OVG-HQ: Online Video Grounding with Hybrid-modal Queries

Runhao Zeng[1*], Jiaqi Mao[1*], Minghao Lai[1], Minh Hieu Phan[2],
Yanjie Dong[1], Wei Wang[1], Qi Chen[2†], Xiping Hu[1†]
[1]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, [2]University of Adelaide
zengrh@smbu.edu.cn, maojiaqi2324@gmail.com, huxp@smbu.edu.cn

Abstract

Video grounding (VG) task focuses on locating specific moments in a video based on a query, usually in text form. However, traditional VG struggles with some scenarios like streaming video or queries using visual cues. To fill this gap, we present a new task named Online Video Grounding with Hybrid-modal Queries (OVG-HQ), which enables online segment localization using text, images, video segments, and their combinations. This task poses two new challenges: limited context in online settings and modality imbalance during training, where dominant modalities overshadow weaker ones. To address these, we propose OVG-HQ-Unify, a unified framework featuring a Parametric Memory Block (PMB) that retain previously learned knowledge to enhance current decision and a cross-modal distillation strategy that guides the learning of non-dominant modalities. This design enables a single model to effectively handle hybrid-modal queries. Due to the lack of suitable datasets, we construct QVHighlights-Unify, an expanded dataset with multi-modal queries. Besides, since offline metrics overlook prediction timeliness, we adapt them to the online setting, introducing oR@n, IoU=m, and online mean Average Precision (omAP) to evaluate both accuracy and efficiency. Experiments show that our OVG-HQ-Unify outperforms existing models, offering a robust solution for online, hybrid-modal video grounding. Source code and datasets are available at `https://github.com/maojiaqi2324/OVG-HQ`.

1. Introduction

Video grounding [11, 61, 70] is a crucial research task that identifies the start and end times of target segments in untrimmed videos based on a given query. However, the current setting suffers from two critical limitations in real-world applications. **First**, it considers an offline setting, imposing strict requirements on complete video accessibility, which is insufficient for immediate detection in streaming media. For example, in surveillance, we need to continuously analyze live feeds and instantly ground queries, such as *"group of people gathering near the front door"*, so that security teams can respond immediately, rather than waiting to process a lengthy offline recording. **Second**, the current video grounding task predominantly relies on natural language queries, limiting its application to multi-modal scenarios. As an example, a text-only system might demand a detailed description such as *"a group of individuals congregating near the front door, frequently looking around, and making brief contact before dispersing in different directions"*. In contrast, with multi-modal queries, security staff could directly upload a past surveillance clip illustrating similar suspicious behavior. With this consideration, we introduce an extended task called Online Video Grounding with Hybrid-modal Queries (OVG-HQ).

Unlike the conventional offline video grounding task that only considers text queries as inputs, our OVG-HQ task ac-

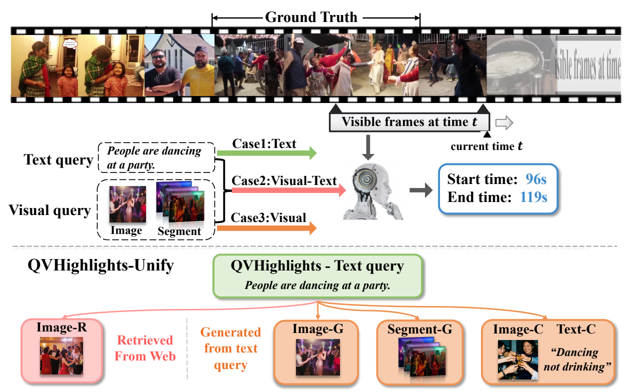

Figure 1. Illustration of our proposed online video grounding with hybrid-modal queries task, with two distinguishing characteristics: online video input and various query configurations. Beyond text query, it accepts visual queries (images, video segments) and their combination with text. We also construct a new QVHighlights-Unify dataset by augmenting QVHighlights dataset with images and video segments and complementary image-text pairs.

*Equal contribution
†Corresponding authors

commodates multiple query modalities (*e.g.*, text, image, video) in online video streaming, as shown in Figure 1. This setup requires the model to dynamically process and integrate information from diverse sources, adapting to evolving queries throughout the video. OVG-HQ emphasizes online inference and cross-modal interactions, challenging the model to ground relevant moments accurately across varying contexts and query types.

The new task poses new challenges. **First**, the video content can vary significantly over time, with changing scenes, lighting, and objects. Models must adapt to this variability in a streaming video, accounting for concept drift without losing prior learned knowledge. **Second**, as noted by [71], modality imbalance poses a significant challenge in hybrid-modal queries, as different modalities (such as text, image, or video) contribute unevenly to the task. Stronger modalities with more informative signals often dominate, overshadowing weaker ones. This imbalance causes the model to rely heavily on stronger modalities, leading to underutilization of the weaker ones, which reduces their contribution and ultimately impacts the model's overall accuracy in integrating diverse information. Consequently, it becomes difficult to use a single unified model to handle all modalities effectively. To tackle the challenging OVG-HQ task, we propose a unified yet flexible model called OVG-HQ-Unify, which supports hybrid-modal query inputs (*i.e.*, both uni- and multi-modal queries) and enables online localization of moments. It mainly has two parts. First, since each streaming video can be regarded as a sequence, to retain previously learned knowledge, we embed a **Parametric Memory Block** (PMB) instantiated with Test-Time Training layer (TTT) [42] that uses the network's parameters as dynamic memory for sequence modeling. With a self-supervised reconstruction loss, PMB encodes historical feature and prediction information, allowing the model to "memorize" past context for better decisions rather than directly saving historical data. In online video streams, PMB's ability to update parameters during inference enables continuous improvement and adaptability to new scenarios. Second, to alleviate the impact of modality imbalance, we design a hybrid distillation strategy that introduces a teacher model to guide the learning of non-dominant modalities, thus enhancing the model's performance consistency across different query modalities.

