Bridging the Semantic Granularity Gap Between Text and Frame Representations for Partially Relevant Video Retrieval

WooJin Jun, WonJun Moon, Cheol-Ho Cho, MinSeok Jung, and Jae-Pil Heo*

Sungkyunkwan University
{junwoojinjin, wjun0830, hoonchcho, minseokjung0328, jaepilheo}@gmail.com

Abstract

Partially Relevant Video Retrieval (PRVR) addresses the challenges of text-to-video retrieval in real-world scenarios where untrimmed videos are prevalent. Traditional PRVR methods encode videos at two feature scales: (1) frame-level to capture fine details, and (2) clip-level to recognize broader content. However, these approaches align both scales with a single sentence representation, leading to suboptimal performance. In particular, we point out the level mismatch in aligning frame-level video features with a sentence representation, as the entire meaning of a sentence contains broader and more diverse content than what frame-level features can encode. This misalignment causes frame-level features to capture broader contexts and overlook local fine details. To tackle this issue, we propose a framework that represents a sentence as a set of multiple components, where each component aligns with frame-level semantics. Specifically, we introduce Semantic-Decomposed Matching (SDM) to adjust the granularity of the text description to match them with frame-level video features. In addition to the matching process, we develop the Adaptive Local Aggregator (ALA) to enhance video encoding in capturing finer local details, ensuring precise text-video alignment at the frame level. ALA adaptively integrates multi-scale local details within short temporal spans obtained by enforcing a strict temporal aggregation range. Finally, we reinforce detailed encoding at the frame level with newly designed objectives for both modalities. Extensive experiments integrating our framework with existing clip branches demonstrate its effectiveness and applicability, highlighting significant improvements in PRVR performance.

1 Introduction

With the growing demand and supply of videos, text-to-video retrieval has drawn significant attention in research communities. Given a text query as input, text-to-video retrieval aims to retrieve the video most relevant to the query in the video database (Luo et al. 2022; Deng et al. 2023; Wang et al. 2024a; Tian et al. 2024; Wang et al. 2023; Gorti et al. 2022; Chen et al. 2020; Dong, Li, and Snoek 2018; Miech et al. 2019; Liu et al. 2019a). The conventional paradigm of text-to-video retrieval operates under the assumption that videos are trimmed so that every frame is relevant to the query. However, real-world videos are often lengthy and untrimmed, containing diverse content. Consequently, typical text-to-video retrieval methods struggle to achieve optimal performance when applied to untrimmed videos, as they are ill-equipped to identify the relevant segments.

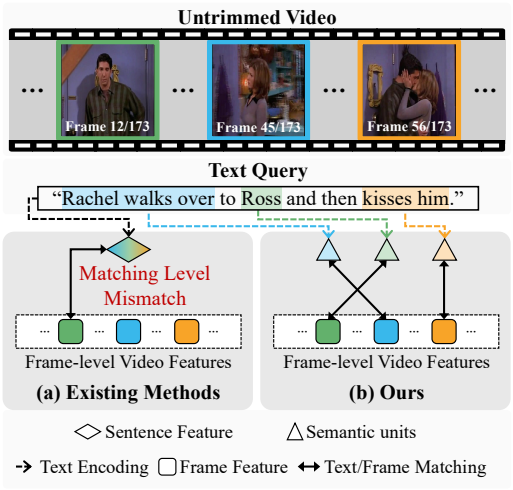

Figure 1: Comparison of the matching process in the frame-level branch. (a) Existing methods match the entire sentence query to frame features, leading to a mismatch in content granularity. (b) Our approach decomposes the broad semantics of a sentence into smaller units that each frame feature can fully encompass. This aligns the content granularity in the matching process, allowing the frame-level branch to focus on capturing fine details.

To address this, Partially Relevant Video Retrieval (PRVR) has been proposed (Dong et al. 2022a). PRVR aims to retrieve an untrimmed video from the storage that contains the most relevant segment corresponding to a given text query. The core challenges in PRVR lie in 1) precisely encoding all varying lengths of contents in long untrimmed videos and 2) aligning the features of the text query with the video features in a joint representation space. Traditional PRVR approaches (Dong et al. 2022a; Wang et al. 2024c,b) primarily focus on the former challenge by implementing dual encoding strategies to handle varying

*Corresponding author

contextual lengths within untrimmed videos. Typically, dual branches are organized with frame-level and clip-level branches. Frame-level features are designed to capture content at short temporal intervals, while clip-level features encapsulate broader content within the videos. These features are then aligned with a single-sentence query representation to match the video with the text.

Although the multi-stream strategy has shown promising progress, we argue that the role of the frame branch is ambiguously designed. Specifically, the matching process between video features and a sentence representation may not be suitable for the frame level, as sentences often encompass a broader context than what frames can describe. This matching process then causes frame-level encoding to capture more extensive content than intended. In Fig. 1, we illustrate the motivation with an example. As shown, we point out that semantics within a sentence query may span multiple frames at different temporal locations, thereby a single frame features may not encompass all given contents. Consequently, the level mismatch during the matching process incurs unexpected comprehensiveness in the frame feature.

