Enhancing Partially Relevant Video Retrieval with Robust Alignment Learning

Long Zhang[1], Peipei Song[1]*, Jianfeng Dong[3], Kun Li[4] and Xun Yang[1,2]*
[1]University of Science and Technology of China
[2]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[3]Zhejiang Gongshang University [4]ReLER, CCAI, Zhejiang University
dragonzhang@mail.ustc.edu.cn, beta.songpp@gmail.com
dongjf24@gmail.com, kunli.hfut@gmail.com, xyang21@ustc.edu.cn

Abstract

Partially Relevant Video Retrieval (PRVR) aims to retrieve untrimmed videos partially relevant to a given query. The core challenge lies in learning robust query-video alignment against spurious semantic correlations arising from inherent data uncertainty: 1) query ambiguity, where the query incompletely characterizes the target video and often contains uninformative tokens, and 2) partial video relevance, where abundant query-irrelevant segments introduce contextual noise in cross-modal alignment. Existing methods often focus on enhancing multi-scale clip representations and retrieving the most relevant clip. However, the inherent data uncertainty in PRVR renders them vulnerable to distractor videos with spurious similarities, leading to suboptimal performance. To fill this research gap, we propose Robust Alignment Learning (RAL) framework, which explicitly models the uncertainty in data. Key innovations include: 1) we pioneer probabilistic modeling for PRVR by encoding videos and queries as multivariate Gaussian distributions. This not only quantifies data uncertainty but also enables proxy-level matching to capture the variability in cross-modal correspondences; 2) we consider the heterogeneous informativeness of query words and introduce learnable confidence gates to dynamically weight similarity. As a plug-and-play solution, RAL can be seamlessly integrated into the existing architectures. Extensive experiments across diverse retrieval backbones demonstrate its effectiveness.

1 Introduction

Text-to-Video Retrieval (T2VR) has been a long-standing challenge in vision and language research, allowing humans to associate textual concepts with video entities (Wang et al., 2025; Jin et al., 2023; Bogolin et al., 2022; Yang et al., 2022, 2024b). However, the mainstream T2VR methods (Li et al., 2024; Wu et al., 2023; Wang et al., 2023) assume

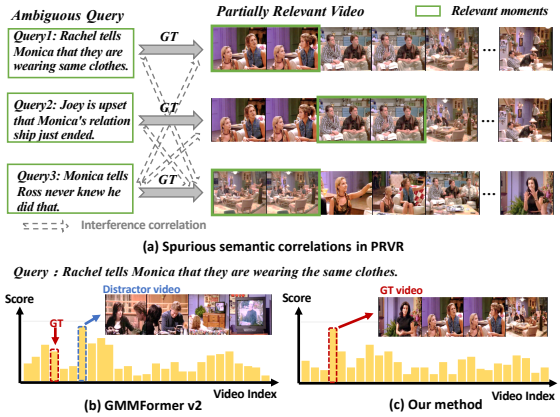

Figure 1: (a) Toy examples of spurious semantic correlations. (b–c) Retrieval scores of our method *vs*. GMMFormer v2. GMMFormer v2 fails to handle uncertainty, assigning the highest score to a distractor video.

that videos are pre-trimmed and text queries fully correspond to the videos (Dong et al., 2023). In real-world scenarios, videos are often untrimmed and the given queries can be incomplete and ambiguous, describing only a portion of the target video. This realistic demand leads to the emergence of Partially Relevant Video Retrieval (PRVR) task (Wang et al., 2024e; Dong et al., 2022), which aims to find untrimmed videos that are only partially relevant to a given text query.

PRVR presents a fundamental challenge of spurious semantic correlations due to the query ambiguity and partial video relevance. As illustrated in Figure 1 (a), such spurious correlations manifest in two aspects: the query " Monica tells Ross never knew he did that" relates to multiple video segments featuring similar actions in different contexts (*query ambiguity*), while the target video contains diverse content described by multiple sentences (*video partial relevance*). These factors make it difficult to establish a robust query-video alignment. Existing PRVR methods primarily attempt to mitigate query ambiguity by learning multi-scale clip representations, thereby maximizing the query-clip similarity within positive query-video pairs (Dong

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4615–4629
November 4-9, 2025

et al., 2023; Wang et al., 2024e). However, they implicitly assume a deterministic query-clip mapping and overlook the inherent data uncertainty in PRVR, thereby reducing inherently complex semantic mappings to deterministic pointwise alignments. Besides, without moment-level annotations, these methods struggle to learn optimal clip representations, leading to performance bottlenecks. Furthermore, they may be influenced by distractor videos with similar segments and provide incorrect retrieval results, as shown in Figure 1 (b).

To address the above issues, we propose Robust Alignment Learning (RAL), which explicitly models and utilizes uncertainty in data to enhance retrieval robustness. Our RAL builds upon the insight that PRVR should not be treated as pointwise query-clip feature alignment but rather as a probabilistic alignment problem that accounts for uncertainty. Inspired by probabilistic distributional representations (Jin et al., 2022), we model both video and query embeddings as Gaussian distributions, where the variance quantifies the inherent aleatoric uncertainty in each instance. Based on the distributional representations, we naturally construct Gaussian-based text and video proxies, which serve as multiple potential alignment candidates, enabling the model to capture diverse cross-modal relationships. Furthermore, most retrieval methods compute similarity scores by applying mean-pooling over words in the word-frame similarity matrix (Zhang et al., 2023, 2025b). We find this approach exacerbates retrieval bias as not all words contribute equally to retrieval, meaningless words (*e.g.*, "a") can distort the similarity estimation. To address this, we introduce confidence-aware alignment that dynamically assigns confidence weights to query words.

As shown in Figure 2, our RAL consists of two key components: (1) **Multimodal Semantic Robust Alignment (MSRA)** quantifies the semantic distribution in each modality by representing samples as multivariate Gaussian distributions. Given video and query embeddings, we first employ multi-granularity aggregation to obtain holistic semantics with sufficient contexts before estimating Gaussian parameters. Considering the incompleteness of the query relative to the video, we construct text distribution from a query support set that combines all video-related queries. Then, we conduct cross-modal learning with these distributional representations to joint video and text domains. To be specific, MSRA is optimized with two losses: a distribution alignment loss $\mathcal{L}_{\mathrm{DA}}$ enforcing probabilistic alignment between video and text distributions for robust cross-modal consistency, and a proxy matching loss $\mathcal{L}_{\mathrm{PM}}$ leveraging multiple alignment candidates to capture diverse semantic relationships. (2) **Confidence-aware Set-to-Set Alignment (CSA)** is to enhance query-video matching by dynamically adjusting the contribution of each query word. Instead of treating all words equally, CSA predicts a confidence score for each word and uses it to weight the word-frame similarity matrix. This effectively mitigates the influence of meaningless words and improves video retrieval.

Our contribution can be summarized as follows:

- We propose a novel robust alignment learning method for PRVR. It explicitly models and utilizes the data uncertainty and considers multiple potential matching relationships to enhance retrieval robustness.
- We propose a confidence-aware dynamic weighting mechanism for query words, which effectively mitigates the matching noise brought by meaningless words, improving retrieval precision.
- Extensive experiments on benchmark datasets (*i.e.*, TVR (Lei et al., 2020) and ActivityNet (Krishna et al., 2017)) demonstrate that our RAL significantly improves existing methods, achieving state-of-the-art results on PRVR.

