# EVIDENT: Routing MLLM Adaptation through Entity-Grounded Visual Evidence for Cross-Domain Video Temporal Grounding

**Geo Ahn**[*]
Kyung Hee University
ahngeo11@khu.ac.kr

**Jiwook Han**[*]
Kyung Hee University
mreraser@khu.ac.kr

**Youngrae Kim**[*]
University of Southern California
youngrae@usc.edu

**Joonseok Lee**
Seoul National University, Hanwha Aerospace
joonseok@snu.ac.kr

**Jinwoo Choi**[†]
Kyung Hee University
jinwoochoi@khu.ac.kr

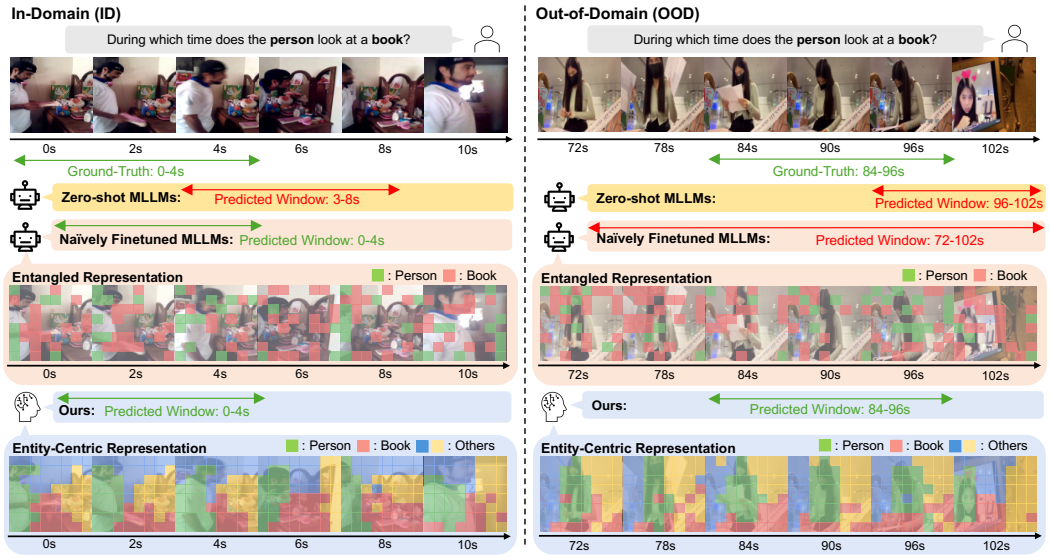



Figure 1: **Teaser.** In this work, we tackle the domain generalization problem of MLLM-based VTG. We observe that naïvely fine-tuned MLLMs learn entangled visual representations that rely on dataset-specific scene context rather than the actual visual content, indiscriminately responding to memorized patterns across frames in OOD videos. In contrast, our method learns an *entity-centric* visual representation that regularizes VTG fine-tuning to ground its predictions in the *visual evidence* of the video, yielding transferability to OOD domains. As a result, ours successfully captures both *person* and *book* consistently across frames and enables correct grounding regardless of the domain.

## Abstract

Fine-tuning MLLMs for Video Temporal Grounding (VTG) often improves in-domain performance but degrades sharply under domain shift. In this work, we find that this failure is primarily driven not just by unseen query concepts, but by *visual domain shift*, which prevents the model from coupling its learned temporal localization knowledge with its inherent entity-attention capability. To address this, we introduce EVIDENT, a parameter-efficient adaptation framework that anchors temporal grounding in the inherent entity-attention of pre-trained MLLMs by routing

[*]Equal contribution.
[†]Corresponding author.

VTG adaptation through *explicit visual entity evidence*. EVIDENT consists of three components: (i) an Entity Bottleneck Adapter that transforms dense visual tokens into compact entity-level slots, (ii) an Entity-Binding Distillation loss that instills objectness priors into the semantically unstructured MLLM visual space, guiding each slot to bind to a coherent entity, and (iii) an Entity-to-eVidence gating mechanism that leverages the captured entities as evidence, steering the model to localize moments containing query-relevant entities. Together, these components enable VTG fine-tuning to rely on entity-grounded evidence rather than brittle dataset shortcuts. Experiments on cross-domain VTG benchmarks show that EVIDENT consistently improves out-of-domain robustness while preserving competitive in-domain performance with modest parameter overhead. These results suggest that entity-level grounding is an effective inductive bias for generalizable temporal localization. The code is available at https://github.com/KHU-VLL/EVIDENT.

## 1 Introduction

Video Temporal Grounding (VTG) requires localizing the temporal extent of an event described by a natural language query in an untrimmed video. While the task has been studied extensively with specialist temporal localization models such as DETR-based architectures [27, 40, 34, 49], Multimodal Large Language Models (MLLMs) have recently emerged as a compelling alternative [19, 46, 15, 52, 38, 60]. The appeal of MLLMs is not merely their scale. Through large-scale image-text and video-text pre-training, they acquire broad visual-language alignment that can associate language queries with relevant visual entities across diverse video domains.

However, VTG exposes a critical *adaptation paradox*. Before VTG-specific fine-tuning, pretrained MLLMs exhibit coarse but domain-agnostic *entity-attention capability*: given a query, they can often attend to semantically relevant visual content across different visual distributions (Figure 2 (a)). Yet this zero-shot attention is too coarse for VTG, which demands exact start and end timestamps, making task-specific fine-tuning necessary [19, 46, 52, 60, 38]. However, the extremely high cost of VTG annotations limits the scale and diversity of supervision, leaving source-domain fine-tuning vulnerable to shortcut learning—including temporal location bias [9, 44, 16], query text bias [9, 30, 21], and appearance bias [7, 45]. These shortcuts cause fine-tuning to rely on source-specific patterns rather than the underlying entity-attention, so the learned localization knowledge fails to inherit the MLLM's domain-agnostic entity-attention. As a result, fine-tuning improves In-Domain (ID) localization but harms generalization under Out-of-Domain (OOD) scenarios where shortcuts no longer hold (Figure 1, Figure 2 (b)). We call this phenomenon *attention-localization decoupling* in OOD VTG.

In this work, we find that this decoupling is primarily driven by *visual domain shift*. We decompose the OOD gap into two factors: a *concept domain gap*, which captures whether the query involves subjects or actions unseen during source training, and a *visual domain gap*, which measures how visually distant the OOD videos are from the source distribution. Splitting OOD samples by query-concept overlap yields only a marginal performance difference (Figure 3 (a)). In contrast, splitting them by visual similarity to the source domain yields a substantially larger gap (Figure 3 (b)). These observations indicate that OOD degradation stems primarily from *visual domain shift*.

These findings point to a different *design principle* for cross-domain MLLM-based VTG: robust adaptation should anchor temporal localization in the *inherent entity-attention* of the pre-trained MLLM, rather than merely fitting dense visual tokens to source-domain scene context. To this end, we propose **EVIDENT**, a parameter-efficient framework for **EVID**ence-grounded **ENT**ity-attentive adaptation for cross-domain VTG. Our goal is to learn transferable VTG knowledge by deriving temporal grounding predictions from *query-relevant visual entity evidence*. Realizing this principle requires three questions: (i) how to route adaptation through entities, (ii) what makes the entities meaningful, and (iii) how to leverage this entity information as evidence for temporal localization.

EVIDENT answers these questions with three key components. First, we propose an **Entity Bottleneck Adapter (EB Adapter)**, which routes intermediate MLLM dense visual tokens through a compact set of entity slots to obtain domain-invariant representations. Unlike prior work [55, 10, 29] that re-runs instruction tuning to reshape MLLM representations into an entity-level form, our adapter design enables VTG learning grounded in entity-based visual representations at minimal cost. Second, we incorporate an **Entity-Binding Distillation (EB Distillation)** loss to encourage each slot to bind to a *coherent* entity. We find that the pre-trained MLLM visual space alone has entangled representations and thus struggles to encode distinct entity information (Figure 1, Figure 5). In contrast, EVIDENT

trained with our EB Distillation loss captures semantically coherent entities even under domain shift, as illustrated in Figure 1. Third, to leverage the entity information captured by EB Adapter with EB Distillation as *explicit visual evidence* guiding temporal localization predictions, we introduce an **Entity-to-eVidence (E2V) gating** mechanism. Using the entity-attention of the pre-trained MLLM, we score each frame for the presence of query-relevant entities and modulate EB Adapter's per-frame contribution accordingly. This injects an inductive bias that the target moment is where the query entities co-occur, enabling evidence-grounded localization under OOD shifts.

We evaluate EVIDENT under cross-domain VTG settings, training on a single source dataset and evaluating on different target datasets [14, 27, 17]. Across standard benchmarks, EVIDENT consistently improves OOD robustness while maintaining competitive ID performance with modest parameter and memory overhead. These results support our central claim: for robust MLLM-based VTG under visual domain shift, what is needed is evidence-grounded adaptation that fully leverages the pre-trained entity-attention, not direct fitting to source-domain context.

We summarize our contributions as follows:

- Through systematic analyses, we find that naïvely fine-tuning MLLMs causes *attention-localization decoupling*—the learned temporal localization fails to inherit the pre-trained entity-attention capability—and show that it is primarily driven by *visual domain shift*.
- We propose EVIDENT, a parameter-efficient framework for evidence-grounded entity-attentive adaptation in cross-domain VTG. Our entity bottleneck adapter with entity-binding distillation reshapes MLLM features into *entity-centric* representations without re-running instruction tuning, while the proposed entity-to-evidence gating leverages these entities as *explicit visual evidence for* temporal localization.
- We show that our evidence-grounded VTG adaptation yields consistent OOD robustness gains on diverse cross-domain VTG settings, while maintaining competitive in-domain performance with a modest number of parameters.

## 2 Related Work

**Video Temporal Grounding** Video Temporal Grounding (VTG) [61] aims to localize the temporal segments in an untrimmed video that correspond to a natural language query. Recent work [19, 46] reformulates VTG as a generative task on MLLMs that produce temporal boundaries as text tokens via task-specific instruction tuning. Follow-up work refines this formulation through interleaved frame–timestamp representations [38, 33], causal event modeling [15], grounded tuning for long videos [60], chain-of-LoRA reasoning [35], and reinforcement learning with verifiable temporal rewards [51], all primarily relying on large-scale instruction tuning. While these efforts enhance the temporal capability of MLLMs, the fine-tuning step that is inevitably required for accurate localization remains an unexamined source of risk: these methods do not directly address the OOD generalization failure that this fine-tuning induces, which is the main issue we study.

**Object-centric Learning** Object-centric learning represents a scene as a composition of entity-level parts, yielding transferable representations under domain shifts. Slot attention [36] is a representative mechanism, where learnable slots iteratively compete to explain the input, and has been extended to real-world images [47, 26] and videos [23, 53, 48, 54]. Recent work brings this idea into MLLMs: Slot-VLM [55] splits video tokens into object- and event-centric slots, Slot-MLLM [10] combines a Q-Former with slot attention for unified multimodal generation, and CORE [29] compresses visual tokens into a compact set of slots for efficient MLLM inference, using segmentation masks from an off-the-shelf model as an object-centric prior. However, all of these approaches require training the entire vision-language pipeline from scratch, whereas EVIDENT introduces a parameter-efficient EB Adapter attachable to pre-trained MLLMs, bringing entity-centric structure while preserving the pretrained vision-language alignment.

