Denoise-then-Retrieve: Text-Conditioned Video Denoising for Video Moment Retrieval

Weijia Liu[1], **Jiuxin Cao**[1], **Bo Miao**[2], **Zhiheng Fu**[3], **Xuelin Zhu**[3], **Jiawei Ge**[1], **Bo Liu**[1], **Mehwish Nasim**[4] and **Ajmal Mian**[4]
[1]Southeast University
[2]The University of Adelaide
[3]The Hong Kong Polytechnic University
[4]The University of Western Australia
{weijia-liu, jx.cao}@seu.edu.cn

Abstract

Current text-driven Video Moment Retrieval (VMR) methods encode all video clips, including irrelevant ones, disrupting multimodal alignment and hindering optimization. To this end, we propose a denoise-then-retrieve paradigm that explicitly filters text-irrelevant clips from videos and then retrieves the target moment using purified multimodal representations. Following this paradigm, we introduce the Denoise-then-Retrieve Network (DRNet), comprising Text-Conditioned Denoising (TCD) and Text-Reconstruction Feedback (TRF) modules. TCD integrates cross-attention and structured state space blocks to dynamically identify noisy clips and produce a noise mask to purify multimodal video representations. TRF further distills a single query embedding from purified video representations and aligns it with the text embedding, serving as auxiliary supervision for denoising during training. Finally, we perform conditional retrieval using text embeddings on purified video representations for accurate VMR. Experiments on Charades-STA and QVHighlights demonstrate that our approach surpasses state-of-the-art methods on all metrics. Furthermore, our denoise-then-retrieve paradigm is adaptable and can be seamlessly integrated into advanced VMR models to boost performance.

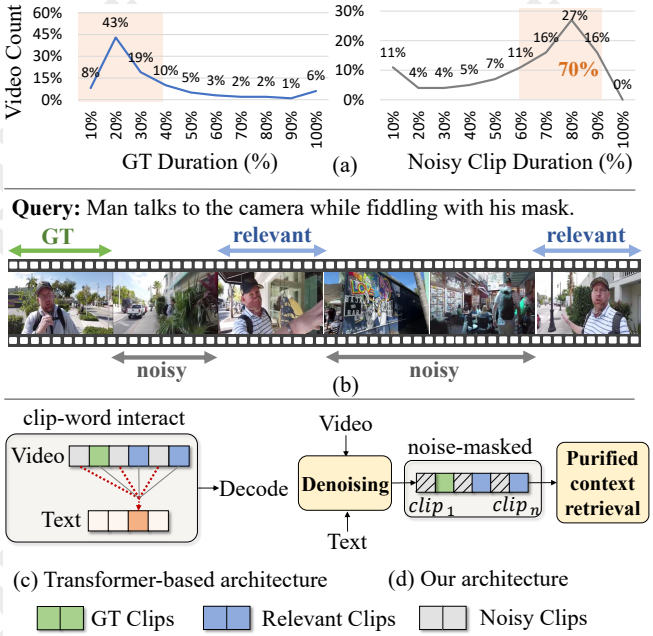

Figure 1: (a) Distribution of Ground Truth Moments (left) and noisy clips (right) relative to video duration in QVHighlights. (b) Our Text-Conditioned Denoising effectively identifies noisy and relevant video clips. (c) Previous transformer-based VMR methods use all video clips for multimodal encoding, including irrelevant ones. (d) Our DRNet approach explicitly excludes noisy video clips, enhancing purified multimodal modeling.

1 Introduction

Text-driven Video Moment Retrieval (VMR) aims to localize moments in untrimmed videos that semantically match the given query text. Unlike conventional temporal action localization constrained by predefined action categories, VMR flexibly localizes the moment through free-form linguistic expressions. VMR streamlines video analysis and benefits downstream applications [Miao *et al.*, 2024b; Miao *et al.*, 2024a], including video semantic segmentation, user-friendly video editing, and mass surveillance.

Traditional two-stage VMR methods [Xiao *et al.*, 2021; Zhang *et al.*, 2020; Xu *et al.*, 2019] aim to extract a set of moment proposals using proposal generation networks, treating the task as a matching and ranking problem between the proposals and the query text. However, to achieve high recall, they generate numerous proposals of varying durations and locations, reducing efficiency and complicating matching

In recent years, Transformer architectures have become prevalent in multimodal understanding tasks [Miao *et al.*, 2023; Miao *et al.*, 2024c], including VMR [Lin *et al.*, 2023; Liu *et al.*, 2022; Moon *et al.*, 2023b; Xu *et al.*, 2023; Liu *et al.*, 2024a], due to their strong feature interaction and representation capabilities. These methods utilize transformer encoders to perform fine-grained word-clip interac-

tion and integrate multimodal representations for moment retrieval. However, they treat all video clips equally, inevitably introducing semantically irrelevant noisy clips into the multimodal representations (see Fig. 1 (c)) which eventually leads to suboptimal performance.

In VMR, input videos include both text-relevant and text-irrelevant clips. Relevant clips are frames semantically aligned with the text query, including ground-truth frames (perfect alignment) and challenging non-target frames (partial alignment). Irrelevant (noise) clips are frames with little or no alignment. For example, in Fig. 1 (b), the clips where the "man talks to the camera" are text-relevant, which can provide useful context for prediction. In contrast, the scenery and walking clips are noise. We argue that relevant clips typically occupy only a small portion of a video, while noisy clips dominate. To validate this, we analyze all videos in the QVHighlights dataset. As shown in Fig. 1 (a), ground truth (GT) clips occupy less than 30% of the duration in most videos, while noisy clips account for over 60% clips in most videos. With excessive noisy clips throughout the video, generating abundant proposals with noisy clips or using all clips for transformer-based multimodal interaction can hinder the VMR model from focusing on the text-relevant clips that are more likely to be the target.

