Towards Unified Vision-Language Models with Incomplete Multi-Modal Inputs

Xiang Fang[1], Wanlong Fang[2], Changshuo Wang[3]*, Keke Tang[4], Daizong Liu[5], Siyi Wang[2], Wei Ji[6]

[1]School of Software Engineering, Huazhong University of Science and Technology
[2]Nanyang Technological University, Singapore
[3]University College London
[4]Guangzhou University
[5]Wuhan University
[6]Nanjing University
xfang9508@gmail.com, wanlongfang@gmail.com, wangchangshuo1@gmail.com, tangbohutbh@gmail.com, daizongliu@whu.edu.cn, siyi002@e.ntu.edu.sg, weiji0523@gmail.com

Abstract

Video-Language Models (VLMs) have demonstrated impressive multi-modal reasoning capabilities across diverse computer vision applications. However, these VLMs are task-specific and assume that both video and language inputs are complete. However, real-world VLM applications might face challenges due to deactivated sensors (e.g., cameras are unavailable due to data privacy), yielding modality-incomplete data and leading to inconsistency between training and testing data. While straightforward incomplete input can boast training generalization-ability and lead to training failure, its potential risks to VLMs regarding safety and trustworthiness have been largely neglected. To this end, we make the first attempt to propose a unified incomplete video-language model to process the incomplete multi-modal inputs. Extensive experimental results show that our method can serve as a plug-and-play module for previous works to improve their performance in various multi-modal tasks.

Introduction

Video-Language Models (VLMs) (Momeni et al. 2023; Fang et al. 2025a, 2023d, 2022) have achieved significant success and demonstrated promising capabilities in various multi-modal downstream applications, such as text-to-video retrieval (Ventura, Schmid, and Varol 2024; Fang et al. 2023c; Fang, Fang, and Wang 2025; Fang, Easwaran, and Genest 2025) and video question-answering (Xiao et al. 2024; Fang and Fang 2026; Fang, Fang, and Wang 2026; Fang et al. 2026, 2025c, 2024b, 2025d,b, 2024a,c, 2023b, 2021b; Fang, Easwaran, and Genest 2025; Fang et al. 2020, 2021a; Fang, Easwaran, and Genest 2024; Fang and Hu 2020). However, with the exponential expansion of downstream applications in the real world, VLMs might face network instability or data loss (Bordes et al. 2024; Wang et al. 2025a,b; Wang, Fang, and Tiwari 2025; Wang et al. 2026, 2025c; Li et al. 2025b,a), posing incomplete multi-modal inputs (Jang, Wang, and Kim 2024). In real-world applications, some frames are missing in the videos, while some words are unavailable in the texts/sentences. As shown in Figure 1, the "surveillance video analysis" application could contain missing frames and words. 1) Missing frames: In security monitoring, cameras may lose frames due to network lag or low bandwidth. 2) Missing words: Transcripts from operators may contain missing or unclear words due to noise or overlapping speech. Missing frames may cause crucial actions (*e.g.*, theft) to be lost, making retrieval harder. Incomplete texts make it challenging to match key moments accurately based on limited words, leading to a failed retrieval. Besides, different modalities have various incompleteness rates[1], which leads to unbalanced incompleteness.

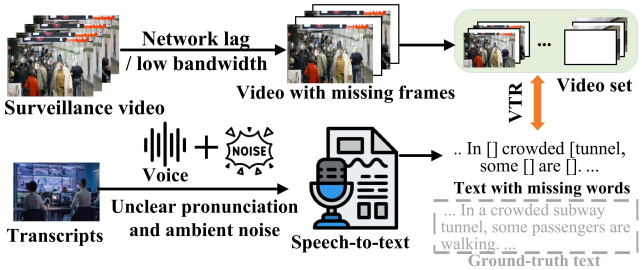


Figure 1: Real-world "surveillance video analysis" application with missing frames and words for video-text retrieval.

The downstream VLM-based methods (Fang, Zhang, and Chan 2026; Zhang et al. 2025) focus on the multi-grained information alignment between video and text. Recently, these VLM-based methods have achieved significant success by first projecting the video and text features into a common feature space and then introducing a loss for cross-modal alignment. Unfortunately, these VLM-based methods rely heavily on the complete video-text pairs during training and inference. In fact, during multi-modal data acquisition and processing, data missing and corruption will inevitably oc-

*Corresponding Author.

[1]Definition for incompleteness rate: incomplete(video) = #(missing frames)/#(total frames) and incomplete(text) = #(missing words)/#(total words). Balanced incompleteness: incomplete(video) = incomplete(text); unbalanced incompleteness: incomplete(video) $\neq$ incomplete(text).

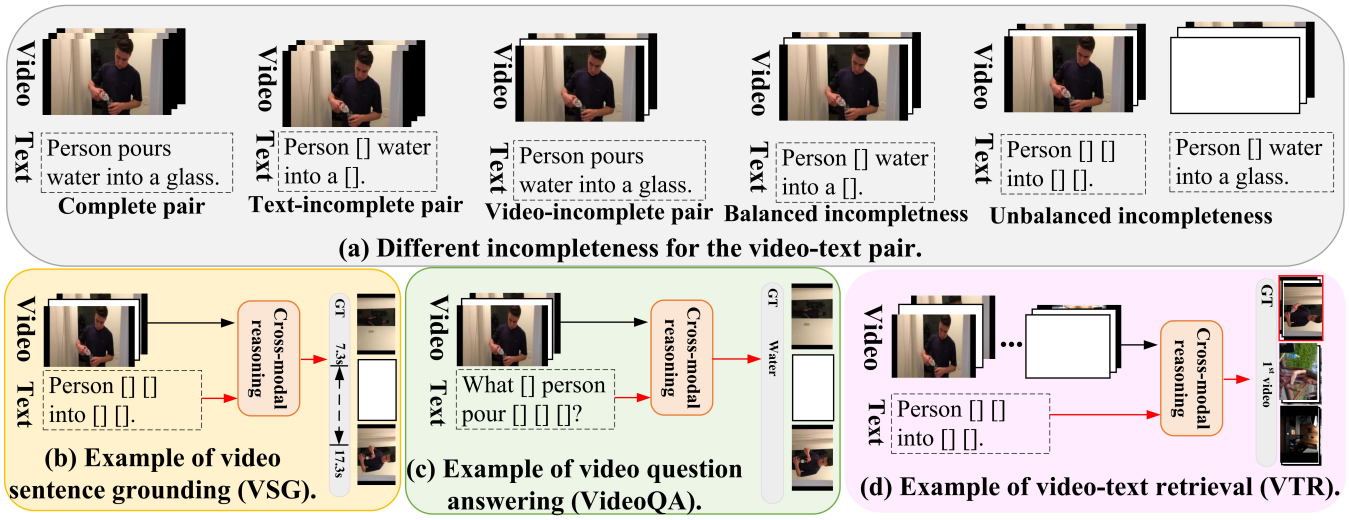


Figure 2: Incomplete multi-modal inputs for different multi-modal tasks. (a) Different incompleteness for the video-text pair. Incomplete pair for different tasks: VSG (b), VideoQA (c), VTR (d). "[]" means that the corresponding word is missing.

