SKIMVTG: VIDEO TEMPORAL GROUNDING VIA PROGRESSIVE BOUNDARY REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Temporal Grounding (VTG) aims to localize the temporal segment in an untrimmed video that corresponds to a natural language query. Despite recent progress, precise boundary localization remains challenging because clip-query relevance scores often vary smoothly around the target moment, providing limited cues for identifying exact start and end boundaries. To address this, we propose SkimVTG, a progressive boundary refinement framework that first localizes the target moment from a temporally compressed representation and then refines its boundaries toward the original temporal resolution. At each temporal resolution, Boundary Cross-Attention (BCA) combines clip–query relevance with its temporal difference to extract boundary-sensitive cues, while the Temporal Bridging Network (TBN) propagates refined coarse-level representations to finer temporal resolutions through Temporal Distance Cross-Attention (TDCA). By alternating BCA and TBN, SkimVTG progressively sharpens coarse localization into precise temporal boundaries. Experiments on `QVHighlights`, `Charades-STA`, and `TACoS` demonstrate strong performance in Moment Retrieval and Highlight Detection. *The code will be released upon publication.*

1 INTRODUCTION

The rapid growth of video content on online platforms has made retrieving specific moments from videos an important research problem. Video Temporal Grounding (VTG) (Jang et al., 2023; Lei et al., 2021; Liu et al., 2022) aims to localize the temporal moment in an untrimmed video that corresponds to a natural language query. Recent unified VTG settings commonly consider two complementary objectives: Moment Retrieval (MR), which predicts the start and end boundaries of the target moment, and Highlight Detection (HD), which estimates clip-level saliency scores indicating each clip's relevance to the query. MR provides explicit boundary localization, while HD provides clip-level relevance cues; therefore, recent works address both tasks within a unified framework.

Existing VTG methods generally follow two main strategies. Encoder–decoder methods (Jang et al., 2023; Lee & Byun, 2024; Lei et al., 2021; Moon et al., 2023b) adopt a DETR-like structure (Carion et al., 2020), where learnable decoder queries serve as candidate moments and predict their temporal center and duration. Encoder-only methods (Lin et al., 2023; Liu et al., 2024; Sun et al., 2025; Yan et al., 2023) instead regress the distances from each temporal position to the start and end boundaries of the target moment, without using a separate decoder. Despite their architectural differences, both strategies still struggle with precise boundary localization.

This difficulty arises because existing methods rely on a single video–query representation for both coarse moment localization and fine-grained boundary estimation. Moreover, boundary prediction is largely driven by clip–query relevance scores, which measure the semantic alignment between temporal positions and the query. These scores often vary smoothly around the target moment, providing limited discrimination near the exact start and end boundaries.

To address this, we propose SkimVTG, a coarse-to-fine framework that performs rough moment localization and precise boundary refinement at different temporal resolutions. SkimVTG first obtains coarse localization of the target moment from a temporally compressed representation and then progressively refines its boundaries at finer temporal resolutions until reaching the original

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

temporal resolution. We further introduce Boundary Cross-Attention (BCA), which combines clip–query relevance with its temporal difference to capture boundary-sensitive cues. While relevance indicates how strongly each clip matches the query, the temporal difference highlights transitions into and out of the target moment. Finally, the Temporal Bridging Network (TBN) propagates refined coarse-level representations to finer temporal resolutions through Temporal Distance Cross-Attention (TDCA), while preserving local temporal coherence. Across QVHighlights, Charades-STA, and TACoS, SkimVTG improves localization accuracy most at high IoU thresholds, where boundary precision matters.

The main contributions of this work are summarized as follows.

- We propose SkimVTG, a coarse-to-fine framework that uses different temporal resolutions for coarse moment localization and precise boundary refinement, progressively refining temporal boundaries from a coarser temporal resolution to the original temporal resolution.
- We propose Boundary Cross-Attention (BCA), which combines clip–query relevance with its temporal difference to extract boundary-sensitive cues for localizing target start and end boundaries.
- We introduce the Temporal Bridging Network (TBN), which propagates boundary-aware representations across adjacent temporal resolutions while preserving local temporal coherence.
- Experiments on three VTG benchmarks show that progressive refinement mainly improves boundary precision rather than coarse moment localization.

2 Related Works

Encoder–Decoder Methods for VTG. Encoder–decoder VTG methods are largely inspired by DETR (Carion et al., 2020), using learnable decoder queries to represent candidate moments. Moment-DETR (Lei et al., 2021) introduced a unified framework for moment retrieval and highlight detection, followed by methods that improve query–video interactions (Jang et al., 2023; Liu et al., 2022; Moon et al., 2023a;b; Sun et al., 2024). BAM-DETR (Lee & Byun, 2024) further adopts a boundary-oriented formulation. Unlike these single-stage localization approaches, SkimVTG progressively refines temporal boundaries from coarse to fine resolutions.

Proposal-based and Encoder-only Methods for VTG. Early proposal-based methods (Anne Hendricks et al., 2017; Gao et al., 2017; Zhang et al., 2020b) generate and rank predefined temporal segments, often resulting in redundant proposals and coarse boundaries. Encoder-only methods (Lin et al., 2023; Liu et al., 2024; Sun et al., 2025; Yan et al., 2023) instead directly regress start and end distances at each temporal position. Recent works improve representations through large-scale pre-training (Lin et al., 2023), parameter-efficient vision–language adaptation (Liu et al., 2024), or discrete vector quantization (Sun et al., 2025). However, they still mainly rely on smoothly varying clip–query relevance scores for boundary prediction. SkimVTG complements these scores with their temporal variation to provide more boundary-sensitive cues.

3 Proposed Methods

3.1 Motivations and Method Overview

SkimVTG is designed for progressive boundary refinement in video temporal grounding. The key motivation is that coarse moment localization and precise boundary estimation require different temporal evidence. Coarser temporal representations provide a global view of the video and are suitable for identifying the approximate extent of the target moment, whereas finer temporal representations preserve detailed temporal variations needed to determine the exact start and end boundaries. However, clip–query relevance, although important for localizing the target moment, often varies smoothly around moment boundaries (Fig. 4), making boundary transitions ambiguous when relevance magnitude alone is used. To obtain boundary-sensitive cues, SkimVTG jointly models clip–query relevance and its temporal difference $\Delta\alpha$, which captures relevance changes over time.

As illustrated in Fig. 1, SkimVTG constructs multi-scale temporal video representations through temporal subsampling, ranging from the original temporal resolution to coarser temporal resolu-

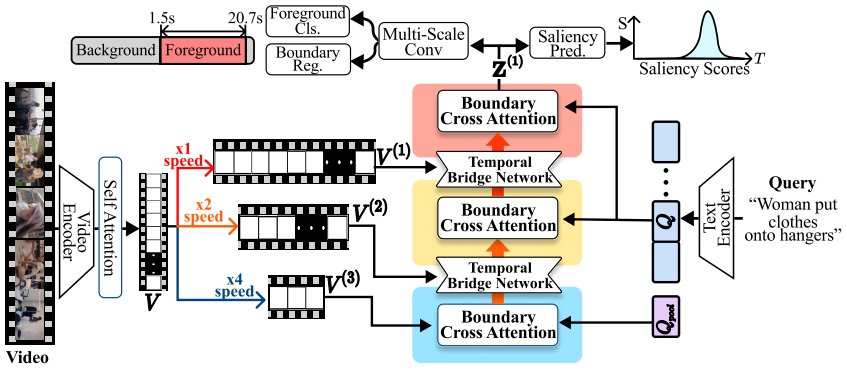


Figure 1: Overview of SkimVTG, which builds multi-resolution temporal video representations via self-attention and hierarchical subsampling. BCA introduces boundary-sensitive cues from relevance variations, while TBN propagates refined coarse representations to finer resolutions for progressive boundary refinement.

tions. Starting from the coarsest temporal representation, SkimVTG progressively refines boundary-aware representations toward the original temporal resolution. At each temporal resolution, Boundary Cross-Attention (BCA) uses clip–query relevance and its temporal difference $\Delta\alpha$ to produce boundary-aware representations. The Temporal Bridging Network (TBN) propagates the refined coarse-level representation to the next finer temporal resolution through cross-resolution feature propagation. This coarse-to-fine process allows SkimVTG to resolve boundary ambiguity progressively and achieve precise temporal localization beyond a single temporal scale.