## 3 Problem Analysis

To understand why naïvely fine-tuned MLLMs suffer under domain shift, we conduct a series of empirical analyses. We fine-tune Qwen2.5-VL-3B [6] and InternVL3-2B [63] on Charades-STA (Cha.) [14] and evaluate on QVHighlights (QVH) [27]. See the appendix for the InternVL3 [63] results and more details.

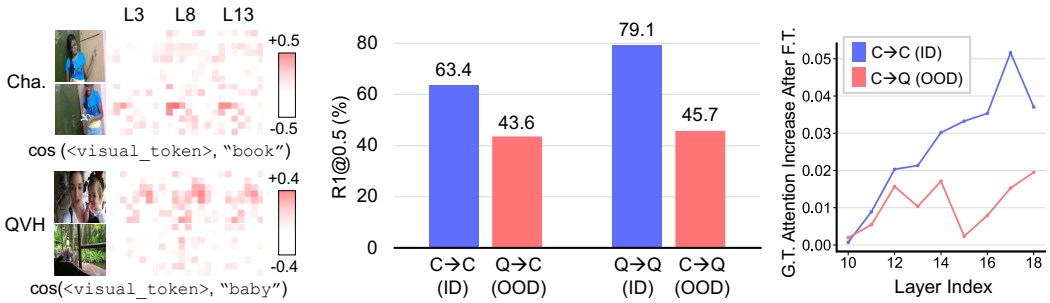



Figure 2: **Naïve fine-tuning breaks generalization.** (a) Cosine similarity between visual patches and object text tokens from Qwen2.5-VL-3B [6] on Charades-STA (Cha.) [14] and QVHighlights (QVH) [27]. Even zero-shot, the MLLM exhibits strong visual-text alignment regardless of video domain. (b) When fine-tuned on a specific domain, the model suffers severe OOD drops. (c) The model's attention on GT-interval frames relative to all frames increases substantially on the source domain (ID) but only marginally on the unseen domain (OOD), indicating that the learned VTG capability fails to transfer to other domains.

### 3.1 OOD Failure Observations

**Pre-trained MLLMs exhibit strong entity-attention capability.** We first measure the visual-text alignment of the pre-trained Qwen2.5-VL-3B [6] by computing the cosine similarity between visual patches and object text tokens. As shown in Figure 2 (a), the pre-trained model exhibits strong visual-text alignment across visually distinct domains (Cha. [14] and QVH [27]), with high similarity between object text tokens (e.g., 'book') and the corresponding visual patches at the frame level. This indicates that the inherent *entity-attention* ability of pre-trained MLLMs to associate language queries with relevant visual content already transfers across diverse video distributions.

**Naïve fine-tuning causes severe OOD performance degradation.** While the pre-trained model shows transferable alignment, naïvely fine-tuning it on a single source dataset breaks this transferability (Figure 2 (b)). Fine-tuning on QVH [27] produces a 19.8-point ID-OOD gap on Cha. [14], and fine-tuning on Cha. [14] produces a 33.4-point gap on QVH [27]. The fine-tuned model overfits to source-domain patterns, failing to generalize to unseen domains.

**Acquired grounding capability fails to transfer.** To probe the source of this collapse, we measure how strongly the fine-tuned model's visual-to-query cross-attention shifts toward ground-truth-interval frames, relative to the zero-shot baseline (Figure 2 (c)). We focus on layers 10–18, following prior findings that visual-language interaction in VideoLLMs primarily occurs at intermediate layers [22]. Fine-tuning shifts attention toward ground-truth segments substantially on the trained domain (ID) but only marginally on the unseen domain (OOD), revealing that the learned localization does not transfer across domains.

### 3.2 Domain Gap Diagnosis

We next ask: *what is the primary cause of this transfer failure?* We decompose the domain gap into two sources, a *concept domain gap* (whether OOD samples involve query concepts unseen during training) and a *visual domain gap* (how visually distant OOD samples are from the source distribution), and analyze which one dominates.

**Concept gap does not explain OOD failure.** We split the OOD test set (QVHighlights [27]) by query-concept overlap with ID (Charades-STA [14]), separating samples whose query concepts appear in the source dataset from those that do not. As shown in Figure 3 (a), the performance gap between the two splits is marginal (3.3 points), indicating that pre-trained MLLMs already cover a broad range of concepts and the concept gap is not the dominant factor in OOD failure.

**Visual gap is the primary cause of OOD degradation.** We instead split OOD samples by their visual feature similarity to the training set, where features are extracted from the vision encoder. As shown in Figure 3 (b), the most similar 20% of OOD samples achieve 52.8 R1@0.5, while the most dissimilar 20% drop to 39.1—a 13.7-point gap that scales with visual distance. This identifies the visual domain gap as the primary cause of OOD performance degradation.

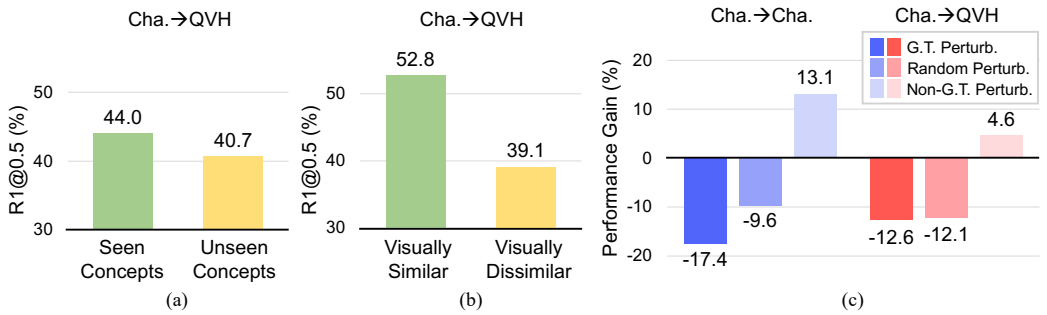



Figure 3: **Visual domain gap dominates.** We decompose the domain gap into *concept domain gap* and *visual domain gap*. (a) Splitting OOD samples by query-concept overlap with ID yields only a marginal gap. (b) Ranking OOD samples by visual similarity to ID, the top-20% most similar samples substantially outperform the bottom-20%, identifying the *visual domain gap* as the primary cause of degradation. (c) On ID, perturbing the G.T. interval causes a much larger drop than perturbing a random non-G.T. interval of the same length, while on OOD the two yield nearly identical drops.

**The model ignores visual content under domain shift.** To verify whether the model genuinely grounds its predictions in visual content, we inject Gaussian noise into either the ground-truth (G.T.) interval or a random non-G.T. interval of the same length and measure R1@0.7 drop. As shown in Figure 3 (c), on ID, G.T. perturbation causes a much larger drop than random (17.4% vs. 9.6%), while on OOD the two are nearly identical (12.6% vs. 12.1%)—the model no longer attends to the actual visual content under distribution shift. Conversely, injecting noise into the entire non-G.T. interval improves performance on both ID and OOD (+13.1% and +4.6%). It suggests that explicitly steering the model toward the visual content of the G.T. interval recouples its learned localization with the pre-trained entity-attention disrupted by visual domain shift, yielding gains even under OOD.

# 4 Method

We introduce **EVIDENT**, a parameter-efficient framework for cross-domain VTG that anchors temporal grounding in the inherent entity-attention of a pre-trained MLLM. EVIDENT builds entity-level structure in the early decoder layers, so that the later layers can apply this entity-attention without relying on domain-specific scene context. To this end, the Entity Bottleneck Adapter (**EB Adapter**, Section 4.2) routes dense visual tokens into entity slots, and the Entity-Binding Distillation (**EB Distillation**, Section 4.3) loss supplies the entity-level prior that the MLLM's visual space lacks. Since slot attention operates only on visual tokens, the slots focus on separating visual content at the entity level, while query relevance is mainly handled by the pre-trained entity-attention through two paths: Entity-to-eVidence gating (**E2V gating**, Section 4.4), which scores each frame by visual-text similarity to the query entities and modulates the EB Adapter output accordingly, and the vision-language integration in the middle and later layers [22]. Figure 4 provides an overview.

## 4.1 VTG Task Formulation

**Interleaved Timestamps.** We follow the generative Video Temporal Grounding (VTG) paradigm, where an MLLM directly generates target timestamps as text tokens. Given $T$ uniformly sampled video frames and a natural language query, we encode each frame into $N$ visual tokens $\mathbf{f}_t \in \mathbb{R}^{N\times D}$ via a frozen vision encoder and a linear projector, where $D$ is the hidden dimension of the LLM decoder, and tokenize its timestamp into a short text sequence $\tau_t$. The input $\mathbf{X}$ to the LLM decoder is constructed by interleaving each frame's visual tokens with its timestamp tokens, followed by the query tokens $\mathbf{q}$:

$$\mathbf{X} = [\mathbf{f}_1, \tau_1, \mathbf{f}_2, \tau_2, \ldots, \mathbf{f}_T, \tau_T, \mathbf{q}], \tag{1}$$

This interleaved layout has been shown to be effective for temporal grounding [38, 33, 62]. The model autoregressively decodes the target temporal window $[t_{\text{start}}, t_{\text{end}}]$.

## 4.2 Entity Bottleneck Adapter

We propose Entity Bottleneck Adapter (EB Adapter), an adapter inserted into the MLLM decoder layers that reorganizes dense visual tokens into entity-based representations using a compact set

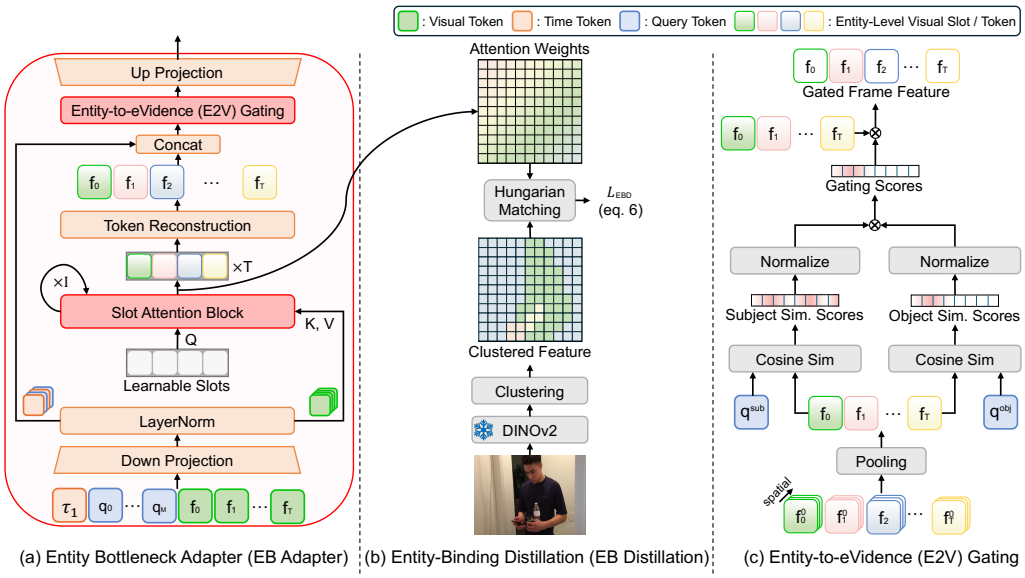



Figure 4: **Overview. (a) Entity Bottleneck Adapter (EB Adapter).** A lightweight adapter in the early LLM decoder layers decomposes visual tokens into entity-level slots via slot attention. **(b) Entity-Binding Distillation (EB Distillation).** Entity cluster maps derived from DINOv2 [43] patch features are distilled into slot assignment maps via BCE loss, encouraging semantically coherent slot formation. **(c) Entity-to-eVidence (E2V) Gating.** A per-frame gating score, reflecting the co-occurrence of query-relevant entities, modulates the adapter's influence.

of abstract slots. By leaving the original vision-language inference path and instruction alignment intact, EB Adapter *preserves the pre-trained entity-attention of the MLLM* while replacing only the domain-fragile dense visual representations with robust slot-based ones at minimal cost.