As there is no off-the-shelf dataset suitable for the OVG-HQ task, we construct a new dataset called QVHighlights-Unify, which expands the QVHighlights dataset [19] by adding image and segment queries[1]. This expansion enables the model to handle not only text queries but also visual modality inputs, validating its adaptability and consistency across various query types. Besides, as the offline metrics fail to capture the timeliness of predictions, we adapt them to the online setting called oR@n, IoU=m and online mean Average Precision (omAP) to evaluate both accuracy and efficiency. Experiments on QVHighlights-Unify, ANet-Captions, TACoS, MAD datasets show that our OVG-HQ-Unify framework achieves superior performance compared to existing methods, particularly in handling hybrid-modal queries. Our main contributions are as follows:

- We introduce a new task, Online Video Grounding with Hybrid-modal Queries (OVG-HQ), enabling multi-modal queries and requiring online segment localization in video streams, which is suited for practical applications.
- We propose a unified framework, called OVG-HQ-Unify, supporting hybrid-modal queries as inputs and enabling online localization of video clips. In detail, we introduce a Parameter Memory Block (PMB) to keep previously learned knowledge and a cross-modal distillation strategy to mitigate imbalances during multi-modal training.
- We construct a new dataset, QVHighlights-Unify, which includes multiple query modalities. Experiments on 4 datasets show that our OVG-HQ-Unify framework outperforms existing models, demonstrating its superiority in the online setting across various query types.

2. Related Work

2.1. Video Grounding with Text Query

Offline Setting. Offline Video Grounding methods [7, 8, 11, 15, 16, 24, 28, 32–34, 45, 48, 54, 55, 61, 70] involve identifying time intervals within a video that are semantically aligned with a given sentence. Proposal-based methods typically follow a two-stage pipeline: the first stage generates proposals, and the second ranks these proposals based on their relevance to the input query. Early techniques generate proposals using sliding windows [11, 12, 66] or predefined temporal anchors [3, 49, 57, 62, 68]. Later methods [20, 23, 45, 47, 67] explore all possible pairs of start and end points or use 2D temporal maps to process multiple candidates at once. Proposal-free methods [13, 33, 58, 64] aim to predict the target moment directly without the need for explicit proposals. They learn the interaction between video and sentence by applying techniques like attention mechanisms [13, 33, 36, 58, 64] and dense regression [5, 27, 60] from individual frames. In addition, efforts have been made to integrate temporal sentence grounding with other video understanding tasks into unified frameworks [22, 53]. Recent query-based models [2, 16, 18, 19, 21, 25, 30, 31, 41, 52] have simplified the process by removing the need for handcrafted components. Training-free methods [29, 51] have been introduced to address challenges in supervised learning, such as biases from

[1]We first expand the QVHighlights for its well-annotated moment retrieval data, enabling systematic evaluation of hybrid-modal queries. In the future, we will collect more complex datasets (*e.g.*, surveillance videos) to further validate and enhance our model in more practical scenarios.

annotations and limited generalization. They avoid relying on annotated data and instead leverage pre-trained models to assess the similarity between video segments and textual queries. Some methods [29] use vision-language models, while others [51] utilize large language models to compare video frame captions with the query. However, all these methods assume full access to the video in advance, which is not feasible for streaming applications where predictions must be made in streaming videos.

Online Setting. Recently, [10] proposed video grounding in an online setting, which involves retrieving relevant moments given a language query during video streaming. However, this setting overlooks the inherent flexibility of the query itself, as users may require inputs from multiple modalities beyond text, such as image, video segments, or any combination of these modalities. In this paper, we propose a task that is more aligned with real-world application scenarios, called Online Video Grounding with Hybrid-modal Queries. This task enables online segment localization in video streams using hybrid-modal queries, accommodating various input modalities to better meet user needs.

2.2. Video Grounding with Multi-modal Query

Video grounding tasks involve localizing specific events or activities within videos based on a given query. Most methods [24, 33, 45, 55] use natural language as the query. [69] was the first to utilize image queries to localize unseen activities in videos. More recently, [14] proposed grounding videos spatio-temporally using images or texts. [63] attempts to localize events in videos using multimodal semantic queries, but image-text pairs in this dataset are complementary and cannot be used independently as queries, neglecting that users may input different types of queries in practical settings. In this paper, we unify multiple modalities and various combinations of queries and additionally introduce the concept of video segment queries, which enables segment localization in video streams using queries comprising any combination of modalities—including images, text, and video segments.

3. Proposed Method

3.1. Problem Definition

Offline Video Grounding with Text Query. This conventional task requires a machine to process an untrimmed video $V = \{x_i\}_{i=1}^{T}$, where x_i denotes the i-th frame, and subsequently identify M relevant moments $\mathcal{M} = \{\mathcal{M}_m = (s_m, e_m)\}_{m=1}^{M}$ that correspond to a text query $\mathcal{Q}$. Each moment $\mathcal{M}_m$ is defined by its start and end frames s_m and e_m. However, this offline setting has two primary limitations in practical applications: 1) videos are often streamed, rendering it impractical to wait until all frames have been processed before predicting moments; 2) users may require inputs from multiple modalities beyond text, such as images or video segments.

Online Video Grounding with Hybrid-modal Queries (OVG-HQ). In this paper, we propose to study a more practical setting, which aims to understand an input multi-modal query $\mathcal{Q} \subseteq \{q_t, q_i, q_s\}$—where q_t, q_i, and q_s represent text, image, and video segment queries, respectively—and retrieve relevant moments from streaming video. In this setting, at each timestamp t ($1 \leq t \leq T$), the model only has access to a sliding window of frames[2] $V_{t-k+1:t} = \{x_i\}_{i=t-k+1}^{t}$, with $k \geq 1$. Using this partial video segment and multi-modal query $\mathcal{Q}$, the model should identify events (sometimes more than one) relevant to $\mathcal{Q}$. Importantly, once predictions are made at any timestamp, they cannot be modified or removed in future steps. Current methods rely on Non-Maximum Suppression (NMS) and future frame predictions to adjust past frames, which is impractical in streaming settings.

3.2. General Scheme

The challenge of online video grounding (OVG) lies in how to efficiently model and utilize historical information to enhance current predictions. To address this, we propose a simple yet effective sequence modeling module, namely **parametric memory block** (M_{PMB}). Inspired by TTT [42], our **parameter-as-memory layer** f_{PML} in M_{PMB} compresses sequential information (e.g., input frame sequences) into the neural network parameters. Based on M_{PMB}, we design an OVG-HQ-Unify model capable of handling various input configurations, including text, text + image, and text + segment, as shown in Figure 2.