cur. Besides, most methods use uniform or fixed-rate frame sampling to understand motion information in videos. For incomplete video, they cannot reconstruct the correct motion due to missing frames, leading to performance degradation or even model failure. Consequently, previous methods cannot effectively understand incomplete videos/texts for cross-modal alignment.

In this paper, we pose a more practical setting called an incomplete video-language alignment, where only incomplete video-text pairs are available during training and inference. However, the realistic task faces the following essential challenges of incomplete multi-modal inputs among various downstream multi-modal tasks: 1) Existing VLM-based methods severely rely on the complete multi-modal inputs, which limits their applications with incomplete multi-modal inputs since most users do not upload all the information to the target multi-modal applications. In this case, these state-of-the-art methods will suffer severe performance degradation. 2) VLMs have great capabilities of handling multiple vision-language tasks with different prompts. However, existing VLM methods only deceive a specific task. When compromising different downstream tasks, we have to design a distinct multi-modal fusion method, which incurs significant time and resource expenditure. To make the attack more robust with high generalization-ability, we target to design a unified completion strategy for various incomplete multi-modal inputs across different downstream tasks.

To tackle the above issues and increase the robustness of incomplete video-language models for real-world applications. To this end, we pose a brand-new setting for VLMs, unbalanced incomplete VLM, where videos and texts are incomplete and have different incompleteness rates. In this work, we define a novel task termed unbalanced incomplete video-language model and construct many datasets to benchmark the challenging settings in various downstream multi-modal tasks (video-text retrieval, video question answering and video sentence grounding). To handle the challenging and realistic setting, we make the first attempt to explore a task-agnostic modality completion method for different video-language models. Especially, our proposed framework consists of three modules: multi-modal feature approximation, multi-modal knowledge distillation and multi-granularity multi-modal integration. Our main contributions are summarized as follows: 1) As far as we know, we make the first attempt to pose a brand-new and realistic setting, incomplete video-text alignment for unbalanced incomplete multi-modal inputs. We propose a unified completeness network to address the modality-incomplete challenges in various downstream multi-modal tasks. 2) We design a multi-modal feature approximation module, which can approximate more reliable completion features for the incomplete modalities. Also, we propose a multi-modal knowledge distillation module to reduce over-reliance on the complete modality. In the multi-granularity multi-modal integration module, we integrate semantics-similar video-text pairs by mapping them more compactly in the common feature space. 3) Extensive experimental results on several benchmarks with different incompleteness rates amply demonstrate that our proposed method can serve as a plug-and-play module for various state-of-the-art task-specific to improve their performance in various multi-modal tasks.

Related Works

Incomplete multi-modal inputs. Real-world multi-modal applications always suffer modality incompleteness since the sample collection in some modalities is very labor-intensive and time-consuming (Hu et al. 2024). Recently, some works focus on improving the model robustness on modality-incomplete data across various multi-modal tasks (Zhao, Liu, and Fu 2016). Some methods aim to optimize the multi-modal fusion strategy (Ma et al. 2022), while other methods try to conduct data augmentation (McKinzie

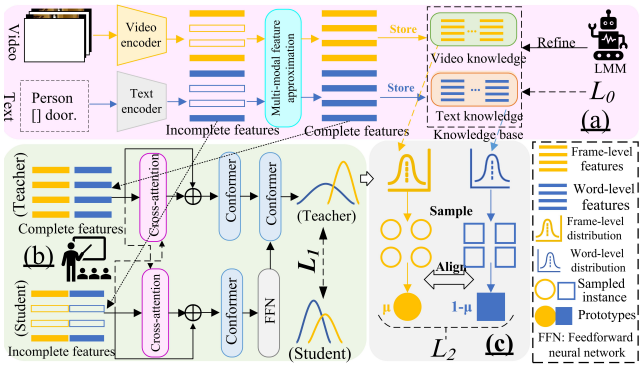


Figure 3: Overview of the proposed architecture for the incomplete video-text pair, where (a) is the "Multi-modal Feature Approximation" module, (b) is the "Multi-modal Knowledge Distillation" module, (c) is the "Multi-granularity Multi-modal Integration" module.

et al. 2023) or regularize objectives (McKinzie et al. 2023) to complete the missing samples. Unfortunately, these methods only perform well on simple classification tasks. When facing some complex multi-modal tasks (*e.g.*, VSG), they often achieve unsatisfactory performance.

Multi-modal learning. Multi-modal learning leverages multiple types of data (*e.g.*, text, visual) to create models that better understand complex, multi-faceted information (Huang et al. 2021). Multi-modal methods (Zhu et al. 2024) have gained traction as they are more aligned with real-world data, which is rarely unimodal. Due to remarkable success, multi-modal learning has achieved attracted more and more attention (Tian et al. 2024). State-of-the-art multi-modal methods (Sun et al. 2024) achieve performance improvement on some tasks under the strict assumption of complete modalities. In many real-world applications, only a subset of modalities are available during training and inference, limiting the performance of these methods.

