provides more confidence to Q^j_m, and higher $p(Q^j_m, c)$ makes Q^j_m more acceptable. An intuitive observation is that if a sentence is semantically consistent with other sentences, the sentence is more convincing and more representative.

Similar to positive text augmentation, we generate negative texts by changing their words and sentence structure. Thus, based on the multi-level language rewriting, we can conduct coarse-grained language alignment for text augmentation.

Attribute-based Text Reasoning

In fact, Section only considers the semantics of the sentence itself, ignoring the latent information of the sentence. For example, "a person is driving a car" contains two significant objects: "person" and "car". "person" corresponds to the following attributes: a head, two eyes, two arms, etc, while the attributes of "car" include: four wheels, a steering wheel, etc. There attributes will assist VLMs to understand videos and texts for bridging the visual and textual gap.

Attribute generation. For some semantically similar sentences, they always have similar attributes. Therefore, we generate the attributes for all positive and negative texts. Although embedding attributes can help us to understand the sentence, current VLM models cannot fully understand the latent semantics. For example, "a person is driving a car" and "a car is running on the road" have similar semantics. Therefore, rather than directly using the original sentence, we design a model with high confidence in visual attributes. Two intuitions are considered in this model: 1) different from the original sentence, aligning explicitly with visual attributes can push the deigned model to mine the inherent semantics in the given sentence. 2) visual attributes contain more fine-grained features, which can provide more details for cross-modal reasoning.

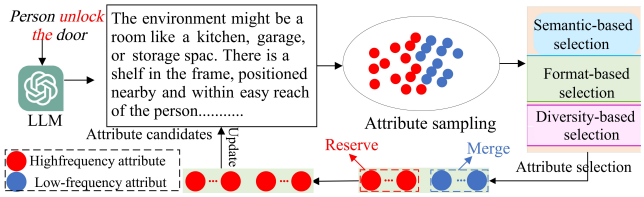

Figure 3: Our attribute selection module.

Firstly, we utilize video and text encoders to extract the video and text features. Since our model is plug-and-play, it does not depend on specific feature encoders. For the fair comparison, we adopt the same video and text encoders with compared methods. For the text Q with J words, we denote word-level text feature as $f^W = \{f^w_j\}^J_{j=1} \in \mathbb{R}^{J\times d}$ and the sentence-level text feature as $f^q \in \mathbb{R}^d$, where d is feature dimension. Similarly, we denote the extracted video features as $f^V = \{f^v_i\}^{N_v}_{i=1} \in \mathbb{R}^{N_v\times d}$, where N_v is the frame number.

Attribute sampling. We find that some generated attributes have a stronger semantic correlation with visual features than others, and some attributes have less significance (even may be hallucination information), which will lead to high computational cost. Therefore, removing some low significance can not only decrease the computational cost but also improve the model generalization. As shown in Figure 3, we address the problem by selecting effective attributes from an attribute pool. Two main criteria are utilized during the attribute selection: Firstly, we prioritize attributes that are both representative and non-redundant. Secondly, we seek attributes with the highest semantic relevance to the images when compared to other attributes. Finally, we use the following steps for attributes: 1) For the attributes a_m associated with sentence Q, we partition them into N_c clusters based on their feature similarity. This clustering strategy aims to ensure that each cluster represents a distinct aspect, *e.g.*, color or shape, in the descriptions. 2) In each cluster, we rank the attributes by assessing their similarity to visual features, and select the one with the highest relevance. By the above strategy, the following attributes will be filtered out: non-visual attributes and incorrect visual attributes that are semantics-unrelated to the videos. To obtain optimal attributes, we introduce three attribute selection strategies:

Semantic-based selection. Firstly, we want to make the sentence text has the similar semantics with its generated attributes. Since the natural language inference (NLI) model (Chen et al. 2017) can mine the relationship between texts and the attributes by inferring the logical entailment, we introduce an NLI-based binary filter (f_{nli}) as a critic, and discard the pairs which do not achieve the entailment score over the threshold γ_1: $O_1(x, y) = \mathbb{1}\{f_{nli}(x \Rightarrow y) \geq \gamma_1\}$, where x denotes the input, and y means the output.