3.2 Multi-scale Temporal Video Representation

SkimVTG takes a video and a natural language query as input, and constructs video representations at multiple temporal resolutions for progressive boundary refinement. The input video is encoded by a pre-trained video encoder (Feichtenhofer et al., 2019; Radford et al., 2021) into a sequence of T video clip features $\mathbf{V} = \{\mathbf{v}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$, where each clip corresponds to a fixed number of consecutive frames. The natural language query is encoded by a pre-trained text encoder (Radford et al., 2021) into token embeddings $\mathbf{Q} \in \mathbb{R}^{W\times d}$, along with a pooled query representation $\mathbf{Q}_{\text{pool}} \in \mathbb{R}^{d}$ obtained via adaptive pooling. We obtain the resulting video and query representations as $\mathbf{V}$, $\mathbf{Q}$, and $\mathbf{Q}_{\text{pool}}$, which are used as inputs to the subsequent modules.

As shown in the left panel of Fig. 1, we construct multi-resolution video representations from the input sequence $\mathbf{V}$. We first apply self-attention to contextualize each clip with global video information:

$$\mathbf{V}^{(1)} = \text{SelfAttn}(\mathbf{V}). \tag{1}$$

Starting from $\mathbf{V}^{(1)}$, we iteratively apply stride-2 temporal average pooling to construct K representations $\{\mathbf{V}^{(k)}\}_{k=1}^{K}$, where $\mathbf{V}^{(k)} \in \mathbb{R}^{(T/2^{k-1})\times d}$. Here, $k = 1$ denotes the original temporal resolution, whereas $k = K$ denotes the coarsest representation. Since the hierarchy is constructed from the same contextualized sequence without resolution-specific encoders, all temporal resolutions share the same feature dimensionality and representation space.

3.3 Boundary Cross-Attention (BCA)

After constructing multi-resolution video representations, we apply Boundary Cross-Attention (BCA) at each temporal resolution to produce boundary-aware video representations by incorporating both query relevance and its temporal variation. As illustrated in Fig. 2-(a), BCA takes the video representations $\{\mathbf{V}^{(k)}\}$ and the text representations $(\mathbf{Q}, \mathbf{Q}_{\text{pool}})$ as input, and models query–video alignment via cross-attention (Vaswani et al., 2017). The resulting attention scores serve as clip–query relevance scores, which help identify query-relevant regions. However, these scores typically vary smoothly around moment boundaries, making it difficult to distinguish whether a clip lies near the start boundary, within the moment, or near the end boundary.

To provide a more boundary-sensitive cue, BCA augments the query-aligned video representation with the temporal variation of the relevance scores, which captures how the relevance changes over time. An increasing trend in the relevance scores indicates a transition into the relevant moment, whereas a decreasing trend indicates a transition out of it. By combining the relevance scores and their temporal variation, BCA produces representations that encode both semantic relevance and boundary-related transitions.

Cross-Attention Design in BCA. BCA is applied at all temporal resolutions, performing cross-attention with video features as Query and text features as Key and Value. At each temporal resolution, BCA adapts the Query–Key/Value configuration to match the characteristics of the corresponding video representation. At the coarsest resolution ($k = K$), the Query is the temporally subsampled video features $\mathbf{V}^{(K)}$, while the Key/Value is the pooled query representation $\mathbf{Q}_{\text{pool}}$, providing a global semantic reference for coarse matching. At finer resolutions ($k < K$), the Query is the refined video features at the current resolution, obtained by propagating information from coarser levels (Sec. 3.4), and the Key/Value is the token embedding sequence $\mathbf{Q}$, enabling fine-grained alignment with individual words. Given the Query, Key, and Value, attention scores are normalized via softmax along the video clip axis rather than the text axis (Locatello et al., 2020), so that clips compete with one another for relevance to the query. This produces the attention scores $\alpha^{(k)}$ and the query-aligned video representation $\bar{\mathbf{V}}^{(k)}$ at each resolution.

Boundary-Aware Feature Refinement. Cross-attention produces a query-aligned video representation $\bar{\mathbf{V}}^{(k)}$ and attention scores $\alpha^{(k)}$, where $\alpha^{(k)}$ represents the relevance between video clips and the text query at resolution k. While $\alpha^{(k)}$ indicates where the relevant moment is likely to occur, its values often change gradually over time, making precise boundary localization ambiguous. To capture boundary-related transitions, BCA models the temporal variation of relevance scores rather than relying only on relevance magnitude. As illustrated in Fig. 4, increasing and decreasing trends in the relevance scores provide cues for transitions into and out of the relevant moment. BCA therefore derives a clip-level relevance sequence $\bar{\alpha}^{(k)}$ from $\alpha^{(k)}$, computes its temporal difference $\Delta\alpha^{(k)}$, and incorporates this boundary-aware signal into $\bar{\mathbf{V}}^{(k)}$.

To compute temporal differences, we first obtain a single relevance score for each clip. At the coarsest resolution ($k = K$), BCA uses the pooled query representation $\mathbf{Q}_{\text{pool}}$, so $\alpha^{(K)}$ is already defined at the clip level. We therefore directly set $\bar{\alpha}_t^{(K)} = \alpha_t^{(K)}$. At finer resolutions ($k < K$), BCA uses the token-level embeddings $\mathbf{Q}$, resulting in a clip–word attention matrix $\alpha^{(k)} \in \mathbb{R}^{(T/2^{k-1}) \times W}$. We then average $\alpha^{(k)}$ along the word axis, resulting in a clip-level relevance sequence $\bar{\alpha}^{(k)}$. Using the resulting relevance sequence, we compute the temporal difference as

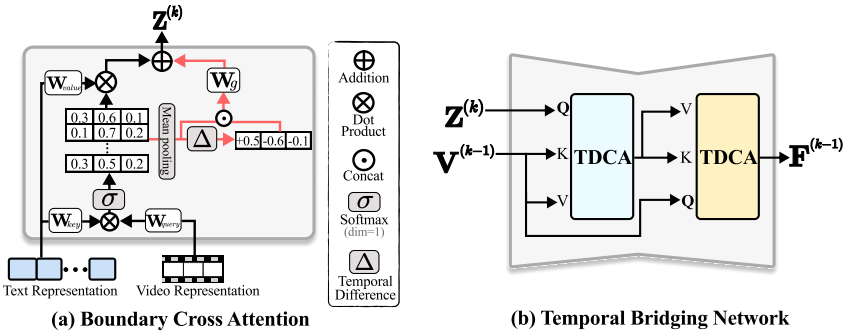
$$\Delta\alpha_t^{(k)} = \bar{\alpha}_{t+1}^{(k)} - \bar{\alpha}_t^{(k)}, \quad t = 1, \ldots, T/2^{k-1} - 1, \tag{2}$$

with $\Delta\alpha_{T/2^{k-1}}^{(k)} = 0$ for the last video clip. Here, $\Delta\alpha_t^{(k)}$ represents the relevance transition between the t-th and $(t+1)$-th clips. Accordingly, a positive value corresponds to an increasing relevance transition, whereas a negative value corresponds to a decreasing relevance transition.