Our Proposed Method

Problem statement. Given an incomplete multi-modal set $\{\mathbb{V}_n, \mathbb{Q}_n, \mathbb{Y}_n\}_{n=1}^N$, each untrimmed video $\mathbb{V}_n$ is represented as $\mathbb{V}_n = \{v_{n,t}\}_{t=1}^T$ frame-by-frame, where $v_{n,t}$ is the t-th frame of n-th video and T is the number of total frames. Similarly, the sentence text $\mathbb{Q}_n$ with M words is denoted as $\mathbb{Q}_n = \{q_{n,m}\}_{m=1}^M$ word-by-word. For missing frame or word, we treat it as null and denote as "[]". Previous VLM-based methods (Weng et al. 2025) fail to address the incompleteness challenge since they cannot deal with these incomplete inputs. When directly using these incomplete inputs for the downstream multi-modal tasks, their performance will drop significantly. To address the above challenges about incomplete multi-modal inputs, we propose a novel plug-and-play multi-modal feature approximation to approximate the missing frame/word features. These approximated features will provide significant semantics for multi-modal alignment. The proposed approach is application-agnostic and can be adopted successfully in the multi-modal task.

Pipeline. Our pipeline is summarized in Figure 3. Given an incomplete video-text pair, we first design a multi-modal feature approximation module to construct relational multi-modal graphs based on available cross-modal high semantic similarity features, which can approximate more reliable completion features for the missing modalities. Then, we propose a multi-modal knowledge distillation module to reduce over-reliance on the complete modality and to balance performance and robustness. Finally, we propose a prototype-based weighted multi-modal integration module to map semantically similar video-text pairs more compactly in the common embedding space.

Multi-modal Feature Approximation

Given incomplete video-text inputs, we first utilize the feature encoder networks to extract the the visual and textual features, where the video is encoded frame-by-frame and the text is encoded word-by-word. For the given incomplete video-text pair $\{V_n, T_n\}_{n=1}^N$, we utilize the feature encoder to obtain the initial features (v_i', t_i'), where v_i' and t_i' denote video and text features, respectively. To conduct the cross-modal integration and semantic alignment, we need to project multi-modal features into a joint feature space. Thus, we introduce the prototype learning strategy to conduct fine-grained multi-modal alignment by constructing the shared prototypes. For convenience, we denote the shared prototypes across videos and texts as $P \in \mathbb{R}^{N_p \times d}$, where N denotes the total number of video-text pairs, and N_p and d denote the prototype number and the feature dimension, respectively. Firstly, we randomly initialize these prototypes, and then update the prototypes during training. To align videos and texts for cross-modal fusion, we treat the shared prototype P as the query in the transformer's cross-attention operation, while the original video features v_i as the key W_K and value W_V. We utilize the video feature as an example, and the same is true for text features. Therefore, we can obtain the reconstructed features: $v_i = v_i' + FFN(v_i' + MCA(P, v_i'))$, $t_i = t_i' + FFN(t_i' + MCA(P, t_i'))$, where v_i and t_i denote the corresponding reconstructed features, $MCA(\cdot)$ denotes multi-head cross-attention, and $FFN(\cdot)$ denotes the feed-forward network.

Due to the continuity of the video frames, we can approximate missing frames using features of neighboring frames. Thus, we target to complete the fine-grained features (word and frame features). Especially, we first introduce a Jaccard distance function to compute the distance between two nearest neighbor samples. Then, we choose the most reliably K-reciprocal nearest neighbors from cross-modality and self-modality. For the missing frame feature v_a, we can obtain the semantics-relevant text feature t_a. Then, we compute the cosine similarity between t_a and all the frame features $\{v_i\}_{i=1}^T$, where T is the total frame number in the given video. After that, we rank and identify the K most similar frame features to the word feature t_a. We denote the K most similar frame features as $N_K(t_a) = \{v_1, ..., v_K\}$. Similarly, for any $v_i \in N_K(t_a)$, we compute the cosine similarity between v_i and all the word features. Thus, we can have the K most similar word features are denoted as $N_K(v_i) = \{t_1, ..., t_K\}$. For the word feature t_a, we can obtain the

cross-modality K-reciprocal nearest neighbors $\mathcal{O}_K(t_a)$ as follows: $\mathcal{O}_K(t_a) = \{v_i|(t_a \in N_K(v_i)) \cap (v_i \in N_K(t_a))\}$. Besides the cross-modal semantics between videos and texts, we can explore the semantic relationship within videos or texts. Therefore, for any two frames $v_i, v_j \in N_K(t_a)$, we compute intra-modal cosine similarity between v_i, v_j and all existing frame features to get the K-nearest neighbor sets $N_K(v_i)$ and $N_K(v_j)$ for v_i and v_j. In the single-modal setting, the K-reciprocal nearest neighbor is computed by: $\mathcal{O}_K(v_i) = \{v_j|(v_j \in N_K(v_i)) \cap (v_i \in N_K(v_j))\}$. Combining multi- and single-modal K-reciprocal nearest neighbors, we introduce the following Jaccard distance $J(t_a, v_i)$ to compute the distance between t_a and v_i: $J(v_i, t_a) = \frac{|\mathcal{O}_K(v_i) \cup \mathcal{O}_K(t_a)| - |\mathcal{O}_K(v_i) \cap \mathcal{O}_K(t_a)|}{|\mathcal{O}_K(v_i) \cup \mathcal{O}_K(t_a)|}$. Thus, to generate more accurate nearest neighbors, we try to search K_0-reciprocal nearest neighbors. Thus, the high semantic similarity neighbor generation set is $N_{K_0}(t_a) = \{v_1, v_2, ..., v_{K_0}\}$. The same applies to the missing text features as well.