Format-based selection. When we rewrite the given sentence, we need to make the format of the given sentence various, and preserve its original meaning. Thus, we want to filter the origin-variance pair to learn the format-free dissimilarity. Especially, two metrics are used to evaluate the dissimilarity: 1) the token overlap between different sen-

tences and 2) their syntactic difference. For the first, we filter the pairs with a higher Rouge-L (Lin 2004) than a threshold γ_3. As for the syntactic difference, we first parse the constituency tree of the origin and variety, and then filter the pairs based on their tree edit distance: $O_2(x, y) = \mathbb{1}\{D_t(x, y) \geq \gamma_2 \ \wedge f_{rou}(x, y) \leq \gamma_3\}$, where $D_t(\cdot, \cdot)$ denotes the tree edit distance. The two dimensions of dissimilarity complement each other. On the one hand, $f_{rou}(\cdot, \cdot)$ promotes lexical divergence in each pair. On the other hand, $D_t(\cdot, \cdot)$ can be used to preempt "hacking" the word-overlap metric by simply switching a few words in the source sentence with corresponding synonyms.

Diversity-based selection. For sentence rewriting, we need a diverse range of generated sentences since the diversity of attributes can directly affect the robustness of the trained model. Therefore, we introduce a critic O_3 for the diversity. We define two pairs (x_1, y_1) and (x_2, y_2) to be duplicates when one pair entails another, either on the input side ($x_1 \Rightarrow x_2$) or on the output side ($y_1 \Rightarrow y_2$). In the diversity filter, we first cluster all entailing pairs, and then discard all but one with the largest entailment score. Thus, we can utilize the graph traversal for the diversity filter.

Based on the above critics, we can filter the attribute candidate pool $\mathcal{A}$ into an updated pool $\mathcal{U}$: $\mathcal{U} = \{(x, y)|(x, y) \in \mathcal{A}, O_1 \wedge O_2 \wedge O_3(x, y) = 1\}$.

Video-guided Self-weighted Cross-modal Bridging

In fact, different words (*e.g.*, noun, verb, and adjective) have distinct significance in text understanding. For instance, some adjectives are more important for video grounding in some cases, while some verbs are more significant for distinguishing different target moments. Previous VLM methods treat all sentence components equally, which might limit these methods to fully understand the entire sentence. For example, if there is no adjective in the anchor text, the negative text with adjectives cannot contribute to our model since the adjective is not discriminative for the text. Thus, we aim to analyze the relative significance of each word to adaptively integrate different words, where we adaptively predict the salience of sentence components for each anchor text. Without any supervision, we can obtain the significance score which means which word is more significant for text understanding. Therefore, we can find an optimal integration strategy of sentence components, which makes VLM selectively understand different sentence components for a query.

Cross-modal bridging. Based on these positive and hard negative samples, we can encourage the designed VLM models to distinguish the difference between different words in each sentence part. For supervising the VLM model to understand the text input, we introduce the following loss based on three types of text input:

$$\mathcal{L}_{cl}^{i} = -\log \frac{\beta \cdot \exp[1/\tau \cdot \cos(f^V, g_i^n)]}{(1-\beta) \cdot \exp[1/\tau \cdot \cos(f^V, g_i^{p_i})] + \beta \cdot \exp[1/\tau \cdot \cos(f^V, g_i^n)]},$$

where $\beta \in (0, 1)$ is a parameter; g_i denotes the i-th text; $g_i^{n_{i,j}}$ and $g_i^{p_i}$ denote the the negative text and the positive text, respectively; τ denotes the temperature parameter. By $\mathcal{L}_{cl}^{i}$, we can enhance the effectiveness of the designed model by these generated auxiliary texts.