2 Related Work

Partially Relevant Video Retrieval PRVR aims to retrieve untrimmed videos partially relevant to a given query. Compared to traditional T2VR, this task is more aligned with real-world application scenarios. Existing research (Dong et al., 2022; Wang et al., 2024e,d; Jiang et al., 2023; Nishimura et al., 2023; Dong et al., 2023; Song et al., 2025; Cho et al., 2025; Zhang et al., 2025b) primarily tackled PRVR by constructing multi-scale clip representations. Specifically, MS-SL (Dong et al., 2022) applies sliding windows to form clip representations and performs similarity calculations at both clip and frame levels. GMMFormer (Wang et al., 2024e) uses multiple Gaussian windows to constrain inter-frame interactions, thereby implicitly generating multi-scale clip features. Its improved version, GMMFormer v2 (Wang et al., 2024d), introduces a learnable feature fusion mechanism to aggregate multi-scale clips. Despite promising advancements, these methods suffer

from performance bottlenecks due to ignoring the spurious semantic correlations caused by data uncertainty and simplifying the complex semantic alignment, which motivates our robust alignment learning method.

Uncertainty in Multimodal Learning Uncertainty modeling has been widely explored in multimodal learning (Gao et al., 2024). HIB (Oh et al., 2019) first introduces probabilistic embeddings to capture the uncertainty in image representations. Similar ideas have been applied to tasks such as sentiment analysis (Gao et al., 2024) and instance segmentation (Zhang and Wonka, 2021). In the field of cross-modal retrieval, PCME (Chun et al., 2021) pioneers the use of probabilistic embeddings to capture the uncertainty of visual concepts. UATVR (Fang et al., 2023) further combines deterministic and probabilistic embeddings to explore optimal matching granularity in T2VR. T-MASS (Wang et al., 2024a) introduces a text-mass-based method, treating text embeddings as stochastic variables. *However*, these methods are typically designed for trimmed videos and exhibit limited effectiveness in PRVR. Inspired by this, we propose robust alignment learning specifically designed for PRVR.

3 Method

3.1 Preliminaries

In this paper, we tackle the task of PRVR. Given a text query q and a gallery of untrimmed videos $\mathcal{V}$, the goal of PRVR is to rank all videos $v \in \mathcal{V}$ so that the video partially corresponding to the text query q is ranked as high as possible. Existing methods primarily rely on multi-scale clip modeling to capture one-to-one correspondences between queries and untrimmed videos implicitly (Wang et al., 2024e; Dong et al., 2023). Here, we first review the common retrieval pipeline. For a query-video pair (q, v), unimodal encoders extract frame features $\mathbf{V} \in \mathbb{R}^{N_f \times d}$ and word features $\mathbf{Q} \in \mathbb{R}^{L \times d}$, where N_f and L denote the number of frames and words, respectively. Both features are projected into d-dimension feature space for cross-modal retrieval. Then, the clip modeling module (*e.g.*, sliding windows (Dong et al., 2022) and Gaussian windows (Wang et al., 2024e)) is applied on $\mathbf{V}$ to form the clip embeddings $\{\mathbf{c}_1, ..., \mathbf{c}_{N_c}\}$. Meanwhile, attention pooling summarizes $\mathbf{Q}$ into a sentence embedding $\mathbf{q}$. The final retrieval score $S(q, v)$ is computed as the maximum cosine similarity between sentence and clip embeddings:

$$S(q, v) = \max(\cos(\mathbf{q}, \mathbf{c}_1), \ldots, \cos(\mathbf{q}, \mathbf{c}_{N_c})). \tag{1}$$

To enforce cross-modal alignment, existing methods typically optimize a combination of InfoNCE contrastive loss $\mathcal{L}_{nce}$ (Miech et al., 2020) and triplet ranking loss $\mathcal{L}_{trip}$ (Dong et al., 2022):

$$\mathcal{L}_{base} = \lambda_1 \mathcal{L}_{nce} + \lambda_2 \mathcal{L}_{trip}, \tag{2}$$

where λ_1 and λ_2 are hyperparameters to balance the losses. $\mathcal{L}_{base}$ encourages high query-clip similarity $S(q, v)$ within positive query-video pairs while pushing apart negatives.

Motivation. In other words, this common pipeline implicitly assumes a deterministic mapping between a query and a video clip. However, this assumption is problematic given the query ambiguity and partial video relevance, *i.e.*, uncertainty in data. To address limitations, we are devoted to explicitly modeling the data uncertainty and leveraging it to augment the query and video representations, thereby improving the robustness of retrieval.

3.2 Multimodal Semantic Robust Alignment

Considering the query ambiguity and partial video relevance, we first propose an MSRA module to quantify the aleatoric uncertainty within both modalities. By modeling this uncertainty, we can better capture the variability in cross-modal correspondences and leverage it to strengthen cross-modal learning, enabling more robust representations of text and video.

(1) Uncertainty Modeling. According to uncertainty estimation theories (Chun et al., 2021; Gao et al., 2024), the aleatoric uncertainty can be predicted with deep learning models as the Gaussian variance. Inspired by this, we model the uncertainty in PRVR by treating feature representations as Gaussian distributions. Given the preliminary embeddings $\mathbf{X}^m$ of input m ($m \in \{q, v\}$), we estimate the mean vector $\boldsymbol{\mu}^m \in \mathbb{R}^d$ and variance vector $\boldsymbol{\sigma}^{m2} \in \mathbb{R}^d$ through two fully connected layers:

$$\boldsymbol{\mu}^m = h_\mu^m(g^m(\mathbf{X}^m)), \ \boldsymbol{\sigma}^m = h_\sigma^m(g^m(\mathbf{X}^m)), \tag{3}$$

where $h_\mu^m(\cdot)$ and $h_\sigma^m(\cdot)$ are the mean and variance estimators for input m, and $g^m(\cdot)$ serves as the feature aggregator. Furtherly, we define the probabilistic representation $\mathbf{z}^m$ as a multivariate Gaussian distribution with d variable:

$$p(\mathbf{z}^m|\mathbf{X}^m) \sim \mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\sigma}^{m2}\mathbf{I}), \tag{4}$$

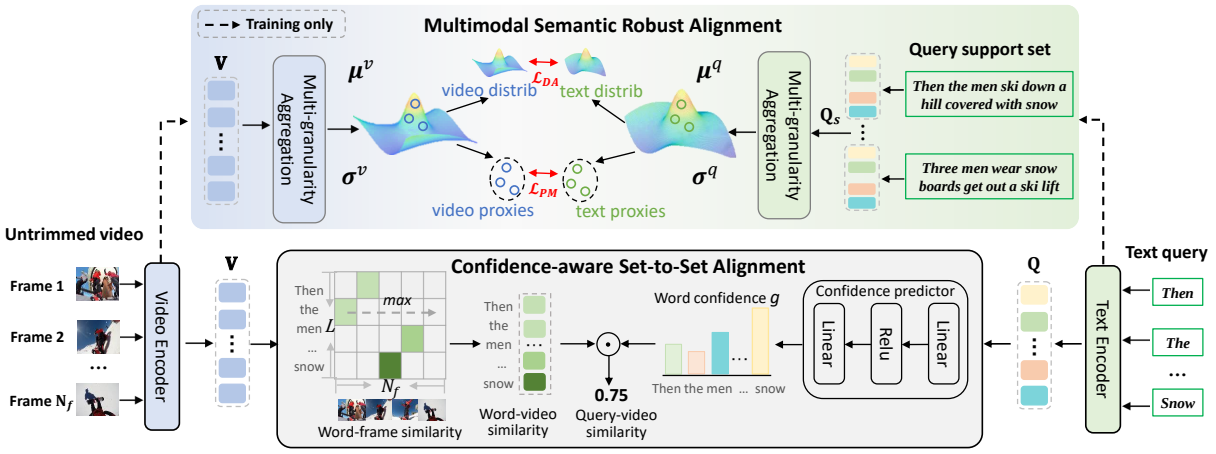

Figure 2: Overview of the proposed framework. It mainly consists of two components: (1) Multimodal Semantic Robust Alignment (MSRA) and (2) Confidence-aware Set-to-Set Alignment (CSA). Given an untrimmed video and a query, we first extract the frame features $\mathbf{V}$ and word features $\mathbf{Q}$ by video and text encoder, respectively. For MSRA, we collect a query support set containing all queries related to the video, obtaining its features $\mathbf{Q}_s$ with rich contexts. Then, we apply multi-granularity aggregation to obtain holistic semantics, and generate distributional representations parameterized by mean vector $\boldsymbol{\mu}$ and variance vector $\boldsymbol{\sigma}$. A proxy matching loss $\mathcal{L}_{\text{PM}}$ and a distribution alignment loss $\mathcal{L}_{\text{DA}}$ are used to unify the video and text domains. For CSA, we adopt a confidence predictor to assign confidence weights to each word, which is used to adjust the word-frame similarity matrix for video retrieval.

where $\mathbf{I}$ is the identity matrix. The uncertainty-aware representation $p(\mathbf{z}^m|\mathbf{X}^m)$ allows the model to capture variability in semantic alignment.