Since the calculation of multi-modal fusion is feature-based, we complete the incomplete modality from the feature level. For missing features (frame feature v_a and word feature t_b), we can obtain K most relevant nearest neighbor sets for each modality: $N_K(t_a) = \{v_1, v_2, ..., v_{K_0}\}$ and $N_K(v_b) = \{t_1, t_2, ..., t_{K_0}\}$. Thus, we can obtain the approximated features (v_a of V_a and t_b of T_b): $\hat{v}_a = M_v \cdot [t_a, N_{K_0}(t_a)]$, $\hat{t}_b = M_t \cdot [v_b, N_{K_0}(v_b)]$, where M_v and M_t denote the affinity matrices of $[t_a, N_{K_0}(t_a)] = [t_a, v_1, v_2, ..., v_{K_0}] = [h_1, h_2, ..., h_{K_0+1}]$ and $[v_b, N_{K_0}(v_b)] = [v_b, t_1, t_2, ..., t_{K_0}]$. Each value denotes a semantic similarity score between two instances (frame or word). Then, we can construct the video memory $\hat{V}_a = \{\hat{v}_a^1, \hat{v}_a^2, ..., \hat{v}_a^{K_0}\}$ and the text memory $\hat{T}_b = \{\hat{t}_b^1, \hat{t}_b^2, ..., \hat{t}_b^{K_0}\}$. To ensure that our feature approximation strategy is sensible in the real world, we introduce a pre-trained large multi-modal model (LMM) to refine the approximated features: $[\hat{V}_a, \hat{T}_b] = LMM([\hat{V}_a, \hat{T}_b])$. Also, we have $M_v = M_L^{-1} \cdot M_1$, where M_L^{-1} denotes the normalized Laplacian matrix of M_1, and each element $M_{1_{ij}} \in M_1$ is obtained by $M_{1_{ij}} = \exp(\cos(h_i, h_j))$, where $\cos(\cdot, \cdot)$ denotes the cosine similarity function. Similarly, we can conduct a similar process to M_t.

Please note that the above method is essentially equivalent to constructing graph relationships, where we can transmit information across different samples within the graph and enhance the feature completion. In the graph, the affinity matrix M_v can be treated as the edges and the feature $[t_a, N_K(t_a)]$ serves as the nodes. Thus, we can formulate our multi-modal feature approximation module as follows:

$$\mathcal{L}_0 = \frac{1}{N_m^v} \sum\nolimits_{a=1}^{N_m^v} ||t_a - \hat{v}_a||_2^2 + \frac{1}{N_m^t} \sum\nolimits_{a=1}^{N_m^t} ||v_a - \hat{t}_a||_2^2, \quad (1)$$

where N_m^v and N_m^t denote the numbers of missing frames and missing words, respectively. By Eq. (1), we can mitigate the modal discrepancy between the approximated features and the original features.

Multi-modal Knowledge Distillation

Completing all the missing frames and words by the multi-modal feature approximation module is time-consuming, especially for long videos. To reduce over-reliance on the complete modality and to balance performance and robustness, we design a novel multi-modal knowledge distillation module to train an efficient student model that does not require the expensive feature approximation.

Different from previous knowledge distillation methods, the difference between teacher and student models in our method is the modal gap, not data size. Especially, we train the teacher model on the complete video-text pairs and the student model on incomplete pairs. Compared with the student model, the teacher model is relatively unbiased with a higher rate of modality-general decisive features f^c in the shared space. When we train the student model, we treat the teacher model as an anchor point, which can prevent the student model from shifting towards an unimodal distribution in the text modality. In our module, we conduct the knowledge at the hidden layer, not the logistic outputs, which can minimize the distances between the decision distribution samples of the teacher and student models. Besides, our knowledge distillation module constrains the intermediate representation subspace distribution of the student model. Therefore, we can take the knowledge from the intermediate representation of the cross-modal encoder layers for the downstream multi-modal task.

In our model, the samples from original feature space $\mathcal{S}^v \times \mathcal{S}^t \times \mathcal{Y}$ can be denoted as triples (v, t, y). For the teacher model, we first train the teacher model $\mathcal{T}(\theta)$ on a complete multi-modal data (v, t, y) model with parameters θ. Then, we can obtain the model's decisions ($P_1(y|t, v)$ and $P_2(f^c|t, v_{k_i})$) in a Bayesian decision problem. Since our model is expected as a unified network for multiple downstream tasks, we train the teacher model by minimizing the following loss function for multi-task learning:

$$\mathcal{T}(\theta) = \min_{\theta} \mathcal{L}_{\text{MLT}}(g(v, t; \theta), y), \quad (2)$$

$$\mathcal{L}_{\text{MLT}}(v, t; \theta) = \mu \log P_{\text{CTC}}(y|t, v) + (1-\mu) \log P_{\text{Att}}(y_i|t, v), \quad (3)$$

where $\mu \in (0, 1)$ is a parameter to balance different losses. During training the student model, we leverage the dropout strategy (McKinzie et al. 2023) on the video modality v, while we freeze the teacher model with complete video-text pairs as multi-modal inputs. Please note that the student and teacher models share a similar network architecture. Besides, we divide the whole decision process of the multi-modal model into a hidden feature generation step and a decision step for better interpretability. We have $P_2(y|t, v_{k_i}) = P_2(y|f^c)P_2(f^c|t, v_{k_i}), P_1(y|t, v) = P_1(y|f^c)P_1(f^c|t, v)$, where $f^c \in \mathbb{R}^d$ denotes the combined feature of modality-specific decisive features $f^t, f^v \in \mathbb{R}^d$, and modality-general decisive features $f^c \in \mathbb{R}^d$. The tuple (f^t, f^v, f^c) denotes a sample drawn from the shared features space, which denotes $\mathcal{S}^t \times \mathcal{S}^v \times \mathcal{S}^c$.

During training, besides initializing the parameter of the teacher model, we utilize an additional loss for constraining the dynamic process of the student model's feature distribution. To conduct the frame-level knowledge distillation, we approximate the difference of distribution by the distance between batch samples from the student and teacher models. The loss is as defined as $\mathcal{L}_{\text{KD}}(v, t, v_k) = \text{KL}(S_t, S_s)$, where

$S_t = \delta_\sigma(\mathcal{F}_s(P_1(f^c|t,v)))$ and $S_s = \delta_\sigma(\mathcal{F}_s(P_2(f^c|t,v_{k_i})))$, where $\mathcal{F}_s(\cdot)$ and $\delta_\sigma(\cdot)$ denote the sample function and the SoftMax function with temperature σ, respectively.