Method	Without text augmentation			With text augmentation		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Text-to-video retrieval						
CLIP-ViT-B/32						
X-Pool	46.9	72.8	82.2	40.1	68.2	76.5
+Ours	**47.8**	**74.9**	**83.5**	**45.9**	**72.3**	**81.4**
CLIP-ViP	50.1	74.8	84.6	42.3	69.4	77.8
+Ours	**51.7**	**75.3**	**85.9**	**50.4**	**73.6**	**84.5**
T-MASS	50.2	75.3	85.1	42.1	68.9	79.2
+Ours	**52.3**	**77.9**	**87.6**	**51.4**	**70.2**	**81.3**
CLIP-ViT-B/16						
X-Pool	48.2	73.7	82.6	39.7	68.5	78.4
+Ours	**50.7**	**76.2**	**85.2**	**48.9**	**75.3**	**84.0**
CLIP-ViP	54.2	77.2	84.8	51.2	73.9	80.4
+Ours	**56.8**	**79.4**	**85.9**	**53.6**	**77.8**	**84.2**
T-MASS	52.7	77.1	85.6	49.2	70.5	83.9
+Ours	**54.9**	**82.6**	**86.8**	**53.4**	**81.0**	**86.2**
Video-to-text retrieval						
CLIP-ViT-B/32						
X-Pool	44.4	73.3	84.0	41.2	68.5	80.4
+Ours	**45.8**	**76.4**	**87.3**	**42.7**	**74.5**	**86.0**
UATVR	46.9	73.8	83.8	43.0	67.9	78.3
+Ours	**49.7**	**75.6**	**86.4**	**47.8**	**74.0**	**83.9**
T-MASS	47.7	78.0	86.3	42.9	73.5	82.6
+Ours	**51.5**	**79.9**	**89.8**	**49.5**	**78.1**	**87.5**
CLIP-ViT-B/16						
X-Pool	46.4	73.9	84.1	42.8	72.0	81.6
+Ours	**50.2**	**77.4**	**86.3**	**48.7**	**76.0**	**84.2**
UATVR	48.1	76.3	85.4	41.6	73.0	81.9
+Ours	**50.9**	**77.4**	**90.5**	**48.9**	**76.3**	**87.9**
T-MASS	50.9	80.2	88.0	48.3	75.6	84.9
+Ours	**53.7**	**84.2**	**91.5**	**50.8**	**82.7**	**90.4**

Table 1: Video text retrieval comparisons on MSR-VTT.

Method(# Frames)	w/o text augmentation				w/ text augmentation			
	Int↑	Seq↑	Pre↑	Fea↑	Int↑	Seq↑	Pre↑	Fea↑
All-in-One (32)	47.5	50.8	47.7	44.0	42.9	48.5	44.0	40.2
+Ours (32)	**48.3**	**51.9**	**49.6**	**45.7**	**47.9**	**51.3**	**48.7**	**44.3**
InternVideo (8)	62.7	65.6	54.9	51.9	55.6	61.0	50.3	47.2
+Ours (8)	**63.8**	**67.7**	**58.9**	**55.2**	**61.8**	**64.9**	**57.4**	**54.3**
BLIP-2 (4)	65.4	69.0	59.7	54.2	60.9	66.3	54.3	50.1
+Ours (4)	**67.8**	**72.5**	**61.4**	**56.8**	**66.2**	**71.6**	**58.7**	**55.3**

Table 2: Comparison Results on STAR VideoQA.

Self-weighted supervision. Since the visual features have higher computational complexity, we generate the positive and negative texts only by the original text (*i.e.*, anchor text) without considering the video input. Since different words contribute variously to sentence understanding, we target to find the most discriminative word for better text understanding by the following loss:

$$\mathcal{L}_{CL}^{i} = \max(\mathcal{L}_{cl}^{i,1}, \mathcal{L}_{cl}^{i,2}, \ldots, \mathcal{L}_{cl}^{i,C}). \quad (1)$$

For C losses $(\mathcal{L}_{cl}^{i,1}, \ldots, \mathcal{L}_{cl}^{i,C})$, each $\mathcal{L}_{cl}^{i}$ corresponds to a specific negative text, where the corresponding sentence component is changed. In Eq. (1), the maximum of these decomposed losses corresponds to the sentence component that is most clearly identified. Considering the significance score, we can obtain the finally video-guided self-weighted cross-

Method(# Frames)	w/o text augmentation			w text augmentation			Method(# Frames)	w/o text augmentation			w text augmentation		
	Tem↑	Cau↑	Des↑	Tem↑	Cau↑	Des↑		Tem↑	Cau↑	Des↑	Tem↑	Cau↑	Des↑
All-in-One(32)	48.6	48.0	63.2	40.2	37.9	53.8	InternVideo(8)	58.5	62.5	75.8	52.9	57.4	70.3
+Ours(32)	**50.1**	**51.9**	**64.7**	**48.6**	**50.2**	**61.3**	**+Ours(8)**	**62.5**	**66.3**	**76.4**	**61.8**	**59.7**	**74.5**
Just Ask(20)	51.4	49.6	63.1	42.7	40.1	54.0	BLIP-2(4)	67.2	70.3	79.8	64.0	61.9	72.3
+Ours(20)	**54.3**	**52.9**	**67.8**	**50.9**	**49.3**	**62.7**	**+Ours(4)**	**70.1**	**72.9**	**80.4**	**69.2**	**70.1**	**78.4**
MIST(32)	56.6	54.6	66.9	51.9	48.2	55.3	SeViLA(4)	67.7	72.1	82.2	64.0	66.8	76.9
+Ours(32)	**60.3**	**56.9**	**69.8**	**57.2**	**55.4**	**67.9**	**+Ours(4)**	**72.4**	**74.9**	**85.3**	**70.5**	**72.7**	**83.9**
HiTeA(16)	58.3	62.4	75.6	52.2	57.6	59.3	VideoQA-TA(4)	58.9	63.2	76.0	52.1	57.9	58.5
+Ours(16)	**62.8**	**65.7**	**77.3**	**60.4**	**63.9**	**74.9**	**+Ours(4)**	63.4	66.1	77.9	61.8	64.7	76.5