**Down Projection.** We first project the entire input sequence $\mathbf{X}$ into a lower-dimensional bottleneck space: $\mathbf{X}_{\text{down}} = \mathbf{X}\mathbf{W}_{\text{down}} \in \mathbb{R}^{L\times d}$, where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D\times d}$ and $d \ll D$. From $\mathbf{X}_{\text{down}}$, we denote the visual portion as $\mathbf{X}^{\text{vis}}_{\text{down}} = [\mathbf{f}'_1, \ldots, \mathbf{f}'_T] \in \mathbb{R}^{T\times N\times d}$ and the text portion as $\mathbf{X}^{\text{txt}}_{\text{down}} = [\mathbf{t}'_1, \ldots, \mathbf{t}'_T, \mathbf{q}']$.

**Slot Attention Block.** A set of $N_s$ learnable slot queries $\mathbf{S}^{(0)} \in \mathbb{R}^{N_s\times d}$ attend to the down-projected visual tokens $\mathbf{f}'_t \in \mathbb{R}^{N\times d}$ of each frame independently through $I$ iterations. At iteration $i \in \{1, \ldots, I\}$, we project the slots and the per-frame visual tokens into a common space: $\mathbf{Q}^{(i)}_t = \mathbf{S}^{(i-1)}_t \mathbf{W}_Q \in \mathbb{R}^{N_s\times d}$, $\mathbf{K}_t = \mathbf{f}'_t \mathbf{W}_K \in \mathbb{R}^{N\times d}$ , and $\mathbf{V}_t = \mathbf{f}'_t \mathbf{W}_V \in \mathbb{R}^{N\times d}$, where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the projection matrices shared across frames. Next, the attention scores for frame $t$ at iteration $i$ are computed as:

$$\mathbf{M}^{(i)}_t = \mathbf{K}_t (\mathbf{Q}^{(i)}_t)^\top / \sqrt{d} \in \mathbb{R}^{N\times N_s}. \tag{2}$$

We normalize $\mathbf{M}^{(i)}_t$ along the *slot axis* via softmax, fostering competitive assignment of visual tokens to slots within each frame:

$$\mathbf{A}^{(i)}_t(n,k) = \frac{\exp(\mathbf{M}^{(i)}_t(n,k))}{\sum_{j=1}^{N_s} \exp(\mathbf{M}^{(i)}_t(n,j))}. \tag{3}$$

We then normalize $\mathbf{A}^{(i)}_t$ along the token axis such that $\hat{\mathbf{A}}^{(i)}_t(\cdot,k)$ sums to one. The aggregated update is computed as a weighted mean $\mathbf{Z}^{(i)}_t = (\hat{\mathbf{A}}^{(i)}_t)^\top \mathbf{V}_t \in \mathbb{R}^{N_s\times d}$, and the slots are then updated via a Gated Recurrent Unit (GRU) [11] based recurrence: $\mathbf{S}^{(i)}_t = \text{GRU}(\mathbf{S}^{(i-1)}_t, \mathbf{Z}^{(i)}_t)$. This competition mechanism encourages each slot to specialize in a distinct semantic entity.

**Token Reconstruction.** Since the LLM decoder expects the original token sequence length, we reconstruct the visual tokens by reusing the slot attention's assignment in a parameter-free manner. This design enforces that every reconstructed visual token is expressed only as a combination of slot features, thereby constraining the representations consumed by the LLM decoder to strictly adhere to *disentangled entity-level semantics*.

Given the final slot representations $\mathbf{S}_t^{(I)} \in \mathbb{R}^{N_s \times d}$ for frame $t$, we reconstruct token-level features as a weighted sum of slot features based on the slot attention's assignment weights:

$$\tilde{\mathbf{f}}'_t = \hat{\mathbf{A}}_t^{(I)} \mathbf{S}_t^{(I)} \in \mathbb{R}^{N \times d}. \tag{4}$$

We empirically find that using $\hat{\mathbf{A}}$ instead of $\mathbf{A}$ performs slightly better (see Section D.1). We then concatenate the reconstructed visual portion $\tilde{\mathbf{X}}^{\text{vis}}_{\text{down}} = [\tilde{\mathbf{f}}'_1, \ldots, \tilde{\mathbf{f}}'_T]$ with the unchanged text portion $\mathbf{X}^{\text{txt}}_{\text{down}}$ and project the result back to the original dimension via an up projection. The adapter output is added to the original input via a residual connection:

$$\mathbf{X}_{\text{out}} = \mathbf{X} + \texttt{concat}(\tilde{\mathbf{X}}^{\text{vis}}_{\text{down}}, \mathbf{X}^{\text{txt}}_{\text{down}})\mathbf{W}_{\text{up}}, \tag{5}$$

where $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d \times D}$ is initialized to zero, so the adapter acts as an identity mapping at the start of training. This ensures training stability while gradually steering the representations toward entity-level decomposition [18, 57].

**Early-Layer Insertion.** We attach the EB Adapter only to the early decoder layers. Recent findings [22] show that cross-frame interactions occur in these early layers, while deeper layers handle language integration and answer generation. By inserting the EB Adapter at this stage, each slot captures temporally coherent semantics across frames rather than frame-independent decompositions. The deeper layers, fine-tuned with LoRA [18, 57], then reason over these disentangled representations.

### 4.3 Entity-Binding Distillation

While the EB Adapter encourages decomposition through its bottleneck structure, the bottleneck alone is insufficient: the entangled visual space of the naïve MLLM lacks the entity-level structure for slots to bind to. As a result, slots fail to specialize and produce near-uniform attention weights over visual tokens, as shown in Section 5.4. To address this, we employ an Entity-Binding Distillation (EB Distillation) loss. Inspired by prior works on object-centric learning [47, 2], our loss distills the objectness prior from a self-supervised vision model to encourage semantically coherent slot formation, as illustrated in Figure 4 (b).

**Entity-level Matching.** We extract DINOv2 [43] patch features from $T$ frames, apply $2 \times 2$ average pooling to match the MLLM's spatial resolution, and $\ell_2$-normalize along the feature dimension. We then perform $K$-means clustering ($K = N_s$) on each frame independently, yielding a binary cluster map $\mathbf{C} \in \{0,1\}^{T \times N_s \times N}$, where $\mathbf{C}_{t,k,n} = 1$ indicates that patch $n$ in frame $t$ belongs to cluster $k$. Since slot and cluster indices are arbitrary, we perform per-frame Hungarian matching [24] between the normalized slot attention maps $\hat{\mathbf{A}} \in [0,1]^{T \times N_s \times N}$ and $\mathbf{C}$, using patch-averaged BCE as the cost, to obtain an optimal one-to-one matching $\pi^{(t)} : \{1, \ldots, N_s\} \to \{1, \ldots, N_s\}$.

**Loss.** Given the matching, we define the alignment loss as the BCE between matched slot–cluster pairs, averaged over all frames, slots, and patches:

$$L_{\text{EBD}} = -\frac{1}{TN_sN} \sum_{t=1}^{T} \sum_{k=1}^{N_s} \sum_{n=1}^{N} \left[ \mathbf{C}^{(t)}_{\pi^{(t)}(k),n} \log \hat{\mathbf{A}}^{(t)}_{k,n} + \left(1 - \mathbf{C}^{(t)}_{\pi^{(t)}(k),n}\right) \log \left(1 - \hat{\mathbf{A}}^{(t)}_{k,n}\right) \right]. \tag{6}$$

### 4.4 Entity to Evidence Gating

We introduce an Entity-to-eVidence (E2V) gating mechanism to leverage the entities captured by EB Adapter as explicit evidence for temporal grounding, injecting a VTG inductive bias that *temporal segments where query-relevant entities co-occur are more likely to be the target moment*. Specifically, we use the MLLM's pretrained entity-attention to score per-frame entity presence and modulate EB Adapter's output accordingly, as shown in Figure 4 (c).

**Entity Extraction from Query.** For each natural language query, we extract the subject and object concept words as query-relevant entities. We perform this offline using a pre-trained LLM (Qwen3-4B [56]) as a preprocessing step. We denote the extracted indices in the query token sequence as $i_{\text{sub}}$ and $i_{\text{obj}}$. See Section C.1 for more details. Using these indices, we select the corresponding query tokens $\mathbf{q}'_{\text{sub}}, \mathbf{q}'_{\text{obj}} \in \mathbb{R}^d$ from the down-projected text portion $\mathbf{X}^{\text{txt}}_{\text{down}}$.