In the following, we first introduce the design of M_{PMB} in Sec. 3.3. We then illustrate how M_{PMB} is employed in multi-modal fusion and prediction in Sec. 3.4 and 3.5, respectively. Lastly, we describe the approach for training a unified model with hybrid-modal queries in Sec. 3.6.

3.3. Parametric Memory Block

To enable memory retention in models, one common approach is to use a memory bank and integrate current inputs with stored memory via self-attention [46]. However, this introduces extra storage overhead and results in increased computational costs as the amount of historical data grows. In contrast, LSTMs store historical information in a fixed-size hidden state, whose expressive capacity is limited [42]. Unlike the above approaches, we propose a learnable parametric memory block M_{PMB} instantiated with TTT [42] that can compress the historical information within network parameters, which have much stronger expression power as

[2] Accessing all past frames is ideal but impractical for long video streams due to computational and memory constraints. A sliding window offers a balanced trade-off between efficiency and accuracy.

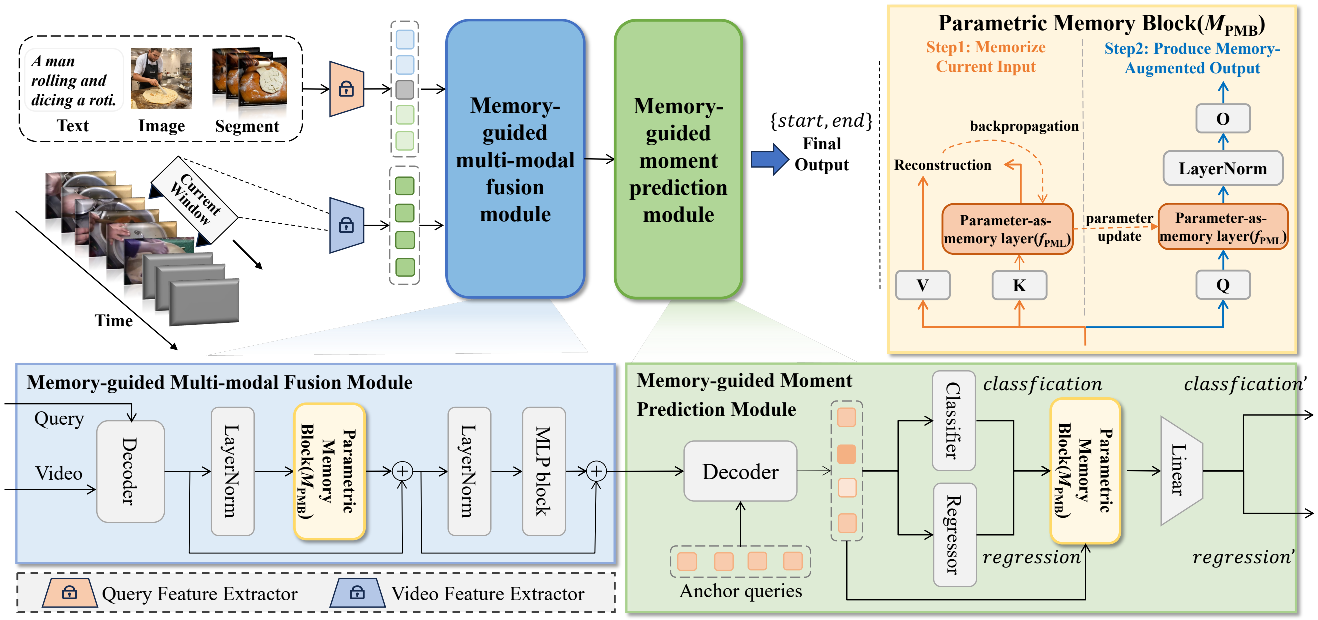

Figure 2. Overview of our OVG-HQ-Unify model. At time t, features extracted from video and query are processed via the memory-guided multi-modal fusion module (Sec. 3.4), where query-aware features are extracted via a transformer decoder and enhanced by parametric memory block M_{PMB} (Sec. 3.3). Then, the memory-guided moment prediction module (Sec . 3.5) decodes anchor features, which, along with the current predictions are fed to M_{PMB} for moment prediction. In M_{PMB}, the parameter-as-memory layer (f_{PML}) first memorizes current input by updating its parameters via self-supervised reconstruction loss and then predicts based on the historical information.

neural networks have a larger capacity than the hidden states of LSTM. It operates in two steps, as shown in Figure 2.

Step 1: Memorize Current Input. The core component of M_{PMB} is the *parameter-as-memory layer* $f_{\text{PML}}(\cdot; W^m)$. To compress the current input r_t into W^m, we employ a reconstruction loss as a form of self-supervision. This approach is akin to how language models utilize reconstruction or masked prediction loss to embed knowledge from training data into the parameters of neural networks through gradient descent. Formally, the reconstruction loss can be defined as follows:

$$\mathcal{L}_{\text{PML}}(r_t; W^m) = \| f_{\text{PML}}(W_K r_t; W^m) - W_V r_t \|^2, \quad (1)$$

where W_K and W_V are two learnable projection matrices. We then update W^m by

$$W^m \leftarrow W^m - \eta_{\text{PML}} \cdot \nabla \mathcal{L}_{\text{PML}}(r_t; W^m), \quad (2)$$

where $\eta_{\text{PML}} = \sigma(W_{lr} \cdot r_t)$ is an adpative learning rate following [42], W_{lr} is a learnable vector and σ is the sigmoid function. At this point, W^m holds information from both prior and current time step, enabling the network parameters to retain the current representation effectively.

Step 2: Produce Memory-Augmented Output. With the updated memory capturing both current and historical information, we can now augment r_t with memory. The current input r_t is first processed through a projection layer W_Q, then passed through the updated function $f_{\text{PML}}(\cdot; W^m)$, followed by layer normalization and another projection layer W_O. Mathematically, the process can be defined as

$$\hat{r_t} = f_{\text{PML}}(r_t; W) = W_O \cdot \text{LN}(f_{\text{PML}}(W_Q r_t; W^m)), \quad (3)$$

where LN denotes a LayerNorm layer. Then, this memory-augmented $\hat{r_t}$ is forwarded to the consequent modules.