Query Support Set: For text modality, a single query provides an incomplete description of the video, limiting the reliability of its probabilistic representation. To this end, we replace the standalone query embedding $\mathbf{Q}$ with an enriched query support set embedding $\mathbf{Q}_s$, for better textual uncertainty modeling. Specifically, for each video v, we construct a query support set $\mathcal{D}^v$ by aggregating all associated queries q_n. The $\mathbf{Q}_s$ is obtained by concatenating the embeddings of all q_n in $\mathcal{D}^v$:

$$\mathbf{Q}_s = ||_{q_n \in \mathcal{D}^v}(\mathbf{Q}_n),\ \mathcal{D}^v = \{q_n | q_n \Leftrightarrow v\}, \quad (5)$$

where $\mathbf{Q}_n$ denotes text embedding of query q_n, $||$ denotes row-wise concatenation, and $\Leftrightarrow$ indicates labeled correspondence between query and video. Therefore, $\mathbf{X}^q = \mathbf{Q}_s$ and $\mathbf{X}^v = \mathbf{V}$ in Eq. (3).

Multi-granularity Aggregation: Estimating Gaussian distributions requires an effective aggregator $g^m(\cdot)$ to extract holistic features. To ensure representation fidelity, we introduce multi-granularity aggregation for the sequential $\mathbf{V}$ and $\mathbf{Q}_s$, which preserves local-global contextual cues before projecting them into a probabilistic space. Specifically, we apply mean pooling and linear mapping to obtain a global feature $\mathbf{x}^{m,g}$ and gated attention (Lin et al., 2017; Vaswani, 2017) to extract fine-grained local semantics $\mathbf{x}^{m,l}$. Formally,

$$\begin{cases} \mathbf{x}^{m,g} = \text{FC}^m(\text{MeanPool}(\mathbf{X}^m)), \\ \mathbf{x}^{m,l} = \text{Softmax}(\mathbf{w}_2\text{Tanh}(\mathbf{W}_1\mathbf{X}^m)) \cdot \mathbf{X}^m, \end{cases} \quad (6)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{w}_2 \in \mathbb{R}^d$ are trainable parameters. Then, we integrate local and global information, obtaining the multi-granularity holistic representations of $\mathbf{X}^m$ as:

$$g^m(\mathbf{X}^m) = \text{LayerNorm}(\mathbf{x}^{m,g} + \mathbf{x}^{m,l}). \quad (7)$$

(2) Joint Video and Text Domain. After obtaining the probabilistic distributions for video and text, we use two complementary loss functions to enforce a structured joint embedding space.

Distribution Alignment Loss: To establish consistency between video and text distributions, we introduce a distribution alignment loss $\mathcal{L}_{\text{DA}}$, which minimizes the Kullback-Leibler (KL) divergence between their probabilistic representations. Additionally, an auxiliary KL regularization item is used to encourage both distributions to approach a standard normal prior $\mathcal{N}(0, I)$ (Wang et al., 2024b). $\mathcal{L}_{\text{DA}}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{DA}} =& \text{KL}\left(p(\mathbf{z}^q|\mathbf{x}^q)\|p(\mathbf{z}^v|\mathbf{x}^v)\right) \\ &+ \sum_{m\in\{q,v\}} \text{KL}\left(p(\mathbf{z}^m|\mathbf{x}^m)\|\mathcal{N}(0,\mathbf{I})\right). \end{aligned} \quad (8)$$

Proxy Matching Loss: In PRVR, multiple semantic relationships exist between queries and

untrimmed videos, making one-to-one matching insufficient. We therefore adopt a proxy matching loss $\mathcal{L}_{\text{PM}}$, which considers multiple candidate alignments to enhance robustness in representation learning. Using the reparameterization technique (Kingma, 2013), we generate K proxy embeddings from the learned distributions as:

$$\hat{\mathbf{z}}_k^m = \boldsymbol{\mu}^m + \boldsymbol{\sigma}^m \cdot \epsilon_k, \quad k = \{1, .., K\}, \tag{9}$$

where $\epsilon^k \sim \mathcal{N}(0, \mathbf{I})$ and $\hat{\mathbf{z}}_k^m$ is the k-th proxy embedding for input m. $\boldsymbol{\mu}^m$, $\boldsymbol{\sigma}^m$ are the mean and standard deviation calculated by Eq. (3). This allows the model to sample diverse but semantically related embeddings, promoting the robustness of semantic alignment.

For each text proxy $\hat{\mathbf{z}}_k^q$, the positive video set $\mathcal{P} = \{\hat{\mathbf{z}}_k^v\}_{k=1}^K$ consists of K video proxies from v, and the negative video set $\tilde{P} = \{\hat{\mathbf{z}}_k^{\tilde{v}}\}_{\tilde{v},k}, \tilde{v} \neq v$ includes proxies from other videos in the batch. We then employ a multi-instance InfoNCE loss (Miech et al., 2020; Fang et al., 2023) to maximize the similarity between positive pairs while pushing apart negatives:

$$\mathcal{L}_{\text{PM}} = -\frac{1}{|\mathcal{B}|} \sum_{(q,v)\in\mathcal{B}} \log \frac{\sum_{\hat{\mathbf{z}}_k^v \in \mathcal{P}} e^{\cos(\hat{\mathbf{z}}_k^q, \hat{\mathbf{z}}_k^v)/\tau}}{\sum_{\hat{\mathbf{z}}_k^v \in \{\mathcal{P} \cup \widetilde{\mathcal{P}}\}} e^{\cos(\hat{\mathbf{z}}_k^q, \hat{\mathbf{z}}_k^v)/\tau}}, \tag{10}$$

where τ is a temperature factor and $\mathcal{B}$ is mini-batch.

3.3 Confidence-aware Set-to-Set Alignment

With query and video representations, $\mathbf{V} = \{\mathbf{v}_j\}_{j=1}^{N_f}$ and $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{L}$, we can obtain similarity matrix $\mathbf{S} \in \mathbb{R}^{L \times N_f}$ via dot product, where each element represents similarity between the i-th query word and the j-th video frame. First, we capture the most relevant frame for each query word through max-pooling, and take the cosine similarity between them as the word-video similarity s_i:

$$s_i = \max(\cos(\mathbf{q}_i, \mathbf{v}_1), \ldots, \cos(\mathbf{q}_i, \mathbf{v}_{N_f})), \tag{11}$$

To further obtain query-video similarity scores, existing methods often apply mean-pooling over $\{s_i\}_{i=1}^L$. However, some words (*e.g.*, function words) can introduce noise to cross-modal alignment. To overcome these limitations, we propose to dynamically assign word-level confidence scores $\mathbf{g} = \{g_i\}_{i=1}^L \in \mathbb{R}^L$ through a learnable predictor. By using the predicted $\mathbf{g}$, we weight the similarities s_i to compute the final query-video similarity:

$$S(q, v) = \sum_{i=1}^{L} g_i s_i, \quad \mathbf{g} = \text{MLP}(\mathbf{Q}), \tag{12}$$

where MLP consists of two linear layers and an activation function. The $S(q, v)$ is directly supervised by the basic retrieval loss $\mathcal{L}_{base}$ (Dong et al., 2022; Wang et al., 2024e). Therefore, the full model including MSRA and CSA modules is jointly end-to-end optimized by the total loss:

$$\mathcal{L} = \mathcal{L}_{base} + \lambda_3 \mathcal{L}_{DA} + \lambda_4 \mathcal{L}_{PM}, \tag{13}$$

where λ_3 and λ_4 are hyperparameters to balance the losses.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We adopt two large-scale video datasets, *i.e.*, ActivityNet Captions (ActivityNet) (Krishna et al., 2017) and TV Show Retrieval (TVR) (Lei et al., 2020). Notably, timestamp annotations are unavailable for PRVR. **TVR** contains 21,793 videos collected from six television shows. Each video is associated with five natural language sentences describing different moments. The average video length is about 76 seconds. **ActivityNet** contains 20,000 YouTube videos, with an average duration of about 118 seconds. Each video has about 3.7 moments with corresponding sentence descriptions. We abide by the popular data partition used in (Dong et al., 2022).