**Frame-level Co-occurrence Score.** We compute the cosine similarity between each frame's averaged visual token and the subject/object query tokens: $s_t^{\text{sub}} = \cos(\bar{\mathbf{f}}_t, \mathbf{q}'_{\text{sub}})$, $\quad s_t^{\text{obj}} = \cos(\bar{\mathbf{f}}_t, \mathbf{q}'_{\text{obj}})$, where $\bar{\mathbf{f}}_t \in \mathbb{R}^d$ is a per-frame visual representation obtained by averaging the $N$ down-projected visual tokens

Table 1: **Performance comparison on cross-domain video temporal grounding benchmarks.** We evaluate EVIDENT on Charades-STA [14], QVHighlights [27], and DiDeMo [17], reporting both In-Domain (ID) (source=target) and Out-of-Distribution (OOD) (source≠target) settings. † denotes our implemented baseline fine-tuned with LoRA [18]. Our results are highlighted; **bold** marks the best under each source→target setting. Zero-shot models are reported for reference, with cells left blank (-) for target datasets seen during their training.

| Source dataset | Method | LLM size | Target dataset: Charades-STA R1@0.3 | Charades-STA R1@0.5 | Charades-STA R1@0.7 | Charades-STA mIoU | QVHighlights R1@0.3 | QVHighlights R1@0.5 | QVHighlights R1@0.7 | QVHighlights mIoU | DiDeMo R1@0.3 | DiDeMo R1@0.5 | DiDeMo R1@0.7 | DiDeMo mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-Shot | GPT-5 [42] | - | 39.7 | 18.3 | 6.2 | 28.4 | - | - | - | - | - | - | - | - |
| | Gemini-2.0-Flash [13] | - | 55.6 | 29.0 | 9.5 | 35.1 | 72.1 | 58.1 | 41.9 | 54.9 | - | - | - | - |
| | HawkEye [52] | 7B | 50.6 | 31.4 | 14.5 | 33.7 | 42.1 | 20.8 | 6.6 | 28.7 | 37.8 | 18.4 | 7.8 | 24.7 |
| | TimeSuite [60] | 7B | 69.9 | 48.7 | 24.0 | - | - | 12.3 | 9.2 | 21.3 | - | - | - | - |
| | UniTime [33] | 7B | - | 59.1 | 31.9 | 52.2 | - | 41.0 | 31.5 | 43.7 | - | - | - | - |
| | VideoMind [35] | 7B | 73.5 | 59.1 | 31.2 | 50.2 | - | - | - | - | - | - | - | - |
| Charades-STA | EaTR [20] | - | 67.7 | 55.2 | 33.1 | 47.7 | 31.7 | 17.0 | 6.4 | 21.5 | 32.7 | 20.1 | 7.8 | 21.2 |
| | CG-DETR [39] | - | 69.7 | 57.6 | 35.1 | 49.5 | 37.4 | 22.8 | 10.5 | 25.2 | 30.3 | 21.8 | 11.0 | 20.7 |
| | Chrono-BLIP [38] | 4B | 79.1 | 68.9 | 48.7 | 58.6 | 67.2 | 44.5 | 23.7 | 44.3 | 56.0 | 43.2 | 23.1 | 37.7 |
| | Qwen2.5-VL† [6] | 7B | 80.3 | 67.7 | 43.9 | 58.1 | 75.2 | 54.3 | 29.7 | 51.2 | 68.2 | 53.8 | 30.3 | 47.4 |
| | EVIDENT | 7B | 80.3 | 69.1 | 49.2 | 59.9 | **77.5** | **59.8** | **36.9** | **55.2** | **69.2** | **55.4** | **32.3** | **48.5** |
| QVHighlights | EaTR [20] | - | 40.8 | 27.2 | 13.0 | 28.0 | 70.3 | 59.6 | 40.3 | 53.1 | 36.0 | 17.8 | 6.5 | 24.1 |
| | CG-DETR [39] | - | 42.8 | 25.5 | 12.2 | 28.5 | 77.5 | 65.6 | 52.1 | 61.3 | 36.6 | 21.2 | 8.9 | 24.6 |
| | Chrono-BLIP [38] | 4B | 63.3 | 39.3 | 20.4 | 42.7 | 86.1 | 76.6 | 62.7 | 70.7 | 43.0 | 24.4 | 11.8 | 33.7 |
| | Qwen2.5-VL† [6] | 7B | 71.1 | 48.1 | 26.2 | 47.3 | 86.9 | 78.2 | 64.8 | 72.0 | 59.4 | 34.1 | 16.9 | 38.2 |
| | EVIDENT | 7B | **73.3** | **50.2** | **28.0** | **49.0** | 89.5 | 82.2 | 69.1 | 75.3 | **62.4** | **36.0** | **17.4** | **40.6** |

$\mathbf{X}^{\text{vis}}_{\text{down}}$. We then apply min-max normalization, yielding $\hat{s}^{\text{sub}}_t$ and $\hat{s}^{\text{obj}}_t$. The final gating score for frame $t$ is the product of the two normalized scores:

$$g_t = \hat{s}^{\text{sub}}_t \cdot \hat{s}^{\text{obj}}_t \in [0,1]. \tag{7}$$

This product captures the co-occurrence of the subject and object: $g_t$ is high only when both entities are present in frame $t$.

**Gated Reconstruction.** We modulate the reconstructed visual tokens by broadcasting the per-frame gating scores $\mathbf{g} = [g_1, \ldots, g_T]$ across the $N$ tokens within each frame, and replace the up projection step in the EB Adapter accordingly: $\mathbf{X}_{\text{out}} = \mathbf{X} + \texttt{concat}(\mathbf{X}^{\text{vis}}_{\text{down}} \odot \mathbf{g}, \mathbf{X}^{\text{txt}}_{\text{down}})\mathbf{W}_{\text{up}}$, where $\odot$ denotes element-wise multiplication with broadcasting along the token and feature dimensions. As a result, the EB Adapter's contribution is concentrated on frames where query-relevant entities co-occur.

### 4.5 Training Objective

The framework is trained end-to-end with the vision encoder frozen. The EB Adapter and LoRA [18] parameters are updated jointly. We train EVIDENT with the combined objective $L_{\text{total}} = L_{\text{CE}} + \lambda L_{\text{EBD}}$, where $\lambda$ controls the strength of the objectness prior.

## 5 Experiments

### 5.1 Experimental Setup

**Training & Evaluation.** We build on Qwen2.5-VL-Instruct [6] (7B) as the backbone MLLM and fine-tune it on two source datasets, Charades-STA (Cha.) [14] and QVHighlights (QVH) [27]. We evaluate the models on three target datasets: Cha., QVH, and DiDeMo [17]. Thanks to the parameter-efficient design of EVIDENT, only the EB Adapter and LoRA modules are trainable, amounting to approximately 201M parameters (*2.42% of the 7B backbone*). We denote each setting by its source-target pair (*e.g.*, Cha.→ QVH), reporting In-Domain (ID) performance when source equals target and Out-of-Domain (OOD) performance otherwise. All results are reported using the standard moment retrieval metrics R1@0.3, R1@0.5, R1@0.7, and mIoU.

**Baselines.** We compare EVIDENT against four categories of methods. (1) Proprietary MLLMs: GPT-5 [42] and Gemini-2.0-Flash [13], whose results are taken from Zhang et al. [62]. (2) Large-scale instruction-tuned MLLMs for VTG, evaluated in a zero-shot manner: HawkEye [52], TimeSuite [60],

Table 2: **Ablation study.** To validate the effect of each component in EVIDENT, we show the results on the Cha.→QVH setting. The best numbers are **highlighted**.

(a) Effects of EVIDENT components.

| EB Adapter | EB Distillation | E2V gating | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|---|---|
| | | | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| | | | 67.7 | 43.9 | 54.3 | 29.7 |
| ✓ | | | 68.1 | 49.5 | 54.9 | 33.1 |
| ✓ | ✓ | | **69.1** | **50.4** | 58.8 | 36.7 |
| ✓ | ✓ | ✓ | **69.1** | 49.2 | **59.8** | **36.9** |

(b) Effects of adapter design.

| Slot Attention Placement | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| After vision encoder | 34.6 | 18.8 | 19.5 | 12.7 |
| Adapter-style (ours) | **68.1** | **49.5** | **54.9** | **33.1** |

(c) Effects of entity-level decomposition.

| Bottleneck Design | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| LoRA | 67.7 | 43.9 | 54.3 | 29.7 |
| Adapter w/ self-attn. | **68.8** | **49.5** | 42.4 | 21.0 |
| EB Adapter | 68.1 | **49.5** | **54.9** | **33.1** |

(d) Effects of query-dependent E2V gating.

| Entity Source | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| Query-agnostic | **69.3** | **49.4** | 58.4 | 36.0 |
| Query-dependent | 69.1 | 49.2 | **59.8** | **36.9** |

UniTime [33], and VideoMind [35]. (3) DETR-based specialists trained on a single source dataset: EaTR [20] and CG-DETR [39]. (4) MLLM-based methods fine-tuned on a single source dataset: Chrono [38] with a BLIP-2 [31] backbone, and Qwen2.5-VL [6] baseline fine-tuned with LoRA [18] under the same training setup as ours.

### 5.2 Cross-Domain Temporal Grounding Results

**Comparison with Existing Methods.** In Table 1, EVIDENT consistently improves OOD performance across all source-target configurations while maintaining competitive ID performance. Among MLLM-based methods, we observe two failure modes that support our motivation presented in Section 1. First, zero-shot MLLMs, even those that are large-scale instruction-tuned, remain substantially below fine-tuned methods in the ID setting, confirming that task-specific adaptation is necessary for precise temporal localization. Second, naïve source-domain fine-tuning suffers severely under domain shift: although Qwen2.5-VL [6] reaches strong ID performance, its OOD R1@0.5 drops by over 19.6 points on Cha.→QVH and over 23.9 points on QVH→Cha. EVIDENT bridges these two regimes, delivering consistent OOD gains over the Qwen2.5-VL baseline (*e.g.*, +5.5 R1@0.5 and +7.2 R1@0.7 on Cha.→QVH). These results demonstrate that entity-grounded visual evidence provides a robust pathway to fully leverage the entity-attention capabilities of pretrained MLLMs under domain shift.

Table 3: **Performance with other backbones.** † denotes the fine-tuned baseline. **Bold** marks the best OOD result under each backbone.

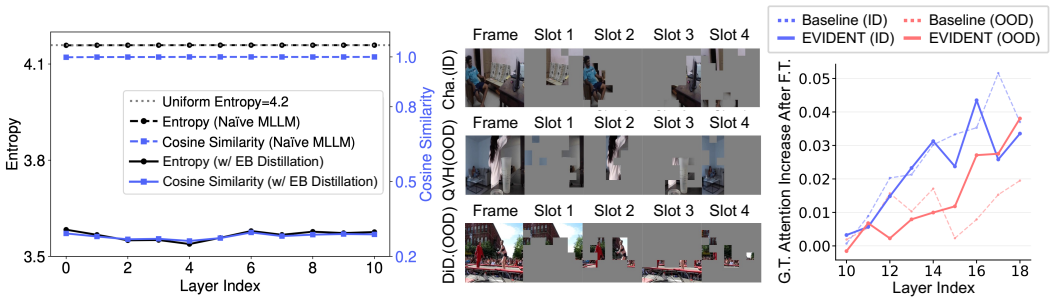
| Backbone | Cha. (ID) | | QVH (OOD) | | DiDeMo (OOD) | |
|---|---|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| Qwen3-VL-8B† | 70.4 | 50.1 | 61.4 | 44.2 | 58.6 | **30.8** |
| +EVIDENT | 70.8 | 50.4 | **64.3** | **46.7** | **60.4** | 30.5 |
| InternVL3.5-8B† | 66.7 | 44.3 | 40.7 | 20.5 | 45.2 | 24.0 |
| +EVIDENT | 72.4 | 49.5 | **48.7** | **30.8** | **55.3** | **29.6** |

**Results with Other MLLM Backbones.** In Table 3, we further apply EVIDENT to Qwen3-VL-8B [5] and InternVL3.5-8B [50]. EVIDENT improves OOD performance on both backbones while maintaining ID performance, with especially large gains on InternVL3.5-8B (+8.0 and +10.1 R1@0.5 on QVH and DiDeMo). This shows that EVIDENT is not tied to a specific MLLM architecture.