In this regard, we propose a novel framework for the frame-level branch in PRVR. The key objective of our framework is to ensure that the frame-level representations capture finer details for both text and video modalities. For the text modality, we first propose Semantic-Decomposed Matching (SDM) that breaks down the semantics in the sentence into frame-level query representations and matches them with video frame features. Specifically, each word in a sentence is encoded into learnable query vectors to form decomposed semantic units, referred to as frame-level query representations. For query encoding, we utilize slot attention (Locatello et al. 2020), which iteratively performs similarity-based operations. By aggregating word semantics based on these similarity scores, each frame-level query representation encompasses specific semantics within the text sentence likely to appear at the frame level of the video. These semantic units are then pushed apart with Semantic Diversity loss to reinforce the diversity within the semantic decomposition process. By aligning the levels of frame representations of both modalities, we enable the frame branch models to focus on more granular details.

For the video modality, we further concentrate on encoding local events in the frame-level branch to align with specific content in a sentence. Recognizing that Gaussian-based attention alone is insufficient for local feature encoding (Please see Fig. 4), we introduce Adaptive Local Aggregator (ALA). This ALA employs a multi-scale window attention process to address the variable lengths of local events. Subsequently, these multi-scale local details are integrated into the frame representation based on their similarity, ensuring a reliable local context for frame-level matching. Finally, to encourage the disentanglement between the frame representations within the same video, we apply Frame-Semantic Alignment loss that discourages the similarity between the frames that are matched to different semantic units derived from the sentence.

Lastly, we performed extensive experiments by integrating our proposed method into existing PRVR methods across two large-scale video datasets: TVR and ActivityNet Captions. Our results demonstrate that the proposed design increases performance by large margins, showing its specialty for frame-level encoding. Additionally, our analysis shows that local fine details are appropriately exploited for both modalities. For detailed results, we refer to Sec. 4.

Overall, our key contributions are as follows: (1) We propose a novel frame-level framework to address the matching-level mismatch between sentence representations and video frame features in PRVR. This clarifies the role of the frame branch in capturing fine details. (2) We propose Semantic-Decomposed Matching (SDM) and semantic diversity loss to construct frame-level text representations. (3) We introduce an Adaptive Local Aggregator (ALA) with Frame-Semantic Alignment loss to enhance locality in frame-level video features. (4) Extensive experiments on two large-scale datasets (i.e., TVR and ActivityNet Captions) demonstrate the effectiveness of our approach.

2 Related Work

Text-to-Video Retrieval. Text-to-video retrieval (T2VR) is a task that searches for the most relevant video based on text queries (Huang et al. 2023; Ma et al. 2023; Pei et al. 2023; Wu et al. 2023; Wang et al. 2024a; Tian et al. 2024; Li et al. 2019; Dong et al. 2019, 2021, 2022b). Recently, the mainstream has been to employ large-scale image-language pre-trained models, such as CLIP (Contrastive Language-Image Pretraining) (Radford et al. 2021), to mitigate the high training costs associated with raw video data (Fang et al. 2023; Li et al. 2024; Jin et al. 2023c). However, since the initial work relied solely on representative vectors of the text and video for similarity matching (Luo et al. 2022), a multi-scale similarity matching paradigm between text and video (e.g., patch-word and video-sentence level) has been introduced (Wang et al. 2023; Li et al. 2023a; Guan et al. 2023; Jin et al. 2023b,a; Li et al. 2023b). Subsequently, another popular approach has focused on maintaining retrieval efficiency while retaining the effectiveness of multi-scale matching (Liu et al. 2022; Deng et al. 2023). These works propose temporal feature aggregation architectures to streamline the retrieval process. Nevertheless, T2VR relies on the relatively strong assumption that trimmed clips highly relevant to the text descriptions are often available, leading to reduced effectiveness and higher costs on long videos.

Partially Relevant Video Retrieval. Partially relevant video retrieval (PRVR) aims to retrieve a video even when only partial clips are relevant to the search query. Initially, MS-SL (Dong et al. 2022a) introduced a dual encoding strategy to encode video features at both the clip and frame levels, accommodating the varying contextual ranges of textual queries. This approach addresses the fact that the context within each query may span short to long temporal ranges. GMMFormer (Wang et al. 2024c) focused on developing an efficient retrieval framework, effectively reducing the number of video clips to be stored by assuming that contexts often form sequentially. Subsequently, GMMFormer v2 (Wang et al. 2024b) enhanced its predecessor by adding more attention blocks along with a consolidation module,

developing uncertainty-aware loss coefficients, and adopting Hungarian matching. However, their limitations lie in overlooking the granularity in the text-video matching process by projecting frame features to align with the sentence feature. In this work, we focus on a frame encoding strategy by decomposing the sentence into multiple segments that match the contextual level of individual frame features.

3 Method

3.1 Preliminary

Typical PRVR methods employ a dual video-encoding strategy to handle untrimmed videos, considering long and short clips. They consist of three phases: video encoding, text encoding, and retrieval score measurement.

Untrimmed video encoding. Given an untrimmed video $\mathbf{v}$, pre-trained 2D or 3D CNNs are exploited to obtain frame-wise video representations $V' \in \mathbb{R}^{L_v \times d_v}$, where L_v is the sequence length and d_v is video feature dimension (for clip branch, mean pooling is implemented on temporal axis to construct initial clips with V'). Subsequently, video features are processed through a fully connected layer and transformer encoder (Vaswani et al. 2017) to incorporate the temporal information. Note that frame and clip branches employ the identical architectural design but with different weights. Consequently, the outputs of frame and clip branches are $V_f \in \mathbb{R}^{L_v \times d} = \left\{v_i^f\right\}_{i=1}^{L_v}$ and $V_c \in \mathbb{R}^{L_c \times d} = \{v_i^c\}_{i=1}^{L_c}$, where L_c is the length of the clip-level video features and d is projected feature dimension in the text-video joint space.