To address this issue, we propose a denoise-then-retrieve paradigm that explicitly removes noisy clips to narrow the retrieval range and strengthens purified multimodal representations for moment retrieval, as shown in Fig. 1 (d). Specifically, we design a Text-Conditioned Denoising (TCD) module to filter out noisy clips by dynamically generating noise masks. It integrates cross-attention and structured state space models for text-video interaction, and generates dynamic kernels to produce noise masks for purified multimodal representations. To provide direct feedback on denoising quality, we introduce a Text-Reconstruction Feedback (TRF) module, which aligns the generated query from purified video features with the input text, serving as auxiliary supervision for TCD during training. Finally, the decoder performs purified multimodal interaction between noise-masked video features and text embeddings, enabling accurate retrieval from text-relevant clips. Additionally, when applied to other methods, our denoise-then-retrieve paradigm leads to notable performance improvements, showcasing the generalization capability. For example, UniVTG [Lin *et al.*, 2023] achieves an increase of 2.75% points on the mAP@Avg metric. Our contributions are summarized as follows:

- We propose the Denoise-then-Retrieve Network (DRNet) with Text-conditioned Denoising and Text-reconstruction Feedback. Our DRNet effectively extracts purified visual representations to enhance text-clip alignment, achieving top-tier performance on popular benchmarks.
- We propose a text-conditioned denoising approach that integrates cross-attention and structured state space blocks for effective multi-level multimodal fusion, generating dynamic kernels for accurate noise identification.
- We introduce a text-reconstruction feedback mechanism that aligns the generated query from purified video features with the input text, providing auxiliary supervision for denoising during training.
- We demonstrate that the denoise-then-retrieve paradigm integrates seamlessly into current VMR models, yielding significant improvements across all metrics.

Experiments on the Charades-STA and QVHighlights benchmarks show that our approach significantly outperforms existing state-of-the-art methods on all metrics. On Charades-STA, we surpass the nearest competitor MESM [Liu *et al.*, 2024b] by 4.36% points on the mAP@0.7 metric.

2 Related Work

Two-stage VMR Methods [Zhang *et al.*, 2021; Wang *et al.*, 2021; Chen *et al.*, 2020a; Qu *et al.*, 2020; Yuan *et al.*, 2019] extract a set of moment proposals through multi-scale sliding windows or proposal-generating networks and treat the task as a matching and ranking problem between proposal candidates and the text query. However, sliding window-based methods [Liu *et al.*, 2018; Jiang *et al.*, 2019; Ge *et al.*, 2019] suffer from inefficient computation due to the re-computation of many overlapping areas in the densely sampled process with predefined multi-scale sliding windows. To reduce the number of candidates, proposal-based methods [Xu *et al.*, 2019; Chen and Jiang, 2019; Xiao *et al.*, 2021] devise various proposal-generating networks. For instance, QSPN [Xu *et al.*, 2019] generates proposals by introducing query representations as guidance for video encoding, while SAP [Chen and Jiang, 2019] pre-trains a visual concept detection CNN with paired query-clip training data to calculate the visual-semantic correlation score for clips, grouping high-scoring clips to form proposals. Differently, BPNet [Xiao *et al.*, 2021] directly utilizes VSLNet [Zhang *et al.*, 2020] to generate moment proposals. To ensure high recall, these methods generate numerous proposals of varying durations and locations, which reduces efficiency and complicates matching.

Transformer-based VMR Approaches [Lin *et al.*, 2023; Liu *et al.*, 2022; Moon *et al.*, 2023b; Lei *et al.*, 2021; Xu *et al.*, 2023; Moon *et al.*, 2023a; Yang *et al.*, 2024] use Transformer encoders to perform word-clip level interactions between text and videos to establish a shared embedding space and regress the temporal span based on the aligned visual-text features. [Lei *et al.*, 2021] introduces detection Transformer (DETR) into the VMR task and models the task as a temporal moment detection problem. To fully exploit the information of a given query, [Moon *et al.*, 2023b] uses cross-attention layers in the encoding stage to explicitly inject the context of text into video representation. Unlike works that enforce text engagement in each clip, CG-DETR [Moon *et al.*, 2023a] carefully controls the degree of query text engagement in cross-modal interaction to enhance multimodal representations. However, text-referred moments occupy only a small portion of videos while noisy clips occupy a significant part and are spread throughout the video. As a result, treating all video clips equally in multimodal modeling can lead to suboptimal multimodal representations. In this work, we explicitly remove noisy clips to narrow localization range

and perform masked context aggregation and decoding to enhance retrieval.

3 Method

Given an untrimmed video V containing L_v clips and a text query T with L_t words, our objective is to localize the target moment as (m_c, m_σ), where m_c and m_σ denote the central temporal coordinate and the span of the moment, respectively. Fig. 2 illustrates the overview of our DRNet and its modules.

Our overall architecture is described in Fig. 2. Given a video and text representation extracted from fixed backbones, DRNet first identifies noisy clips and generates purified video representations by masking them. To further enhance the denoising, we regenerate query embeddings from purified video representations and aligns it with the input text, providing auxiliary supervision for denoising process during training. Finally, the decoder performs multimodal interaction on the purified representations for accurate retrieval.

Video and Text Encoders. We represent video features using the concatenation of CLIP [Radford *et al.*, 2021] and SlowFast [Feichtenhofer *et al.*, 2019], and extract text features using the CLIP text encoder, consistent with previous works [Lei *et al.*, 2021; Liu *et al.*, 2022; Moon *et al.*, 2023b]. The input video V and text T are encoded with frozen encoders and projected to the same dimension D via two Feed-Forward Networks (FFN). The resulting video embeddings $\mathbf{V} = [v_1, v_2, ..., v_{L_v}]$ and text embeddings $\mathbf{T} = [t_1, t_2, ..., t_{L_t}]$ are then fed into the TCD module.

3.1 Text-Conditioned Denoising (TCD)

As shown in Fig. 2 (b), to dynamically identify noisy clips, we leverage the query text to guide video representation learning, multimodal interaction, and noise mask generation. Specifically, we first inject textual context into video clips via cross-attention, obtaining text-aware video representations. These representations, along with text features, are then processed by state space models to propagate intra- and inter-modal context, generating refined multimodal representations. Finally, text-driven dynamic convolution kernels are constructed to identify and filter out noisy clips.