### 5.3 Ablation Study

**Effects of EVIDENT Components.** We investigate the effect of each component of EVIDENT in Table 2 (a). Comparing the results with and without EB Distillation, we observe that applying the EB Adapter alone to the semantically unstructured MLLM visual space fails to yield effective entity-centric representations, as discussed in Section 4.3. Only when our EB Distillation strategy is incorporated does the model learn transferable entity-level visual representations, leading to a gain in R1@0.5 by 3.9 points under the OOD setting. On top of this, the VTG-oriented inductive bias injected by E2V gating provides an additional 1.0-point improvement in OOD R1@0.5, confirming that the two components are complementary.

**Effects of Adapter Design.** In Table 2 (b), we show that attaching the learnable module to the MLLM decoder in an adapter style is the most effective placement for entity-level restructuring while preserving the pre-trained entity-attention. Compared to a variant that inserts a slot attention block between the vision encoder and the LLM decoder, which suffers a catastrophic performance collapse, our adapter-style design avoids disrupting the MLLM's pre-trained alignment and still enables effective entity-level decomposition.




(a) (b) (c)

Figure 5: **Additional analysis on EVIDENT.** (a) Entropy and pairwise cosine similarity of normalized slot attention weights in EB Adapter across layers. (b) Slot attention maps of EVIDENT on both ID and OOD samples, visualized by masking each frame with its highest-attending slot. (c) Attention increase on G.T. frames relative to all frames after fine-tuning, compared between the baseline and EVIDENT.

**Effects of Entity-level Decomposition.** In Table 2 (c), an adapter with standard self-attention improves ID R1@0.7 over LoRA [18] but suffers a severe OOD drop (-11.9 R1@0.5), indicating overfitting. In contrast, our EB Adapter maintains the ID gain and improves OOD by +3.4 R1@0.7, demonstrating that entity-level decomposition learns domain-invariant visual representations.

**Effects of Query-dependent E2V gating.** In Table 2 (d), we compare two strategies for selecting the entities used to compute the gating score in E2V gating: extracting per-sample subject/object words from each query (query-dependent) versus using the fixed generic words, *e.g.*, 'person' and 'object', for all samples (query-agnostic). The query-dependent variant yields a gain in OOD R1@0.5 by 1.4 points, demonstrating that our gating actively exploits the MLLM's inherent entity-attention.

### 5.4 Additional Analysis on EVIDENT

**How Does EB Distillation Affect Entity-Binding?** In Figure 5 (a), we measure the entropy of each slot's attention over visual tokens and the pairwise cosine similarity between slots across all EB Adapter layers. When EB Adapter is trained directly on the pre-trained MLLM visual space, the attention is nearly uniform (entropy $\approx 4.2$, the uniform bound) with near-identical slots (cosine similarity $\approx 1.0$), indicating that this space does not afford entity-level binding on its own. With the EB Distillation loss, however, both metrics drop substantially (entropy $\approx 3.5$, cosine similarity $\approx 0.3$), showing that each slot binds to a distinct, localized entity.

**What Do Slots Learn?** In Figure 5 (b), we visualize the slot attention maps of EVIDENT. Across both ID and OOD samples, the slots decompose scenes into semantically coherent regions such as people, objects, and backgrounds. Importantly, this decomposition generalizes to unseen domains without any domain-specific supervision, confirming that EVIDENT learns *transferable entity-level representations* rather than source-biased patterns.

**Does EVIDENT Resolve Attention-Localization Decoupling?** In Figure 5 (c), we measure EVIDENT's attention increase on G.T. frames over to all frames, following the analysis in Figure 2 (c). In the later layers, EVIDENT's OOD increase becomes comparable to its ID increase, showing that EVIDENT performs localization grounded in entity-attention even on unseen domains.

## 6 Conclusions

We presented **EVIDENT**, a parameter-efficient framework for cross-domain Video Temporal Grounding that anchors temporal localization in the inherent entity-attention of pre-trained MLLMs. Our analysis identifies *attention-localization decoupling* as a key failure in OOD VTG, which is primarily driven by visual domain shift rather than unseen query concepts. EVIDENT addresses this through three coupled components: an EB Adapter for routing visual tokens into entity-level slots, an EB Distillation loss for coherent slot bindings, and an E2V gating mechanism for entity-attention–based evidence grounding. Experiments on cross-domain VTG benchmarks show that EVIDENT consistently improves OOD robustness while maintaining competitive ID performance. Our results highlight that evidence-grounded entity-attentive adaptation, rather than direct fitting to source-domain context, is the key to robust cross-domain VTG.

## Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) under grant RS-2024-00353131 (10%) and Information & Communications Technology Planning & Evaluation (IITP) - Information Technology Research Center (ITRC) grant funded by the Korea government (MSIT) (IITP-2025-RS-2023-00259004, 10%). Additionally, it was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-02216217; 50%, RS-2026-25616405; 5%). This work was also supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2023-0-00042, 5%) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2026.

This work was also supported by Youlchon Foundation, National Research Foundation of Korea (NRF) grants, RS-2021-NR05515 (10%), RS-2024-00336576 (3%), and RS-2023-0022663 (3%), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grant, RS-2022-II220264 (3%). Computing resources were partially supported by a National IT Industry Promotion Agency (NIPA) grant, RS-2026-25621677 (1%), funded by the Korean government.

## References

[1] Geo Ahn, Inwoong Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Jinwoo Choi. Why can't I open my drawer? mitigating object-driven shortcuts in zero-shot compositional action recognition. In *ECCV*, 2026. i

[2] Adil Kaan Akan and Yücel Yemez. Slot-guided adaptation of pre-trained diffusion models for object-centric learning and compositional generation. In *ICLR*, 2025. 7

[3] Kyungho Bae, Geo Ahn, Youngrae Kim, and Jinwoo Choi. DEVIAS: Learning disentangled video representations of action and scene. In *ECCV*, 2024. i

[4] Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. MASH-VLM: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *CVPR*, 2025. i

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv:2511.21631*, 2025. 9

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv:2502.13923*, 2025. 3, 4, 8, 9, ii, iii, iv

[7] Peijun Bao and Yadong Mu. Learning sample importance for cross-scenario video temporal grounding. In *ICMR*, 2022. 2, i

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. i

[9] Jinyeong Chae, Donghwa Kim, Kwanseok Kim, Doyeon Lee, Sangho Lee, Seongsu Ha, Jonghwan Mun, Wooyoung Kang, Byungseok Roh, and Joonseok Lee. Towards a complete benchmark on video moment localization. In *AISTATS*, 2024. 2, i

[10] Donghwan Chi, Hyomin Kim, Yoonjin Oh, Yongjin Kim, Donghoon Lee, Daejin Jo, Jongmin Kim, Junyeob Baek, Sungjin Ahn, and Sungwoong Kim. Slot-MLLM: Object-centric visual tokenization for multimodal llm. *arXiv:2505.17726*, 2025. 2, 3

[11] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014. 6

[12] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't I dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. i

[13] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*, 2025. 8

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 3, 4, 8, i, ii, iv, v, vii, viii

[15] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. TRACE: Temporal grounding video llm via causal event modeling. *arXiv:2410.05643*, 2024. 2, 3

[16] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *ECCV*, 2022. 2, i

[17] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 3, 8, iv, v, vii

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 7, 8, 9, 10, v, vi

[19] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower llm to grasp video moments. In *CVPR*, 2024. 2, 3, xi

[20] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, 2023. 8, 9, i

[21] Hao Jiang, Yang Yizhang, and Yadong Mu. Transferable video moment localization by moment-guided query prompting. In *AAAI*, 2024. 2, i

[22] Minji Kim, Taekyung Kim, and Bohyung Han. Map the flow: Revealing hidden pathways of information in videollms. In *ICLR*, 2026. 4, 5, 7, vi

[23] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *ICLR*, 2022. 3

[24] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955. 7

[25] Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. Curriculum multi-negative augmentation for debiased video grounding. In *AAAI*, 2023. i

[26] Hyun Lee, Hyemin Jeong, Yejin Kim, Hyungwook Choi, Hyunsoo Cho, Soo Kyung Kim, and Joonseok Lee. A more word-like image tokenization for mllms. In *CVPR*, 2026. 3

[27] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 2, 3, 4, 8, i, ii, iv, v, vii, viii

[28] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, 2023. i

[29] Jingyu Lei, Gaoang Wang, and Der-Horng Lee. CORE: Compact object-centric representations as a new paradigm for token merging in lvlms. In *CVPR*, 2026. 2, 3

[30] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, 2022. 2, i

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 9

[32] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. i

[33] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *NeurIPS*, 2025. 3, 5, 8, 9, i

[34] 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 *ICCV*, 2023. 2

[35] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-lora agent for temporal-grounded video reasoning. In *ICLR*, 2026. 3, 8, 9, i, xi

[36] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 3

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. v

[38] Boris Meinardus, Hector Rodriguez, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. Chrono: A simple blueprint for representing time in mllms. In *ICCVW*, 2025. 2, 3, 5, 8, 9, i, v, xi

[39] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv:2311.08835*, 2023. 8, 9, i

[40] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, 2023. 2, i

[41] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *CVPR*, 2021. i

[42] OpenAI. Introducing gpt-5, 2025. Available from OpenAI announcement, August 7, 2025. 8

[43] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 6, 7, iii, iv, v

[44] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv:2009.00325*, 2020. 2, i

[45] Zhaobo Qi, Yibo Yuan, Xiaowen Ruan, ShuHui Wang, Weigang Zhang, and QingMing Huan. Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In *AAAI*, 2024. 2, i

[46] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024. 2, 3, xi

[47] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023. 3, 7

[48] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. 2022. 3

[49] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. TR-DETR: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *AAAI*, 2024. 2

[50] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv:2508.18265*, 2025. 9

[51] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *NeurIPS*, 2025. 3

[52] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. HawkEye: Training video-text llms for grounding text in videos. *arXiv:2403.10228*, 2024. 2, 8, i

[53] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *ICLR*, 2023. 3

[54] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. SlotDiffusion: Object-centric generative modeling with diffusion models. 2023. 3

[55] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-VLM: Object-event slots for video-language modeling. In *NeurIPS*, 2024. 2, 3

[56] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv:2505.09388*, 2025. 7, iii, iv, xv

[57] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video understanding. In *ICLR*, 2023. 7

[58] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *ACM SIGIR*, 2021. i

[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 *ACM MM Workshop*, 2021. i

[60] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *ICLR*, 2025. 2, 3, 8, i, xi

[61] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv:2011.09046*, 2020. 3

[62] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal LLMs. In *CVPR*, 2026. 5, 8

[63] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2504.10479*, 2025. 3, ii, iii

# Appendix

In this appendix, we provide comprehensive analyses, method details, and additional quantitative results to complement the main paper. We organize the appendix as follows:


- **Section A: Extended Related Work**
- **Section B: Further Analysis of the Domain Gap**
  - Section B.1: Constructing Seen/Unseen Concept Splits
  - Section B.2: Measuring Cross-Domain Visual Similarity
  - Section B.3: Diagnosing on a Different MLLM
- **Section C: Further Method Details**
  - Section C.1: Extracting Subject-Object Concepts for E2V gating
  - Section C.2: Visualizing DINOv2 Cluster Maps as Supervision Signal
  - Section C.3: Implementation Details
- **Section D: Additional Experimental Results**
  - Section D.1: Additional Ablation Study
  - Section D.2: Qualitative Analysis of EVIDENT
- **Section E: Limitations**
- **Section F: Broader Impact**


## A Extended Related Work

**DETR-based VTG Specialists** Prior VTG works follow the DETR paradigm [8] dominantly, where Moment-DETR [27] first cast moment retrieval as a set prediction problem and established the QVHighlights [27] benchmark. Subsequent models extend this paradigm by injecting VTG-specific inductive biases into the decoder, such as query-dependent representations [40], event-aware attention [20], and correlation-guided calibration [39]. More recently, MLLM-based VTG methods [52, 60, 33, 35, 38] have emerged as a compelling alternative, outperforming DETR-based specialists. In this work, we show that they still struggle to generalize to unseen domains, identify the visual domain gap as the primary cause, and address it with EVIDENT.