The temporal difference $\Delta\alpha^{(k)}$ captures the direction and magnitude of relevance changes, complementing the relevance sequence $\bar{\alpha}^{(k)}$. For example, when the relevance is low but increasing (i.e., $\Delta\alpha_t^{(k)} > 0$), it indicates a transition into the relevant moment, whereas when the relevance is high but decreasing (i.e., $\Delta\alpha_t^{(k)} < 0$), it indicates a transition out of it. To encode this distinction, we concatenate the relevance score $\bar{\alpha}_t^{(k)}$ and its temporal difference $\Delta\alpha_t^{(k)}$, project them into the d-dimensional space, and add the result to $\bar{\mathbf{V}}^{(k)}$:

$$\mathbf{g}_t^{(k)} = \mathbf{W}_g^{(k)} \begin{bmatrix} \bar{\alpha}_t^{(k)} \\ \Delta\alpha_t^{(k)} \end{bmatrix}, \qquad \mathbf{Z}_t^{(k)} = \bar{\mathbf{V}}_t^{(k)} + \mathbf{g}_t^{(k)}. \tag{3}$$

where $\mathbf{W}_g^{(k)} \in \mathbb{R}^{d \times 2}$ is a learnable projection. This additive update preserves the query-aligned representation while incorporating the temporal-difference cue into the feature.



(a) Boundary Cross Attention (b) Temporal Bridging Network

Figure 2: **Architecture of SkimVTG.** (a) BCA aligns video features with the query and adds boundary cues from $\Delta\alpha$, the temporal difference of clip–query relevance (pink arrows). (b) TBN propagates coarse-level features to finer resolutions via TDCA for progressive boundary refinement.

3.4 Temporal Bridging Network (TBN)

After BCA produces a boundary-aware representation at a coarser temporal resolution, the Temporal Bridging Network (TBN) transfers and adapts it to the next finer resolution for further refinement. TBN operates between two adjacent temporal resolutions, k and $k-1$, where k denotes the coarser resolution and $k-1$ denotes the finer one, as illustrated in Fig. 2-(b). Directly propagating the coarser representation to the finer resolution may be insufficient because the two representations describe the same video with different temporal granularities and may not be directly aligned. TBN therefore performs a two-step cross-resolution interaction between adjacent temporal resolutions. In the first step, the coarser representation at resolution k is updated by incorporating local context from the finer representation at resolution $k-1$. In the second step, the refined coarser representation is propagated to the finer resolution to serve as the input for the next stage of progressive refinement.

Temporal Distance Cross-Attention (TDCA). The two steps in TBN are implemented with Temporal Distance Cross-Attention (TDCA), which is designed to incorporate temporal distance into the attention logits. When features from different temporal resolutions interact, standard cross-attention determines attention weights solely based on feature similarity. As a result, semantically similar clips can receive high attention weights even when they are far apart in time. TDCA mitigates this issue by introducing a distance-based penalty for temporally distant clips, encouraging stronger interaction between clips that are both temporally nearby and semantically relevant.

To compute temporal distance, TDCA uses clip indices at the original temporal resolution. Since all temporal resolutions are derived from the same original clip sequence, each clip at any temporal resolution can be mapped to its corresponding index at the original temporal resolution. Let i and j denote clip indices at temporal resolutions k and $k-1$, respectively. We map these indices to the original temporal resolution and denote the corresponding indices as $r_i^{(k)}$ and $r_j^{(k-1)}$, respectively. The normalized temporal distance is defined as

$$\mathrm{dist}^{(k)}(i,j) = \left| r_i^{(k)} - r_j^{(k-1)} \right| \; / \; T, \tag{4}$$

where T is the number of clips at the original temporal resolution. Since the distance is measured in the original-resolution index space and normalized by T, the resulting distance remains comparable across different temporal resolutions.

Let $\mathbf{q}_i$ and $\mathbf{k}_j$ denote the Query and Key vectors associated with clips i and j, respectively. TDCA computes the attention weights as

$$A_{ij}^{(k)} = \mathrm{Softmax}_j \left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h}} - \lambda \, \mathrm{dist}^{(k)}(i,j) \right), \tag{5}$$

where d_h is the feature dimensionality of each attention head, and λ controls the strength of the temporal-distance term. The first term measures feature-based semantic compatibility, whereas the second term reduces the attention logits for cross-resolution clip pairs that are temporally distant on the original-resolution timeline. When TDCA is applied in the reverse direction, the same distance definition is used by swapping the roles of the two temporal resolutions.

Table 1: Comparison with state-of-the-art methods for MR and HD on the QVHighlights (Lei et al., 2021) test set. The best and second-best results are highlighted in bold and underlined, respectively.

Method	MR					HD	
	R1@0.5	R1@0.7	mAP Avg.	mAP@0.5	mAP@0.75	mAP	HIT@1
MCN (Anne Hendricks et al., 2017)	11.41	2.72	10.67	24.94	8.22	-	-
CAL (Escorcia et al., 2019)	25.49	11.54	9.89	23.40	7.65	-	-
XML (Lei et al., 2020)	41.83	30.35	32.14	44.63	31.73	34.49	55.25
Moment-DETR (Lei et al., 2021)	52.89	33.02	30.73	54.82	29.40	35.69	55.60
UMT (Liu et al., 2022)	56.23	41.18	36.12	53.83	37.01	38.18	59.99
MomentDiff (Li et al., 2023)	58.21	41.48	36.84	54.57	37.21	-	-
QD-DETR (Moon et al., 2023b)	62.40	44.98	39.86	62.52	39.88	38.94	62.40
MH-DETR (Xu et al., 2024)	60.05	42.48	38.38	60.75	38.13	38.22	60.51
UniVTG (Lin et al., 2023)	58.86	40.86	35.47	57.60	35.59	38.20	60.96
TR-DETR (Sun et al., 2024)	64.66	48.96	42.62	63.98	43.73	39.91	63.42
CG-DETR (Moon et al., 2023a)	65.43	48.38	42.86	64.51	45.77	40.33	**66.21**
R^2-Tuning (Liu et al., 2024)	<u>68.03</u>	49.35	46.17	<u>69.04</u>	<u>47.56</u>	**40.75**	64.20
CausalVTG (Wang et al., 2026)	66.06	**51.42**	<u>47.40</u>	-	-	39.60	-
SkimVTG (Ours)	**69.47**	<u>50.81</u>	**48.80**	**69.08**	**50.77**	<u>40.61</u>	<u>65.91</u>

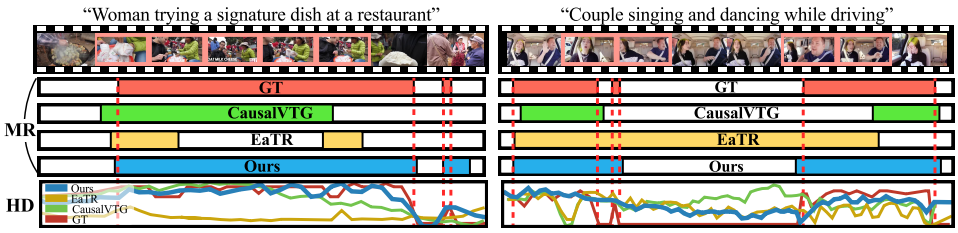


Figure 3: Qualitative comparison of Moment Retrieval on QVHighlights (Lei et al., 2021). Red boxes indicate ground-truth (GT) moments, and red timeline lines denote their boundaries.