Following (Wang et al., 2024e; Jiang et al., 2023), we use rank-based metrics to evaluate the model, namely R@M (M = 1, 5, 10, 100). R@M measures the proportion of queries that correctly retrieve the target videos in the top M results. We also report the sum of all R@M scores (SumR) for overall comparisons. All metrics are reported as percentages (%).

Implementation Details. Following existing methods (Dong et al., 2022), we use ResNet (He et al., 2016) and I3D (Carreira and Zisserman, 2017) for visual feature extraction and RoBERTa (Liu et al., 2019) for text feature extraction on ActivityNet and TVR. In $\mathcal{L}_{PM}$, we set the number of proxies to $K = 6$. The loss coefficients are set to λ_1=0.05, λ_2=1, λ_3=0.001, and λ_4=0.004. We use the Adam optimizer with a learning rate of $1e$-4, a batch size of 128, and train for 100 epochs. An

Table 1: Performance comparison. Models are sorted in ascending order in terms of SumR on TVR.

Model	Venue	TVR					ActivityNet				
		R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
T2VR models:											
DE++ (Dong et al., 2021)	TPAMI'21	8.8	21.9	30.2	67.4	128.3	5.3	18.4	29.2	68.0	121.0
CLIP4Clip (Luo et al., 2022)	ArXiv'21	9.9	24.3	34.3	72.5	141.0	5.9	19.3	30.4	71.6	127.3
Cap4Video (Wu et al., 2023)	CVPR'23	10.3	26.4	36.8	74.0	147.5	6.3	20.4	30.9	72.6	130.2
UMT-L (Li et al., 2023)	ICCV'23	13.7	32.3	43.7	83.7	173.4	6.9	22.6	35.1	76.2	140.8
InternVideo2 (Wang et al., 2024c)	ECCV'24	13.8	32.9	44.4	84.2	175.3	7.5	23.4	36.1	76.5	143.5
VCMR models w/o moment localization:											
XML (Lei et al., 2020)	ECCV'20	10.0	26.5	37.3	81.3	155.1	5.3	19.4	30.6	73.1	128.4
ReLoCLNet (Zhang et al., 2021)	SIGIR'21	10.7	28.1	38.1	80.3	157.1	5.7	18.9	30.0	72.0	126.6
QCLPL(Zhang et al., 2025a)	TCSVT'25	11.0	28.9	39.6	81.3	160.8	6.5	20.4	31.8	74.3	133.1
JSG (Chen et al., 2023)	ACM MM'23	11.3	29.1	39.6	80.9	161.0	6.7	22.5	34.8	76.2	140.3
PRVR models:											
MS-SL (Dong et al., 2022)	ACM MM'22	13.5	32.1	43.4	83.4	172.4	7.1	22.5	34.7	75.8	140.1
PEAN (Jiang et al., 2023)	ICME'23	13.5	32.8	44.1	83.9	174.2	7.4	23.0	35.5	75.9	141.8
GMMFormer (Wang et al., 2024e)	AAAI'24	13.9	33.3	44.5	84.9	176.6	8.3	24.9	36.7	76.1	146.0
BGM-Net (Yin et al., 2024)	TOMM'24	14.1	34.7	45.9	85.2	179.9	7.2	23.8	36.0	76.9	143.9
DL-DKD (Dong et al., 2023)	ICCV'23	14.4	34.9	45.8	84.9	179.9	8.0	25.0	37.4	77.1	147.6
ARTVL (Cho et al., 2025)	AAAI'25	15.6	36.3	47.7	86.3	185.9	8.3	24.6	37.4	78.0	148.3
GMMFormer v2 (Wang et al., 2024d)	ArXiv'24	16.2	37.6	48.8	86.4	189.1	8.9	27.1	40.2	78.7	154.9
MGAKD (Zhang et al., 2025b)	TOMM'25	16.0	37.8	49.2	87.5	190.5	7.9	25.7	38.3	77.8	149.6
MS-SL + RAL	-	14.5	34.3	45.8	84.5	179.1	7.4	23.4	35.4	76.7	143.0
GMMFormer + RAL	-	15.8	36.4	47.9	86.0	186.1	8.4	25.1	37.2	77.0	147.7
GMMFormer v2 + RAL	-	**18.2**	**40.4**	**52.1**	**88.0**	**198.8**	**8.9**	**27.7**	**40.4**	**79.1**	**156.1**

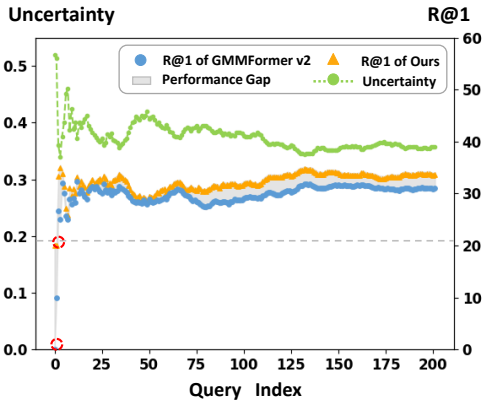

Figure 3: Performance comparison between our model and GMMFormer v2 under different levels of uncertainty in queries. Our model consistently outperforms GMMFormer v2, especially under extreme uncertainty.

early stopping strategy is applied, terminating training if SumR does not improve within 10 epochs. All experiments are conducted on a single A800 GPU. To gain insight into the effectiveness and generalization ability of our proposed approach, we integrate MSRA and CSA modules into three baselines: MS-SL (Dong et al., 2022), GMMFormer (Wang et al., 2024e), and GMMFormer-v2 (Wang et al., 2024d). More implementation details are provided in the supplementary.

4.2 Performance Comparison

Effectiveness on PRVR Task. Our method allows seamless integration into the various baseline models. As shown in Table 1, we apply it to three advanced PRVR models, MS-SL, GMMFormer, and GMMFormer v2. The experimental results reveal two key findings: (1) our method consistently enhances all baseline models and achieves substantial performance gains across two datasets; (2) our method sets a new state-of-the-art performance for PRVR, with a SumR of 198.8 on TVR, remarkably surpassing the previous best model (GMMFormer v2) by 9.7. These findings validate the effectiveness of our approach across the different architectures. In the following parts, we adopt GMMFormer v2 as our default benchmark model for further analysis and comparisons.