Table 3: VideoQA performance comparison on NExT-QA, where the value means the accuracy of providing the right answer.

Method	Type	ActivityNet Captions								Charades-STA							
		w/o text augmentation				w/ text augmentation				w/o text augmentation				w/ text augmentation			
		R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
2D-TAN	FS	59.45	44.51	85.53	77.13	48.32	29.38	71.36	62.30	39.81	23.25	79.33	52.15	20.18	11.35	47.05	33.82
+Ours	**FS**	**60.46**	**45.29**	**87.94**	**77.43**	**51.86**	**32.64**	**72.98**	**63.75**	**40.27**	**24.95**	**82.96**	**53.28**	**23.99**	**14.75**	**49.22**	**34.18**
MMN	FS	65.05	48.59	87.25	79.50	55.30	31.76	74.88	71.52	47.31	27.28	83.74	58.41	25.33	18.80	45.97	35.08
+Ours	**FS**	**66.05**	**49.31**	**89.75**	**81.27**	**58.76**	**33.08**	**75.33**	**73.59**	**49.07**	**29.32**	**85.06**	**60.13**	**26.87**	**22.48**	**46.03**	**37.85**
G2L	FS	-	51.68	-	81.32	55.75	33.01	75.25	70.89	47.91	28.42	84.80	59.33	26.54	19.85	48.06	36.70
+Ours	**FS**	**66.34**	**54.26**	**91.77**	**84.29**	**60.90**	**46.86**	**84.39**	**80.62**	**55.77**	**32.97**	**91.38**	**60.39**	**34.85**	**27.96**	**74.28**	**46.70**
BM-DETR	FS	-	49.62	-	-	56.24	33.52	75.34	70.26	54.42	33.84	-	-	28.10	21.43	47.26	35.29
+Ours	**FS**	**66.85**	**54.35**	**92.13**	**85.13**	**61.98**	**47.93**	**85.27**	**81.38**	**56.20**	**33.51**	**92.15**	**60.92**	**35.75**	**28.74**	**75.13**	**47.82**
VCA	WS	50.45	31.00	71.79	53.83	31.74	25.37	46.98	42.76	38.13	19.57	78.75	37.75	17.87	12.39	45.70	22.13
+Ours	**WS**	**51.72**	**33.19**	**72.85**	**55.11**	**32.99**	**28.56**	**48.31**	**44.07**	**40.95**	**20.31**	**80.42**	**39.26**	**18.63**	**15.72**	**46.17**	**23.88**
WSTAN	WS	52.45	30.01	79.38	63.42	33.72	25.74	49.30	45.88	29.35	12.28	76.13	41.53	8.15	5.43	35.27	11.86
+Ours	**WS**	**53.10**	**31.56**	**80.24**	**65.77**	**35.20**	**27.99**	**51.84**	**48.69**	**30.24**	**14.06**	**77.35**	**42.99**	**10.77**	**6.92**	**37.40**	**13.88**
CNM	WS	55.68	33.33	-	-	35.72	28.95	50.06	48.72	35.15	14.95	-	-	14.34	9.65	43.88	18.79
+Ours	**WS**	**56.11**	**34.08**	**81.09**	**67.34**	**39.56**	**31.77**	**52.88**	**51.99**	**35.72**	**16.33**	**76.52**	**43.18**	**16.83**	**12.05**	**45.60**	**21.64**
MCMT	WS	58.82	33.97	-	-	36.21	29.52	50.78	48.21	36.23	15.29	-	-	15.75	9.84	44.21	19.67
+Ours	**WS**	**56.75**	**35.12**	**82.43**	**68.72**	**40.25**	**32.42**	**53.40**	**52.74**	**36.81**	**17.00**	**78.59**	**45.27**	**17.53**	**13.24**	**46.82**	**22.87**

Table 4: VSG performance comparison, where "FS" is "fully-supervised" and "WS" is "weakly-supervised".

modal bridging loss:

$$\mathcal{L}_{weighted} = \sum\nolimits_{i,j,m,c} S_1(q_i, Q_m, c) \cdot S_2(Q_m^i, Q_m^j, c) \cdot \mathcal{L}_{CL}^i. \quad (2)$$

Since our method is plug-and-play, we borrow the cross-modal fusion module from an open-source works into our framework, which is the base version of our method.