Two-Step Cross-Resolution Feature Propagation. TBN takes the BCA output $\mathbf{Z}^{(k)}$ at temporal resolution k and the video representation $\mathbf{V}^{(k-1)}$ at the next finer temporal resolution, and bridges the two resolutions through two TDCA operations. Throughout this section, we write $\mathrm{TDCA}(\cdot,\cdot,\cdot)$ so that the first argument serves as the query and the second and third arguments serve as the key and the value, respectively. The first operation incorporates local temporal details from the finer resolution into the coarser representation, while the second propagates the refined coarser representation to the finer resolution.

In the first TDCA, the coarser representation $\mathbf{Z}^{(k)}$ attends to the finer video representation $\mathbf{V}^{(k-1)}$:

$$\mathbf{M}^{(k)\leftarrow(k-1)} = \mathrm{TDCA}\big(\mathbf{Z}^{(k)},\ \mathbf{V}^{(k-1)},\ \mathbf{V}^{(k-1)}\big). \tag{6}$$

This step refines $\mathbf{Z}^{(k)}$ by incorporating local temporal details from resolution $k-1$. In the second TDCA, the finer video representation $\mathbf{V}^{(k-1)}$ attends to the refined coarser representation $\mathbf{M}^{(k)\leftarrow(k-1)}$:

$$\mathbf{F}^{(k-1)} = \mathrm{TDCA}\big(\mathbf{V}^{(k-1)},\ \mathbf{M}^{(k)\leftarrow(k-1)},\ \mathbf{M}^{(k)\leftarrow(k-1)}\big). \tag{7}$$

This step propagates the refined coarse-level information to resolution $k-1$, yielding $\mathbf{F}^{(k-1)}$.

Through these two TDCA operations, TBN performs cross-resolution propagation: the coarser representation first incorporates local details from the finer resolution, and the finer representation is then updated with refined coarse-level guidance. The resulting $\mathbf{F}^{(k-1)}$ is passed to BCA at resolution $k-1$ for the next stage of progressive boundary refinement.

Table 2: MR performance comparison on Charades-STA (Gao et al., 2017) and TACoS (Regneri et al., 2013).

Method	Charades-STA				TACoS			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN (Zhang et al., 2020b)	58.76	46.02	27.40	41.25	40.01	27.99	12.92	27.22
VSLNet (Zhang et al., 2020a)	60.30	42.69	24.14	41.58	35.54	23.54	13.15	24.99
Moment-DETR (Lei et al., 2021)	65.83	52.07	30.59	45.54	37.97	24.67	11.97	25.49
QD-DETR (Moon et al., 2023b)	-	57.31	32.55	-	52.39	36.77	21.07	35.76
UniVTG (Lin et al., 2023)	70.81	58.01	35.65	50.10	51.44	34.97	17.35	33.60
MomentDiff (Li et al., 2023)	-	55.57	32.42	-	46.64	28.92	12.37	30.36
BAM-DETR (Lee & Byun, 2024)	72.93	59.95	39.38	52.33	56.69	41.54	26.77	39.31
EAMAT (Yang & Wu, 2022)	74.25	61.18	41.72	54.53	62.20	**41.60**	24.14	44.15
SkimVTG (Ours)	**76.52**	**63.48**	**43.48**	**56.57**	**64.59**	41.44	**27.70**	**46.25**

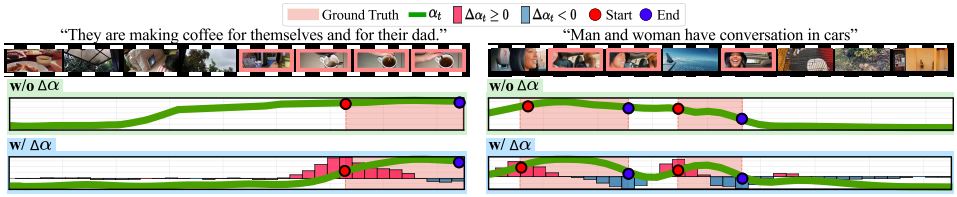


Figure 4: Qualitative comparison of attention scores with and without the temporal-difference cue $\Delta\alpha$ (line plots and bar plots, respectively). The cue highlights start and end regions, yielding attention scores more closely aligned with the ground-truth moments.

3.5 Prediction Heads and Training Objective

After refinement reaches the original temporal resolution, SkimVTG applies HD and MR heads to the boundary-aware representation $\mathbf{Z}^{(1)}$. For HD, per-clip saliency scores are computed by cosine similarity between each clip representation in $\mathbf{Z}^{(1)}$ and the pooled query representation $\mathbf{Q}_{\text{pool}}$, and trained with the SampledNCE loss (Yang et al., 2020) $\mathcal{L}_{hd}$, which contrasts foreground and background clips.

For MR, following (Lin et al., 2023; Liu et al., 2024), we stack 1D convolutional layers on $\mathbf{Z}^{(1)}$ to construct a temporal feature pyramid for moments of varying durations. At each level, parallel foreground and regression heads predict the foreground confidence and the left/right distances to the target start/end boundaries, respectively. The foreground head uses focal loss (Lin et al., 2017) $\mathcal{L}_{fg}$ to address foreground–background imbalance, while the regression head uses L1 loss $\mathcal{L}_{reg}$ only at foreground locations.

The overall objective is

$$\mathcal{L} = \lambda_{hd}\mathcal{L}_{hd} + \lambda_{fg}\mathcal{L}_{fg} + \lambda_{reg}\mathcal{L}_{reg}, \tag{8}$$

where λ_{hd}, λ_{fg}, and λ_{reg} balance the three losses.

4 Experimental Results

4.1 Experimental Setup

Datasets. We evaluate SkimVTG on three widely used VTG benchmarks. QVHighlights (Lei et al., 2021) contains 10,148 vlog and news videos and 10,310 queries, uniquely supporting joint moment retrieval (MR) and highlight detection (HD), with an average of 1.8 relevant moments per query and clip-level saliency annotations. Charades-STA (Gao et al., 2017) contains 9,848 indoor videos and 16,128 query–moment pairs, with average video and moment durations of 30.6 and 8.1 seconds. TACoS (Regneri et al., 2013) contains 18,818 pairs from 127 cooking videos, whose long videos (4.8 minutes) and short moments (6.1 seconds) make precise boundary localization particularly challenging.

Table 3: Component-wise ablation study on QVHighlights (Lei et al., 2021).