Text query encoding. Given a text query $\mathbf{q}$ with L_q words, a pre-trained text encoder extracts initial word features $W' \in \mathbb{R}^{L_q \times d_w}$, where d_w is the feature dimension. These features are projected through an FC layer and transformer encoder to obtain contextualized word features $W \in \mathbb{R}^{L_q \times d} = \{w_i\}_{i=1}^{L_q}$. Then, a simple attention module is applied to integrate word features into a single sentence representation $q \in \mathbb{R}^d$ as:

$$q = \sum_{i=1}^{L_q} \alpha_i^q \times w_i, \alpha^q = \text{softmax}(\mathbf{u}W^T), \tag{1}$$

where $\alpha^q \in \mathbb{R}^{1 \times L_q}$ denotes attention vector yielded with the learnable vector $\mathbf{u} \in \mathbb{R}^{1 \times d}$.

Retrieval score measuring. The retrieval score between each text and video is determined through similarity matching. Specifically, the similarity between the video segments within the sequences V_f and V_c and the text query q is calculated. For each branch, the maximum similarity is considered as the retrieval score. This process is expressed as:

$$S_f(\mathbf{v}, \mathbf{q}) = \max(\cos(v_1^f, q), \ldots, \cos(v_{L_v}^f, q)), \tag{2}$$

$$S_c(\mathbf{v}, \mathbf{q}) = \max(\cos(v_1^c, q), \ldots, \cos(v_{L_c}^c, q)), \tag{3}$$

where S_f and S_c represent the frame-level and clip-level scores, and $\cos(,)$ denotes the cosine similarity function. The overall retrieval score S for each text-video pair is calculated by combining these scores as follows:

$$S(\mathbf{v}, \mathbf{q}) = \beta \times S_f(\mathbf{v}, \mathbf{q}) + (1 - \beta) \times S_c(\mathbf{v}, \mathbf{q}), \tag{4}$$

where β is a hyperparameter used to balance the two scores.

Training objective. To align video and text features, infoNCE $\mathcal{L}^{\text{nce}}$ (Faghri et al. 2017; Dong et al. 2022a) and triplet ranking $\mathcal{L}^{\text{trip}}$ (Miech et al. 2020; Luo et al. 2022) losses are popularly employed. These losses ensure that positive text-video pairs retain higher similarity than negative pairs. Formally, the typical objectives are expressed as:

$$\mathcal{L}^{\text{base}} = \mathcal{L}_c^{\text{nce}} + \mathcal{L}_c^{\text{trip}} + \mathcal{L}_f^{\text{nce}} + \mathcal{L}_f^{\text{trip}}, \tag{5}$$

where $\mathcal{L}_c^*$ and $\mathcal{L}_f^*$ represent the losses using the clip-level score S_c and frame-level score S_f, respectively.

3.2 Method Overview

As illustrated in Sec. 3.1, existing works treat two different levels of video branches equally; aligning both levels with a single sentence query and employing identical architectural designs. Particularly, this alignment with sentence-level query limits the ability of the frame branch to capture fine details, as the full meaning of a sentence encompasses broader contents than frame-level features can encode. This results in overlapping roles and reduces the diversity of video branches. To address these issues, we propose a novel framework for the frame branch that clearly defines its role in both text and video encoding streams. In Sec. 3.3, we initially illustrate Semantic Decomposed Matching (SDM), which involves text encoding at the frame level. SDM decomposes a sentence into multiple semantic units, aligning them with video frames to obtain frame-level scores. To ensure the generated units capture diverse semantics, we also introduce the Semantic Diversity loss. In Sec. 3.4, we present an Adaptive Local Aggregator (ALA) to strengthen the frame-level encoding for the video modality. ALA refines frame features by modeling multi-scale local events with window attention, allowing them to encompass adjacent contexts. Then, we supplement the diverse frame encoding with Frame-Semantic Alignment loss.

3.3 Semantic-Decomposed Matching (SDM)

Decomposed representation from query sentence. Typical frame-level matching involves aligning a single sentence representation with video frame features, leading video frame features to capture broader content than intended at the frame level. Thus, we aim to encode varied semantic units from a sentence to align more effectively with video frame features through our semantic decomposition approach. Our semantic decomposition process consists of three layers: a projection layer, a transformer, and a slot encoder. First, the word features $W_f \in \mathbb{R}^{L_q \times d}$ are obtained via processing initial word features W' through projection and transformer encoder. Then, we employ the slot encoder (Locatello et al. 2020) to group semantically relevant word features W_f within the learnable vectors, as shown in Fig. 2. To illustrate, n learnable vectors $L_0 = \left\{l_i^0\right\}_{i=1}^n \in \mathbb{R}^{n \times d}$