Three main purposes are considered in the distribution approximation: 1) by the dual cross-attention design, the process complements the information extracted from x^a, which can effectively address the condition of missing frames and promote out-of-distribution generality. 2) when the student network encounters a missing modality feature v_{k_i} during training, the convergence of the student's decisive feature $z^u = g(t, v_{k_i}; \theta_s)$ towards the teacher's decisive feature $z^u = g(t, v; \theta_t)$ encourages the utilization of contextual information from v_{k_i}. 3) we can utilize the KD loss to maximize the similarity between the distributions of the teacher and student models, which can prevent the student model from converging to trivial solutions. Finally, we train the student model jointly with a weighted sum of the standard training loss and distillation loss:

$$\mathcal{L}_1(v,t,x_k^v) = \beta\mathcal{L}_{\text{KD}}(v,t,x_k^v) + (1-\beta)\mathcal{L}_{\text{MLT}}(v,t_k). \quad (4)$$

Multi-granularity Multi-modal Integration

In real-world multi-modal applications, there are various data granularities in different tasks. For example, we need fine-grained multi-modal understanding for the video sentence grounding task since we need to localize the target segment based on the language sentence. Unlikely, we only need the coarse-grained multi-modal understanding for the video text retrieval since we can directly finish the retrieval based on the global video and text features. To handle the multi-granularity multi-modal inputs, we design a Multi-granularity Multi-modal Integration module. Especially, we introduce different weights based on the matching probability between different instances, and then adjust the frame-word alignment in the shared space. Also, an empirical observation is that noun phrases consistently share either the same or synonymous attributes within two textual descriptions from the same video. Given the sentence T_i, we extract relevant noun phrases as $P(T_i) = Z_i = \{z_1, z_2, ..., z_{N_p}\}$, where P and N_p denote the noun phrase extractor and the number of noun phrases, respectively. Besides, we can calculate matching probability weights between instance i and instance j as follows: $W_{i,j} = \frac{|Z_i \cap Z_j|/|Z_i \cup Z_j|}{\sum_{k=1}^{N} |Z_i \cap Z_k|/|Z_i \cup Z_k|}$, where $|Z_i \cap Z_j|$ denotes the count of synonymous noun phrases shared between Z_i and Z_j. Similarly, $|Z_i \cup Z_j|$ denotes the number of noun phrases in the union between Z_i and Z_j. By introducing the weight α to balance the cross-modal alignment of different samples, we can obtain:

$$\mathcal{L}_2^{v2t} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} L(v_i,t_j)\cdot(\alpha W_{i,j} + (1-\alpha)I_{i,j}), \quad (5)$$

$$\mathcal{L}_2^{t2v} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} L(t_i,v_j)\cdot(\alpha W_{i,j} + (1-\alpha)I_{i,j}), \quad (6)$$

where $L(v_i,t_j) = -log\frac{exp(\cos(v_i,t_j)/\sigma)}{\sum_{k=1}^{N} exp(\cos(v_i,t_k)/\sigma)}$ and $L(v_i,t_j) = -log\frac{exp(\cos(v_i,t_j)/\sigma)}{\sum_{k=1}^{N} exp(\cos(v_i,t_k)/\sigma)}$, and $\alpha \in [0,1]$ denotes the prior probability that frame v_i is matched with its paired text t_j. When $\alpha = 1$, we need to utilize the one-hot labels I_{ij} for cross-modal contrastive learning. However, to better align unpaired text feature t_j with frame feature v_i, $\alpha W_{i,j}$ supervises the unpaired samples, while $(1-\alpha)I_{i,j}$ provides supervision for paired frame-text samples. Finally, we can obtain the final loss:

$$\mathcal{L}_2 = \mu\mathcal{L}_2^{v2t} + (1-\mu)\mathcal{L}_2^{t2v}, \quad (7)$$

where μ is a parameter to balance the importance between different modalities.

Thus, our model is trained by the following loss:

$$\mathcal{L} = \mathcal{L}_0 + \alpha_1\mathcal{L}_1 + \alpha_2\mathcal{L}_2, \quad (8)$$

where α_1 and α_2 are parameters to balance the significance between different losses.

Method	Complete video-text pair			Incomplete video-text pair		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Text-to-video retrieval						
CLIP-ViT-B/32						
X-Pool (Gorti et al. 2022)	46.9	72.8	82.2	27.5	43.6	50.8
+Ours	**48.1**	**73.6**	**84.0**	**36.3**	**63.7**	**70.4**
DiffusionRet (Jin et al. 2023)	49.0	75.2	82.7	27.7	44.0	51.2
+Ours	**50.8**	**78.2**	**86.3**	**36.8**	**64.2**	**69.5**
CLIP-ViP (Xue et al. 2023)	50.1	74.8	84.6	31.4	44.7	52.0
+Ours	**52.3**	**77.1**	**86.0**	**41.2**	**69.5**	**73.1**
T-MASS (Wang et al. 2024)	50.2	75.3	85.1	30.8	45.1	52.4
+Ours	**51.9**	**76.8**	**86.3**	**42.6**	**70.3**	**72.5**
CLIP-ViT-B/16						
X-Pool (Gorti et al. 2022)	48.2	73.7	82.6	28.0	44.2	51.3
+Ours	**50.2**	**75.1**	**84.9**	**37.4**	**65.8**	**72.7**
CLIP-ViP (Xue et al. 2023)	54.2	77.2	84.8	28.2	44.7	51.5
+Ours	**55.9**	**79.8**	**86.3**	**38.2**	**66.3**	**73.4**
T-MASS (Wang et al. 2024)	52.7	77.1	85.6	28.5	45.3	52.4
+Ours	**53.8**	**80.5**	**86.9**	**43.6**	**71.3**	**74.9**
Video-to-text retrieval						
CLIP-ViT-B/32						
X-Pool (Gorti et al. 2022)	44.4	73.3	84.0	23.3	42.8	50.1
+Ours	**46.3**	**75.9**	**86.2**	**35.0**	**62.4**	**71.2**
UATVR (Fang et al. 2023a)	46.9	73.8	83.8	26.4	43.8	51.7
+Ours	**48.0**	**77.2**	**85.9**	**36.6**	**63.7**	**72.4**
T-MASS (Wang et al. 2024)	47.7	78.0	86.3	29.7	45.4	52.9
+Ours	**51.2**	**80.3**	**88.2**	**38.7**	**65.2**	**73.6**
CLIP-ViT-B/16						
X-Pool (Gorti et al. 2022)	46.4	73.9	84.1	26.3	43.5	51.7
+Ours	**48.9**	**76.2**	**87.5**	**36.8**	**64.7**	**75.2**
UATVR (Fang et al. 2023a)	48.1	76.3	85.4	26.2	44.3	52.0
+Ours	**49.2**	**78.0**	**88.9**	**37.4**	**66.9**	**74.0**
T-MASS (Wang et al. 2024)	50.9	80.2	88.0	28.5	44.7	52.9
+Ours	**51.8**	**83.4**	**92.3**	**41.0**	**69.5**	**75.8**