We first apply cross-attention between video clips and text embeddings to enhance the target awareness. Here, we use italicized letters to denote the *query*, *key*, and *value* in the cross-attention layer. Specifically, the video representations $\mathbf{V}$ serve as the *query*, while the textual features $\mathbf{T}$ are used as the *key* and *value*:

$$\mathrm{Attn}(Q_{\mathbf{V}}, K_{\mathbf{T}}, V_{\mathbf{T}}) = \mathrm{softmax}(\frac{Q_{\mathbf{V}}(K_{\mathbf{T}})^T}{\sqrt{D}})V_{\mathbf{T}} \quad (1)$$

The updated video representations are obtained by computing a weighted sum of the text features $\mathbf{T}$, where the attention scores are projected through a Multi-layer Perceptron (MLP) and integrated into the original video representations, resulting in text-aware video representations $\hat{\mathbf{V}} = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_{L_v}]$.

For text-clip interaction and context integration, we design the Context Interaction Operator (CIO) based on state space models, which have demonstrated significant success in visual understanding tasks[Zhu *et al.*, 2024; Chen *et al.*, 2024], including Mamba [Gu and Dao, 2023]. The gating mechanisms and linear complexity of these models facilitate efficient modeling of long sequences. Specifically, each CIO consists of two separate Mamba blocks [Gu and Dao, 2023] that propagate the context of feature sequences in both forward and backward temporal directions.

We concatenate the text-aware video representations $\hat{\mathbf{V}}$ with the text features to form a multimodal sequence, and add learnable global tokens at the end of the sequence to aggregate global features. These global tokens are denoted as $\mathbf{G} = [g_1, g_2, ..., g_{L_g}]$, where L_g denotes the number of global tokens. To preserve positional and modality-specific information during cross-modal interaction, learnable position embeddings $\mathbf{E}^p$ and modality-type embeddings $\mathbf{E}^m$ are incorporated into each modality. The input multimodal sequence $\mathbf{F}$ to the CIOs is then represented as

$$\tilde{\mathbf{V}} = \hat{\mathbf{V}} + \mathbf{E}^p_V + \mathbf{E}^m_V, \quad (2)$$

$$\tilde{\mathbf{T}} = \mathbf{T} + \mathbf{E}^p_T + \mathbf{E}^m_T, \quad (3)$$

$$\tilde{\mathbf{G}} = \mathbf{G} + \mathbf{E}^p_G + \mathbf{E}^m_G, \quad (4)$$

$$\mathbf{F} = [\tilde{\mathbf{T}}, \tilde{\mathbf{V}}, \tilde{\mathbf{G}}] \quad (5)$$

where $\mathbf{F} \in \mathbb{R}^{(L_g+L_t+L_v)\times D}$. After encoding with the bi-directional Mambas in the CIO, the output of the backward Mamba block is flipped and added to the corresponding output of the forward Mamba block along the temporal channel, completing one round of context interaction. The multimodal sequence $\mathbf{F}$ undergoes intra- and inter-modal context integration by stacking three CIOs.

The interactions occur at three levels: 1) Intra-modal contextual interaction: Tokens within each modality learn contextual semantics from surrounding tokens through forward and backward Mamba propagation. 2) Cross-modal contextual interaction: The forward Mamba block propagates textual semantics into visual features, while the backward Mamba block integrates visual semantics into textual features. 3) Global context integration: Through information propagation of bi-directional Mambas, the global tokens at the end of the sequence integrate semantics from both text and video features. By stacking multiple CIOs, we generate a context-aware multimodal sequence, where $\boldsymbol{f_t}$, $\boldsymbol{f_v}$, and $\boldsymbol{f_g}$ represent text, video, and global features, respectively.

Dynamic Denoising. To purify the visual context, we employ text features to create dynamic kernels [Chen *et al.*, 2020b], which performs point-wise convolutions to identify noisy clips. The word-level text feature $\boldsymbol{f_t}$ is first pooled to a fixed length of L_k to handle varying text lengths, followed by a fully-connected layer to generate dynamic kernels $\Theta = \{\theta_i\}_{i=1}^{N_k}$, where N_k is the number of kernels, and $\theta_i \in \mathbb{R}^{D\times 1}$. These dynamic kernels are then applied to $\boldsymbol{f_v}$ to update the visual features:

$$\boldsymbol{f_v}' = (\varphi(\theta_1 \boldsymbol{f_v} \oplus \cdots \oplus \theta_{N_k} \boldsymbol{f_v}) + \boldsymbol{f_v}) \quad (6)$$

where $\oplus$ denotes concatenation along the channel dimension, and $\varphi(\cdot)$ represents a 1×1 convolution for dimensionality reduction.

After applying the Sigmoid function σ, we obtain the text-clip alignment scores $\mathbf{S} = [s_1, s_2, \ldots, s_n]$, which indicate