**Bias in Video Understanding** Dataset bias has been widely studied in video understanding. In action recognition, prior works have shown that models exploit static appearance cues, such as scenes and objects, as shortcuts [12, 32, 3, 1]. In multimodal video understanding, models have similarly been found to over-rely on background context or a single frame [28, 4].

In VTG, Otani et al. [44] show that blind baselines without video input can rival trained models by exploiting the moment location distribution of annotations. This finding motivated benchmarks for evaluating generalization, including OOD re-splits [59], compositional splits with novel query compositions [30], and a comprehensive benchmark across multiple datasets [9]. To mitigate temporal location bias, prior works apply causal intervention [58, 41], perturb or augment the temporal structure of videos [16, 25], or synthesize bias-conflict samples [45]. Others address cross-scenario transfer via sample reweighting [7] or query prompting [21]. However, these methods mainly target non-MLLM grounding models, whereas we study the visual domain shortcuts that arise when fine-tuning MLLMs, and address them by preserving the pre-trained entity-attention rather than by explicit debiasing.

## B Further Analysis of the Domain Gap

### B.1 Constructing Seen/Unseen Concept Splits

For the concept domain gap analysis shown in Figure 3 (a), we partition the QVHighlights (QVH) [27] test set into *seen* and *unseen* concept subsets with respect to the Charades-STA (Cha.) [14] training distribution. Specifically, we first build a Cha.-centric concept vocabulary $\mathscr{C}$ by taking the 50 most frequent content lemmas across all Cha. training queries, after standard lowercasing, lemmatization, and stop-word removal. The resulting set covers the dominant indoor activity, object, and motion vocabulary of Cha. (e.g. *glass, walk, sneeze, room, door, cup, shelf, sit, pillow, refrigerator, blanket*). Each QVH test query is then processed with the identical pipeline, and labeled *seen* if at least one of

its lemmas appears in $\mathscr{C}$, otherwise *unseen*. This rule yields 635/915 seen/unseen samples out of 1,550; to remove the class-size imbalance, we draw a fixed random subset of 600 samples from each group as our final evaluation splits. We use only word-level lemma matching, without expanding to a hand-crafted concept ontology, so that the split reflects lexical familiarity to the Cha.-trained model without additional human bias.

### B.2 Measuring Cross-Domain Visual Similarity

For the visual domain gap analysis shown in Figure 3 (b), we embed each video as a single global descriptor by passing 20 uniformly sampled 224×224 frames through a frozen Qwen2.5-VL-3B [6] vision encoder, mean-pooling visual tokens within each frame and then averaging the resulting per-frame vectors. Taking Charades-STA (Cha.) [14] training videos as the reference distribution, we define a centroid $\mathbf{c}_{\text{Cha.}}$ as the mean of their global descriptors, and characterize each test video by its cosine similarity $s(v) = \cos(\mathbf{v}, \mathbf{c}_{\text{Cha.}})$ to this centroid. While the Cha. training set is internally tight (intra-set similarity $0.894 \pm 0.035$ over 5,335 videos), QVHighlights (QVH) [27] test videos drift toward lower similarities ($0.831 \pm 0.061$, range $[0.527, 0.942]$ over 1,519 videos), indicating a measurable train→test visual shift (Figure I).

We sort the QVH test set by $s(v)$ and partition it into five equally sized quantile bins (Figure I). Reporting performance on the top and bottom quantiles—denoted **Visually similar** (Q1, $s \geq 0.881$) and **Visually dissimilar** (Q5, $s \leq 0.789$)—isolates the impact of the visual domain gap on temporal grounding without conflating it with other dataset-specific factors. Note that the cosine values themselves are meaningful only as relative scores within the Qwen2.5-VL [6] embedding space; we use them solely to induce an ordering over test videos.

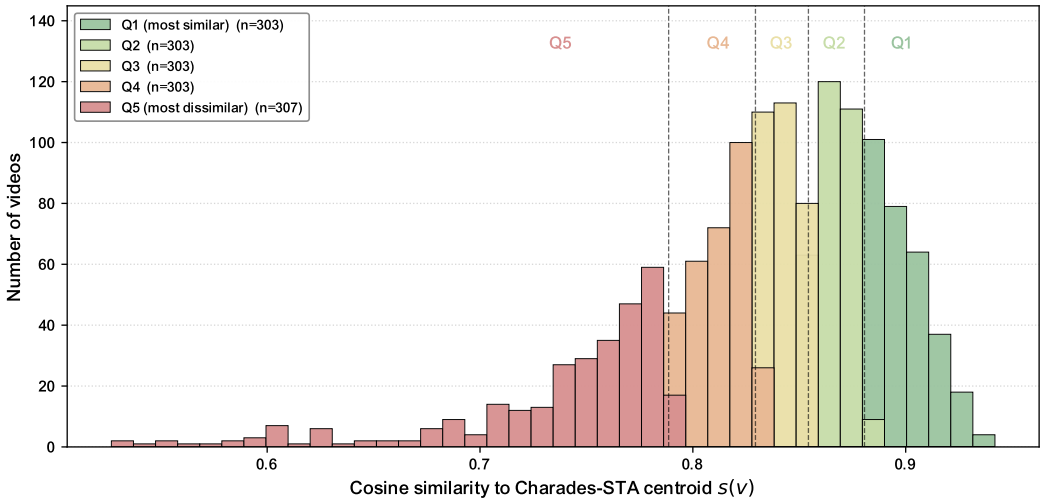
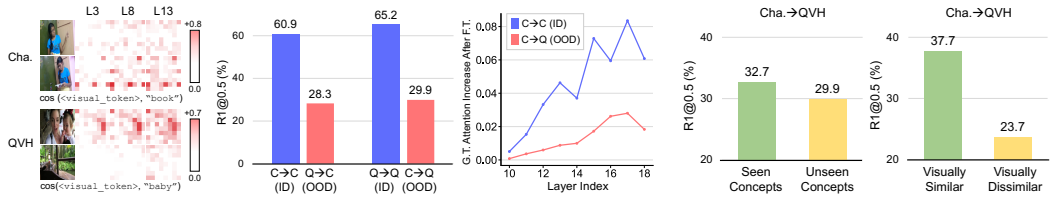



Figure I: **Visual similarity distribution and quintile partitioning.** Distribution of cosine similarities $s(v) = \cos(\mathbf{v}, \mathbf{c}_{\text{Cha.}})$ between QVHighlights [27] test video descriptors $\mathbf{v}$ and the Charades-STA [14] training centroid $\mathbf{c}_{\text{Cha.}}$, computed in a frozen Qwen2.5-VL-3B [6] embedding space. The 1,519 test videos (1,550 query samples) are sorted by $s(v)$ and partitioned into five equally sized quantile bins (dashed lines). We evaluate temporal grounding on the highest- and lowest-similarity bins, **Q1** (*Visually similar*, $s \geq 0.881$) and **Q5** (*Visually dissimilar*, $s \leq 0.789$), to isolate the effect of visual domain shift. Cosine values are relative scores within this encoder's embedding space and are not directly comparable to similarities computed under a different encoder.

### B.3 Diagnosing on a Different MLLM

To verify that our problem analysis is not specific to Qwen2.5-VL [6], we repeat the analyses of Figure 2 and Figure 3 on a different MLLM, InternVL3-2B [63]. As shown in Figure II, InternVL3-2B exhibits the same pattern as Qwen2.5-VL reported in Section 3.

**Entity-attention analysis (Figure II (a)).** The zero-shot InternVL3-2B exhibits strong visual-text alignment across both Charades-STA [14] and QVHighlights [27], but naïve fine-tuning suffers




(a) Entity-attention analysis

(b) OOD gap breakdown

Figure II: **Problem analysis on InternVL3-2B [63].** We repeat the analysis of Figure 2 and Figure 3 on a different MLLM, InternVL3-2B, and observe the same pattern as reported for Qwen2.5-VL [6] in Section 3. **(a) Entity-attention analysis.** (Left) Zero-shot entity-attention exhibits strong visual-text alignment regardless of the video domain. (Middle) Naïve fine-tuning suffers severe ID-OOD gaps. (Right) After fine-tuning, the attention on ground-truth-interval frames substantially increases on the trained domain (ID) but only marginally on the unseen domain (OOD), indicating that the learned visual-text alignment ability fails to transfer to other domains. **(b) OOD gap breakdown.** (Left) Splitting the OOD test set by query-concept overlap with ID yields only a marginal performance gap, indicating that the concept gap is not the dominant factor. (Right) Splitting OOD samples by visual feature similarity to ID yields a much larger gap, identifying the visual domain gap as the primary cause of OOD degradation.

severe ID-OOD performance gaps. The fine-tuned model's attention on ground-truth-interval frames substantially increases on the trained domain (ID) but only marginally on the unseen domain (OOD), indicating that the learned visual-text alignment ability fails to transfer to other domains.

**OOD gap breakdown (Figure II (b)).** Splitting the OOD test set by query-concept overlap with ID yields only a marginal performance gap, while splitting by visual feature similarity to ID yields a much larger gap. This identifies the visual domain gap as the primary cause of OOD degradation, consistent with our observations on Qwen2.5-VL. Together, these results suggest that entity-attention bypass under visual domain shift is a general phenomenon across MLLMs rather than a model-specific artifact.

# C Further Method Details

## C.1 Extracting Subject-Object Concepts

Entity-to-eVidence (E2V gating, Section 4.4) takes a subject-object concept pair as textual input. For each training query, we extract this pair by prompting Qwen3-4B-Instruct-2507 [56] in a zero-shot (Figure III) to return a JSON object with two entries: a `Subject:` (primary actor) and an `Object:` (tangible interaction target), each restricted to a 1–2 word noun. Given the extracted concept words, we identify their token positions in the original query token sequence as $i_{\text{sub}}$ and $i_{\text{obj}}$, which are then used to select the corresponding query token features $\mathbf{q}'_{\text{sub}}$ and $\mathbf{q}'_{\text{obj}}$ for E2V gating. This schema deliberately excludes verbs, scenes, and abstract descriptions so that the resulting concepts correspond to *visually grounded* entities that align with frame-level evidence during E2V gating.