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

Experiment

Datasets. For a fair comparison, we use the following open-source video-language datasets to evaluate the effectiveness of our proposed framework in various tasks. 1) For the VTR task, we adopt two datasets: MSRVTT (Xu et al. 2016) and LSMDC (Liu et al. 2019). 2) For the VSG task, we utilize three datasets: ActivityNet Captions (Caba Heilbron et al. 2015), and Charades-STA (Sigurdsson et al. 2016) and TACoS (Regneri et al. 2013). 3) For the VideoQA task, we use two datasets: NExT-QA (Xiao et al. 2021) and STAR (Wu et al. 2021). Unless otherwise specified in this paper,

Method	# Frames	Complete pair			Incomplete pair		
		Tem	Cau	Des	Tem	Cau	Des
All-in-One (Wang et al. 2023)	32	48.6	48.0	63.2	29.8	31.3	41.7
+Ours	**32**	**51.2**	**52.9**	**65.0**	**38.9**	**40.3**	**52.6**
MIST (Gao et al. 2023)	32	56.6	54.6	66.9	31.4	32.9	43.7
+Ours	**32**	**58.7**	**57.2**	**68.9**	**40.3**	**43.0**	**54.7**
HiTeA (Ye et al. 2022)	16	58.3	62.4	75.6	32.0	33.4	43.1
+Ours	**16**	**61.3**	**66.0**	**78.2**	**41.8**	**44.5**	**55.3**
InternVideo (Wang et al. 2022)	8	58.5	62.5	75.8	33.1	34.2	45.9
+Ours	**8**	**63.0**	**64.8**	**78.9**	**42.5**	**46.3**	**57.0**
BLIP-2 (Li et al. 2023a)	4	67.2	70.3	79.8	36.8	39.7	49.6
+Ours	**4**	**71.2**	**73.5**	**82.3**	**49.0**	**53.1**	**62.4**

Table 2: VideoQA performance comparison on NExT-QA.

Method (# Frames)	Complete pair				Incomplete pair			
	Int	Seq	Pre	Fea	Int	Seq	Pre	Fea
All-in-One (Wang et al. 2023) (32)	47.5	50.8	47.7	44.0	25.3	30.2	28.4	23.7
+Ours (32)	**49.1**	**52.3**	**48.5**	**47.5**	**32.7**	**40.8**	**37.6**	**34.9**
MIST (Gao et al. 2023) (32)	55.5	54.2	54.2	44.4	30.4	35.7	31.0	32.3
+Ours (32)	**57.5**	**58.2**	**59.3**	**48.7**	**38.6**	**43.1**	**39.6**	**38.4**
InternVideo (Wang et al. 2022) (8)	62.7	65.6	54.9	51.9	35.2	37.9	33.0	35.8
+Ours (8)	**64.2**	**68.1**	**57.4**	**56.3**	**43.1**	**49.6**	**43.2**	**48.1**
SeViLA (Yu et al. 2023) (4)	63.7	70.4	63.1	62.4	35.9	38.1	32.4	34.9
+Ours (4)	**65.2**	**73.0**	**65.8**	**65.7**	**44.2**	**50.9**	**45.8**	**50.7**
BLIP-2 (Li et al. 2023a) (4)	65.4	69.0	59.7	54.2	36.7	39.2	41.4	37.5
+Ours (4)	**68.1**	**73.1**	**60.8**	**57.4**	**46.3**	**52.7**	**46.9**	**53.2**

Table 3: Comparison Results on STAR VideoQA dataset, where "Int" is "Interaction", "Seq" is "Sequence", "Pre" is "Prediction", and "Fea" is "Feasibility".

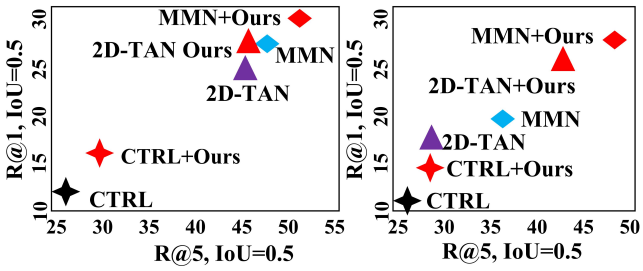


Figure 4: VSG performance on TACoS, where the left one is complete pair (incompleteness rate is 0%) and the right one is balanced incomplete pair (incompleteness rate is 50%).

we default that incomplete pairs refer to a 30% missing rate (*i.e.*, incomplete(video) = incomplete(text) = 30%).

Evaluation metrics. For the VTR task, we utilize Recall at rank $\{1, 5, 10\}$ (R@1, R@5, and R@10) for evaluating the retrieval performance. For the VSG task, 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. We use $n \in \{1, 5\}$ for all datasets, $m \in \{0.5, 0.7\}$ for ActivityNet Captions and Charades-STA, $m \in \{0.3, 0.5\}$ for TACoS. As for the VideoQA task, we introduce the following metrics: temporal (Tem), causal (Cau), description (Des), interaction (Int), sequence (Seq), prediction (Pre) and feasibility (Fea). Bold value denotes the best performance.