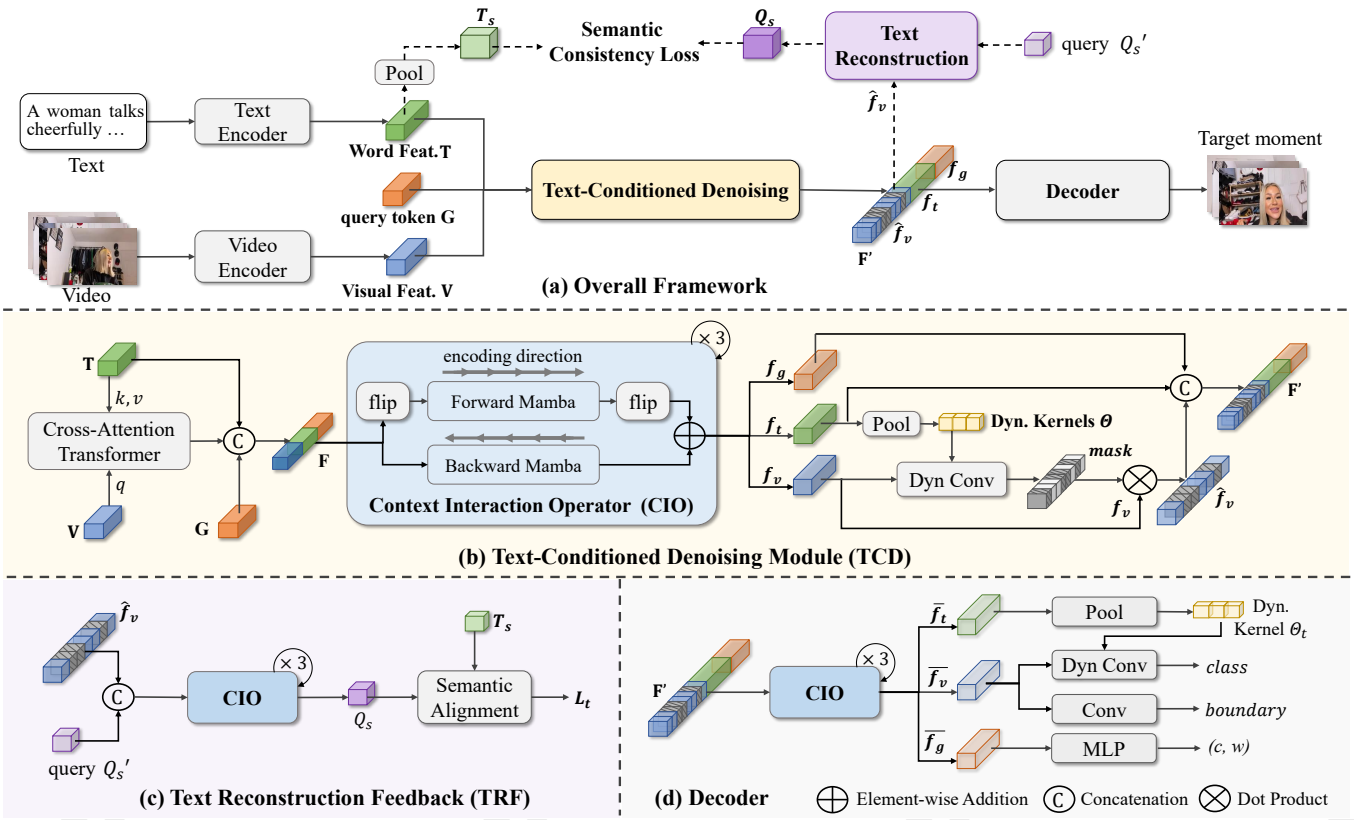

Figure 2: Overview of our DRNet. Dashed lines represent components used only during training. TCD identifies noisy clips and generates purified video representations by masking them. TRF provides feedback on TCD's denoising quality using regenerated query embeddings from purified representations. The decoder performs multimodal interaction on the purified representations for accurate retrieval.

the degree of semantic relevance to the text. By applying a threshold μ, we generate the noise mask vector $\mathbf{M} = [m_1, m_2, \ldots, m_n]$,

$$\mathbf{S} = \sigma(\boldsymbol{f_v}') \tag{7}$$

$$m_i = \begin{cases} 1 & \text{if } s_i > \mu \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, \ldots, n \tag{8}$$

where $m_i = 0$ means that the i-th clip is semantically irrelevant to the text (a noisy clip), and $m_i = 1$ means relevant. Finally, $\mathbf{M}$ is applied to mask out noisy clips and produce purified visual representations $\hat{\boldsymbol{f}}_{\boldsymbol{v}}$. Fig. 3 visualizes the text-clip alignment scores $\mathbf{S}$ and the noise mask $\mathbf{M}$ for a given case, demonstrating that our TCD module effectively identifies relevant clips and filters out noise. The updated textual, visual, and global features are then concatenated as $\mathbf{F}' = [\boldsymbol{f_t}, \hat{\boldsymbol{f}}_{\boldsymbol{v}}, \boldsymbol{f_g}]$ and fed into the decoder.

3.2 Text-Reconstruction Feedback (TRF)

TRF further enhances denoising by regenerating query embeddings from the purified visual representations. It aligns the regenerated query with the input text in textual semantic space, providing auxiliary supervision for denoising process during training. This feedback not only improves denoising quality but also strengthens the purified video representations.

Specifically, as shown in Fig. 2 (c), we distill a query embedding from the purified video representations $\hat{\boldsymbol{f}}_{\boldsymbol{v}}$ using context interaction operators (CIO) (see Section 3.1), projecting $\hat{\boldsymbol{f}}_{\boldsymbol{v}}$ into the text semantic space. To achieve this, we introduce a learnable query embedding $\mathbf{Q}'_{\mathbf{s}} \in \mathbb{R}^{D\times 1}$ and construct a mapping network $\mathcal{D}$ by stacking three CIOs. The query embedding $\mathbf{Q}'_{\mathbf{s}}$ interacts with the purified video representations $\hat{\boldsymbol{f}}_{\boldsymbol{v}}$ to generate the reconstructed sentence-level embedding:

$$\hat{\mathbf{Q}}_{\mathbf{s}} = \mathcal{D}(\hat{\boldsymbol{f}}_{\boldsymbol{v}}, \mathbf{Q}'_{\mathbf{s}}) \tag{9}$$

where $\hat{\mathbf{Q}}_{\mathbf{s}} \in \mathbb{R}^{D\times 1}$. The sentence-level embedding of the input text, $\mathbf{T_s}$, is obtained by average pooling its word-level embeddings, $\mathbf{T}$. We then compute the semantic consistency loss $\mathcal{L}_t$ between $\mathbf{T_s}$ and $\hat{\mathbf{Q}}_{\mathbf{s}}$ using cosine similarity:

$$\mathcal{L}_t = \lambda_t\left(1 - \frac{\mathbf{T_s} \cdot \hat{\mathbf{Q}}_{\mathbf{s}}}{||\mathbf{T_s}||\,||\hat{\mathbf{Q}}_{\mathbf{s}}||}\right), \tag{10}$$

where $\lambda_t (\lambda_t = 2)$ is a hyperparameter balancing the loss terms. Since the regenerated query embeddings from text-relevant clips capture key visual semantics, minimizing $\mathcal{L}_t$ forces the purified visual representations to maximally reflect the input text semantics, enhancing the denoising process.