Model Robustness on Uncertain Samples. To verify the robustness and stability of our proposed method, we conduct more comparisons on queries with different uncertainty levels and observe the R@1 score. For clarity, we select a subset of TVR test set (*i.e.*, queries with M/V ratio $\in$[0.2,0.4] (Dong et al., 2022)) and group every 5 queries into a set. The uncertainty level of each query set is quantified using the geometric mean (Gao et al., 2024) of $\boldsymbol{\sigma}^q$ in Eq. (3). By observing the experimental results in Figure 3, we can find that: (1) our method consistently outperforms GMMFormer v2 across different uncertainty levels; (2) the performance gap between our method and GMMFormer v2 widens as uncertainty increases; (3) under extreme uncertainty, GMMFormer v2 collapses, with R@1 approaching zero, while our model remains

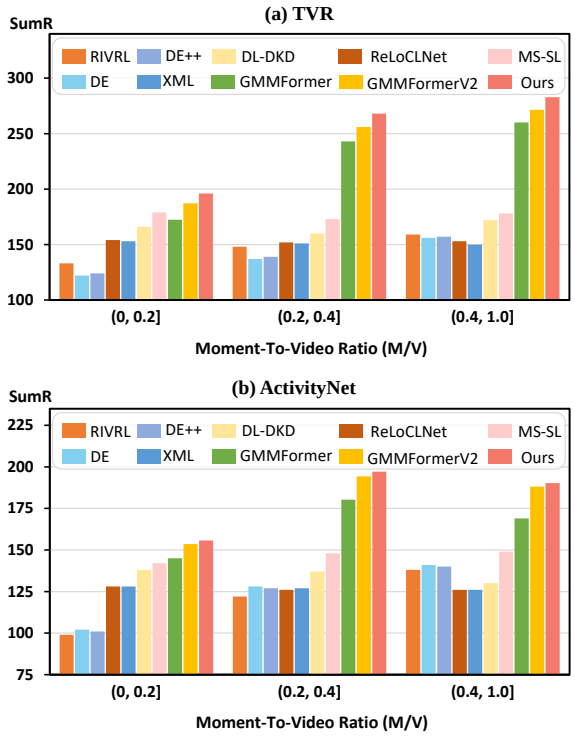

Figure 4: Performance on different types of queries. Queries are grouped according to their M/V ratio r. The smaller r indicates less relevant content while more irrelevant content to the query.

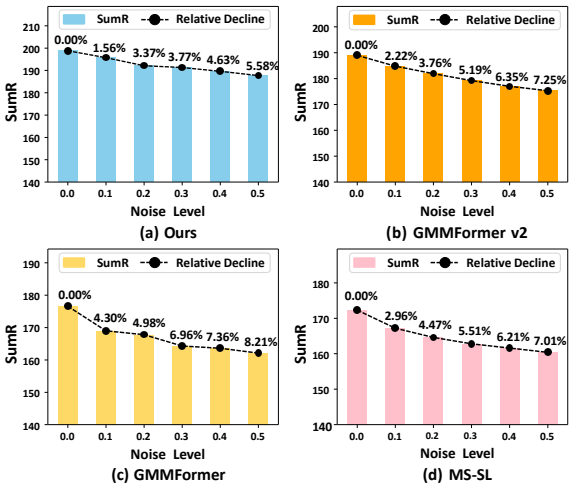

Figure 5: Performance of different methods under different levels of noise on TVR. Our model exhibits **the smallest performance drop** as noise level increases.

Table 2: Ablation studies on model structure on TVR.

MSRA	CSA	R@1	R@5	R@10	R@100	SumR	ΔSumR
		16.2	37.6	48.8	86.4	189.1	-
✓		17.5	39.2	50.7	87.4	194.8	+5.7
	✓	17.0	38.5	51.0	88.1	194.5	+5.4
✓	✓	**18.2**	**40.4**	**52.1**	**88.0**	**198.8**	+9.7

stable, achieving an R@1 of nearly 20. These findings demonstrate the effectiveness of our method in mitigating the impact of data uncertainty, ensuring robust query-video alignment even in *highly ambiguous cases*.

M/V Performance Analysis. In PRVR, queries capture only partial aspects of the video content. Here, we analyze the performance across queries with different M/V ratios r, namely the proportion of query-relevant moment to the total video length. A smaller r indicates that the target video contains less relevant content. This semantic imbalance between the query and the video makes retrieval more challenging. Following (Dong et al., 2022), we categorize the test queries into three groups: short ($r \in (0, 0.2]$), medium ($r \in (0.2, 0.4]$), and long ($r \in (0.4, 1.0]$). As shown in Figure 4, our model consistently outperforms others, demonstrating its effectiveness and robustness across queries with varying relevance levels.

4.3 Further Analysis

Model Robustness Under Noise. Performance under noisy conditions poses a greater challenge to model robustness (Yang et al., 2024a; Pan et al., 2024). Following (Yang et al., 2021), we insert a randomly generated segment with a duration of $h \times p$ seconds at the beginning of the test video, where h represents the duration of the test video and p denotes the noise level. As shown in Figure 5, our model consistently outperforms comparison methods under different noise levels and exhibits the smallest performance drop as noise intensity increases. This highlights the superior resilience of our uncertainty-aware alignment strategy to noisy inputs.

Analysis on Model Structure. We provide an ablation study on TVR in terms of uncertain learning (*w.t.f.* MSRA) and confidence-aware alignment (*w.t.f.* CSA) in Table 2. Firstly, we show the baseline GMMFormer v2 (top row). Based on it, we introduce the MSRA module (2nd row), obtaining 5.7 boost at R@1. This shows the superiority of introducing multimodal learning on distributional representations over plain semantic features. We also evaluate the effect of the CSA module (3rd row). By comparison, considering the word-level confidence of the query significantly improves the performance. This is because meaningless words in the query can capture unrelated background frames, misleading retrieval. By jointly using the designed MSRA and CSA, our method acquires an improvement of 9.7 on SumR (4th row). These ablations demonstrate the effectiveness of each component

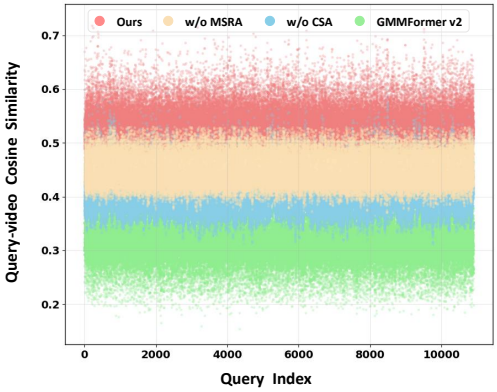

Figure 6: Query-video cosine similarities between test queries on TVR and their top-10 retrieved videos .

of our method in improving PRVR baselines.

In Figure 6, we show the cosine similarities between queries on TVR and the top 10 retrieved videos given by different models. By and large, our model produces similarities above 0.5, while other models range from 0.2 to 0.5. Our complete model not only demonstrates superior retrieval performance but also retrieves videos with higher similarities, indicating that the model can achieve more stable and confident query-video alignment.

Analysis on Distribution Optimization. In Table 3, we conduct ablation studies concerning the learning objectives on the MSRA module. $\mathcal{L}_{DA}$ minimizes the KL distance between the distributions of each query-video pair. $\mathcal{L}_{PM}$ promotes the semantic similarity of random video and text proxies in a contrastive learning framework. By observing Table 3, we have drawn the following conclusions: (1) although removing any loss term leads to a performance decline, both variants are still superior to the baseline without distribution optimization. (2) the joint usage of $\mathcal{L}_{DA}$ and $\mathcal{L}_{PM}$ achieves the best performance, showing the complementarity and effectiveness of constraints on the multimodal distributions and random proxies.

We further discuss the number of sampling proxies for $\mathcal{L}_{PM}$ in Table 3. For "w/o sampling", we directly use the mean of Gaussian distribution as a proxy during the training. This gives a sub-optimal performance due to fixing features rather than exploiting data uncertainty. As the number of proxies K increases from 2 to 6, our method enables better performance by gradually augmenting the data representation based on uncertainty. Considering the trade-off between the performance and computational cost, we choose K=6 in our final model.