## C.2 Visualizing DINOv2 Cluster Maps as Supervision Signal

Our EB Distillation relies on DINOv2 [43] $K$-means cluster maps as pseudo-ground-truth supervision for slot attention. To verify the reliability of this supervision signal, we examine whether DINOv2 features produce semantically meaningful clusters.

We visualize DINOv2 cluster maps on videos from our evaluation benchmarks. Concretely, we extract DINOv2-large patch features for 20 uniformly sampled 224×224 frames, apply 2×2 average pooling to reduce each frame to an 8×8 spatial grid (64 patches), $\ell_2$-normalize the features, and run $K$-means with $K$=4. Figure IV shows the resulting cluster assignments: for each cluster, we display only the patches assigned to that cluster.

The results demonstrate that DINOv2 $K$-means clusters naturally partition frames into semantically coherent regions—consistently separating foreground actors, interaction targets, background scene elements, and motion-salient regions. This confirms that using DINOv2 cluster maps as pseudo-GT

## User

You are an expert in semantic decomposition for video temporal grounding. Extract the most important Subject and Object from the query for frame-level grounding.

**Roles & Prefixes you MUST use:**

- `"Subject:[noun]"` → The primary living actor or main entity.
- `"Object:[physical noun]"` → The primary tangible target object being interacted with. If none exists, omit this.

**Rules:**

1. Return a JSON list containing up to 2 strings.
2. Each string MUST start with `"Subject:"` or `"Object:"`.
3. If the Subject is interacting with a physical item, you MUST extract it as an Object.
4. Ignore verbs, scenes, abstract concepts, and meta-descriptions.
5. Keep the nouns extremely concise (1–2 words).
6. Return ONLY a valid JSON list.

**Examples:**

Query: "a person opens the refrigerator in the kitchen"
JSON list: `["Subject: person", "Object: refrigerator"]`

Query: "The girl dances around the room."
JSON list: `["Subject: girl"]`

Query: {query}
JSON list:

## Assistant

JSON list of extracted concepts

Figure III: Zero-shot prompt used with Qwen3-4B-Instruct-2507 [56] for key concept extraction. Blue text represents variables.

for EB Distillation provides a semantically grounded training signal rather than an arbitrary one, justifying our design choice. As shown in the bottom row of Figure IV, slots trained with our EB Distillation mirror this DINO structure, binding each slot to a distinct entity consistently across both ID and OOD domains.

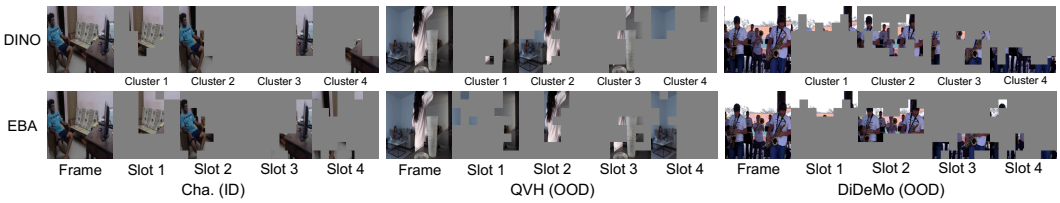



Figure IV: **DINOv2 cluster maps vs. EB Adapter slot assignments.** Each group shows a frame and four isolated patch views across Charades-STA (Cha.) [14] (ID), QVHighlights (QVH) [27](OOD), and DiDeMo [17] (OOD). **Top (DINO):** DINOv2 [43] $K$-means naturally separates semantically coherent regions, validating its use as pseudo-ground-truth in EB Distillation. **Bottom (EB Adapter):** Slots trained with EB Distillation mirror the DINO structure, binding each slot to a distinct entity consistently across domains.

### C.3 Implementation Details

We build upon Qwen2.5-VL-Instruct [6] (7B) as the backbone MLLM, where the LLM decoder has a hidden dimension of $D$=3584. The vision encoder processes each 224×224 frame into $N$=64 visual tokens (8×8 spatial grid). For the EB Adapter, we set the bottleneck dimension to $d$=896, the

Table I: **Fine-tuning setting of EVIDENT**.

| Category | Hyperparameter | Charades-STA [14] | QVHighlights [27] |
|---|---|---|---|
| Input | # Frames | 20 | 60 |
| | Resolution | $224 \times 224$ | |
| Optimization | Optimizer | AdamW [37] | |
| | Learning rate | $5 \times 10^{-5}$ | |
| | LR scheduler | Cosine, 5% warmup | |
| | Weight decay | 0.1 | |
| | Epochs | 2 | |
| | Global batch size | 32 | |
| LoRA | Applied layers | 11–27 | |
| | Rank $r$ | 16 | |
| | $\alpha$ | 64 | |
| | Dropout | 0.05 | |
| EB Adapter | Inserted layers | 0–10 | |
| | # Slots $K$ | 4 | |
| | Bottleneck dim | 896 | |
| | # Iterations | 3 | |
| | Tokens per frame | 64 | |
| EB Distillation | Loss weight $\lambda$ | 0.1 | 1.0 |

number of slots to $K$=4, and use 8 attention heads with $I$=3 iterations of GRU-based refinement. We inject the EB Adapter into layers 0–10, while LoRA [18] (rank 16, $\alpha$=64) is applied to the remaining deeper layers; the vision tower, the LLM's frozen layers, and the vision–language merger are kept fixed throughout training. The E2V gating module is applied at every EB Adapter layer (0–10) and modulates the slot residual using the cosine similarity between subject/object queries and per-frame mean-pooled visual tokens. The EB Distillation derives pseudo-GT cluster maps by $K$-means clustering DINOv2-large [43] patch features, and aligns slot-attention maps to these clusters via Hungarian matching with BCE loss at every EB Adapter layer (0–10) with $\lambda$=0.1.

**Video sampling and timestamps.** We uniformly sample 20 and 60 frames for the models trained on Charades-STA [14] and QVHighlights [27], respectively. For cross-domain evaluation, we follow each *target* dataset's native protocol and sample 20 frames for Charades-STA, 60 frames for QVHighlights, and 20 frames for DiDeMo [17]. For Charades-STA training, whose clips are confined to a narrow 0–30 s range, we further adopt the *relative-integer* timestamp encoding of Chrono-BLIP [38], representing each timestamp as an integer in $[0, 100]$ relative to clip duration; this mitigates the text bias toward a narrow band of numeric tokens that the dataset's limited duration range would otherwise induce.

**Gating under partial concept queries.** Not every query specifies both a subject and an object: queries such as *"a person is dancing"* contain a subject but no object. For such samples, we pass the absent concept's score through as 1 rather than 0: since the per-frame gate is the multiplicative product of the two scores, 1 is its identity element, so the absent concept exerts no influence on gating while the present one continues to drive it—whereas 0 would collapse the gate regardless of the present concept's signal.

**Optimization.** We train for 2 epochs with AdamW [37] (learning rate $5\times10^{-5}$, weight decay 0.1) under a cosine schedule with 5% linear warmup, using a global batch size of 32 on 8 NVIDIA 3090/4090 GPUs. Full hyperparameter details are summarized in Table I. In total, the trainable parameters amount to approximately 201M (2.42% of the backbone) for the 7B model.

# D Additional Experimental Results

## D.1 Additional Ablation Study

We provide further ablations on the design choices of EVIDENT in Table II.

Table II: **Additional Ablation Study.** Further ablations on the formulation of EB Distillation and the detailed configurations of E2V gating.

(a) Effects of EB Adapter insertion layers.

| Layer Index $i$ | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| 0-6 | **69.2** | **49.5** | 56.8 | 34.9 |
| 0-10 | 69.1 | 49.2 | **59.8** | **36.9** |
| 0-14 | 69.1 | 49.4 | 58.3 | 33.7 |

(b) Effects of token reconstruction design.

| Method | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| Cross attention | 69.0 | **49.3** | 57.7 | 35.1 |
| Attention reuse | **69.1** | 49.2 | **59.8** | **36.9** |

(c) Effects of the number of slots.

| # of slots $N_s$ | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| 4 | 69.1 | 49.2 | **59.8** | **36.9** |
| 8 | **69.2** | **49.4** | 56.6 | 36.6 |
| 16 | **69.2** | 49.1 | 53.9 | 33.7 |

(d) Effects of feature source in E2V gating.

| Visual feature | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| Before down proj. | 68.4 | 49.0 | 55.8 | 34.1 |
| After down proj. | **69.1** | **49.2** | **59.8** | **36.9** |
| After token reconstruction | 69.2 | 50.3 | 51.9 | 29.7 |

(e) Effects of EB Distillation scale and placement.

| Loss scale $\lambda$ | Layer index $l$ | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|---|
| | | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| 1.0 | 0-10 | 68.3 | 49.2 | **60.4** | 36.0 |
| 0.5 | 0-10 | 68.7 | **49.3** | 58.4 | 34.9 |
| 0.1 | 0-10 | **69.1** | 49.2 | 59.8 | **36.9** |
| 0.1 | 10 | 68.7 | 49.0 | 57.9 | 36.5 |

(f) Effects of E2V gating applied layers.

| Layer Index $i$ | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| 10 | 68.9 | **49.8** | 55.8 | 36.2 |
| 0-10 | **69.1** | 49.2 | **59.8** | **36.9** |

(g) Effects of attention normalization.

| Attention Map | Cha. (ID) | | QVH (OOD) | |
|---|---|---|---|---|
| | R1@0.5 | R1@0.7 | R1@0.5 | R1@0.7 |
| $\mathbf{A}$ | 68.7 | 48.9 | 56.5 | 33.4 |
| $\hat{\mathbf{A}}$ | **69.1** | **49.2** | **59.8** | **36.9** |

**Effects of EB Adapter insertion layers.** In Table II (a), we show that inserting the EB Adapter into decoder layers 0–10 yields the best OOD performance. As discussed in Section 4.2, this choice places our EB Adapter in the early layers where cross-frame interactions occur [22], allowing each slot to capture temporally coherent disentangled semantics. Meanwhile, the mid-to-later layers responsible for vision-language integration [22] are left to lightweight LoRA [18] fine-tuning, preventing excessive overfitting to the source domain.

**Effects of token reconstruction design.** In Table II (b), we demonstrate that reusing the slot attention weights to reconstruct the original token sequence, mentioned in Section 4.2, substantially improves OOD performance (+2.1 points in R1@0.5). This attention reuse design routes VTG adaptation through disentangled entity-level representations and eliminates the additional learnable parameters that could overfit to the source domain.