3.3 Decoder

After masking noisy clips, we perform multimodal interaction on the purified visual and textual features to capture fine-

grained differences between text-relevant clips. The resulting purified multimodal representations are then decoded for accurate retrieval.

As shown in Fig. 2 (d), we use context interaction operators (CIO) (see Section 3.1) to build a multimodal encoder $\mathcal{E}$ with three layers. The encoder takes the noise-masked multimodal sequence $\mathbf{F}'$ from the TCD module as input. Through cross-modal context encoding, the purified multimodal representation $\mathbf{F} = \mathcal{E}(\mathbf{F}')$ is generated and passed to different decoding heads for target moment retrieval. We denote the global, textual, and visual features within the $\mathbf{F}$ sequence as $\bar{\boldsymbol{f}}_{\boldsymbol{g}}$, $\bar{\boldsymbol{f}}_{\boldsymbol{t}}$, and $\bar{\boldsymbol{f}}_{\boldsymbol{v}}$, respectively.

Moment retrieval. 1) Global retrieval. We use the global features $\bar{\boldsymbol{f}}_{\boldsymbol{g}}$ to directly regress the central temporal coordinate m_c via an MLP, and regress the moment span m_σ via a fully connected layer. The global localization loss combines an L1 loss and a generalized IoU loss $\mathcal{L}_{gIoU}(\cdot)$ following [Moon *et al.*, 2023b].

$$\mathcal{L}_g = \lambda^g_{L1}||m - \hat{m}|| + \lambda^g_{iou}\mathcal{L}_{iou}(m, \hat{m}), \tag{11}$$

where m is the ground-truth moment and $\hat{m}$ is the corresponding prediction, each containing the center coordinate and span.

2) Boundary prediction. Following [Lin *et al.*, 2023], we apply three 1×3 Conv layers with N_k filters and ReLU activation to the output $\bar{\boldsymbol{f}}_{\boldsymbol{v}} \in \mathbb{R}^{L_v \times D}$ from the multimodal encoder $\mathcal{E}$. The final layer has two output channels, representing the left and right offsets $\hat{d}_i \in \mathbb{R}^{2\times L_v}$ for each clip. The predicted boundaries $\hat{b}_i$ are then calculated, and the boundary loss $\mathcal{L}_b$, which includes smooth L1 and IoU losses, is used to supervise the predictions.

$$\mathcal{L}_b = \lambda^b_{L1}\mathcal{L}_{SmoothL1}(\hat{d}_i, d_i) + \lambda^b_{iou}\mathcal{L}_{iou}(\hat{b_i}, b_i). \tag{12}$$

3) Text-conditioned Foreground Classification. As illustrated in Fig. 1 (d), we pool the text features $\bar{\boldsymbol{f}}_{\boldsymbol{t}}$ to a fixed length N_k and apply a fully connected layer to generate convolution kernels $\Theta_t \in \mathbb{R}^{D\times N_k}$. The decoder head for clip classification (foreground/background) is the same as the boundary prediction head. However, in this case, we use Θ_t as the convolution kernels in the first layer, and the final layer outputs a single channel to classify each clip $\hat{c}_i$. Binary cross-entropy loss is applied for classification.

$$\mathcal{L}_c = -\lambda_c(c_i log\hat{c_i} + (1 - c_i)log(1 - \hat{c_i})). \tag{13}$$

4) Contrastive Learning. Following prior works [Lei *et al.*, 2021; Lin *et al.*, 2023; Moon *et al.*, 2023b], we incorporate intra-video and inter-video contrastive learning during training. Intra-video contrastive learning treats clips within the ground truth moment as positive pairs and those outside as negative pairs, while inter-video contrastive learning uses text from other samples in the batch as negative pairs. The relevance between a clip embedding and a sentence-level text embedding $\mathbf{Q_s}$ is quantified by cosine similarity r_i.

$$\mathcal{L}^{intra}_r = -log\frac{\exp(r_p/\tau)}{\exp(r_p/\tau) + \sum_{j\in\Omega}\exp(r_j/\tau)} \tag{14}$$

$$\mathcal{L}^{inter}_r = -log\frac{\exp(r_p/\tau)}{\sum_{k\in\Omega'}\exp(r_k/\tau)}, \tag{15}$$

Model	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	@Avg.
$\text{MDETR}_{\text{21neurips}}$	52.89	33.02	54.82	29.4	30.73
$\text{UMT}_{\text{22cvpr}}$	56.23	41.18	53.38	37.01	36.12
$\text{MomentDiff}_{\text{24neurips}}$	57.42	39.66	54.02	35.73	35.95
$\text{UniVTG}_{\text{23iccv}}$	58.86	40.86	57.60	35.59	35.47
$\text{QD-DETR}_{\text{23cvpr}}$	62.4	44.98	62.52	39.88	39.86
$\text{MESM}_{\text{24aaai}}$	62.78	45.2	62.64	41.45	40.68
$\text{UVCOM}_{\text{24cvpr}}$	63.55	47.47	63.37	42.67	43.18
$\text{TR-DETR}_{\text{24aaai}}$	64.66	48.96	63.98	43.73	42.62
Our Model	**66.73**	**50.52**	**64.17**	**45.79**	**43.73**

Table 1: Comparison on the QVHighlights *test* split obtained from the official server. All methods use only video (no audio) data, with Slowfast and CLIP as the visual backbones for fair comparison.