Analysis on Uncertainty Modeling. Table 4 highlights the effect of our key design choices in uncertainty modeling. First, we examine the impact of our query support sets. By reducing the query support set to a single query, the text distribution fails to capture broader contextual semantics. This results in a severe semantic mismatch between the text and video distributions, disrupting the optimization process and significantly degrading performance. Next, we explore the role of multi-granular feature aggregation in quantifying data uncertainty. The global aggregation summarizes the holistic context while local aggregation supplements fine-grained details. The results reveal that the combination of global-local aggregation contributes to robust uncertainty modeling and achieves the best performance.

Table 3: Ablation studies on distribution optimization and proxy number on TVR.

Loss	R@1	R@5	R@10	R@100	SumR
w/o $\mathcal{L}_{DA}$	17.7	39.9	51.8	88.0	197.4
w/o $\mathcal{L}_{PM}$	17.4	39.7	51.6	87.8	196.5
Proxy	R1	R5	R10	R100	SumR
w/o sampling	17.6	39.8	51.7	87.6	196.7
K=2	17.9	40.1	51.8	87.7	197.5
K=4	18.0	40.3	51.8	87.9	198.0
K=**6**	**18.2**	40.4	**52.1**	**88.0**	**198.8**

Table 4: Effect of different uncertainty modeling methods on TVR.

Method	R@1	R@5	R@10	R@100	SumR
$\mathbf{X}^q = \mathbf{Q}$	17.6	39.8	51.5	87.9	196.9
$g^m(\mathbf{X}^m) = \mathbf{x}^{m,l}$	17.9	40.2	51.2	87.4	196.7
$g^m(\mathbf{X}^m) = \mathbf{x}^{m,g}$	18.0	40.1	51.6	88.2	197.9
Ours	**18.2**	**40.4**	**52.1**	**88.0**	**198.8**

4.4 Qualitative Results

Challenging Retrieval Cases. To further investigate the impact of data uncertainty, we analyze two challenging retrieval cases with data uncertainty. As shown in Figure 7, we display two queries that refer to different moments within the ground-truth (GT) video and compare the Top-1 retrieval results of our model and GMMFormer v2. We find GMMFormer v2 fails in both cases, retrieving distractor videos containing similar actions (*spreading out cards* or *shuffling a deck of cards*), while ranking GT video at position 9th and 15th. In contrast, our model effectively excavates semantic relationships in data and retrieves GT video at Rank-1.

Confidence-aware Alignment. Here, we investigate how our proposed CSA improves retrieval performance. Figure 8a shows the word-frame similarity matrix, where words like "to be" are aligned

Figure 7: PRVR results on TVR: top-1 retrieved video by our method and GMMFormer v2 (Wang et al., 2024d). Green and red boxes indicate the ground truth and distractor videos, respectively.

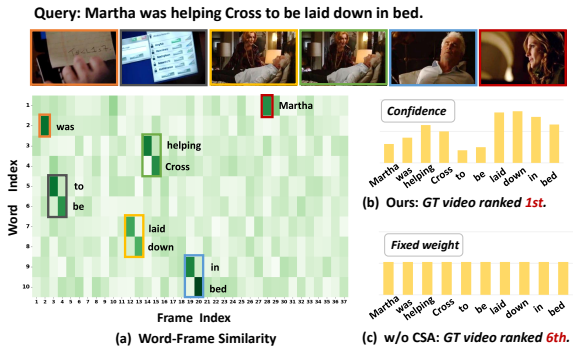

Figure 8: Visualization of CSA mechanism. (a) Word-frame similarity shows uninformative words (*e.g.*, "to be") align with irrelevance frames (gray box). (b) With CSA, uninformative words receive lower confidence, improving retrieval. (c) Fixed average weight causes performance drop, with GT video ranked 6th.

with frames of low relevance to the query (highlighted in gray box). By introducing CSA (Figure 8b), the words "to be" receive a lower confidence, resulting in the correct retrieval of the GT video. In contrast, when confidence is replaced with a fixed average weight (Figure 8c), the GT video drops to the 6th rank, demonstrating the importance of dynamical confidence weighting for precise retrieval.

4.5 Versatility on T2VR

T2VR can be viewed as a simple case of PRVR, where videos are trimmed to correspond to queries. In Table 5, we apply our RAL to T2VR task, comparing it with CLIP4Clip (Luo et al., 2022) under two different visual backbones. It can be found that combining RAL with CLIP4Clip improves the R@1 by about 6.1% and 6.6% under ViT-B/32 and ViT-B/32, respectively. The results further demonstrate the effectiveness and versatility of our framework in enhancing cross-modal alignment.

5 Conclusion

In this paper, we investigate the fundamental challenge of spurious semantic correlations, which arises from query ambiguity and partial video relevance. We propose a novel Robust Alignment Learning (RAL) framework that explicitly models data uncertainty by representing both video and text features as probabilistic distributions, enabling more robust cross-modal alignment. We introduce a query support set that aggregates multiple descriptions of the same video, and multi-granularity feature aggregation to quantify data uncertainty more effectively. Additionally, we design a confidence-aware set-to-set alignment mechanism to assign adaptive weights to query words, improving retrieval precision. Extensive experiments on benchmark datasets demonstrate the effectiveness and versatility of our RAL, achieving significant improvements in both PRVR and T2VR.

Table 5: Text-to-video performance of our method on MSR-VTT dataset for the T2VR task.

Method	R@1	R@5	R@10	MdR↓	MnR↓
CLIP4Clip (ViT-B/32)	44.5	71.4	81.6	2.0	15.3
+RAL	**47.2**	**73.6**	**83.1**	**2.0**	**12.5**
CLIP4Clip (ViT-B/16)	47.1	74.1	81.8	2.0	14.9
+RAL	**50.2**	**76.1**	**85.2**	**1.0**	**12.7**

Limitations

In the validation experiments on the TVR dataset, we conducted an attribution analysis of retrieval failure cases and identified a prominent pattern: *cross-modal alignment bias caused by missing named entities*. For instance, in the query "Beckett confronts a friend at the bar", the discrepancy between the model's retrieved result and the ground-truth (GT) video stems from the model's failure to associate the textual character entity "Beckett" with the corresponding visual representation in the video. Specifically, the GT video contains distinctive visual cues associated with this character—such as a red jacket and curly hair. In contrast, the retrieved distractor video, although set in

a similar bar scene, lacks these fine-grained identity indicators. Our current approach does not explicitly model the correspondence between named entities in the query and specific characters in the video, leading to retrieval ambiguity. This limitation highlights a potential direction for future research: *incorporating identity-aware modeling* to associate textual mentions of characters with their visual counterparts in the video (Song et al., 2024b; Zhou et al., 2025b,a). This could involve integrating entities attribute information from knowledge graphs and using attention mechanisms to guide the model toward identity-relevant visual cues, thereby enhancing its applicability in real-world retrieval scenarios (Zhang et al., 2024; Song et al., 2024a).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62402471, Grant U22A2094, Grant 62472385, and Grant 62272435. We also acknowledge support from the Pioneer and Leading Goose R&D Program of Zhejiang under Grant 2024C01110. This research was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC. We also acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. 2022. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Zhiguo Chen, Xun Jiang, Xing Xu, Zuo Cao, Yijun Mo, and Heng Tao Shen. 2023. Joint searching and grounding: Multi-granularity video content retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 975–983.

Cheol-Ho Cho, WonJun Moon, Woojin Jun, MinSeok Jung, and Jae-Pil Heo. 2025. Ambiguity-restrained text-video representation learning for partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2500–2508.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080.

Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu, Xun Wang, and Baolong Liu. 2023. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11302–11312.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xun Jiang, Zhiguo Chen, Xing Xu, Fumin Shen, Zuo Cao, and Xunliang Cai. 2023. Progressive event alignment network for partial relevant video retrieval. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1973–1978.

Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in neural information processing systems*, 35:30291–30306.

Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2024. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889.