Method	Type	Complete video-text pair				Incomplete video-text pair			
		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
ActivityNet Captions									
MMN	FS	65.05	48.59	87.25	79.50	37.90	28.53	60.34	47.98
+Ours	**FS**	**67.32**	**50.28**	**90.34**	**80.75**	**48.93**	**43.62**	**75.57**	**68.49**
G2L	FS	-	51.68	-	81.32	38.12	31.60	61.49	49.88
+Ours	**FS**	**68.57**	**53.20**	**91.24**	**83.72**	**49.31**	**44.82**	**77.28**	**69.92**
VCA	WS	50.45	31.00	71.79	53.83	27.34	18.33	46.52	32.40
+Ours	**WS**	**52.83**	**34.76**	**73.94**	**56.11**	**35.82**	**25.47**	**54.38**	**40.96**
WSTAN	WS	52.45	30.01	79.38	63.42	28.11	19.04	46.70	33.97
+Ours	**WS**	**53.86**	**32.19**	**81.72**	**66.31**	**36.95**	**26.70**	**56.42**	**43.07**
CNM	WS	55.68	33.33	-	-	28.99	21.34	48.72	34.20
+Ours	**WS**	**57.35**	**35.04**	**82.96**	**68.43**	**38.52**	**28.99**	**58.43**	**46.00**
Charades-STA									
MMN	FS	47.31	27.28	83.74	58.41	21.03	14.20	55.42	22.87
+Ours	**FS**	**48.92**	**28.93**	**86.73**	**59.67**	**29.34**	**19.72**	**68.91**	**34.82**
G2L	FS	47.91	28.42	84.80	59.33	22.43	13.87	56.20	22.34
+Ours	**FS**	**50.82**	**31.27**	**86.95**	**61.83**	**30.84**	**20.39**	**71.15**	**36.82**
WSTAN	WS	29.35	12.28	76.13	41.53	13.83	7.90	42.09	13.08
+Ours	**WS**	**31.06**	**14.29**	**78.13**	**42.80**	**20.65**	**9.34**	**54.26**	**25.67**
CNM	WS	35.15	14.95	-	-	15.29	8.14	44.19	14.27
+Ours	**WS**	**36.87**	**16.95**	**78.66**	**40.15**	**22.96**	**11.43**	**54.88**	**27.39**
VCA	WS	38.13	19.57	78.75	37.75	16.24	9.25	48.10	16.73
+Ours	**WS**	**41.16**	**21.06**	**79.52**	**40.82**	**25.39**	**14.72**	**56.30**	**29.37**

Table 4: Performance comparison for VSG.

Method	Run-Time	Model Size	R@1, IoU=0.5
ACRN (Liu et al. 2018)	5.96s	128M	13.27
CTRL (Gao et al. 2017)	3.58s	**22M**	12.13
TGN (Chen et al. 2018)	0.89s	166M	15.82
2D-TAN (Zhang et al. 2020)	0.71s	232M	19.96
MomentDiff (Li et al. 2023b)	1.85s	248M	21.40
Ours+2D-TAN	**0.63s**	103M	**26.76**

Table 5: Efficiency comparison for VSG on TACoS.

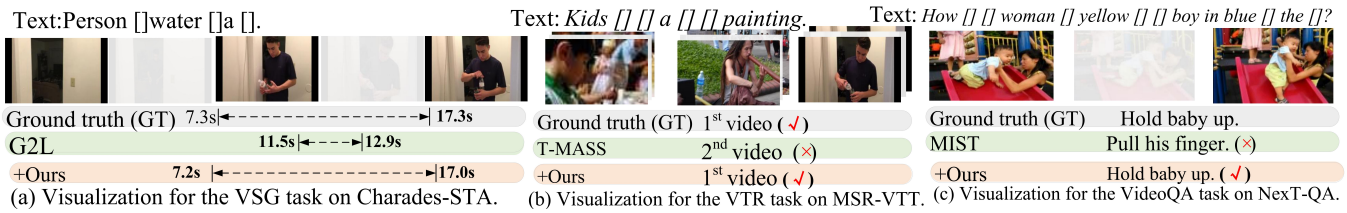
Model	ActivityNet Captions				Charades-STA			
	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Ours(a)	40.85	40.17	65.93	60.47	21.16	13.05	62.40	28.96
Ours(b)	44.99	42.83	68.19	65.30	23.76	16.37	67.12	32.59
Ours(c)	46.25	43.72	70.95	68.45	25.73	18.42	68.35	34.80
Ours(full)	**49.31**	**44.82**	**77.28**	**69.92**	**30.84**	**20.39**	**71.15**	**36.82**

Table 6: Main ablation study for the VSG task with G2L as the base model, where we remove each key individual component to investigate its effectiveness.

Performance Comparison

For a fair comparison, we follow previous open-source methods to directly cite the corresponding results from compared methods. Since our framework is a unified framework, we treat our framework as the plug-and-play module for state-of-the-art models to evaluate its effectiveness.

Performance comparison on the VTR task. In this task, we consider two significant subtasks: 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. When inputting incomplete pairs, all



(a) Visualization for the VSG task on Charades-STA. (b) Visualization for the VTR task on MSR-VTT. (c) Visualization for the VideoQA task on NexT-QA.

Figure 5: Visualization results for different downstream tasks on incomplete multi-modal datasets.

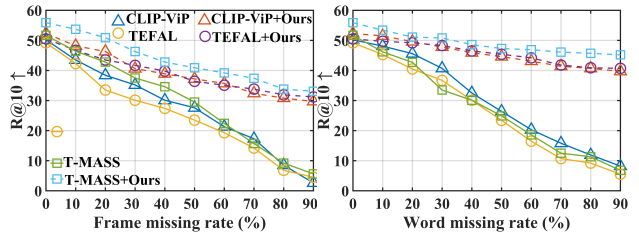


Figure 6: Different incompleteness rates for different modality for the text-to-video retrieval task on the LSMDC dataset (left: incomplete video and right: incomplete text).

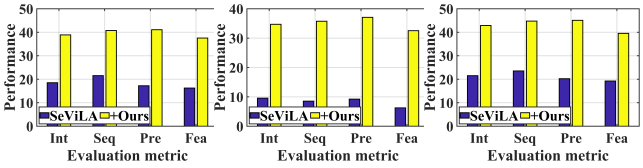


Figure 7: Balanced incomplete (left: incomplete(video) = incomplete(text) = 50%) vs unbalanced incomplete (middle: incomplete(video) = 70% and incomplete(text) = 30%; right: incomplete(video) = 30% and incomplete(text) = 70%) for the ideoQA task on the STAR dataset.

the compared methods suffer significant performance degradation. This is mainly because these VTR methods ignore missing frames and words, and directly concatenate reserved words and frames for cross-modal alignment, which results in these methods not being able to fully understand the whole video and text. With our proposed framework as the plug-and-play module, these VTR methods can obtain significant performance improvement.

Performance comparison on the VideoQA task. Tables 2 and 3 report the experimental results for VideoQA, where the performance of previous methods can perform well on complete video-text pairs but suffer from severe performance degradation on incomplete video-text pairs.

Performance comparison on the VSG task. As for the VSG task, we adopt official train/test splits under both fully-supervised and weakly-supervised setting. Table 4, Table 5 and Figure 4 summarize the quantitative comparison results. We can find that our proposed framework can serve as the plug-and-play module to effectively improve the performance of state-of-the-art VSG methods over all the metrics. The impressive performance of our framework illustrates its superiority.