Experiments

Datasets. For a fair comparison, we utilize the following datasets: 1) For VSG, we utilize three datasets: ActivityNet Captions (Caba Heilbron et al. 2015), and Charades-STA (Sigurdsson et al. 2016) and TACoS (Regneri et al. 2013). 2) For VTR, we adopt two datasets: MSRVTT (Xu et al. 2016) and LSMDC. 3) For VideoQA, we use two datasets: NExT-QA (Xiao et al. 2021) and STAR (Wu et al. 2021).

Compared methods For better reproducibility, we compare some open-source state-of-the-art works from three multi-modal tasks. 1) VTR (text-to-video retrieval and video-to-text retrieval. 2) VideoQA. 3) VSG.

Evaluation metrics. For VTR, we utilize Recall at rank $\{1, 5, 10\}$ (R@1, R@5, and R@10), Median Rank (MdR), and Mean Rank (MnR) for evaluating the retrieval performance. For VSG, we evaluate the grounding performance by "R@n, IoU=m", which means the percentage of queries having at least one result whose Intersection over Union (IoU) with ground truth is larger than m. For VideoQA, we introduce seven metrics: temporal (Tem), causal (Cau), description (Des), interaction (Int), sequence (Seq), prediction (Pre) and feasibility (Fea). Bold denotes the best performance.

Performance Comparison

Following previous open-source methods, we directly cite the corresponding results from compared methods. In this paper, we treat our as the plug-and-play module for state-of-the-art VLM models to improve their performance.

Performance comparison on the VTR task. VTR is a challenging multi-modal task, which requires the designed model can effectively bridge the gap between videos and texts. In this paper, we consider two subtask: text-to-video retrieval and video-to-text retrieval. Table 1 illustrates the effectiveness of our model as the plug-and-play module for previous VTR methods. We can find that when using augmented text, all the compared methods suffer performance degradation. The core reason is that previous VTR methods pay less attention to the language input, and ignore much language information in the sentence query. By using our model as the plug-and-play module, previous method can

Model	ActivityNet Captions				Charades-STA			
	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Ours(a)	53.77	40.28	76.94	72.25	28.51	20.34	67.85	38.71
Ours(b)	55.35	42.03	79.50	74.91	30.88	23.92	70.66	41.58
Ours(c)	57.63	43.86	81.34	77.99	32.50	24.03	71.76	42.92
Ours(full)	**60.90**	**46.86**	**84.39**	**80.62**	**34.85**	**27.96**	**74.28**	**46.70**

Table 5: Main ablation study for VSG with G2L as the base model, where we remove each key individual component to show its effectiveness.

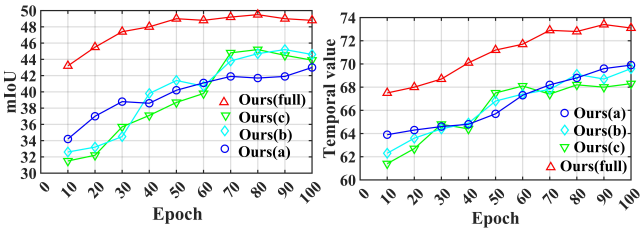

Figure 4: Training performance of each ablation module with text augmentation on the ActivityNet Captions dataset (left, VSG), the NexT-QA dataset (right, VideoQA).

obtain significant performance improvement.

Performance comparison on the VideoQA task. Similar to the VTR task, we conduct performance comparison VideoQA performance comparison. The experimental results are summarized in Tables 2 and 3, where the performance of previous methods was unsatisfactory.

Performance comparison on the VSG task. We conduct VSG performance comparison on all three datasets under both fully- and weakly-supervised settings. Tables 4 reports the quantitative comparison results. Obviously, our proposed model can help state-of-the-art VSG methods for performance improvement over all metrics on three datasets, which demonstrates the superiority of our proposed model.