Taichi Nishimura, Shota Nakada, and Masayoshi Kondo. 2023. Large-scale vision-language models learn super images for efficient and high-performance partially relevant video retrieval. *arXiv preprint arXiv:2312.00414*.

Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. 2019. Modeling uncertainty with hedged instance embeddings. In *7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6-9, 2019*.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multi-modal neurons in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024a. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 33:1122–1135.

Peipei Song, Long Zhang, Long Lan, Weidong Chen, Dan Guo, Xun Yang, and Meng Wang. 2025. Towards efficient partially relevant video retrieval with active moment discovering. *IEEE Transactions on Multimedia*, pages 1–12.

Peipei Song, Yuanen Zhou, Xun Yang, Daqing Liu, Zhenzhen Hu, Depeng Wang, and Meng Wang. 2024b. Efficiently gluing pre-trained language and vision models for image captioning. *ACM Transactions on Intelligent Systems and Technology*, 15(6):1–16.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024a. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16551–16560.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. 2024b. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024c. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Xun Yang, Dapeng Wu, Xiangyu Chen, and Xuelong Li. 2025. Episodic memory representation for long-form video understanding. *arXiv preprint arXiv:2508.09486*.

Yuting Wang, Jinpeng Wang, Bin Chen, Tao Dai, Ruisheng Luo, and Shu-Tao Xia. 2024d. Gmmformer v2: An uncertainty-aware framework for partially relevant video retrieval. *arXiv preprint arXiv:2405.13824*.

Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. 2024e. Gmmformer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5767–5775.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713.

Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. 2024a. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision*, 132(11):4823–4849.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1–10.

Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216.

Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. 2024b. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67(10):202104.

Shukang Yin, Sirui Zhao, Hao Wang, Tong Xu, and Enhong Chen. 2024. Exploiting instance-level relationships in weakly supervised text-to-video retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(10):1–21.

Biao Zhang and Peter Wonka. 2021. Point cloud instance segmentation using probabilistic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8883–8892.

Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. 2023. Multi-event video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22113–22123.

Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695.

Jing Zhang, Dan Guo, Xun Yang, Peipei Song, and Meng Wang. 2024. Visual-linguistic-stylistic triple reward for cross-lingual image captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–23.

Long Zhang, Peipei Song, Zhangling Duan, Shuo Wang, Xiaojun Chang, and Xun Yang. 2025a. Video corpus moment retrieval with query-specific context learning and progressive localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):5659–5670.

Qun Zhang, Chao Yang, Bin Jiang, and Bolin Zhang. 2025b. Multi-grained alignment with knowledge distillation for partially relevant video retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. 2025a. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373.

Sheng Zhou, Junbin Xiao, Xun Yang, Peipei Song, Dan Guo, Angela Yao, Meng Wang, and Tat-Seng Chua. 2025b. Scene-text grounding for text-based video question answering. *IEEE Transactions on Multimedia*.

Example Appendix

This supplementary document includes the following:

(*i*) Additional information about our implementation details (Section A);

(*ii*) Additional experimental results and analysis (Section B), including the variation trend of uncertainty and performance during training (Section B.1), the impact of loss coefficients (Section B.2), and studies on the retrieval efficiency of different PRVR methods (Section B.3);

(*iii*) Additional qualitative examples of our method and discussions on future work (Section C).

A Implementation Details

Figure 9 elaborates on the experimental details for applying the proposed URAL framework on existing baselines, including MS-SL (Dong et al., 2022), GMMFormer (Wang et al., 2024e), and GMMFormer v2 (Wang et al., 2024d). Specifically, URAL is incorporated solely into the frame-level branch of the baselines. This design choice is motivated by considerations of: (1) the frame-level branch provides fine-grained temporal information, which is essential for handling the uncertainty inherent in cross-modal alignment while avoiding

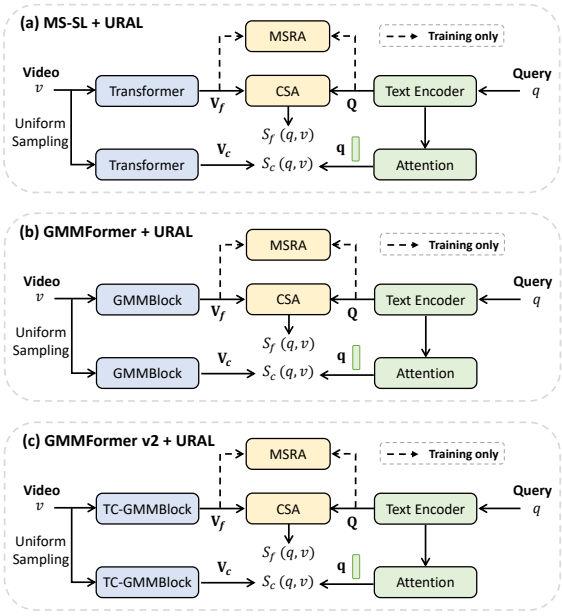

Figure 9: Integration of RAL with PRVR baselines, where MSRA and CSA stand for the proposed multi-modal semantic robust alignment and confidence-aware set-to-set alignment modules, respectively. RAL is integrated into the frame-level branch, and the final retrieval score is a combination of frame-level and clip-level scores.

unnecessary computational overhead; (2) frame-level features are more conducive to integrating our proposed modules, such as the confidence-aware set-to-set alignment module. In our implementation, we retain the original video and text encoders from the baselines to ensure fair comparison. The extracted video frame features $\mathbf{V}_f$ and query word features $\mathbf{Q}$ are fed into the MSRA (Multimodal Semantic Robust Alignment) module, which explicitly models and mitigates uncertainty to enhance cross-modal alignment. The resulting robust $\mathbf{Q}$ and $\mathbf{V}_f$ are subsequently processed by CSA (Confidence-aware Set-to-set Alignment) module for query-video alignment with adaptive confidence weighting. The CSA module generates reliable frame-level retrieval scores $S_f(q, v)$, which are then summed with the clip-level scores $S_c(q, v)$ to produce the final retrieval result.

B More Experimental Results

B.1 Uncertainty Mitigation During Training

To investigate the variation in video and query uncertainty during training and its impact on performance, we quantify the uncertainty every 5 training epochs and report the corresponding SumR performance on the test set. We observe that the uncertainty of both videos and queries decreases

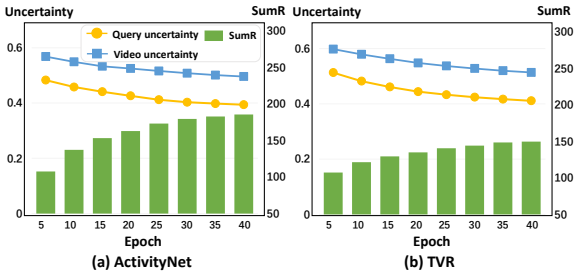

Figure 10: The variation trend of data uncertainty and retrieval performance during training. As training progresses, reduced uncertainty leads to improved retrieval accuracy.

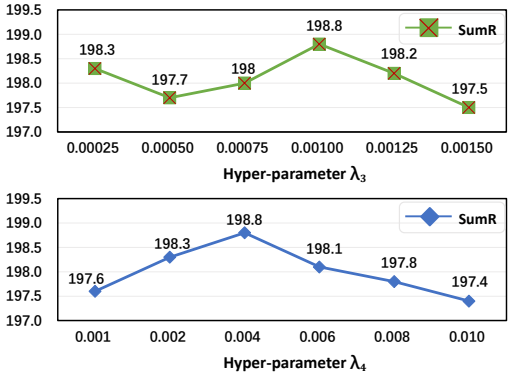

Figure 11: The impact of the loss coefficients, λ_3 and λ_4, of distribution alignment loss $\mathcal{L}_{DA}$ and proxy matching loss $\mathcal{L}_{PM}$.

as training progresses, and the model's retrieval performance improves. This indicates that mitigating uncertainty is crucial for improving retrieval accuracy. Additionally, we find that videos exhibit higher uncertainty than queries. The presence of redundant content in untrimmed videos is a major challenge for PRVR. This highlights an important direction for future research.