Method	MR			HD	
	R1@0.5	R1@0.7	mAP Avg.	mAP	HIT@1
w/o TBN	64.30	48.51	44.78	37.21	64.44
w/o $\Delta\alpha$	64.92	49.12	46.01	38.04	63.26
w/o TDCA	66.21	50.01	47.48	39.50	64.12
w/o $\mathbf{Q}_{\text{pool}}$	<u>67.73</u>	**51.53**	<u>48.48</u>	<u>39.77</u>	<u>64.84</u>
Ours	**69.47**	<u>50.81</u>	**48.80**	**40.61**	**65.91**

Table 4: Effect of the number of temporal resolutions K on QVHighlights (Lei et al., 2021).

K	MR			HD	
	R1@0.5	R1@0.7	mAP Avg.	mAP	HIT@1
1	63.81	47.13	42.11	36.93	64.37
2	<u>67.79</u>	<u>50.06</u>	<u>47.22</u>	<u>39.82</u>	<u>65.74</u>
3	**69.47**	**50.81**	**48.80**	**40.61**	**65.91**
4	66.11	49.02	46.59	37.81	63.73

Evaluation Metrics. We follow the standard protocol for each benchmark. On QVHighlights (Lei et al., 2021), we evaluate both MR and HD. For MR, we report Recall@1 at IoU thresholds of 0.5 and 0.7 (R1@0.5 and R1@0.7), mAP@0.5, mAP@0.75, and mAP Avg. over thresholds from 0.5 to 0.95 in increments of 0.05. For HD, we report mAP and HIT@1, treating clips with the highest saliency label ("Very Good") as positives. On Charades-STA (Gao et al., 2017) and TACoS (Regneri et al., 2013), we report R1 at IoU thresholds of 0.3, 0.5, and 0.7, and the mean IoU (mIoU) of the top-1 prediction.

Implementation Details. Following prior works (Lei et al., 2021; Mun et al., 2020; Zhang et al., 2020a; 2019), we use CLIP (Radford et al., 2021) for text features and concatenate SlowFast (Feichtenhofer et al., 2019) with CLIP ViT-B/32 for video features. Video clips span 2 seconds on QVHighlights and TACoS, and 1 second on Charades-STA. We set $d = 256$, use 8 attention heads and $K = 3$ temporal resolutions, with the choice of K validated in Sec. 4.3. The loss weights are set to $\lambda_{hd} = 0.1$, $\lambda_{fg} = 1.0$, and $\lambda_{reg} = 0.2$. Unless otherwise specified, all models are trained with AdamW on 8 NVIDIA H200 GPUs under the same configuration.

4.2 Comparison with State-of-the-art Methods

Joint MR and HD Evaluation on QVHighlights. Table 1 reports the MR and HD results on QVHighlights (Lei et al., 2021), where SkimVTG ranks first on four of the five reported MR metrics, with the largest margin on mAP@0.75 (+3.21) and consistent gains on mAP Avg. (+1.40) and R1@0.5 (+1.44), while ranking second on R1@0.7 behind CausalVTG by 0.61; for HD, it achieves competitive performance on mAP and HIT@1, showing that the learned representations are also effective for clip-level saliency estimation. Figure 3 further compares its MR predictions with CausalVTG (Wang et al., 2026) and EaTR (Jang et al., 2023) on two examples containing multiple ground-truth moments: existing methods often miss some target moments or predict boundaries that deviate from the ground truth, whereas SkimVTG identifies all target moments with boundaries that closely match the ground truth, suggesting that coarse-to-fine refinement and the $\Delta\alpha$-based boundary cue are effective for localizing moments scattered across the video.

Moment Retrieval (MR) on Charades-STA and TACoS. Table 2 reports the MR results on Charades-STA (Gao et al., 2017) and TACoS (Regneri et al., 2013), where SkimVTG achieves the best performance on most metrics, with particularly notable gains on the strict localization metrics R1@0.7 and mIoU. Unlike QVHighlights, both datasets are annotated with a single ground-truth moment per query, showing that SkimVTG is also effective in the single-moment retrieval setting.

4.3 Ablation Study and Detailed Analysis

Component-wise Ablation Study. Table 3 summarizes the contribution of each component of SkimVTG. Removing the $\Delta\alpha$ branch from BCA (w/o $\Delta\alpha$) reduces BCA to standard cross-attention and causes the reduction in R1@0.5, showing that temporal variation provides information beyond relevance magnitude alone; its effect on start and end boundary localization is examined in the following boundary analysis. Replacing TBN with a simple upsample operation (w/o TBN) degrades all metrics, indicating that cross-resolution propagation requires explicit interaction rather than direct feature reuse, and replacing TDCA with standard cross-attention (w/o TDCA) likewise degrades

Table 5: Boundary localization accuracy on the QVHighlights (Lei et al., 2021) val split. Err. is the mean absolute boundary difference in seconds, and Acc.@5s is the percentage of matched moments whose boundary falls within ± 5 s of the ground truth.

Method	Err. (s) ↓			Acc.@5s (%) ↑		
	Start	End	Avg.	Start	End	Both
w/o TBN	4.47	4.56	4.52	78.32	78.65	67.04
w/o $\Delta\alpha$	4.42	4.08	4.31	79.33	79.42	68.87
SkimVTG	**2.75**	**2.73**	**2.69**	**82.38**	**82.91**	**75.94**

all five metrics, supporting the utility of the explicit temporal distance term in cross-resolution interaction. Replacing $\mathbf{Q}_{\text{pool}}$ with token-level embeddings $\mathbf{Q}$ reduces R1@0.5 and mAP Avg., although it slightly improves R1@0.7; overall, the full model achieves the best R1@0.5 and mAP Avg., demonstrating the overall effectiveness of the proposed design.

Effect of the Number of Temporal Resolutions. Table 4 summarizes the effect of varying the number of temporal resolutions K. The single-resolution baseline ($K = 1$), which performs no temporal subsampling or cross-resolution refinement, yields the lowest overall performance; performance then improves as K increases and peaks at $K = 3$, but decreases at $K = 4$, one possible explanation being that excessive temporal compression leaves too few clips at the coarsest resolution for reliable coarse localization. These results indicate that a moderate hierarchy depth provides an effective balance between temporal compression and fine-grained refinement; therefore, we use $K = 3$ by default.

Boundary Localization Accuracy. To directly evaluate boundary precision beyond IoU-based metrics, we perform one-to-one matching on the QVHighlights *val* split, pairing each ground-truth moment with the highest-IoU prediction and discarding moments with no overlapping prediction. Start Err. and End Err. measure the mean absolute differences in seconds for the corresponding boundaries and Avg. Err. averages the two, while Start Acc.@5s and End Acc.@5s denote the percentages of matched moments within ± 5 s of each boundary, and Both Acc.@5s requires both boundaries to satisfy this tolerance. As shown in Tab. 5, removing either component degrades boundary quality: Avg. Err. increases from 2.69 s to 4.31 s without $\Delta\alpha$ and to 4.52 s without TBN, while Both Acc.@5s drops from 75.94% to 68.87% and 67.04%, respectively. Since both variants share the same prediction heads and differ only in how boundary-sensitive cues are formed and propagated across resolutions, these results isolate the contributions of the temporal-difference cue and of cross-resolution propagation to boundary precision; with the comparable mAP@0.5 but improved mAP@0.75 in Tab. 1, this supports our claim that progressive refinement improves boundary precision rather than coarse localization.

Qualitative Analysis of Temporal Difference. Figure 4 visualizes attention scores with and without the temporal-difference cue $\Delta\alpha$: without the cue, the scores vary smoothly around the target moment, whereas its positive and negative responses appear near the start and end regions, respectively. These examples qualitatively complement the boundary results in Tab. 5.