Efficiency comparison. To comprehensively evaluate our model, we compare the efficiency and effectiveness of our framework (with 2D-TAN as the base model) with state-of-the-art open-source methods. In Table 5, our model achieves much faster processing speeds with relatively fewer learnable parameters than most of these state-of-the-art methods.

Visualization. Figure 5 depicts the visualizations of three challenging tasks on various incomplete video-text datasets. It illustrates that state-of-the-art methods obtain poor performance on incomplete multi-modal datasets.

Ablation Study and Analysis

Main ablation studies. To evaluate the effectiveness of each module in our framework, we conduct ablation studies regarding the modules (*i.e.*, Multi-modal Feature Approximation, Multi-modal Knowledge Distillation, Multi-granularity Multi-modal Integration) in Table 6. In particular, we remove each key individual module while keeping the other modules to investigate its contribution. For convenience, we design four ablation models: 1) Ours(a). We remove the "Multi-modal Feature Approximation" module. 2) Ours(b). We remove the "Generating Positive and Negative texts" module. 3) Ours(c). We remove the "Multi-granularity Multi-modal Integration" module. Besides, we treat our full model as the baseline: Ours(full). In Table 6, all the modules contribute a lot to the final performances on two challenging datasets, showing their effectiveness for VSG.

Influence of the incompleteness rate. To evaluate the influence of different incompleteness rates on each modality, we conduct ablation study on the LSMDC dataset. Figure 6 illustrates the corresponding results. For both video and text, as the miss rate increases, the performance of all the base methods drops severely. Fortunately, our framework can serve as the plug-and-play module for these methods to maintain the satisfactory performance.

Balanced incompleteness vs unbalanced incompleteness. In Figure 7, we further evaluate the impact of the unbalanced incompleteness. Obviously, our framework can work better with unbalanced incompleteness.

Conclusion

In this paper, we target a new task: incomplete video-text alignment. A unified completeness network is proposed to address the modality-incomplete challenges in various downstream multi-modal tasks. Extensive experiments show the effectiveness of our proposed method in various tasks.

References

Bordes, F.; Pang, R. Y.; Ajay, A.; Li, A. C.; Bardes, A.; Petryk, S.; Mañas, O.; Lin, Z.; Mahmoud, A.; Jayaraman, B.; et al. 2024. An introduction to vision-language modeling. *arXiv*.

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

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *EMNLP*.

Fang, B.; Liu, C.; Zhou, Y.; Yang, M.; Song, Y.; Li, F.; Wang, W.; Ji, X.; Ouyang, W.; et al. 2023a. Uatvr: Uncertainty-adaptive text-video retrieval. In *ICCV*.

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*.

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*.

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 Transactions on Artificial Intelligence*, 3(2): 192–206.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2020. V3H: 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. 2023b. 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*.

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

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

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023d. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14773–14783.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5267–5275.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*.

Hu, L.; Shi, T.; Feng, W.; Shang, F.; and Wan, L. 2024. Deep Correlated Prompting for Visual Recognition with Missing Modalities. *arXiv preprint arXiv:2410.06558*.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Jang, J.; Wang, Y.; and Kim, C. 2024. Towards Robust Multimodal Prompting with Missing Modalities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8070–8074. IEEE.

Jin, P.; Li, H.; Cheng, Z.; Li, K.; Ji, X.; Liu, C.; Yuan, L.; and Chen, J. 2023. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. MomentDiff: Generative Video Moment Retrieval from Random to Real. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Li, Q.; Li, X.; Chang, Z.; Zhang, Y.; Ji, C.; and Wang, S. 2025a. Multimodal Knowledge Retrieval-Augmented Iterative Alignment for Satellite Commonsense Conversation. In *IJCAI*.

Li, Q.; Liang, S.; Zhang, Y.; Ji, C.; Chang, Z.; and Wang, S. 2025b. Meta-Knowledge Path Augmentation for Multi-Hop Reasoning on Satellite Commonsense Multi-Modal Knowledge Graphs. In *ACM MM*.

Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018. Attentive moment retrieval in videos. In *Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 15–24.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are Multimodal Transformers Robust to Missing Modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

McKinzie, B.; Shankar, V.; Cheng, J. Y.; Yang, Y.; Shlens, J.; and Toshev, A. T. 2023. Robustness in multimodal learning under train-test modality mismatch. In *International Conference on Machine Learning*, 24291–24303. PMLR.

Momeni, L.; Caron, M.; Nagrani, A.; Zisserman, A.; and Schmid, C. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15579–15591.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

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*.

Sun, C.; Wu, X.; Yang, H.; Han, H.; and Zhao, D. 2024. Multi-Modal Learning-Based Interval Type-2 Fuzzy Neural Network. *IEEE Transactions on Fuzzy Systems*.

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2024. ArGue: Attribute-Guided Prompt Tuning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28578–28587.

Ventura, L.; Schmid, C.; and Varol, G. 2024. Learning text-to-video retrieval from image captioning. *International Journal of Computer Vision*, 1–21.

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

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, 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 *Advances in Neural Information Processing Systems*.

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.; Ge, Y.; Yan, R.; Ge, Y.; Lin, K. Q.; Tsutsui, S.; Lin, X.; Cai, G.; Wu, J.; Shan, Y.; et al. 2023. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6598–6608.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16551–16560.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Weng, Y.; Han, M.; He, H.; Chang, X.; and Zhuang, B. 2025. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, 453–470. Springer.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xiao, J.; Yao, A.; Li, Y.; and Chua, T.-S. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13204–13214.

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

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2023. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In *ICLR*.

Ye, Q.; Xu, G.; Yan, M.; Xu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2022. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546*.

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*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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*.

Zhao, H.; Liu, H.; and Fu, Y. 2016. Incomplete multi-modal visual data grouping. In *IJCAI*, 2392–2398.

Zhu, X.-F.; Xu, T.; Liu, Z.; Tang, Z.; Wu, X.-J.; and Kittler, J. 2024. UniMod1K: Towards a More Universal Large-Scale Dataset and Benchmark for Multi-modal Learning. *International Journal of Computer Vision*, 1–16.