B.2 Hyper-parameter Analysis

In addition to the basic retrieval loss $\mathcal{L}_{base}$ (Dong et al., 2022), our model incorporates auxiliary distribution alignment loss $\mathcal{L}_{DA}$ and proxy matching loss $\mathcal{L}_{PM}$ to enhance alignment. In Figure 11, we study the sensitivity of two loss coefficients, λ_3 and λ_4, on the TVR dataset. The initial settings are $\lambda_3 = 0.001$ and $\lambda_4 = 0.00025$ to keep each loss item at the same magnitude. We adjust these hyper-parameters within a certain range to assess their impact. As shown in Figure 11, our model maintains stable performance and reaches the optimal balance at $\lambda_3 = 0.004$ and $\lambda_4 = 0.001$.

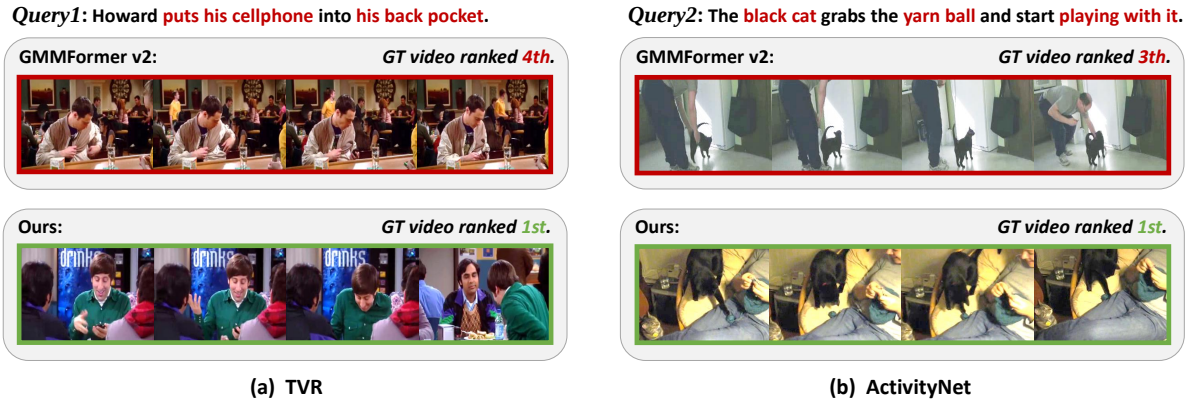

Figure 12: More visualization results on TVR and ActivityNet. Top-1 retrieved videos from our method and GMMFormer v2 (Wang et al., 2024d) are shown. Green and red boxes indicate the ground truth and distractor videos, respectively.

Table 6: Comparison in terms of FLOPs (G) and parameters (M). Δ denotes our relative changes over the baseline (GMMFormer v2) for different metrics.

	MS-SL	GMMFormer	GMMFormer v2	Ours	Δ
FLOPs	1.29	1.95	5.43	5.75	+0.32
Params	4.85	12.85	32.27	35.53	+3.26
SumR	172.4	176.6	189.1	198.8	+9.7

Table 7: Comparisons in terms of runtime (ms) of PRVR models.

Database Size	500	1,000	1,500	2,000	2,500
MS-SL (Dong et al., 2022)	4.89	6.11	8.06	10.42	12.93
GMMFormer (Wang et al., 2024e)	2.68	2.93	3.40	3.94	4.56
GMMFormer v2 (Wang et al., 2024d)	3.95	4.32	5.02	5.81	6.73
Ours	4.61	5.05	5.86	6.79	7.86

B.3 Retrieval Efficiency

To evaluate model efficiency, we compare several PRVR methods in terms of floating-point operations (FLOPs) and model parameters. Our method builds upon GMMFormer v2 as the baseline while introducing uncertainty learning and confidence-aware alignment. As shown in Table 6, while our model increases FLOPs by 0.32G and parameters by 3.26M, it achieves a substantial 9.7% improvement in SumR. This highlights a favorable trade-off between computational cost and performance gain.

For retrieval efficiency in practical situations, we measure the retrieval speed (in milliseconds) as shown in Table 7. Specifically, we construct a video subset from the TVR dataset and measure the average runtime to complete the retrieval process for a single text query under different database size settings. Despite introducing confidence-aware alignment during retrieval, our model's runtime remains comparable to GMMFormer v2. Moreover, as the database size increases, the runtime increases marginally, demonstrating the potential for large-scale applications.

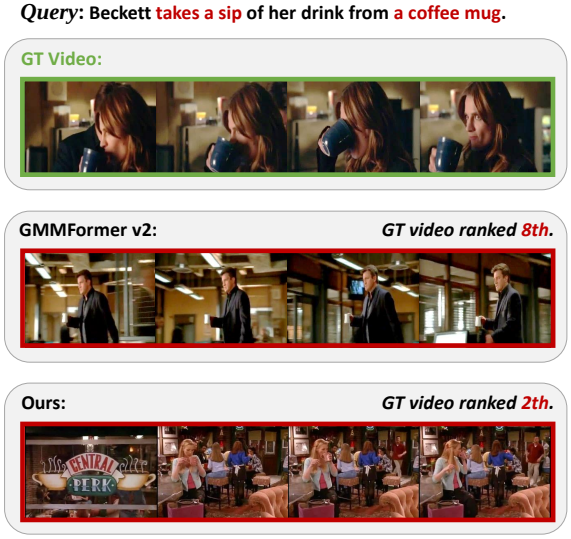

Figure 13: Failure case on TVR. red boxes indicate the top-1 retrieved video by our method and GMMFormer v2 (Wang et al., 2024d). Green box indicates the ground truth video.

C More Visualization Results

C.1 Qualitative Retrieval Results

Figure 12 presents two additional visualization examples from TVR (Lei et al., 2020) and ActivityNet (Krishna et al., 2017) datasets, comparing the top-1 retrieval results of our model against GMMFormer v2 (Wang et al., 2024d). In both cases, GMMFormer v2 fails to retrieve the target videos, instead selecting distractor videos with similar scenes, such as a "*cellphone*" and a "*black cat*", while ranking the ground-truth (GT) videos at the 4th and 3rd positions, respectively. In contrast, our model effectively uncovers semantic relationships and successfully ranks the GT videos at 1st. For example,

in Figure 12 (a), our model is sensitive to the action "*puts his cellphone*", whereas GMMFormer v2 retrieves a distractor video featuring a different action, "*pull out phone*". In Figure 12 (b), the "yarn ball" is a subtle but crucial visual cue that GMMFormer v2 overlooks, whereas our model successfully detects it for accurate retrieval. These qualitative results demonstrate that our approach significantly enhances retrieval accuracy by capturing critical semantic details in query and video.

C.2 Failure Cases and Future Work

Figure 13 presents a failure case from the TVR dataset, comparing the top-1 retrieval results of our model and GMMFormer v2. The query describes a common scenario of drinking coffee. Although both models fail to retrieve the GT video as the top-1 result, our model correctly captures the key phrase in the query (*i.e.*, "*takes a sip of her drink*") and retrieves a highly relevant video, ranking the GT video in 2nd place. In contrast, GMMFormer v2 retrieves a video of "*a man carrying a coffee cup*" and ranks the GT video only at 8th place.

Further analysis reveals that a critical factor distinguishing the GT video from our retrieved video is the presence of "*Beckett*", a named entity in the query. Our approach does not involve the correspondence between named entities in the query and specific individuals in the video, leading to retrieval ambiguity. This limitation highlights a potential direction for future research: incorporating identity-aware modeling to associate textual mentions of people with their visual counterparts in videos, making it better suited for real-world retrieval scenarios.