where Ω and Ω' are negative sets, r_p is the relevance score of positive samples, and τ is a temperature parameter. The overall contrastive learning loss is $\mathcal{L}_r = \lambda_{intra}\mathcal{L}^{intra}_r + \lambda_{inter}\mathcal{L}^{inter}_r$. Finally, after combining the textual reconstruction loss $\mathcal{L}_t$, our total training objective becomes:

$$\mathcal{L} = \frac{1}{L_v}\sum_{i=1}^{L_v}(\mathcal{L}_r + \mathcal{L}_b + \mathcal{L}_c) + \mathcal{L}_g + \mathcal{L}_t. \tag{16}$$

4 Experiments

Datasets. We validate the effectiveness of our method through extensive experiments on two popular datasets: QVHighlights and Charades-STA. **QVHighlights** [Lei *et al.*, 2021] is designed for moment retrieval and highlight detection, comprising over 10,000 videos, each averaging 150 seconds. The dataset includes 10,310 human-written text queries describing relevant segments, with an average segment length of 24.6 seconds, resulting in 18,367 annotated moments. We follow the original data splits, using the training set for model training and the test set for evaluation. **Charades-STA** [Sigurdsson *et al.*, 2016] is focused on temporal sentence grounding, derived from the Charades dataset. It contains 12,408 training and 3,720 testing moment-sentence pairs, with videos averaging 29.8 seconds in length, capturing various human actions and corresponding text queries.

Evaluation Metrics. Following previous VMR work [Li *et al.*, 2023], we use the standard evaluation metric R@n, IoU=m. This metric measures the percentage of queries that have at least one correctly retrieved moment (IoU >m) among the top-n output moments. For QVHighlights, we follow standard metrics [Lei *et al.*, 2020], using Recall@1 with IoU thresholds 0.5 and 0.7, mean average precision (mAP) with IoU thresholds 0.5 and 0.75, and the average mAP over a series of IoU thresholds [0.5:0.05:0.95] for moment retrieval. For Charades-STA, we follow [Lin *et al.*, 2023] and use Recall@1 with IoU thresholds 0.3, 0.5, and 0.7, and mIoU.

Implementation Details. For a fair comparison, we use pre-extracted SlowFast and CLIP video features, and CLIP text features, for both datasets, provided by [Lin *et al.*, 2023]. In our DRNet, all encoders constructed using CIO consist of three CIO layers, each with a hidden size of $D = 1024$. Loss

Method	feat.	R1@0.5	R1@0.7	mAP@0.5	mAP@0.7	mAP@Avg	Method	feat.	R1@0.5	R1@0.7	mIoU
$\text{RaNet}_{21\text{arxiv}}\star$	VGG	42.91	25.82	53.28	24.41	28.55	$\text{MDETR}_{21\text{neurips}}$	SF+C	52.07	30.59	45.54
$\text{MomentDETR}_{21\text{neurips}}\star$	VGG	50.54	28.01	57.39	25.62	29.87	$\text{QD-DETR}_{23\text{cvpr}}$	SF+C	57.31	32.55	-
$\text{UMT}_{22\text{cvpr}}\ddagger$	VGG	48.44	29.76	58.03	27.46	30.37	$\text{VMS}_{24\text{arxiv}}$	SF+C	57.18	36.05	-
$\text{MMN}_{22\text{aaai}}\star$	VGG	46.93	27.07	58.85	28.16	31.58	$\text{UniVTG}_{23\text{iccv}}$	SF+C	58.01	35.65	50.1
$\text{QD-DETR}_{23\text{cvpr}}\star$	VGG	51.51	32.69	62.88	32.6	34.46	$\text{TR-DETR}_{24\text{aaai}}$	SF+C	57.61	33.52	-
$\text{MomentDiff}_{24\text{neurips}}$	VGG	51.94	28.25	59.86	29.11	31.66	$\text{LLMEPET}_{24\text{arxiv}}$	SF+C	-	36.49	50.25
$\text{MESM}_{24\text{aaai}}$	VGG	56.69	35.99	67.94	33.64	37.33	$\text{UVCOM}_{24\text{cvpr}}$	SF+C	59.25	36.63	-
DRNet	VGG	**59.03**	**36.26**	**69.75**	**38**	**39.33**	**DRNet**	SF+C	**60.86**	**39.78**	**52.07**

Table 2: Comparison on Charades-STA *test* split. ‡: methods that use additional audio data. ⋆: results re-implemented under the same training strategies as [Li *et al.*, 2024; Liu *et al.*, 2024b]. SF+C: SlowFast and CLIP features.

	TCD			TRF	Decoder	R1		mAP		
	CA	DK	LGT			@0.5	@0.7	@0.5	@0.75	Avg.
A1				✓	✓	56.45	39.61	56.55	35.94	34.89
A2	✓	✓	✓		✓	66.84	51.87	65.23	46.79	45.18
A3	✓	✓	✓	✓		67.21	52.63	64.87	46.52	44.82
A4	✓	✓	✓			66.84	51.23	65.1	46.19	44.51
A5	Mamba → Transformer					64.58	47.74	61.67	42.45	40.07
B1		✓	✓	✓	✓	63.29	49.68	61.48	43.76	42.62
B2	✓		✓	✓	✓	67.81	52.97	64.51	46.6	44.98
B3	✓	✓		✓	✓	67.16	53.16	64.23	46.78	45.34
Full model	✓	✓	✓	✓	✓	**68.06**	**54.58**	**65.2**	**48.02**	**46.11**

Table 3: Ablation study on QVHighlights *val* split. A1-A4 analyze the modules of DRNet, while B1-B3 analyze the components in TCD. CA, DK, and LGT denote cross-attention, dynamic kernels and learnable global tokens, respectively. A5: replaces Mamba within CIOs with standard Transformer encoder. B2: replaces text-conditioned dynamic convolutions with standard convolutions. B3: removes learnable global tokens used for global retrieval.

weights are set as: $\lambda_t = 2$, $\lambda^g_{L1} = 5$, $\lambda^g_{iou} = 1$, $\lambda^b_{L1} = 10$, $\lambda^b_{iou} = 1$, and $\lambda_c = 10$ for both datasets. For QVHighlights, λ_{intra} and λ_{inter} are set to 2 each, while for Charades-STA, they are set to 1 and 0.5, respectively. All experiments are conducted on a single RTX 3090 GPU.