Ablation Study and Analysis

Main ablation studies. To demonstrate the effectiveness of each component in our model, we conduct ablation studies regarding the components in Table 5. In particular, we remove each key individual module to investigate its contribution. For convenience, we design four ablation models: 1) Ours(a). We remove the "multi-level text augmentation" module while keeping the other modules. 2) Ours(b). We remove the "attribute-based text reasoning" module while keeping the other modules. 3) Ours(c). We remove the "video-guided self-weighted cross-modal bridging" module while keeping the other modules. Besides, we use our full model as the baseline: Ours(full). All four modules contribute a lot to the final performances on all three datasets, demonstrating their effectiveness in the VSG task.

Training process of different ablation models. Following (Lin et al. 2020), we analyze the training process and retrieval performance of different ablation models in Figure 4. We can obtain the following representative observations: (i) During training, Our(full) outperforms other ablation models, which further demonstrates the effectiveness of each module. (ii) Our(full) converges faster than ablation models, showing that our full model is more efficient. For instance, Our(full) converges within 70 epochs, while Our(c) converges after 80 epochs. Thus, our full model can process these challenging datasets more efficiently.

NExT-QA				STAR				
Method	Tem↑	Cau↑	Des↑	Method	Int↑	Seq↑	Pre↑	Fea↑
Back-translation	62.8	62.7	68.9	Contrastive loss	61.4	62.7	54.9	52.6
Paraphrasing	65.3	63.4	72.5	Ours(w/o S_1)	64.8	68.4	56.6	55.0
LLaMA-7B(*)	66.8	65.4	74.8	Ours(w/o S_2)	65.1	67.2	57.9	54.6
Ours	**69.2**	**70.1**	**78.4**	**Ours(full)**	**66.2**	**71.6**	**58.7**	**55.3**

Table 6: Ablation study on different augmentation methods on NExT-QA and $\mathcal{L}_{weighted}$ in Eq. (2) on STAR for VideoQA, where BLIP2 is the base model with # Frames=4. "LLaMA-7B(*)" means that we directly use LLaMA-7B for text augmentation.

Effect of different augmentation methods. In Table 6, we further compare our augmentation method in Section with other methods (Back-translation, Paraphrasing and LLaMA-7B). Obviously, our method achieves the best performance. It is because LLaMA-7B generates some hallucination information during augmentation. Back-translation and paraphrasing only treat each word with the same significance, which means that some unimportant words have a large weight. We can evaluate the significance score of each word to better understand the text. Also, we generate attributes to reason the fine-grained semantics in the given sentence and utilize attribute sampling to purify these semantics-rich attributes for better text understanding.

Influence of $\mathcal{L}_{weighted}$ in Eq. (2). To evaluate the effectiveness of our video-guided self-weighted cross-modal bridging loss, we conduct ablation study in Table 6. Obviously, directly using the contrastive loss will lead to unsatisfactory since it only construct positives and negatives from a level, we construct positives and negatives from different levels (word-level and structure-level). Comparing Ours(full) with Ours(w/o S_1) and Ours(w/o S_2), we can achieve significant performance improvement since our model can fully understand the text inputs based on S_1 and S_2.

Conclusion

In this paper, we rethink the LLM task from the user-friendly language input perspective. We observe that many VLMs cannot fully understand the language texts. Given some texts with similar semantics and a video, these VLMs output various results. To this end, we design a novel plug-and-play framework to improve the generation ability of previous methods on various text templates. Extensive experiments show that our framework can serve as the plug-and-play module for state-of-the-art VLM works to improve their performance on various video-language tasks. In our future work, we will extend our model into image-language model or video-audio model to achieve broader applicability.

References

Abdar, M.; Kollati, M.; Kuraparthi, S.; Pourpanah, F.; McDuff, D.; Ghavamzadeh, M.; Yan, S.; Mohamed, A.; Khosravi, A.; Cambria, E.; et al. 2024. A review of deep learning for video captioning. *IEEE TPAMI*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Carolan, K.; Fennelly, L.; and Smeaton, A. F. 2024. A Review of Multi-Modal Large Language and Vision Models. *arXiv preprint arXiv:2404.01322*.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for Natural Language Inference. In *ACL*, 1657–1668.

Fang, W.; Zhang, T.; and Chan, A. 2026. To Align or Not to Align: Strategic Multimodal Representation Alignment for Optimal Performance. *AAAI*.

Fang, X.; Easwaran, A.; and Genest, B. 2024. Uncertainty-Guided Appearance-Motion Association Network for Out-of-Distribution Action Detection. In *MIPR*.

Fang, X.; Easwaran, A.; and Genest, B. 2025. Adaptive Multi-prompt Contrastive Network for Few-shot Out-of-distribution Detection. In *ICML*.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2025a. Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *ESWA*.