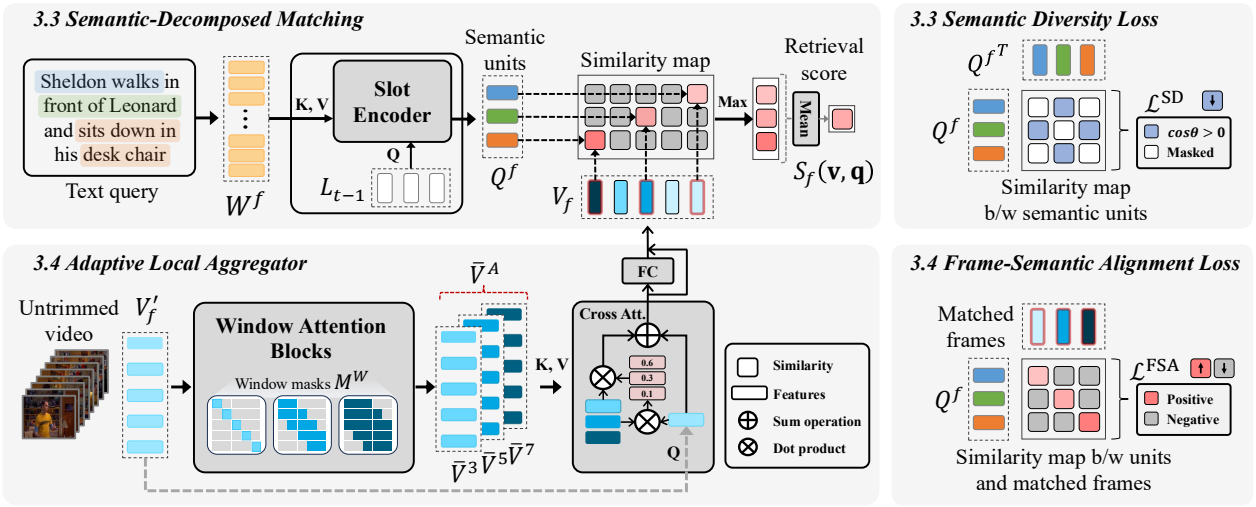

Figure 2: Overview of our frame-level framework. Semantic-Decomposed Matching (SDM) identifies frame-level sub-semantics within a sentence, using a slot encoder to generate semantic units (frame-level queries). These semantic units are then aligned with frame-level video features to compute the retrieval score $S_f(\mathbf{v}, \mathbf{q})$. Video features are extracted using an Adaptive Local Aggregator (ALA), which first generates multi-scale local features via window attention blocks and then integrates these features into frame-level features using a cross-attention layer. On the right side, we present our proposed objectives. The Semantic Diversity loss reduces similarity between frame-level queries within a single sentence, while the Frame-Semantic Alignment loss further discourages the entanglements of frame features that are matched to different semantic units.

are randomly initialized and subsequently refined at each iteration $t = 1 \dots T$ to bind with specific semantics within the word features W_f. At iteration t, we utilize linearly projected features of word features W_f as key and value, and learnable vectors $L_{t-1} = \{l_i^{t-1}\}_{i=1}^{n}$ as the query in the cross-attention layer. Note that the softmax operation in the cross-attention is applied along the query dimension to encode different word information for each learnable vector. Then, the output of t-th iteration is yielded via the refining process composed of GRU (Chung et al. 2014) and MLP layers. The final semantic units after T iterations are denoted as $L_T = Q_f = \left\{q_i^f\right\}_{i=1}^{n}$ where each learnable query encodes specific semantics to match the frame-level features.

Retrieval score with decomposed query representations. As the conventional retrieval process only measures the similarity based on a single text representation, a different strategy is required to process multiple text representations. As each unit captures different semantics within the specified sentence, we aim to discover a video that includes all semantics given in the units. Thus, we average the retrieval scores for each unit to produce the frame-level score as follows:

$$S_f(\mathbf{v}, \mathbf{q}) = \frac{1}{n}\sum_{i=1}^{n} \max(\cos(v_1^f, q_i^f), \dots, \cos(v_{L_v}^f, q_i^f)). \quad (6)$$

By assigning same weight to all semantic units, we expect the model to consider each semantic within the unit equally for video retrieval. We refer to Sec. 4.3 for comparison between different scoring options other than using average.

Semantic Diversity loss. To discourage the correlation between different semantic units, we apply the Semantic Diversity loss. To be specific, the decomposed query representations, referred to as semantic units, $Q_f = \left\{q_i^f\right\}_{i=1}^{n}$ are penalized for their cosine similarity. Given a mini-batch $\mathcal{B}$ containing m queries, the loss is calculated as follows:

$$\mathcal{L}^{\text{SD}} = \frac{1}{mn}\sum_{\mathbf{q}\in\mathcal{B}}\sum_{q_i^f, q_j^f \in Q^f} \cos(q_i^f, q_j^f), \quad (7)$$

where $i \neq j$ and only pairs with $\cos(q_i^f, q_j^f) > 0$ are included.

3.4 Adaptive Local Aggregator (ALA)

In addition to the enhancements for text modality, achieving text-video alignment at the frame level also requires the video frame features to be at the same level of granularity (i.e., focusing on local fine details). To provide reliable local fine details in the video frame branch, we introduce an Adaptive Local Aggregator (ALA). Given that the frame-wise features $V_f' \in \mathbb{R}^{L_v \times d}$ are produced by projecting the frame features V' from the backbone, ALA aims to generate frame-level features V_f that focus on the details within neighboring frames. As shown in Fig. 2, ALA consists of window attention blocks and aggreagation processes. Initially, the window attention blocks encode multi-scale frame features to accommodate events of varying temporal lengths. Within the attention mechanism of each block, a window mask matrix $M^W \in \mathbb{R}^{L_v \times L_v}$ is applied to the attention score matrix using an element-wise sum. Note that M^W is organized with the scalar 0 and $-\infty$ where features positioned at $-\infty$ are masked out. This window attention mechanism is formulated as:

$$\text{Attn}^W(V_f') = \text{softmax}\left(M^W + \frac{V_f'\Theta^Q (V_f'\Theta^K)^T}{\sqrt{d_k}}\right) V_f'\Theta^V, \quad (8)$$