5 Conclusions and Future Works

We proposed SkimVTG, a progressive boundary refinement framework that constructs multi-resolution video representations and refines temporal boundaries from coarse to fine resolutions. Boundary Cross-Attention (BCA) combines clip–query relevance with its temporal difference $\Delta\alpha$ to capture boundary-sensitive transitions, while the Temporal Bridging Network (TBN) propagates refined representations across adjacent resolutions, preserving temporal coherence. Experiments on QVHighlights, Charades-STA, and TACoS demonstrate strong performance across diverse VTG settings, supported by ablations and quantitative and qualitative boundary analyses. However, the fixed hierarchy depth K and temporal subsampling ratio may limit adaptability to videos and target moments of varying durations. Future work will explore adaptive temporal hierarchies and progressive boundary refinement for long-form video.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) solely for writing assistance. Specifically, we used them to improve the readability of the manuscript, condense text to meet the page limit, suggest a structure for the appendix, and adjust LaTeX formatting of tables and figures. We have not used generative AI tools to generate synthetic data, develop the theoretical framework, formulate or prove mathematical claims, design experiments, implement methods, clean or reformat datasets, analyze data, or interpret results. Translation and qualitative/thematic analysis are not applicable to this work. All AI-assisted edits were reviewed by the authors to ensure that the technical content, claims, and numerical results remained unchanged from the original author-written text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that our results are reproducible. The model architecture is fully described in Sec. 3, including the multi-scale representation construction (Sec. 3.2), Boundary Cross-Attention with the temporal-difference cue (Sec. 3.3, Eqs. 2–3), the Temporal Bridging Network and TDCA (Sec. 3.4, Eqs. 4–7), and the prediction heads and training objective (Sec. 3.5, Eq. 8). Datasets, evaluation protocols, feature extractors, clip lengths, hyperparameters (d, number of heads, K, loss weights), and training hardware are specified in Sec. 4.1. The boundary localization metrics used in Tab. 5 are defined in Sec. 4.3, and the sensitivity of the temporal subsampling stride and IoU thresholds is reported in Appendix B. All experiments use the publicly available `QVHighlights`, `Charades-STA`, and `TACoS` benchmarks with their standard splits. Source code, pretrained checkpoints, and configuration files will be released upon publication.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. 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*, pp. 6202–6211, 2019.

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*, pp. 5267–5275, 2017.

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)*, pp. 13846–13856, October 2023.

Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision*, pp. 220–238. Springer, 2024.

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

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.

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*, 36:65948–65966, 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*, pp. 2794–2804, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

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*, pp. 3042–3051, 2022.

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 *European conference on computer vision*, pp. 421–438. Springer, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

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*, pp. 23023–23033, 2023b.

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*, pp. 10810–10819, 2020.

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*, pp. 8748–8763. PMLR, 2021.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

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, pp. 4998–5007, 2024.

Xiaolong Sun, Le Wang, Sanping Zhou, Liushuai Shi, Kun Xia, Mengnan Liu, Yabing Wang, and Gang Hua. Moment quantization for video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20137–20146, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Qiyi Wang, Senda Chen, and Ying Shen. Causalvtg: Towards robust video temporal grounding via causal inference. *Advances in Neural Information Processing Systems*, 38:103609–103633, 2026.

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)*, pp. 1–8. IEEE, 2024.

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*, pp. 13623–13633, 2023.

Shuo Yang and Xinxiao Wu. Entity-aware and motion-aware transformers for language-driven action localization in videos. *arXiv preprint arXiv:2205.05854*, 2022.

Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, and Jie Tang. Understanding negative sampling in graph representation learning. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1666–1676, 2020.

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*, pp. 6543–6554, 2020a.

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*, pp. 1230–1238, 2019.

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*, volume 34, pp. 12870–12877, 2020b.

A Overview

The supplementary material is organized as follows:

- Section B provides additional quantitative analyses of the temporal subsampling stride and IoU thresholds.
- Section C visualizes how attention scores and their temporal differences evolve across temporal resolutions and analyzes the effect of the temporal-difference cue in BCA.
- Section D presents additional qualitative comparisons with CausalVTG and EaTR on single- and multi-moment queries.
- Section E discusses the limitations of SkimVTG and possible future directions.

B Additional Quantitative Analyses

Effect of the subsampling level. We further analyze the effect of the subsampling level in the proposed coarse-to-fine hierarchy. Table 6 reports the performance of SkimVTG when the subsampling level is varied among 2, 3, and 4. The model achieves the best performance when the subsampling level is set to 2, which we adopt as the default setting in our main experiments. This result suggests that a moderate degree of temporal compression is sufficient to provide a coarse global view of the video while still preserving temporal cues necessary for boundary refinement. In contrast, using more aggressive subsampling may over-compress the video representation, making it harder to recover fine temporal transitions at later refinement stages.

Table 6: Effect of the temporal subsampling stride on `QVHighlights` (Lei et al., 2021).

SubSampling	MR					HD	
	R1@0.5	R1@0.7	mAP Avg.	mAP@0.5	mAP@0.75	mAP	HIT@1
2	**69.47**	**50.81**	**48.80**	**69.08**	**50.77**	**40.61**	**65.91**
3	65.73	50.53	46.48	67.49	48.81	40.01	65.32
4	63.14	47.51	45.80	66.67	46.48	39.70	64.20

Performance across IoU Thresholds. To distinguish coarse moment coverage from precise boundary alignment, we evaluate R1 across IoU thresholds from 0.3 to 0.8 on `QVHighlights` (Table 7). Lower thresholds assess target coverage, while higher thresholds require more accurate start and end boundaries, revealing the effects of the temporal-difference cue and cross-resolution refinement on localization precision.

SkimVTG consistently outperforms the single-resolution and ablated variants across a wide range of IoU thresholds. The improvements are maintained even under high IoU requirements, demonstrating that the refinement improves localization without sacrificing coarse moment coverage.

C Analysis of Progressive Boundary Refinement

Evolution of Attention across Temporal Resolutions. To better understand the progressive refinement process, we visualize the clip–query attention score $\bar{\alpha}^{(k)}$ and its temporal difference $\Delta\alpha^{(k)}$ at each temporal resolution. Figure 5 presents the intermediate responses produced by the default model with $K = 3$, proceeding from the coarsest resolution $k = 3$ to a finer resolution$(k = 2)$.

At the coarsest resolution, the attention score provides a compressed global view and identifies the approximate temporal region associated with the query. Its temporal difference captures broad increasing and decreasing transitions around the candidate moment. Consequently, the attention response becomes progressively more concentrated around the target region, while the positive and negative temporal differences become more closely associated with its start and end boundaries.

Table 7: R1 performance across different IoU thresholds on `QVHighlights` (Lei et al., 2021). Higher thresholds require more precise alignment with the ground-truth boundaries.