Update Rule of Parametric Memory Block. Let W^p be the parameters of M_{PMB}, by excluding those of $f_{\text{PML}}(\cdot; W^m)$, we denote the remaining parameters as $W^r = W^p \setminus W^m$. In other words, all these parameters W_Q, W_K, $W_V$and W_O belong to W^r, as illustrated in the upper-right section of Figure 2. **First**, fix the parameters W^r, forward the current input r_t into f_{PML} and use Eqn. (1) to update the f_{PML} parameters W^m. **Second**, with parameters W^m fixed, use Eqn. (3) to produce memory-augmented output. **Third**, update the parameters W^r by minimizing the loss function derived from the video grounding task.

3.4. Memory-guided Multi-Modal Fusion

Query Feature Extraction. For text and image queries, we use the text and image encoder of CLIP [35] to extract features $\mathbf{F}_t$ and $\mathbf{F}_i$, respectively. For segment queries, we use [35] to extract features $\mathbf{F}_s$ at intervals of M seconds.

Video Feature Extraction. We process video sequences as streaming data through a sliding window mechanism with size L, which dynamically emulates the model's temporal receptive field at each time instant t by spanning frames within the interval $[t-L, t]$. The window slides forward with a step size of M seconds, where features of overlapping segments are computed only once and cached for subsequent reuse. Consequently, at each temporal position t, we employ [35] to extract new features from the current video frame. This operational paradigm ultimately yields snippets-level features $\mathbf{F}_v \in \mathbb{R}^{K \times D_v}$ for each sliding window, where K denotes the number of video frames extracted within the window.

Transformer-based Cross-modal Fusion. We transform all unimodal features to a unified dimension D via modality-specific linear layers and use a Transformer decoder with cross-attention to fuse video and query features. Queries may include multiple modalities, so we pad each modality with a specific token $\mathbf{m}_*$, where $* \in \{t, i, s\}$. For example, with text and image queries, the decoder input is a combination of multi-modal features $\mathbf{Q} = [\mathbf{m}_t, \mathbf{F}_t, \mathbf{m}_i, \mathbf{F}_i]$. In the decoder, video snippets' features $\mathbf{F}_v$ in each window serve as queries Q_v, and query features $\mathbf{Q}$ serve as keys K_q and values V_q. The rest of the decoder follows the standard Transformer architecture, resulting in query-aware video representations $\mathbf{F}_{qv}$.

Memory-guided Fusion via f_{PML}. As the query-aware video feature $\mathbf{F}_{qv}$ mainly focuses on information within the current window, to capture long-term video relationships, we further introduce a memory-guided sequence modeling module based on f_{PML} to incorporate historical context. As shown in Figure 2, this module resembles a Transformer encoder but replaces the self-attention layer with our f_{PML} mechanism. At each time step t, the feature vector $\mathbf{F}_{qv}$ is processed by our new module, and produces an output according to the equation in Eqn. (3). This update merges current and historical information, producing a memory-guided feature $\hat{\mathbf{F}}_{qv}$ for subsequent moment prediction.

3.5. Memory-guided Moment Prediction

At time t, our model generates a series of proposals based on predefined anchors, which end at t with lengths $L_n = L_q/2^{n-1}$ for $n = 1, \ldots, N$. For instance, the n-th anchor is represented as $A_n = (t - L_n, t)$. We use a Transformer decoder structure, following [17], to process the learnable anchor query $\mathbf{A} \in \mathbb{R}^{N \times D}$ and features $\hat{\mathbf{F}}_{qv}$ from the Memory-guided Multi-Modal Fusion Module, producing anchor features $\mathbf{F}_a \in \mathbb{R}^{N \times D}$ (see Figure 2). Using $\mathbf{F}_a$, a classification head predicts $\{s_f, s_b\}$ for foreground and background scores, while a regression head predicts $\{\Delta l, \Delta o\}$, indicating the target moment length and offset. Thus, the n-th anchor boundary, $\hat{A}_n = (s_n, e_n)$, is adjusted by:

$$\begin{aligned} s_n &= e_n - L_n \exp(\Delta l_n), \\ e_n &= t + L_n \Delta o_n. \end{aligned} \tag{4}$$

Memory-guided Prediction Refinement. In the online video grounding setting, predictions made at earlier time steps cannot be adjusted later. Thus, we design the model to refine current predictions using past results. As discussed in Sec. 3.3, f_{PML} can retain historical data, inspiring our *Prediction Refinement Module (PRM)*, shown in Fig. 2. First, we concatenate the classification outputs $\{s_f, s_b\}$ and regression outputs $\{\Delta l, \Delta o\}$, passing them through a linear layer to create the prediction feature $\mathbf{F}_p$. This is then combined with anchor features $\mathbf{F}_a$ to form $\mathbf{F}_c$, which is processed through M_{PML}.

Within f_{PML}, two main operations occur: 1) The anchor feature and current prediction $\mathbf{F}_c$ are compressed into parameters to incorporate historical prediction information; 2) The updated f_{PML} generates refined classification results $\{s_f^{\text{r}}, s_b^{\text{r}}\}$ and boundary offsets $\{\Delta l^{\text{r}}, \Delta o^{\text{r}}\}$ based on $\mathbf{F}_c$. Only anchors with $s_f > \theta$ (a predefined threshold) are selected, and their boundaries are calculated using Eqn. (4).