Fang, X.; and Fang, W. 2026. Disentangling Adversarial Prompts: A Semantic-Graph Defense for Robust LLM Security. In *AAAI*.

Fang, X.; Fang, W.; Ji, W.; and Chua, T.-S. 2025b. Turing Patterns for Multimedia: Reaction-Diffusion Multi-Modal Fusion for Language-Guided Video Moment Retrieval. In *ACM MM*.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024a. Not all inputs are valid: Towards open-set video moment retrieval using language. In *ACM MM*, 28–37.

Fang, X.; Fang, W.; and Wang, C. 2025. Hierarchical Semantic-Augmented Navigation: Optimal Transport and Graph-Driven Reasoning for Vision-Language Navigation. In *NeurIPS*.

Fang, X.; Fang, W.; and Wang, C. 2026. Unveiling the Fragility of Vision-Language Models: Multi-Modal Adversarial Synergy via Texture-Constrained Perturbations and Cross-Modal Optimization. In *AAAI*.

Fang, X.; Fang, W.; Wang, C.; Liu, D.; Tang, K.; Dong, J.; Zhou, P.; and Li, B. 2025c. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. In *AAAI*, volume 39, 2915–2923.

Fang, X.; Fang, W.; Wang, C.; Liu, D.; Tang, K.; Dong, J.; Zhou, P.; and Li, B. 2025d. Multi-Pair Temporal Sentence Grounding via Multi-Thread Knowledge Transfer Network. In *AAAI*.

Fang, X.; Fang, W.; Wang, C.; Tang, K.; Liu, D.; Wang, S.; and Ji, W. 2026. Towards Unified Vision-Language Models With Incomplete Multi-Modal Inputs. In *AAAI*.

Fang, X.; and Hu, Y. 2020. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. 2021a. ANIMC: A Soft Approach for Autoweighted Noisy and Incomplete Multiview Clustering. *IEEE TAI*, 3(2): 192–206.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2020. V ́3 H: View variation and view heredity for incomplete multiview clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *TETCI*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024b. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *AAAI*, volume 38, 1735–1743.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-Modal Cross-Domain Alignment Network for Video Moment Retrieval. *IEEE TMM*, 1–16.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *CVPR*, 2448–2460.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023c. Hierarchical local-global transformer for temporal sentence grounding. *TMM*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2024c. Rethinking Weakly-supervised Video Temporal Grounding From a Game Perspective. In *ECCV*. Springer.

Gao, D.; Zhou, L.; Ji, L.; Zhu, L.; Yang, Y.; and Shou, M. Z. 2023. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. In *CVPR*.

Hakim, Z. I. A.; Sarker, N. H.; Singh, R. P.; Paul, B.; Dabouei, A.; and Xu, M. 2023. Leveraging Generative Language Models for Weakly Supervised Sentence Component Analysis in Video-Language Joint Learning. *arXiv*.

Jia, X.; Gao, S.; Qin, S.; Ma, K.; Li, X.; Huang, Y.; Dong, W.; Liu, Y.; and Cao, X. 2025a. Evolution-based region adversarial prompt learning for robustness enhancement in vision-language models. *arXiv preprint arXiv:2503.12874*.

Jia, X.; Gao, S.; Qin, S.; Pang, T.; Du, C.; Huang, Y.; Li, X.; Li, Y.; Li, B.; and Liu, Y. 2025b. Adversarial Attacks against Closed-Source MLLMs via Feature Optimal Alignment. *arXiv preprint arXiv:2505.21494*.

Jia, X.; Pang, T.; Du, C.; Huang, Y.; Gu, J.; Liu, Y.; Cao, X.; and Lin, M. 2024. Improved techniques for optimization-based jailbreaking on large language models. *arXiv*.

Jia, X.; Wei, X.; Cao, X.; and Han, X. 2020. Adv-watermark: A novel watermark perturbation for adversarial examples. In *ACM MM*.

Liang, K.; Liu, Y.; Zhou, S.; Tu, W.; Wen, Y.; Yang, X.; Dong, X.; and Liu, X. 2023a. Knowledge graph contrastive learning based on relation-symmetrical structure. *TKDE*.

Liang, K.; Meng, L.; Li, H.; Wang, J.; Lan, L.; Li, M.; Liu, X.; and Wang, H. 2025. From concrete to abstract: multi-view clustering on relational knowledge. *TPAMI*.