4.1 Comparison to State-of-the-art Methods

We compare our method to many state-of-the-art methods: LLMEPET [Jiang *et al.*, 2024], MomentDiff [Li *et al.*, 2024], MESM [Liu *et al.*, 2024b], UVCOM [Xiao *et al.*, 2024], LMR [Liu *et al.*, 2024a], TR-DETR [Sun *et al.*, 2024], VMS [Chen *et al.*, 2024], UniVTG [Lin *et al.*, 2023], QD-DETR [Moon *et al.*, 2023b], UMT [Liu *et al.*, 2022], RaNet [Gao *et al.*, 2021], MomentDETR [Lei *et al.*, 2021],MDETR [Lei *et al.*, 2021].

QVHighlights. Table 1 compares our method with state-of-the-art (SOTA) approaches. Our method sets new SOTA benchmarks, demonstrating significant improvements on all metrics. Specifically, it outperforms the latest 2024 methods by 9.63% over $\text{MomentDiff}_{24\text{neurips}}$, 3.64% over $\text{MESM}_{24\text{aaai}}$, 2.14% over $\text{UVCOM}_{24\text{cvpr}}$, and 1.4% over $\text{TR-DETR}_{24\text{aaai}}$ on the average of all metrics. Notably, $\text{MomentDiff}_{24\text{neurips}}$ and $\text{UVCOM}_{24\text{cvpr}}$ use diffusion-based generative and general Transformer-based architectures, respectively, while $\text{MESM}_{24\text{aaai}}$ and $\text{TR-DETR}_{24\text{aaai}}$ employ DETR-based Transformer architectures. These substantial performance gains across various VMR architectures underscore the effectiveness and superiority of our method.

Charades-STA. In Table 2, we evaluate our model's performance against the SOTA approaches using both VGG and SF+C backbones. Our method consistently achieves top-tier performance across all metrics and backbones. With the VGG backbone, our method outperforms $\text{MomentDiff}_{24\text{neurips}}$ and $\text{MESM}_{24\text{aaai}}$ by an average of 8.31% and 2.16%, respectively. For the SF+C backbone, our method surpasses the latest SOTA models by 4.76% compared to $\text{TR-DETR}_{24\text{aaai}}$, 2.56% over $\text{LLMEPET}_{24\text{arxiv}}$, and 2.38% over $\text{UVCOM}_{24\text{cvpr}}$. In particular, for the challenging mAP@0.7 and R1@0.7 metrics, which require high semantic alignment and IoU accuracy, our method surpasses the nearest competitors by 4.36% and 3.15% using the VGG and SF+C backbones, respectively. Overall, the superior performance of our method across all metrics on real-world datasets demonstrates that removing noisy clips under text constraints and performing contextual fusion on the purified clips effectively enhances video moment retrieval (VMR).

4.2 Ablation Study

Ablation on DRNet. Table 3 (A1-A5) evaluate the contribution of each module in DRNet. A1-A3 present the results of removing the TCD, TRF, and Decoder modules respectively, each resulting in a performance drop. Notably, removing TCD (A1) leads to an average performance drop of 11.71%, highlighting the critical role of noise filtering. Compared with A3, removing both the TRF and Decoder modules in A4 results in a more significant performance drop than removing only the Decoder. This is because the semantic consistency loss computed in the TRF module provides auxiliary supervision for the denoising process, encouraging the generation of cleaner visual features for subsequent decoding.

In A5, replacing Mamba with standard Transformer encoders as the base in DRNet leads to a significant 5.1% performance

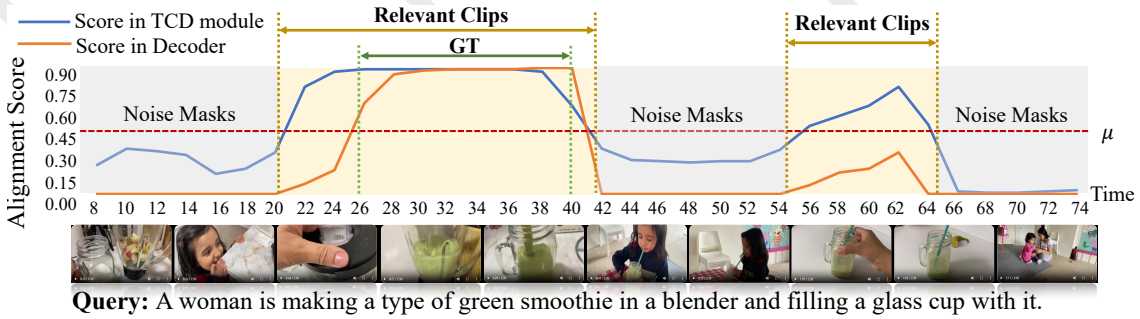

Figure 3: Text-clip alignment scores show that our method effectively filters out noisy clips and accurately localizes the target moment within relevant segments.

Model	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP@Avg.
$\text{UniVTG}_{23\text{iccv}}$	60.52	42.39	59.08	37.12	36.66
$\text{UniVTG}_{23\text{iccv}}$†	62.77 (↑ 2.25)	44.77 (↑ 2.38)	60.83 (↑ 1.75)	39.67 (↑ 2.55)	39.38 (↑ 2.72)
$\text{VMS}_{24\text{arxiv}}$	65.48	50.06	62.92	45.2	43.62
$\text{VMS}_{24\text{arxiv}}$†	66.58 (↑ 1.1)	51.87 (↑ 1.81)	64.51 (↑ 1.59)	45.97 (↑ 0.77)	44.76 (↑ 1.14)
DRNet	**68.06**	**54.58**	**65.2**	**48.02**	**46.11**

Table 4: Comparison of methods with and without video denoising on QVHighlights *val* split. All methods are re-implemented based on their official codes. † indicates results with denoising, with red values showing the performance improvement.