3.6. Unified Multi-modal Training and Inference

We empirically found that directly training a model with hybrid-modal data does not consistently yield strong performance across query types. While models perform well on text queries, performance significantly drops when text is absent (see Fig. 3). To address this, we propose a training strategy called hybrid distillation: **1)** We train using three query types (text, vision, and vision+text), alternating between them in batches. **2)** We apply distillation by first training an expert teacher model on text+segment-g queries, which provide the best multi-modal information. This expert model then guides the unified student model through distillation, applied to classification ($\mathbf{c} = \{s_f, s_b\}$), regression ($\mathbf{r} = \{\Delta l, \Delta o\}$), and anchor features ($\mathbf{F}_{a,i}^t$) with the following loss function:

$$\mathcal{L}_d = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_{\text{KL}}(\mathbf{F}_{a,i}^s, \mathbf{F}_{a,i}^t) + \mathcal{L}_2(\mathbf{r}_i^s, \mathbf{r}_i^t) + \mathcal{L}_2(\mathbf{c}_i^s, \mathbf{c}_i^t)), \tag{5}$$

where $\mathcal{L}_{\text{KL}}$ is KL Divergence and $\mathcal{L}_2$ is MSE loss, with N as the number of anchors, and s and t as the student and teacher outputs, respectively. Additionally, standard video grounding loss functions are applied to train the student model. The classification head's training loss is defined as:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{Focal}}(\mathbf{r}_i, \hat{\mathbf{r}}_i), \tag{6}$$

where we use the Focal loss [39] as $\mathcal{L}_{\text{Focal}}$. The training loss function for the regression head is defined as :

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_1(\mathbf{\Delta o}_i, \mathbf{\Delta \hat{o}}_i) + \mathcal{L}_1(\mathbf{\Delta l}_i, \mathbf{\Delta \hat{l}}_i)), \tag{7}$$

where $\mathcal{L}_1$ is the L1 loss. The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{8}$$

where λ is a hyperparameter, and we have found that $\lambda = 10$ works well across all experiments.

Dynamic Inference Details. During inference, unlike prior video grounding methods that keep the learned neural network fixed, our model's parameters (*i.e*., f_{PML}) are dynamically updated based on the self-supervised loss in Eqn. (1), allowing it to "memorize" and leverage historical information to adapt more effectively to unseen data.

4. Benchmark Creation and Evaluation Metric

We establish a new QVHighlights-Unify by expanding the QVHighlights dataset [19] with image and segment queries.

4.1. QVHighlights Dataset

It covers daily vlogs and news events for both moment retrieval and highlight detection. It contains more than 10,000 videos annotated with free-form queries. Each query is associated with one or multiple variable-length moments in its corresponding video, and a comprehensive 5-point Likert-scale saliency annotation for each clip in the moments.

4.2. Our QVHighlights-Unify Dataset

The QVHighlights dataset includes only text queries. We expand it with the following three types of queries.

1) Image-R: retrieved images based on text query. To simulate users searching online for visual clues, we first use QVHighlights text queries to retrieve ten semantically matching images. Then, we apply the InternVL vision-language model [6] to compute similarity scores and select the top-scoring image as the retrieved query. We did not consider retrieving videos because, compared to images, it is substantially more challenging to find a video that accurately matches the text without including irrelevant content.

2) Text-C+Image-C: complementary text-image pairs. As noted in [63], users may struggle to express unfamiliar or abstract concepts verbally or to find an image that perfectly matches their interests. Providing a simple sketch or sample image alongside a text query can help, as both complement each other semantically to convey the user's intent. Following [63], we modify the text query and generate a complementary image (Image-C) based on the revised text. We also create a corresponding textual description reflecting these modifications (for example, changing "Swimming" to "Dancing" yields "The action is swimming, not dancing."). Please refer to [63] for more details.

3) Image/Segment-G: generated visual queries w.r.t. text query. In practical applications, a visual query may not always be retrievable from the internet using its corresponding text query. To address this, we leverage modern generative models to produce images and videos as visual queries. Following [63], we design four prompt templates reflecting distinct image styles, randomly pair each text query in the QVHighlights dataset with one template, and use Stable Diffusion [38] for image generation. For videos, we employ the text-to-video model CogVideoX-5B [56] to create a six-second clip per text query as a generated segment query. We then manually filter out visually unclear or semantically mismatched samples, iteratively adjusting the textual input until the output meets the desired criteria.

4.3. Evaluation Metrics for Online VG

In online settings, where early and continuous predictions are essential, traditional metrics like mAP fail to account for timeliness. This leads to high scores even when predictions are delayed, making them unrealistic for real-time applications. To bridge this gap, we introduce two evaluation metrics (*i.e*., oR@n, IoU=m and omAP) that enalize delayed responses by incorporating a decay factor β ($0 < \beta < 1$). If a prediction is made on the ground truth's end time, $\beta = 1$; otherwise, β linearly decreases until it reaches zero once the prediction time exceeds the ground truth by a threshold $t_s \in \{1\text{s}, 3\text{s}, 5\text{s}\}$. Although other decay schemes exist (*e.g*., [59]), we adopt linear decay for simplicity. Lastly, we average over these t_s thresholds to obtain the final metrics.

1) oR@n, IoU=m (oR^n_m). We extend the standard R@n, IoU=m metric by introducing the decay factor β. If at least one of the top n retrieved moments have an IoU exceeding m, we set $r(n, m, q_i) = 1$; otherwise, $r(n, m, q_i) = 0$. For moments that match the i-th ground truth, we compute β_i using the method above. Formally, we compute

$$\text{oR@}n, \text{IoU@}m = \frac{1}{N_q} \sum_{i=1}^{N_q} \beta_i \cdot r(n, m, q_i), \tag{9}$$

where N_q is the number of queries.

2) omAP$_m$. We define omAP$_m$ as

$$\text{omAP}_m = \frac{1}{N_q} \sum_{i=1}^{N_q} \text{oAP}_m^{(i)}, \tag{10}$$

$$\text{oAP}_m^{(i)} = \sum_{j=2}^{H_i} (\beta_{i,j} R_{i,j} - \beta_{i,j-1} R_{i,j-1}) \beta_{i,j} P_{i,j}, \tag{11}$$

where H_i is the number of predictions that hit the ground truth corresponding to the i-th query, $P_{i,j}$ and $R_{i,j}$ are the precision-recall pairs obtained at different cutoff values during Average Precision (AP) calculation, $\beta_{i,j}$ is the sum of β values for the true positives used in the calculation of $R_{i,j}$ and $P_{i,j}$. We multiply $R_{i,j}$ and $P_{i,j}$ by $\beta_{i,j}$ to measure timeliness. Source code will be released.