Liang, K.; Meng, L.; Liu, M.; Liu, Y.; Tu, W.; Wang, S.; Zhou, S.; and Liu, X. 2023b. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *ACM SIGIR*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, volume 34, 11539–11546.

Liu, D.; Fang, X.; Hu, W.; and Zhou, P. 2023. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *TMM*.

Liu, D.; Fang, X.; Qu, X.; Dong, J.; Yan, H.; Yang, Y.; Zhou, P.; and Cheng, Y. 2024. Unsupervised domain adaptative temporal sentence localization with mutual information maximization. In *AAAI*.

Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; and King, I. 2024. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv preprint arXiv:2405.14093*.

Qiu, X.; Hao, T.; Shi, S.; Tan, X.; and Xiong, Y.-J. 2024a. Chain-of-lora: Enhancing the instruction fine-tuning performance of low-rank adaptation on diverse instruction set. *IEEE Signal Processing Letters*.

Qiu, X.; Qian, J.; Wang, H.; Tan, X.; and Jin, Y. 2024b. An attentive copula-based spatio-temporal graph model for multivariate time-series forecasting. *Applied Soft Computing*.

Qiu, X.; Shao, S.; Wang, H.; and Tan, X. 2025a. Bio-K-Transformer: A pre-trained transformer-based sequence-to-sequence model for adverse drug reactions prediction. *CMPB*.

Qiu, X.; Shi, S.; Tan, X.; Qu, C.; Fang, Z.; Wang, H.; Gao, Y.; Wu, P.; and Li, H. 2023. Gram-based attentive neural ordinary differential equations network for video nystagmography classification. In *ICCV*.

Qiu, X.; Wang, H.; Tan, X.; and Jin, Y. 2025b. CVDLLM: Automated cardiovascular disease diagnosis with large-language-model-assisted graph attentive feature interaction. *TAI*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL*, 1: 25–36.

Rizve, M. N.; Fei, F.; Unnikrishnan, J.; Tran, S.; Yao, B. Z.; Zeng, B.; Shah, M.; and Chilimbi, T. 2024. VidLA: Video-Language Alignment at Scale. In *CVPR*, 14043–14055.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Wang, C.; Fang, X.; and Tiwari, P. 2025. DyPolySeg: Taylor Series-Inspired Dynamic Polynomial Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *ICML*.

Wang, C.; He, S.; Fang, X.; Han, J.; Liu, Z.; Ning, X.; Li, W.; and Tiwari, P. 2025a. Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding. In *CVPR*, 22182–22192.

Wang, C.; He, S.; Fang, X.; Hu, Z.; Huang, J.; Shen, Y.; and Tiwari, P. 2025b. Reasoning Beyond Points: A Visual Introspective Approach for Few-Shot 3D Segmentation. In *NeurIPS*.

Wang, C.; He, S.; Fang, X.; Wu, M.; Lam, S. K.; and Tiwari, P. 2025c. Taylor Series-Inspired Local Structure Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *AAAI*.

Wang, C.; Hu, Z.; Fang, X.; Yu, Z.; Wu, Y.; Xu, M.; Wang, Y.; Gao, X.; and Tiwari, P. 2026. Biologically-Inspired Evolutionary Domain Symbiosis for Few-shot and Zero-shot Point Cloud Semantic Segmentation. In *AAAI*.

Wang, J.; Sun, G.; Wang, P.; Liu, D.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 2024. Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. In *CVPR*.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2021. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 9777–9786.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. *arXiv preprint arXiv:2305.06988*.

Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2024. Self-chained image-language model for video localization and question answering. *NeurIPS*, 36.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2023. Temporal sentence grounding in videos: A survey and future directions. *IEEE TPAMI*, 45(8): 10443–10465.

Zhang, T.; Fang, W.; Woo, J.; Latawa, P.; Subramanian, D. A.; and Chan, A. 2025. Can LLMs Reason Over Non-Text Modalities in a Training-Free Manner? A Case Study with In-Context Representation Learning. *NeurIPS*.

Zhang, Y. 2018. A better autoencoder for image: Convolutional autoencoder. In *ICONIP17-DCEC*.

Zhang, Y.; Zhu, H.; Song, Z.; Koniusz, P.; and King, I. 2022. COSTA: covariance-preserving feature augmentation for graph contrastive learning. In *KDD*.

Zhu, C.; Jia, Q.; Chen, W.; Guo, Y.; and Liu, Y. 2023. Deep learning for video-text retrieval: a review. *IJMIR*, 12(1): 3.