Model	TVR R@1	TVR R@5	TVR R@10	TVR R@100	TVR SumR	ActivityNet Captions R@1	ActivityNet Captions R@5	ActivityNet Captions R@10	ActivityNet Captions R@100	ActivityNet Captions SumR
VCMR models w/o moment localization:										
XML	10.0	26.5	37.3	81.3	155.1	5.3	19.4	30.6	73.1	128.4
ReLoCLNet	10.7	28.1	38.1	80.3	157.1	5.7	18.9	30.0	72.0	126.6
CONQUER	11.0	28.9	39.6	81.3	160.8	6.5	20.4	31.8	74.3	133.1
PRVR models:										
PEAN	13.5	32.8	44.1	83.9	174.2	7.4	23.0	35.5	75.9	141.8
DL-DKD	14.4	34.9	45.8	84.9	179.9	8.0	25.0	37.4	77.1	147.6
MS-SL	13.5	32.1	43.4	83.4	172.4	7.1	22.5	34.7	75.8	140.1
MS-SL†	12.9	32.0	43.1	83.3	171.3	6.9	22.2	34.6	76.0	139.7
+ Ours	**15.7**	**36.7**	**47.9**	**85.8**	**186.2**	**7.2**	**23.5**	**35.8**	**76.9**	**143.4**
GMMFormer	13.9	33.3	44.5	84.9	176.6	**8.3**	24.9	36.7	76.1	146.0
GMMFormer‡	15.2	35.4	47.2	85.9	183.8	8	24.7	36.9	76.2	145.8
+ Ours	**17.2**	**38.3**	**49.8**	**87.0**	**192.3**	8.1	**25.2**	**37.9**	**77.9**	**149.1**
GMMFormer v2	16.2	37.6	48.8	86.4	189.1	8.9	27.1	40.2	78.7	154.9
GMMFormer v2†	15.3	36.4	48.5	86.7	186.9	8.9	27.1	40.0	78.6	154.6
+ Ours	**17.4**	**39.7**	**51.4**	**87.9**	**196.4**	**9.1**	**27.3**	**40.4**	**79.8**	**156.6**

Table 1: Performance comparison on TVR and ActivityNet Captions. The symbol † indicates the reproduced performance, while ‡ represents a variant of GMMFormer that excludes the pooling layer in the video frame-level branch. We note that the original GMMFormer outputs only a single visual frame features for the frame branch.

where Θ^Q, Θ^K, and Θ^V are learnable matrices to project the query, key, and value, respectively, and d_k is the dimension of queries and keys. The ALA comprises o window attention blocks, with each block producing an output $\bar{V}^{2i+1} \in \mathbb{R}^{L_v \times d}$, where i (=1, . . . , o) represents the window size used in attention mask. These outputs are then concatenated along the window axis to form $\bar{V}^A \in \mathbb{R}^{o \times (L_v \times d)}$, which is leveraged to reinforce the local details within each frame feature while accommodating variations in the lengths of semantically coherent consecutive frames. Specifically, $\bar{V}^A$ is integrated into the video frame feature $V'_f \in \mathbb{R}^{1 \times (L_v \times d)}$ through a cross-attention layer to obtain final frame features V_f. This process with a residual connection is formulated as follows:

$$\hat{V} = \text{CA}(V'_f, \bar{V}^A, \bar{V}^A) + V'_f, \tag{9}$$

$$V_f = \text{FC}(\hat{V}) + \hat{V}, \tag{10}$$

where CA denotes the cross-attention layer and FC stands for a fully connected layer.

Frame-Semantic Alignment loss. Frame-Semantic Alignment loss further aims to suppress the semantic entanglement between frame features. This is achieved by increasing the similarity between each semantic unit and its corresponding frames while penalizing the similarity with frames matched to other queries. Given the frame-level representations of positive video and text pair, each denoted as V_f=$[v_1^f, \ldots, v_{L_v}^f]$ and Q_f=$[q_1^f, \ldots, q_n^f]$, Frame-Semantic Alignment loss is defined as:

$$\mathcal{L}^{\text{FSA}} = -\frac{1}{mn} \sum_{(\mathbf{v},\mathbf{q}) \in \mathcal{B}} \sum_{i=1}^{n} \left[\log \left(\frac{\exp(\cos(v_{k_i}^f, q_i^f))}{\sum_{j=1}^{n} \exp(\cos(v_{k_i}^f, q_j^f))} \right) + \log \left(\frac{\exp(\cos(v_{k_i}^f, q_i^f))}{\sum_{j=1}^{n} \exp(\cos(v_{k_j}^f, q_i^f))} \right) \right] \tag{11}$$

where k_i denotes the index of v^f in frame-level video representation V_f that has the highest cosine similarity with a frame-level query q_i^f. This promotes both semantic units and video features to encapsulate different semantics.

Total loss. Our frame-level framework is a plug-and-play model that can be simply applied by replacing the conventional frame branch. Therefore, we adopt the training losses from each baseline, denoted as $\mathcal{L}^{\text{base}}$. Combined with our losses, the training loss is expressed as:

$$\mathcal{L} = \mathcal{L}^{\text{base}} + \lambda_1 \mathcal{L}^{\text{SD}} + \lambda_2 \mathcal{L}^{\text{FSA}} \tag{12}$$

where λ_1 and λ_2 denote hyperparameters to balance losses. Note that $\mathcal{L}^{\text{base}}$ may include more objectives than illustrated in Sec. 3.1 depending on the baselines' objectives.