5. Main Experiments

We compare our method with state-of-the-art methods on 4 datasets, including our QVHighlights-Unify dataset, ANet-

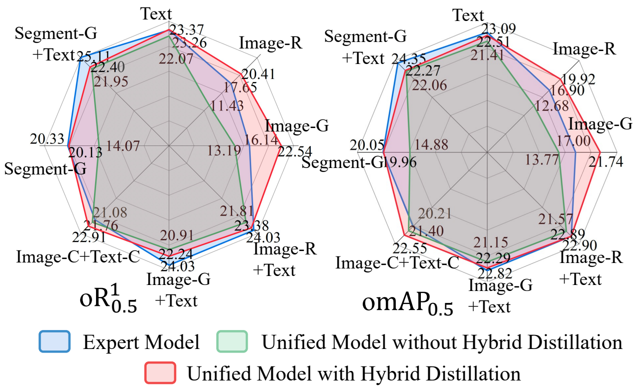

Figure 3. Performance comparisons on our QVHighlights-Unify dataset w.r.t. $oR^1_{0.5}$ and $omAP_{0.5}$.

Captions [1], TACoS [37], and MAD [40] datasets. Due to page limits, we put more details in the supplementary.

5.1. Results on the QVHighlights-Unify Dataset

The hybrid-modal query framework includes eight distinct input configurations, incorporating both single-modal and dual-modal queries. Figure 3 compares three models: 1) an expert model trained for each query type (blue), 2) a unified model directly trained for hybrid-modal queries (green), and 3) a unified model trained with hybrid distillation (red).

Training an expert model for each query type. We observe *1) Segment queries outperform Image queries:* This difference (20.33% vs 16.14%) likely arises as video grounding retrieves dynamic video segments, and Segment queries are better suited than Image queries to accurately describe a user's content of interest. *2) Multi-modal queries outperform single-modal queries:* For example, the Segment-G + Text expert model achieves a performance metric of 25.11%, substantially higher than either the Segment-G (20.33%) or Text (23.26%). This result suggests that multimodal queries provide richer and more comprehensive information about the desired moment.

Training a unified model to handle all query types. Our key findings include *1) Challenges with training a hybrid-modal unified model:* When multiple query types are directly combined into a single unified model, performance generally declines compared with the expert model. As shown in Figure 3, visual queries (e.g., Segment-G 14.07%) are considerably lower than the text query's score (22.07%). This suggests that during training, the model tends to prioritize the dominant modality, suppressing the optimization of other modalities [71]. *2) Improvement with hybrid distillation:* The model's performance improved significantly, especially in cases without text query. Specifically, when only an Image-R query is used, the $oR^1_{0.5}$ metric increases by 8.98% (from 11.43% to 20.41%), demonstrating the effectiveness of our proposed approach.

Comparisons with other VG methods. Following the implementation in [10], we adapt SoTA offline video grounding (VG) algorithms for the online VG task. Additionally, we re-implement TwinNet on our dataset. Notably, all compared methods utilize CLIP features for both video and text modalities. Given that previous approaches exclusively employ text queries during training, we conduct evaluations on the QVHighlights-Unify benchmark with text query as input. The detailed results are shown in Table 1. Our method exhibits notable improvements across various metrics.

Table 1. Comparisons with SoTA models on QVHighlights-Unify.

Setting (Text Query)	Method	$oR^1_{0.5}$	$omAP_{0.5}$
Offline VG (Modified to online)	TaskWeave [54]	7.02	5.96
	TR-DETR [41]	7.37	6.06
	R^2-Tuning [26]	9.30	8.17
Online VG	TwinNet [9]	20.78	19.73
	Ours	**23.26**	**23.09**

Table 2. Results on ANet-Captions, TACoS, and MAD datasets.

Setting	Method	ANet-Captions $R^1_{0.5}$	ANet-Captions $R^1_{0.7}$	TACoS $R^1_{0.5}$	TACoS $R^1_{0.7}$	MAD $R^5_{0.3}$	MAD $R^5_{0.5}$
Online Action Detection (Modified to VG)	OadTR [44]	23.27	10.97	21.12	10.92	2.50	0.90
	LSTR [50]	24.05	11.19	26.02	16.75	3.56	1.43
	GateHUB [4]	23.30	11.31	27.10	17.25	3.38	1.47
Offline VG (Modified to online)	VSLNet [64]	12.89	5.05	25.74	12.60	-	-
	2DTAN [67]	8.39	2.96	6.82	3.32	-	-
	SeqPAN [65]	12.57	4.79	25.07	13.67	-	-
	SMIN [43]	7.47	2.64	6.00	2.92	-	-
	TaskWeave [54]	8.22	3.67	14.93	6.78	-	-
	TR-DETR [41]	10.37	4.31	16.25	7.44	-	-
	R^2-Tuning [26]	9.17	4.16	21.69	11.24	-	-
Online VG	TwinNet [9]	25.48	12.56	29.74	19.07	4.71	2.00
	Ours	**26.57**	**14.36**	**30.98**	**21.17**	**6.32**	**3.27**

Table 3. Computational overhead of PMB and dynamic updates.

Method	Latency(ms)	FPS	FLOPs(M)	MACs(M)
Overall Model	21.76	45.95	5932.42	2966.21
PMB	2.20	454.54	11.43	5.72
Dynamic Update	0.30	3333.30	1.17	0.59

5.2. Results on Text Query-based VG Datasets

For a more comprehensive comparison, we evaluate our method against baselines on existing text query-based datasets. To ensure fairness, we employ the ANet-Captions and TACoS datasets with C3D features, and the MAD dataset with CLIP features. Following [10], we not only compare variants of offline VG modified for online settings but also evaluate several online action detection methods (likewise modified for online VG). These baseline results are directly provided by [10]. Since we do not have access to the models and therefore cannot measure the online metrics, we compare the offline metrics to ensure fairness and consistency. As shown in Table 2, our method substantially outperforms TwinNet and other approaches. Specifically, for $R^1_{0.7}$, our method achieves an improvement of 1.80% over TwinNet on ANet-Captions. These findings further underscore the unique challenges presented by online VG compared to offline VG, indicating the need for specialized strategies to address these challenges.