Method	R1@IoU					
	0.3	0.4	0.5	0.6	0.7	0.8
$K = 1$	74.14	67.63	63.81	55.95	47.13	40.41
w/o TBN	75.58	68.75	64.30	56.81	48.51	41.57
w/o $\Delta\alpha$	76.09	70.75	64.92	56.95	49.12	41.59
SkimVTG	**81.65**	**75.82**	**69.47**	**60.81**	**50.81**	**44.70**

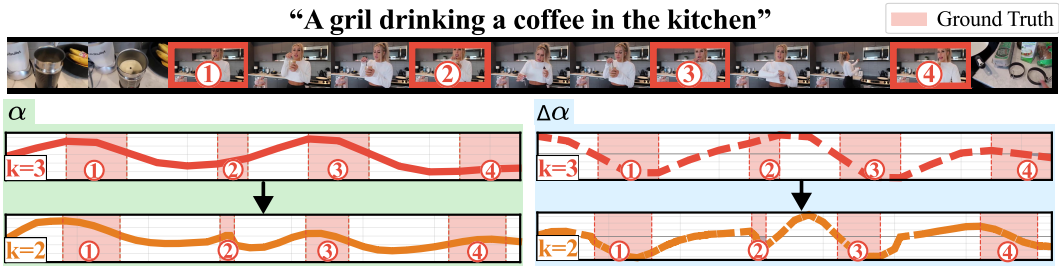


Figure 5: **Evolution of attention scores and temporal differences across temporal resolutions.** We visualize the clip-level attention score and its temporal difference $\Delta\alpha^{(k)}$ from the coarsest resolution to a finer resolution. Coarse representations identify the approximate target region, while subsequent refinement produces attention responses that are more closely aligned with the ground-truth boundaries.

At the original temporal resolution, the refined representation provides the most detailed boundary-related response.

This visualization complements the quantitative analysis of K in the main paper. While increasing the number of temporal resolutions provides additional opportunities for refinement, excessive compression may remove temporal details at the coarsest resolution. The default setting of $K = 3$ provides a balance between global localization and fine-grained boundary refinement.

Effect of the Temporal-Difference Cue on Attention. Building on the resolution-wise analysis above, we isolate the contribution of the temporal-difference cue by comparing BCA with and without $\Delta\alpha$. Figure 6 and 7 visualize the resulting attention scores for the same input videos and queries.

Without $\Delta\alpha$, BCA relies only on the magnitude of clip–query relevance. The resulting attention scores tend to vary smoothly over temporally adjacent clips and often remain broadly activated around the target region. Such responses are sufficient for identifying the approximate moment but provide limited evidence for distinguishing its exact start and end boundaries. When the signed temporal difference is incorporated, increasing and decreasing relevance transitions provide complementary cues for entering and leaving the target moment. The resulting attention scores become more concentrated around the relevant regions and more clearly separated from the surrounding context.

Together with the boundary localization results in the main paper, these examples show that $\Delta\alpha$ does not merely rescale the attention response. Instead, it provides transition-sensitive information that guides BCA toward representations better suited for precise boundary localization.

D Additional Qualitative Results

Fig. 8 provides additional qualitative comparisons with CausalVTG (Wang et al., 2026) and EaTR (Jang et al., 2023). We visualize examples containing both a single target moment and mul-

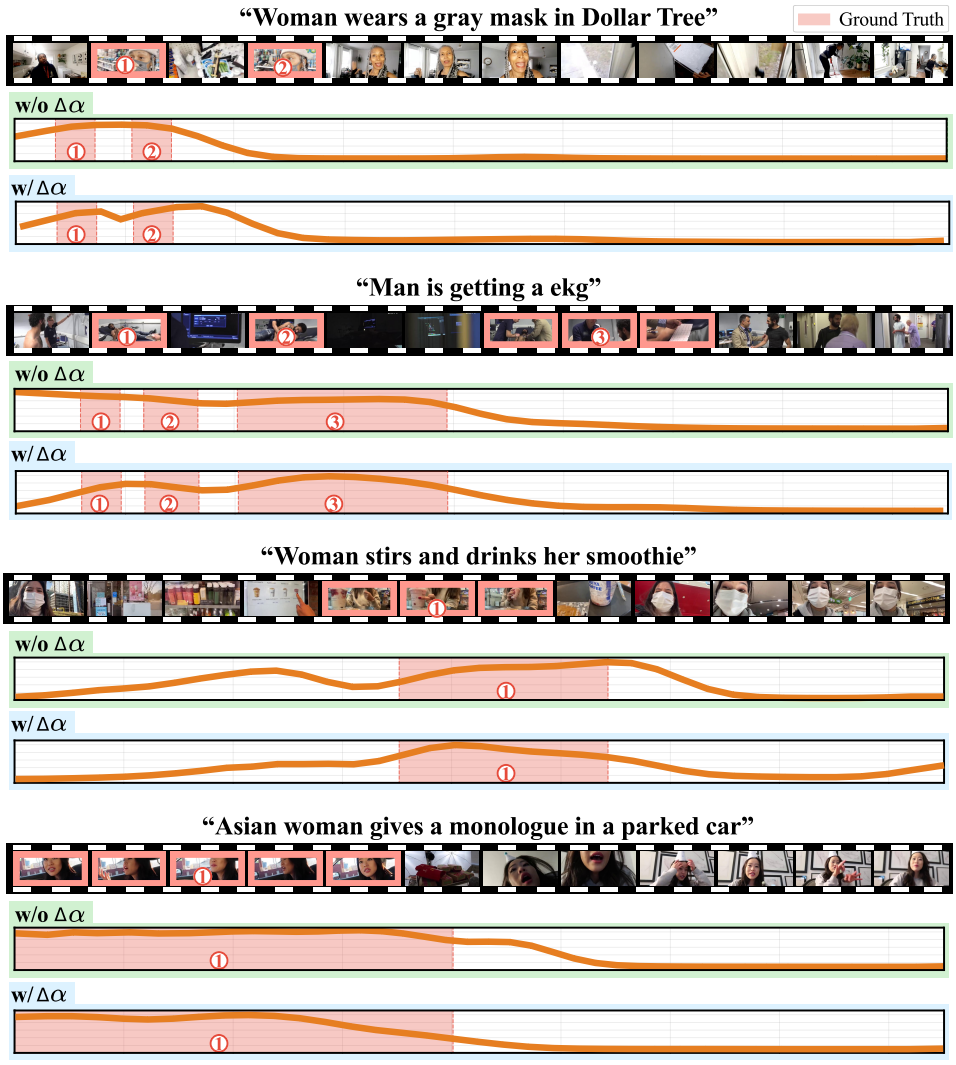


Figure 6: **Qualitative comparison of attention scores with and without the temporal-difference cue.** Compared with the model without temporal difference, incorporating the temporal-difference cue produces attention scores that are more clearly concentrated around the ground-truth moments, indicating its effectiveness for boundary-aware localization.

tiple target moments to evaluate whether each method can accurately localize different temporal structures. Compared with the baselines, SkimVTG produces predictions that are more closely aligned with the ground-truth boundaries in both settings.

For queries associated with a single target moment, SkimVTG localizes the relevant segment more precisely, reducing boundary deviations at the start and end of the moment. For queries with multiple target moments, SkimVTG more consistently identifies the relevant intervals while suppressing irrelevant surrounding clips. These results demonstrate that the proposed coarse-to-fine boundary refinement is effective not only for simple single-moment cases but also for more challenging cases.