4 Experiment

4.1 Experimental Setting

Datasets. We validate our proposed method on two long untrimmed video datasets: TVR (Lei et al. 2020a) and ActivityNet Captions (Krishna et al. 2017). TVR is a challenging video dataset collected from TV shows, containing 21.8K videos and 109K human-annotated query-moment pairs. Each query in this dataset is organized with an average of 13.4 words and typically describes scenes with more than two people performing multiple actions. In contrast, ActivityNet Captions includes 15K videos and 72K text annotations, with each query averaging 14.8 words.

Evaluation Metrics. To evaluate our methods, we adopt rank-based recall (*R@K*). Specifically, we set K to 1, 5, 10, and 100 to measure performance under varying degrees of harshness. Additionally, we report SumR, the sum of all recall values, to compare the overall performance.

Implementation Details. For a fair comparison, we follow existing works to adopt pre-trained feature extractors (e.g., ResNet (He et al. 2016), I3D (Carreira and Zisserman 2017), and Roberta (Liu et al. 2019b)). To integrate our framework with baseline methods, we replaced frame-level branches for

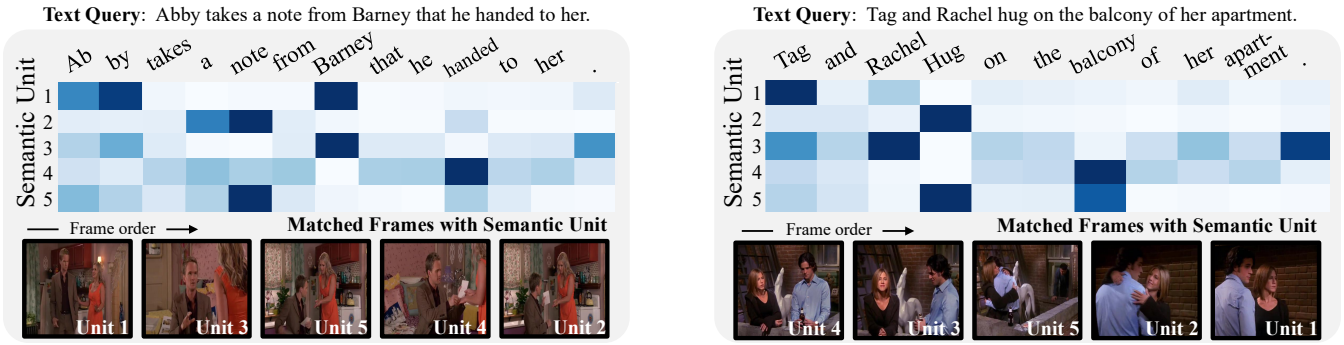

Figure 3: Visualization of word attention scores within our semantic units and their matched frames of semantic units (highest similarity). The top row shows the text query, followed by the attention scores of the semantic units for each query word. The bottom row shows the frames in the video that match each semantic unit with maximum cosine similarity.

processing both text and video modalities with our proposed methods. We set n, the number of learnable vectors (frame-level query representations), to 5 with the slot encoder performing $T = 3$ iterations. Additionally, the number of window attention layers in the ALA o is set to 3. In the appendix, we provide more implementation details.

4.2 Comparison with the State-of-the-Art

We compare our proposed framework with state-of-the-art PRVR and VCMR methods (Dong et al. 2022a, 2023; Jiang et al. 2023; Wang et al. 2024c,b; Lei et al. 2020b; Zhang et al. 2021; Hou, Ngo, and Chan 2021) (Note that VCMR methods are trained without the moment annotations). Specifically, we validate the effectiveness of our proposed framework by replacing the frame branch of previous PRVR works (i.e., MS-SL, GMMFormer, and GMMFormer v2). In Tab. 1, we observe that our framework improves the performances of baselines across all rank-based recall metrics on all datasets, demonstrating its compatibility with existing PRVR methods. Notably, our framework on top of GMMFormer v2 achieves a substantial margin of 7.3%p higher than the previous best method on the TVR dataset. We also find that our framework provides greater performance improvements for methods that lack consideration of locality, like MS-SL. Our method also benefits for AcitivityNet-Captions dataset. Overall, our framework demonstrates substantial performance improvements when applied to baselines, highlighting its compatibility across different dual encoding methods.

4.3 Ablation Studies and Analyses

Ablation study on our proposed components. In Tab. 2, we conduct an ablation study by incrementally integrating each of our proposed components into GMMFormer v2 and evaluating the rank-based performance on the TVR dataset. First, (a) presents the performance of reproduced GMMFormer v2. From (b) to (d), we demonstrate the performance improvements resulting from our enhancements to the query branch. Specifically, in (b), we initialize different query branches for each frame and clip branch, showing that these branches benefit from learning distinct details when corresponding text branches are present. Subsequently, rows (c) and (d) validate the effectiveness of semantic decomposition, achieving substantial improvements over (b). This result underscores that employing diverse semantic queries for video matching enhances retrieval performance. Furthermore, (e) illustrates that replacing the frame-level video encoding in (d) with our ALA results in additional performance gains. However, as shown in (g), we observe performance degradation when ALA is used without our proposed query branch design, SDM. This supports our hypothesis that a single query representation with diverse contexts does not align well with features that emphasize local fine detail. Lastly, (f) shows that incorporating our loss for further semantic decomposition alongside visual features enhances retrieval performance.