5.3. Runtime Analysis of Our Method

As we focus on the online VG problem, runtime efficiency is critical. To evaluate its performance, we test the model

Table 4. **Left:** Effect of parametric memory layer. f_{PML} is replaced by LSTM and self-attention (ATT), respectively. **Right:** Ablation study on the inputs of prediction refinement module: w/o Refine (no prediction refinement module), Pred (prediction only), and Pred+AF (prediction with anchor features).

Query	Variant	$\text{oR}^1_{0.5}$	$\text{omAP}_{0.5}$	Variant	$\text{oR}^1_{0.5}$	$\text{omAP}_{0.5}$
Text	Ours-ATT	13.93	16.41	w/o Refine	17.64	17.43
	Ours-LSTM	22.37	21.66	Pred	18.99	21.07
	Ours	**23.37**	**22.51**	Pred+AF	**23.37**	**22.51**
Image-R	Ours-ATT	10.71	13.3	w/o Refine	16.20	15.95
	Ours-LSTM	18.96	18.92	Pred	16.66	18.68
	Ours	**20.41**	**19.92**	Pred+AF	**20.41**	**19.92**
Image-G	Ours-ATT	11.69	14.35	w/o Refine	16.89	17.00
	Ours-LSTM	19.26	18.69	Pred	18.96	20.82
	Ours	**22.54**	**21.74**	Pred+AF	**22.54**	**21.74**
Image-R+Text	Ours-ATT	13.55	15.89	w/o Refine	17.23	17.69
	Ours-LSTM	22.39	21.41	Pred	18.78	20.96
	Ours	**23.38**	**22.89**	Pred+AF	**23.38**	**22.89**
Image-G+Text	Ours-ATT	14.42	16.17	w/o Refine	18.90	18.61
	Ours-LSTM	21.97	21.46	Pred	20.05	21.89
	Ours	**22.24**	**22.29**	Pred+AF	**22.24**	**22.29**
Image-C+Text-C	Ours-ATT	12.2	15.33	w/o Refine	16.53	17.02
	Ours-LSTM	20.46	20.24	Pred	19.61	21.30
	Ours	**22.91**	**22.55**	Pred+AF	**22.91**	**22.55**
Segment-G	Ours-ATT	11.85	14.2	w/o Refine	15.43	15.88
	Ours-LSTM	17.41	16.93	Pred	17.30	19.28
	Ours	**20.13**	**19.96**	Pred+AF	**20.13**	**19.96**
Segment-G+Text	Ours-ATT	13.32	15.48	w/o Refine	17.69	18.12
	Ours-LSTM	21.95	21.14	Pred	19.76	21.39
	Ours	**22.40**	**22.27**	Pred+AF	**22.40**	**22.27**

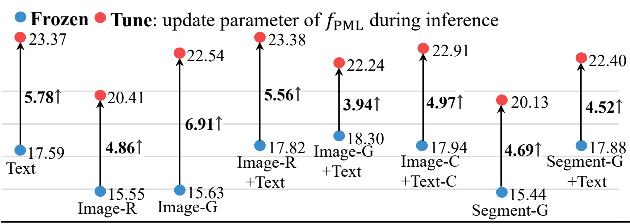

Figure 4. Effectiveness of test-time model updates w.r.t. $\text{oR}^1_{0.5}$.

on a single RTX 4090 GPU. All metrics, including latency, FPS, FLOPs, and MACs, are computed on a per-frame basis. As shown in Table 3, the overall model achieves an FPS of 45.95, satisfying real-time processing requirements. Moreover, the FLOPs and latency of the PMB and Dynamic Update components are significantly lower than those of the entire model, indicating that both the proposed PMB and the dynamic update process exhibit high efficiency.

6. Ablation Studies

6.1. Does f_{PML} Help Online Video Grounding?

In our approach, both feature fusion and prediction refinement are embedded within the proposed f_{PML}. We designed two variants for comparison: **1) Ours-LSTM:** where f_{PML} is replaced with an LSTM, and **2) Ours-ATT**: where f_{PML} is replaced with a self-attention layer of equivalent parameter size. The remaining network structures are identical to our method to ensure a fair comparison. As shown in Table 4, both the LSTM and our f_{PML} consistently outperform the self-attention (ATT) layer, with f_{PML} achieving 23.37% compared to 13.93% for the ATT layer, highlighting the importance of incorporating historical information in online video grounding task. Furthermore, across different query configurations, our method surpasses the LSTM in all cases, notably improving the Text query from 22.37% to 23.37% and the Segment-G query from 17.41% to 20.13%, further highlighting that when modeling historical information, a more expressive neural network—such as f_{PML}—is superior to a fixed-size hidden state, as it provides more effective information for current predictions.

6.2. What Benefits Prediction Refinement?

In our prediction refinement module, the f_{PML} parameter encapsulates both the current prediction and the current anchor feature (AF), compressing the information of the current step. This approach models the historical context of both the prediction and the anchor feature. We progressively removed these two types of information and present the results in Table 4. Removing the anchor feature input leads to a significant decline (from 23.37% to 18.99%) in the $\text{oR}^1_{0.5}$ metric when using a text query. Moreover, when the entire Prediction Refinement Head is eliminated, the performance deteriorates (1.35%) even further. These results highlight the critical role of prediction information memory and demonstrate that including anchor features considerably improves model performance.

6.3. Does Updating f_{PML} in Inference Time Help?

The key feature of our method is that, upon the arrival of each new video, the parameters of our model (i.e., f_{PML}) are reset and dynamically updated with each frame input, based on the self-supervised loss defined in Eqn. (1). To investigate the impact of this strategy on online video grounding performance, we compare two implementation variants: 1) **Frozen**: Parameters of f_{PML} are kept fixed during inference. 2) **Tune**: the parameters of f_{PML} are dynamically updated during inference. As shown in Figure 4, the Tune configuration consistently outperforms Frozen across eight distinct query composition settings. These results indicate that updating f_{PML} during the testing phase enables better adaptation to unseen data.