E LIMITATION

Although SkimVTG achieves strong performance across multiple VTG benchmarks, it still has several limitations. First, the proposed coarse-to-fine refinement relies on temporally subsampled representations to obtain a compressed global view of the video. While this design is effective for

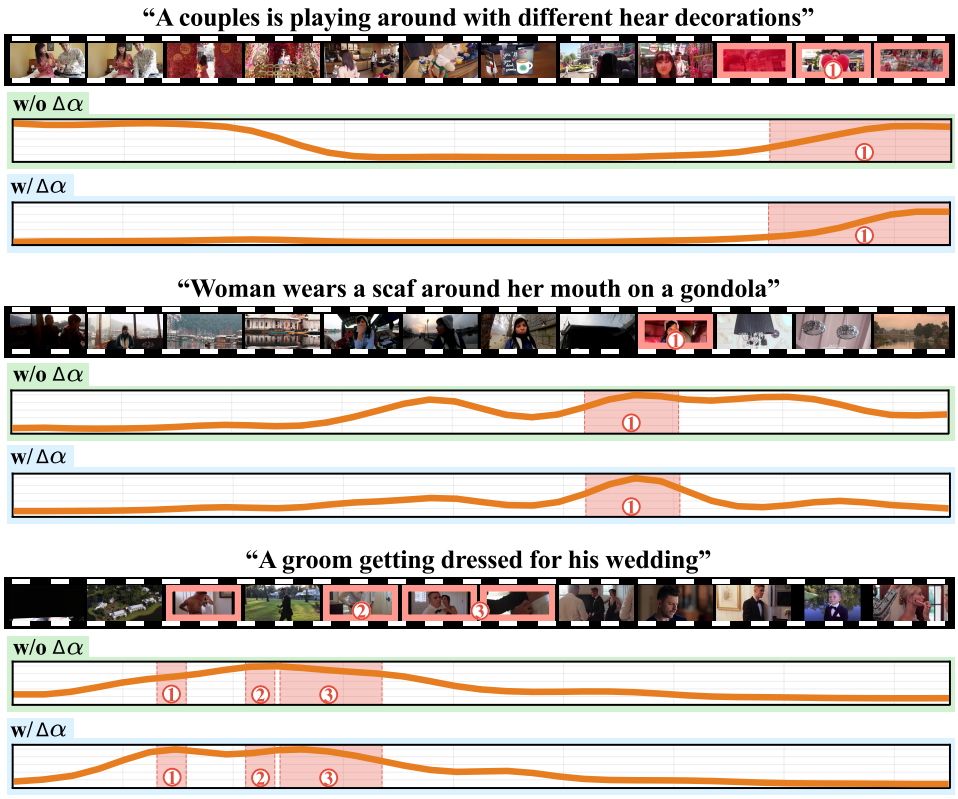


Figure 7: **Qualitative comparison of attention scores with and without the temporal-difference cue.** Compared with the model without temporal difference, incorporating the temporal-difference cue produces attention scores that are more clearly concentrated around the ground-truth moments, indicating its effectiveness for boundary-aware localization.

many cases, aggressive temporal compression may discard subtle visual changes that are important for localizing very short or visually ambiguous moments. As a result, performance may degrade when the target moment is brief or when its boundary is defined by fine-grained motion patterns.

Second, SkimVTG models boundary-sensitive cues using the temporal variation of clip–query relevance scores. This assumes that the relevance profile changes noticeably around the start and end of the target moment. However, in cases where the foreground and surrounding background are semantically similar, the relevance change may be weak, making boundary localization more challenging.

Finally, the proposed framework performs refinement over multiple temporal resolutions, which introduces additional computation compared with single-resolution encoder-only methods. Although the coarse-to-fine design improves boundary localization accuracy, future work may explore more efficient resolution selection or adaptive refinement strategies to reduce computational overhead.

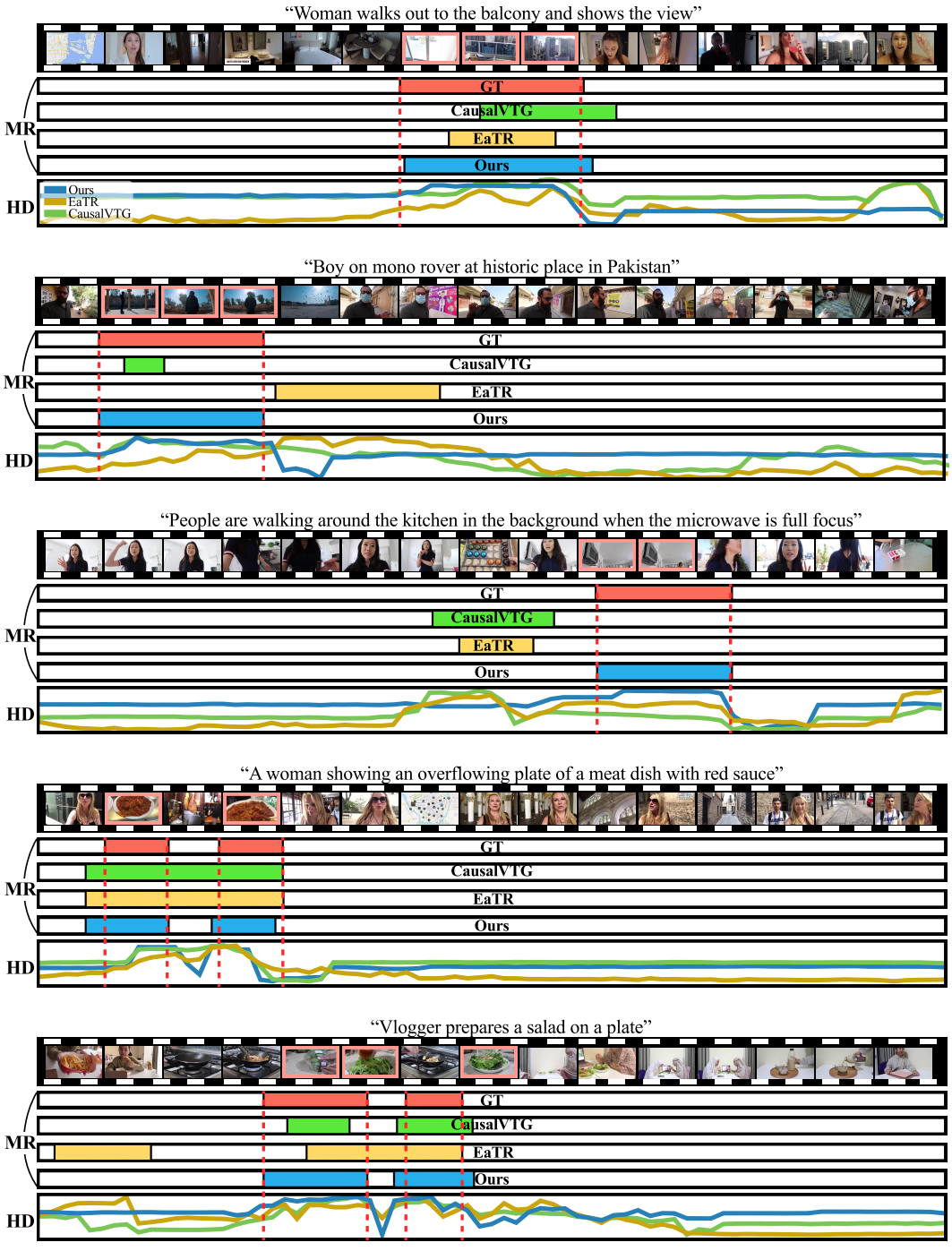


Figure 8: **Qualitative comparison of Moment Retrieval on the QVHighlights (Lei et al., 2021).** Red boxes on the video frames indicate the ground-truth (GT) moments, and red lines on the timeline denote their boundaries.