	Model	R@1	R@5	R@10	R@100	SumR
(a)	$*$ GMMFormer v2	15.3	36.4	48.5	86.7	186.9
(b)	(a) + query branch	16.0	37.7	49.1	86.7	189.5
(c)	(a) + SDM	17.1	38.5	50.2	87.0	192.8
(d)	(c) + $\mathcal{L}^{SD}$	16.8	38.9	50.4	87.4	193.5
(e)	(d) + ALA	17.2	39.4	51.2	87.7	195.6
(f)	(e) + $\mathcal{L}^{FSA}$ (Ours)	17.4	39.7	51.4	87.9	196.4
(g)	(b) + ALA	15.8	37	48.4	87.0	188.2

Table 2: Ablation study of proposed components on TVR dataset. The symbol $*$ indicates the reproduced performance.

Analysis of sentence decomposition. In Fig. 3, we visualize how each semantic units are formed to be matched with visual frames from the corresponding videos. As observed, we claim that each semantic unit is built upon the words that are highly likely to appear in the same frame coincidentally (e.g., "Abby" with "Barney", "a note", and "handed" in the left example.) Moreover, we find that each semantic unit is appropriately matched to the visual frame that fully describes the content within semantic units. These results demonstrate that our proposed SDM effectively breaks down the sentence into frame-level semantic units.

Analysis of the number of semantic units. Selecting the appropriate number of decomposed units requires balancing

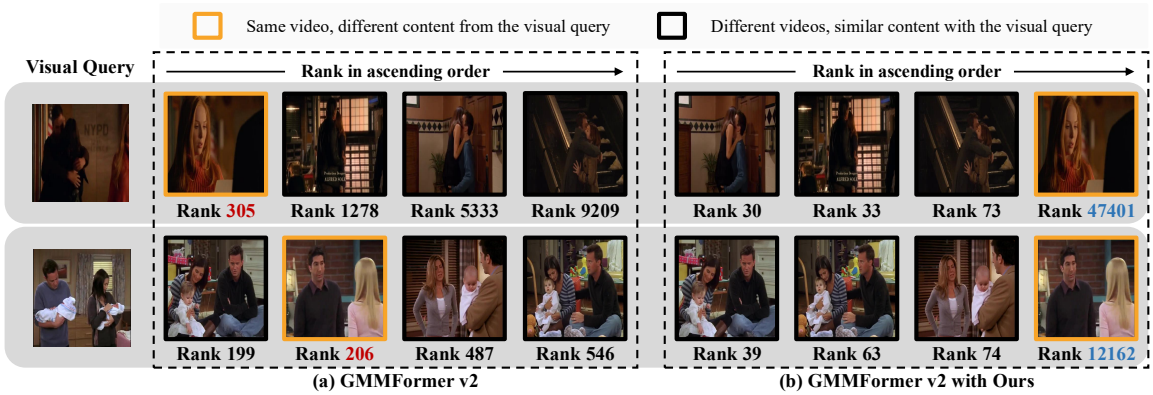

Figure 4: Qualitative results of frame retrieval using visual queries. From left to right, we show the input visual queries, retrieval results from GMMFormer v2, and our ALA-based frame branch. The ranks below each frame indicate the similarity ranking across the entire video frame features for each method. These results are based on the TVR validation set, containing 110,078 frames. Top-ranked frames are often adjacent within the same video, so comparisons for these frames are omitted.

# Units	R@1	R@5	R@10	R@100	SumR	Runtime(ms)
1	16.2	38.2	49.7	86.9	191.0	2.209
3	17.0	38.5	50.4	87.4	193.3	2.211
5	17.4	39.7	51.4	87.9	196.4	2.225
7	17.9	39.5	51.0	87.6	196.0	2.261
9	17.8	39.7	51.2	88.1	196.8	2.274

Table 3: Analysis of the number of semantic units on TVR dataset. Runtime evaluated on RTX 3090 with 2000 videos.

Method	R@1	R@5	R@10	R@100	SumR
Max pooling	16.1	37.8	49.8	86.7	190.5
Softmax weight	16.7	39.1	51.1	88.0	194.8
Average (ours)	17.4	39.7	51.4	87.9	196.4

Table 4: Comparison of scoring methods with units.

the performance with runtime efficiency. In Tab. 3, we examine the impact of varying the number of units. As shown in Tab. 3, increasing the number of units from 1 to 5 improves rank-based recall performance, albeit at the cost of increased runtime. Beyond 5 units, the performance gains become negligible. We posit that decomposing a sentence into five distinct components sufficiently captures alignment with the local features in videos. Based on this, we chose five representations at frame-level for our approach.

Assessment of local details in video features. To validate that our proposed ALA effectively encodes fine local details within each video, we compare the retrieval performance of the frame branch against GMMFormer v2 when visual queries are given as input, as shown in Fig. 4. Note that retrieval is performed using similarity matching. Our analysis reveals that GMMFormer v2 tends to produce high similarity scores between visual queries and frame features that belong to the same video, indicating entanglement within video sequences. In contrast, our method successfully disentangles frames within the same video, resulting in lower retrieval ranks for semantically relevant frames (e.g., those with similar actions or poses) from different videos. This demonstrates that ALA is a well-suited design choice for encoding frame features within video sequences.