Metrics	TCD			TRF			Decoder		
	2	3	4	2	3	4	2	3	4
R1@0.5	67.54	68.06	67.12	67.2	68.06	66.7	67.6	68.06	67.2
R1@0.7	54.25	54.58	53.96	53.8	54.58	53.5	54.1	54.58	53.5

Table 5: Ablation study on CIO layers in each module.

drop. We attribute this to Mamba's selective information propagation, which more effectively integrates key textual information into visual features than the Transformer's global self-attention mechanism, which has information redundancy.

Ablation on TCD. Ablation experiments for removing each component in TCD (see B1-B3 of Table 3) shows a performance drop. The highest drop is observed for removing cross-attention (B1), highlighting the importance of cross-attention for integrating text and visual features after Mamba-based multimodal interaction. Similarly, the dynamic kernels (B2) and adding query tokens at the end of multimodal sequences also contribute to improved retrieval performance.

Ablation on CIO layers. In this ablation study, we investigate the effect of varying the layers of CIO in each module. Specifically, while adjusting the layers in one module, we keep the CIO layers in other modules fixed at three layers to ensure a fair comparison. As shown in Table 5, using three layers of CIO consistently achieves optimal performance across all modules, confirming the effectiveness and simplicity of the current design.

Denoise-then-Retrieve Paradigm Generalization. We apply the generated noise masks to existing Transformer-based (UniVTG) and Mamba-based (VMS) VMR methods without modifying their architectures. As shown in Table 4, this leads to notable performance gains across all metrics for both UniVTG [Lin *et al.*, 2023] and VMS [Chen *et al.*, 2024]. Specifically, by removing noisy clips, UniVTG and VMS see improvements of 2.72% and 1.14% points on mAP@Avg and 2.38% and 1.81% points on the challenging R1@0.7 metric, respectively. These results underscore the effectiveness and generalization of our denoise-retrieve paradigm.

4.3 Qualitative Results

We present visualizations in Fig. 3, where blue and orange curve denote text-clip alignment scores in TCD and Decoder modules, respectively. Red dashed line is the noise filtering threshold μ, with gray background clips indicating those filtered out. The text-conditioned denoising module enables our method to effectively distinguish between text-relevant and noisy clips, while the decoder module can further localize the target clips within the text-relevant clips precisely. Specifically, the blue curve illustrates that in TCD, there is a significant disparity in text-clip alignment scores between noisy and relevant clips, with consistently high alignment scores among the relevant clips. In contrast, the orange curve shows a clear distinction between the relevant clips, accurately identifying the text-referred clips.

5 Conclusion

In this work, we analyzed the importance of denoising in Video Moment Retrieval (VMR) and introduced DRNet, a Text-conditioned Denoising and Text-reconstruction Feedback approach. Our method filters irrelevant clips by generating noise masks and refines the process by aligning re-generated queries, distilled from purified video representations, with the input text. Extensive experiments on Charades-STA and QVHighlights benchmarks validate the effectiveness of denoising, demonstrating substantial improvements over state-of-the-art methods and showcasing our paradigm's adaptability to enhance other VMR models.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants No.62472092, No. 62172089, No.62106045. Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No. 93K-9, Nanjing Purple Mountain Laboratories, Fintech and Big Data Laboratory of Southeast University. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

References

[Chen and Jiang, 2019] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206, 2019.

[Chen *et al.*, 2020a] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 333–351. Springer, 2020.

[Chen *et al.*, 2020b] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020.

[Chen *et al.*, 2024] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[Gao *et al.*, 2021] Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. Relation-aware video reading comprehension for temporal language grounding. *arXiv preprint arXiv:2110.05717*, 2021.

[Ge *et al.*, 2019] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 245–253. IEEE, 2019.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Jiang *et al.*, 2019] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pages 217–225, 2019.

[Jiang *et al.*, 2024] Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiaoyong Wei, Chang Wen Chen, and Qing Li. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. *arXiv preprint arXiv:2407.15051*, 2024.

[Lei *et al.*, 2020] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 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*, pages 447–463. Springer, 2020.

[Lei *et al.*, 2021] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[Li *et al.*, 2023] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12032–12042, 2023.

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

[Lin *et al.*, 2023] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023.

[Liu *et al.*, 2018] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018.

[Liu *et al.*, 2022] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022.

[Liu *et al.*, 2024a] Weijia Liu, Bo Miao, Jiuxin Cao, Xuelin Zhu, Bo Liu, Mehwish Nasim, and Ajmal Mian. Context-enhanced video moment retrieval with large language models. *arXiv preprint arXiv:2405.12540*, 2024.

[Liu *et al.*, 2024b] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3855–3863, 2024.

[Miao *et al.*, 2023] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 920–930, 2023.

[Miao *et al.*, 2024a] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Region aware video object segmentation with deep motion modeling. *IEEE Transactions on Image Processing*, 33:2639–2651, 2024.

[Miao *et al.*, 2024b] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11373–11385, 2024.

[Miao *et al.*, 2024c] Bo Miao, Mingtao Feng, Zijie Wu, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Referring human pose and mask estimation in the wild. In *NeurIPS*, 2024.

[Moon *et al.*, 2023a] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023.

[Moon *et al.*, 2023b] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023.

[Qu *et al.*, 2020] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Sigurdsson *et al.*, 2016] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[Sun *et al.*, 2024] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4998–5007, 2024.

[Wang *et al.*, 2021] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2021.

[Xiao *et al.*, 2021] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021.

[Xiao *et al.*, 2024] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18709–18719, 2024.

[Xu *et al.*, 2019] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019.

[Xu *et al.*, 2023] Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxiang Zhu, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. *arXiv preprint arXiv:2305.00355*, 2023.

[Yang *et al.*, 2024] Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18308–18318, 2024.

[Yuan *et al.*, 2019] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[Zhang *et al.*, 2020] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[Zhang *et al.*, 2021] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.