7. Conclusion

We have introduced Online Video Grounding with Hybrid-modal Queries (OVG-HQ), extending traditional video grounding task to support text, images, video snippets, and their combinations in streaming scenarios. To enable this, we have developed QVHighlight-Unify and introduced two new metrics to jointly evaluate accuracy and timeliness. To benchmark OVG-HQ, we have proposed OVG-HQ-Unify, a unified model featuring a Parametric Memory Block for retaining past context and a hybrid-distillation strategy for training. We hope this work inspires further research in online video grounding, bridging the gap between academic benchmarks and real-world applications.

Acknowledgements. This work was partially supported by National Natural Science Foundation of China (NSFC) under Grants 62202311, the Guangdong Basic and Applied Basic Research Foundation under Grants 2023A1515011512, Key Scientific Research Project of the Department of Education of Guangdong Province 2024ZDZX3012.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 7

[2] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9810–9823, 2021. 2

[3] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)*, pages 162–171, 2018. 2

[4] Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Gatehub: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19925–19934, 2022. 7

[5] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10551–10558, 2020. 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 6

[7] Ioana Croitoru, Simion-Vlad Bogolin, Samuel Albanie, Yang Liu, Zhaowen Wang, Seunghyun Yoon, Franck Dernoncourt, Hailin Jin, and Trung Bui. Moment detection in long tutorial videos. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 2594–2604, 2023. 2

[8] Xiang Fang, Zeyu Xiong, Wanlong Fang, Xiaoye Qu, Chen Chen, Jianfeng Dong, Keke Tang, Pan Zhou, Yu Cheng, and Daizong Liu. Rethinking weakly-supervised video temporal grounding from a game perspective. In *European Conference on Computer Vision. Springer*, 2024. 2

[9] Miquel Ferriol-Galmés, José Suárez-Varela, Jordi Paillissé, Xiang Shi, Shihan Xiao, Xiangle Cheng, Pere Barlet-Ros, and Albert Cabellos-Aparicio. Building a digital twin for network optimization using graph neural networks. *Computer Networks*, 217:109329, 2022. 7

[10] Tian Gan, Xiao Wang, Yan Sun, Jianlong Wu, Qingpei Guo, and Liqiang Nie. Temporal sentence grounding in streaming videos. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 4637–4646, 2023. 3, 7

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 5267–5275, 2017. 1, 2

[12] 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. 2

[13] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, 2019. 2

[14] Raghav Goyal, Effrosyni Mavroudi, Xitong Yang, Sainbayar Sukhbaatar, Leonid Sigal, Matt Feiszli, Lorenzo Torresani, and Du Tran. Minotaur: Multi-task video grounding from multimodal queries. *arXiv preprint arXiv:2302.08063*, 2023. 3

[15] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision (ECCV)*, pages 724–740. Springer, 2022. 2

[16] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13846–13856, 2023. 2

[17] Young Hwi Kim, Hyolim Kang, and Seon Joo Kim. A sliding window scheme for online temporal action localization. In *European Conference on Computer Vision*, pages 653–669. Springer, 2022. 5

[18] Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision (ECCV)*, pages 220–238. Springer, 2025. 2

[19] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:11846–11858, 2021. 2, 6

[20] 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 (ICCV)*, pages 12032–12042, 2023. 2

[21] 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 (NeurIPS)*, 36, 2024. 2

[22] 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 (ICCV)*, pages 2794–2804, 2023. 2

[23] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11235–11244, 2021. 2

[24] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1665–1673, 2022. 2, 3

[25] 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 (CVPR)*, pages 3042–3051, 2022. 2

[26] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R^2-Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 421–438. Springer, 2024. 7

[27] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5144–5153, 2019. 2

[28] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23045–23055, 2023. 2

[29] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Zero-shot video moment retrieval from frozen vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5464–5473, 2024. 2, 3

[30] 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. 2

[31] 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 (CVPR)*, pages 23023–23033, 2023. 2

[32] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940, 2024. 2

[33] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10810–10819, 2020. 2, 3

[34] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1847–1856, 2024. 2

[35] 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. 4, 5

[36] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pages 2464–2473, 2020. 2

[37] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer, 2014. 7

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[39] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2980–2988, 2017. 5

[40] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 7

[41] 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*, pages 4998–5007, 2024. 2, 7

[42] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024. 2, 3, 4

[43] 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 (CVPR)*, pages 7026–7035, 2021. 7

[44] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7565–7575, 2021. 7

[45] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2613–2623, 2022. 2, 3

[46] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13587–13597, 2022. 3

[47] 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*, pages 2986–2994, 2021. 2

[48] 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 (CVPR)*, pages 18709–18719, 2024. 2

[49] 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*, pages 9062–9069, 2019. 2

[50] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 7

[51] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. *Applied Sciences*, 14(5):1894, 2024. 2, 3

[52] Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Youyao Jia, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024. 2

[53] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13623–13633, 2023. 2

[54] 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 (CVPR)*, pages 18308–18318, 2024. 2, 7

[55] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 2, 3

[56] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 6

[57] 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 (NeurIPS)*, 32, 2019. 2

[58] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9159–9166, 2019. 2

[59] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd international workshop on human-centric multimedia analysis*, pages 13–21, 2021. 6

[60] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10287–10296, 2020. 2

[61] Yingsen Zeng, Yujie Zhong, Chengjian Feng, and Lin Ma. Unimd: Towards unifying moment retrieval and temporal action detection. In *European Conference on Computer Vision (ECCV)*, pages 286–304. Springer, 2025. 1, 2

[62] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1247–1257, 2019. 2

[63] Gengyuan Zhang, Mang Ling Ada Fok, Yan Xia, Yansong Tang, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. Localizing events in videos with multimodal queries. *arXiv preprint arXiv:2406.10079*, 2024. 3, 6

[64] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6543–6554, 2020. 2, 7

[65] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Siow Mong Rick Goh. Parallel attention network with sequence matching for video grounding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 776–790, 2021. 7

[66] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, pages 1230–1238, 2019. 2

[67] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, pages 12870–12877, 2020. 2, 7

[68] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664, 2019. 2

[69] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Deng Cai. Localizing unseen activities in video via image query. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4390–4396, 2019. 3

[70] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision (ECCV)*, pages 20–37. Springer, 2025. 1, 2

[71] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. Intra-and inter-modal curriculum for multi-modal learning. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 3724–3735, 2023. 2, 7