Variations for integrating frame-level retrieval scores. In Tab. 4, we present the results of using other operations, such as max pooling and softmax, to integrate the five frame-level retrieval scores calculated with our query representations (semantic units). Here, the values used in the softmax are derived from each unit through an additional linear layer. As shown, the average operation provides the best retrieval performance, as other methods heavily rely on a single unit, thereby not considering all the details within the sentence. These results validate our decision to use the average operation to integrate frame-level retrieval scores, ensuring that all semantic units contribute equally to the retrieval score.

5 Conclusion

Typical works for PRVR learn to match both the clip-level and frame-level visual features with sentence query features. However, this leads to an ambiguous role of the frame-level branch, as sentences typically encompass a broader context than that appears in individual frames. Thus, this paper proposed a novel frame-level framework for PRVR to address the context-level mismatch. Specifically, we designed both text and video encoder designs optimized for extracting frame-level semantics. For text representation, Semantic-Decomposed Matching decomposes sentences into multiple semantic units, capturing partial content that is highly likely to appear in video frames. In parallel, we introduced Adaptive Local Aggregator to ensure that video frame features capture local fine details. This is accomplished by constructing and aggregating multi-scale local features into frame-level representations. Finally, these frame-level representations are further refined to encode varied semantics through proposed objectives. The improvements observed across various baselines and our comprehensive studies validate the effectiveness of our framework.

Acknowledgments

This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2020-0-01821, 2019-0-00421, RS-2024-00459618, RS-2024- 00360227, RS-2024-00437102, RS-2024-00437633), and MSIT/NRF (No. RS-2024-00357729).

References

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

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10638–10647.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Deng, C.; Chen, Q.; Qin, P.; Chen, D.; and Wu, Q. 2023. Prompt switch: Efficient clip adaptation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15648–15658.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially Relevant Video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

Dong, J.; Li, X.; and Snoek, C. G. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12): 3377–3388.

Dong, J.; Li, X.; Xu, C.; Ji, S.; He, Y.; Yang, G.; and Wang, X. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9346–9355.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022b. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE transactions on circuits and systems for video technology*, 32(8): 5680–5694.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11302–11312.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Fang, B.; Liu, C.; Zhou, Y.; Yang, M.; Song, Y.; Li, F.; Wang, W.; Ji, X.; Ouyang, W.; et al. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. *arXiv preprint arXiv:2301.06309*.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5006–5015.

Guan, P.; Pei, R.; Shao, B.; Liu, J.; Li, W.; Gu, J.; Xu, H.; Xu, S.; Yan, Y.; and Lam, E. Y. 2023. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11164–11173.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, Z.; Ngo, C.-W.; and Chan, W. K. 2021. CONQUER: Contextual query-aware ranking for video corpus moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3900–3908.

Huang, S.; Gong, B.; Pan, Y.; Jiang, J.; Lv, Y.; Li, Y.; and Wang, D. 2023. VoP: Text-Video Co-operative Prompt Tuning for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6565–6574.

Jiang, X.; Chen, Z.; Xu, X.; Shen, F.; Cao, Z.; and Cai, X. 2023. Progressive Event Alignment Network for Partial Relevant Video Retrieval. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1973–1978. IEEE.

Jin, P.; Huang, J.; Xiong, P.; Tian, S.; Liu, C.; Ji, X.; Yuan, L.; and Chen, J. 2023a. 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*, 2472–2482.

Jin, P.; Li, H.; Cheng, Z.; Huang, J.; Wang, Z.; Yuan, L.; Liu, C.; and Chen, J. 2023b. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 938–946.

Jin, P.; Li, H.; Cheng, Z.; Li, K.; Ji, X.; Liu, C.; Yuan, L.; and Chen, J. 2023c. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020a. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 447–463. Springer.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020b. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 447–463. Springer.

Li, H.; Song, J.; Gao, L.; Zhu, X.; and Shen, H. 2024. Prototype-based Aleatoric Uncertainty Quantification for

Cross-modal Retrieval. *Advances in Neural Information Processing Systems*, 36.

Li, P.; Xie, C.-W.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023a. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4100–4110.

Li, P.; Xie, C.-W.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4100–4110.

Li, X.; Xu, C.; Yang, G.; Chen, Z.; and Dong, J. 2019. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM international conference on multimedia*, 1786–1794.

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

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, 319–335. Springer.

Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33: 11525–11538.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, W.; Chen, Q.; Zhou, T.; Zhao, S.; and Cai, Z. 2023. Using Multimodal Contrastive Knowledge Distillation for Video-Text Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 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*, 9879–9889.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2630–2640.

Pei, R.; Liu, J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu, J.; and Yan, Y. 2023. CLIPPING: Distilling CLIP-Based Models with a Student Base for Video-Language Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18983–18992.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Tian, K.; Zhao, R.; Xin, Z.; Lan, B.; and Li, X. 2024. Holistic Features are almost Sufficient for Text-to-Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17138–17147.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Sun, G.; Wang, P.; Liu, D.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 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*, 16551–16560.

Wang, Y.; Wang, J.; Chen, B.; Dai, T.; Luo, R.; and Xia, S.-T. 2024b. GMMFormer v2: An Uncertainty-aware Framework for Partially Relevant Video Retrieval. *arXiv preprint arXiv:2405.13824*.

Wang, Y.; Wang, J.; Chen, B.; Zeng, Z.; and Xia, S.-T. 2024c. GMMFormer: Gaussian-Mixture-Model Based Transformer for Efficient Partially Relevant Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5767–5775.

Wang, Z.; Sung, Y.-L.; Cheng, F.; Bertasius, G.; and Bansal, M. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2816–2827.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 685–695.