**Effects of the number of slots.** In Table II (c), ID performance remains stable across $N_s$, whereas OOD performance degrades as $N_s$ increases (-5.9 R1@0.5 from $N_s$=4 to 16). We find that four slots are sufficient for entity-level binding, and additional slots are likely to split each entity into fragmented parts, weakening cross-domain transferability.

**Effects of visual feature source in E2V gating.** In Table II (d), we ablate which visual feature is used to compute the per-frame gating score, comparing tokens *before* down projection, *after* down projection, and *after* slot-based reconstruction. Features *after* down projection give the best overall performance, achieving the highest OOD scores (R1@0.5 of 59.8) while remaining competitive on ID. Pre-projection features underperform on both splits, and post-reconstruction features collapse on OOD (R1@0.5 drops to 51.9), as the slot bottleneck discards fine-grained details needed for entity-level scoring. This indicates that E2V gating should be computed from the LLM's visual stream rather than the bottlenecked slot output.

**Effects of EB Distillation loss scale and placement.** In Table II (e), we ablate the loss weight $\lambda$ and the layer indices at which EB Distillation is applied. Across the three loss scales tested under the all-layer setting (0–10), the differences are minor: ID and OOD metrics remain within roughly 1–2 points across $\lambda \in \{1.0, 0.5, 0.1\}$, indicating that EB Distillation is robust to the choice of loss scale. Our default $\lambda$=0.1 achieves the best ID R1@0.5 (69.1) and OOD R1@0.7 (36.9), while $\lambda$=1.0 slightly favors OOD R1@0.5 (60.4 vs. 59.8). Restricting EB Distillation to only the last EB Adapter layer ($l$=10) leads to a small but consistent drop across metrics, suggesting that progressively applying EB Distillation across all EB Adapter layers (0–10) is mildly more effective than aligning only the final layer.

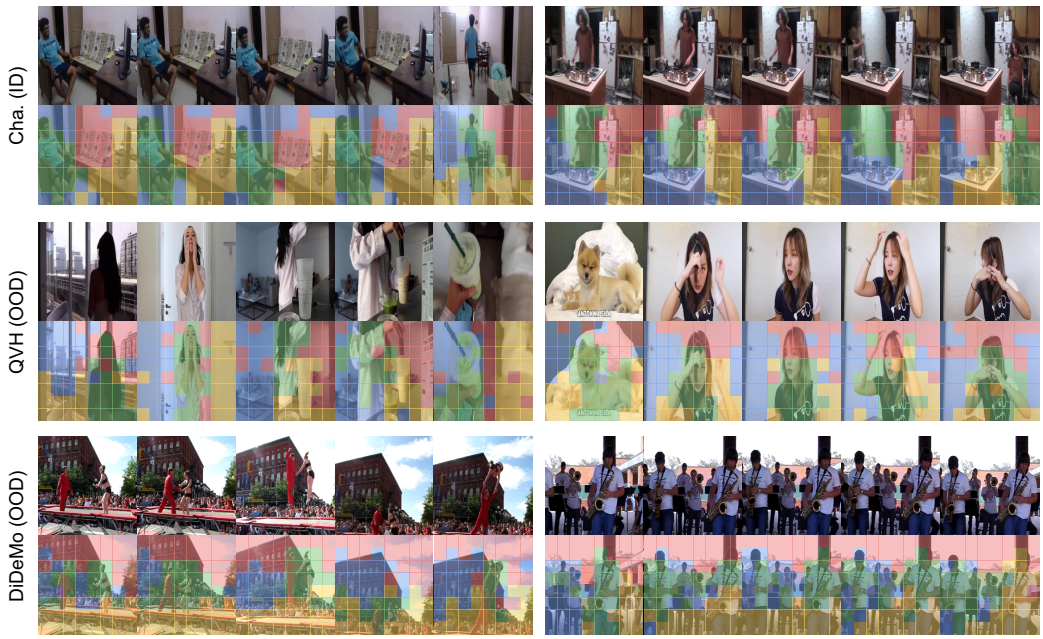



Figure V: **EB Adapter visualization.** We visualize the slot assignments on samples from Charades-STA (Cha.) [14] (ID), QVHighlights (QVH) [27] (OOD), and DiDeMo [17] (OOD) by masking each frame with its highest-attending slot. Frames are arranged in temporal order from left to right, and the same color denotes the same slot.

**Effects of E2V gating applied layers.** In Table II (f), we ablate whether E2V gating should be applied at every EB Adapter layer (L0–10) or only at the final layer (L10). Applying E2V gating across all EB Adapter layers consistently outperforms the last-layer-only variant on OOD R1@0.5 (59.8 vs. 55.8) while maintaining competitive ID performance. This shows that progressively gating the slot residual at every layer better aligns the entity-level representations with query-relevant frames, leading to more robust cross-domain generalization.

**Effects of attention normalization.** In Table II (g), we compare two attention maps for token reconstruction (eq. (4)) and EB Distillation (eq. (6)): the slot-normalized map $\mathbf{A}$ and the token-normalized map $\hat{\mathbf{A}}$ used in the slot update. While both variants perform comparably in the ID setting, $\hat{\mathbf{A}}$ yields higher OOD performance (+3.3 R1@0.5).

### D.2 Qualitative Analysis of EVIDENT

**Slot visualization in temporal order** As shown in Figure V, each slot consistently attends to the same entity across consecutive frames, forming temporally coherent trajectories along the time axis. Regions corresponding to the same slot (i.e., the same color) remain stable as the scene evolves, indicating that EB Adapter binds each slot to a specific entity and tracks it over time rather than re-assigning slots arbitrarily at each frame. This temporal continuity is observed not only on the in-domain Charades-STA [14] samples but also on out-of-domain QVHighlights [27] and DiDeMo [17] samples, suggesting that the learned slot-entity binding generalizes beyond the training distribution.

**Gating score visualization** In Figure VI, we visualize the gating scores of each entity produced by E2V gating on Charades-STA [14] (ID) and QVHighlights [27] (OOD) samples. We observe that the gating scores computed by EVIDENT accurately reflect the temporal intervals in which each entity appears in the visual content.

## E Limitations and Future Work

EVIDENT addresses cross-domain VTG under visual distribution shift, with experiments spanning three diverse benchmarks (Charades-STA, QVHighlights, DiDeMo). Extending our analysis to more

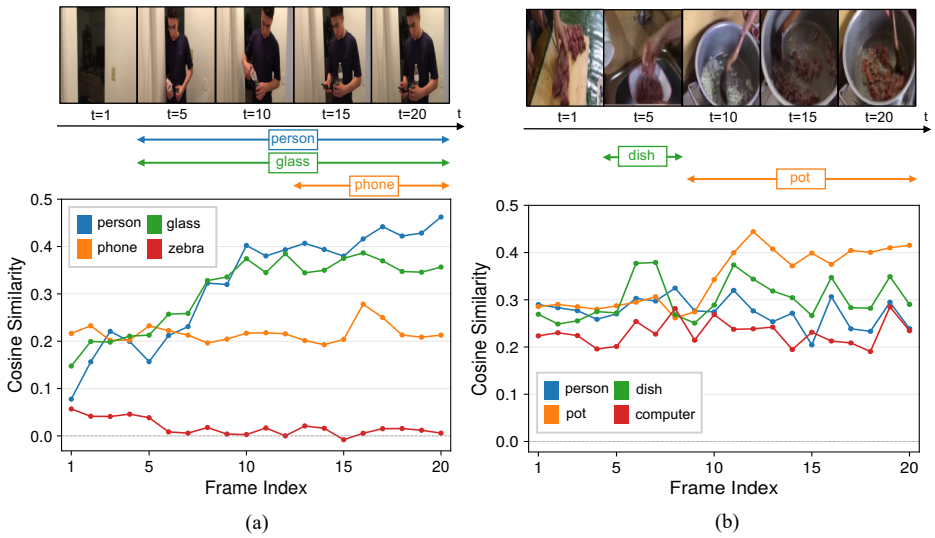



Figure VI: **E2V gating score visualization.** We visualize the per-entity gating scores from E2V gating (see Section 4.4) on (a) Charades-STA [14] (ID) and (b) QVHighlights [27] (OOD) samples.

challenging settings such as long-form videos, egocentric perspectives, or open-vocabulary temporal grounding is a promising direction for future work. While we identify visual domain shift as the dominant failure mode of naïvely fine-tuned MLLMs, jointly addressing other forms of distribution shift, such as temporal granularity or query style, would require complementary advances. We view EVIDENT as a first step toward systematically diagnosing and mitigating distribution shifts in MLLM-based video temporal grounding.

## F Broader Impact

EVIDENT improves the robustness of MLLM-based video temporal grounding under domain shift, with potential positive impacts on applications such as video search, content accessibility (e.g., caption alignment for visually impaired users), and video understanding tools for education and media analysis. By promoting entity-grounded reasoning rather than reliance on dataset-specific shortcuts, our framework may also reduce systematic biases inherited from training data, contributing to more reliable model behavior in unseen scenarios. We do not foresee direct negative societal impacts from this work, as it builds on publicly available pre-trained models and datasets and does not enable new generative or surveillance capabilities. Standard considerations applicable to general video understanding research, such as privacy concerns when applying such systems to personal video content, should be observed by downstream users.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 clearly state our central claim that robust cross-domain VTG requires evidence-grounded entity-attentive adaptation rather than direct fitting to source-domain context, and the contributions (problem analysis, EVIDENT framework, and cross-domain experiments) match the results reported in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations and future directions in Section E, including the scope of cross-domain evaluation and other forms of distribution shift that complement our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents an empirical study and does not include formal theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the architecture, training procedure, datasets, and evaluation protocol in Section 4 and Section 5, and provide complete hyperparameters in Table I along with implementation details in Section C.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code is not released at submission time to preserve anonymity, but we plan to release the code upon acceptance. All datasets used (Charades-STA, QVHighlights, DiDeMo) are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Training and evaluation details, including data splits, hyperparameters, and optimization settings, are provided in Section 5 and Section C.3, with full hyperparameters listed in Table I.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: Following common practice in MLLM-based VTG [19, 46, 38, 60, 35], we report results from a single training run due to the substantial computational cost of fine-tuning 7B-scale MLLMs. Our improvements over baselines are consistent across multiple datasets and source-target configurations (Table 1), suggesting that the gains are not due to seed variance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources (8 NVIDIA 3090/4090 GPUs) and training duration (2 epochs) are specified in Section C.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses publicly available datasets, does not involve human subjects, and conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive impacts (e.g., video accessibility, reduced bias in cross-domain settings) and considerations for downstream use in Section F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release new pre-trained models or scraped datasets that pose a high risk for misuse. Our work builds on publicly available pre-trained MLLMs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all datasets (Charades-STA, QVHighlights, DiDeMo) and pre-trained models (Qwen2.5-VL, InternVL3, DINOv2, Qwen3) used, and respect their respective licenses for academic research.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new datasets or models at submission time.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects and therefore does not require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use Qwen3-4B-Instruct-2507 [56] as a preprocessing component to extract subject-object concept words from natural language queries, which is described in Section 4.4 and detailed in Section C.1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.