Unleashing the Potential of Multimodal LLMs for Zero-Shot Spatio-Temporal Video Grounding

Zaiquan Yang **Yuhao Liu**[†] **Gerhard Hancke** **Rynson W.H. Lau**[†]
Department of Computer Science
City University of Hong Kong
`{zaiquyang2-c, yuhliu9-c}@my.cityu.edu.hk`
`{gp.hancke, Rynson.Lau}@cityu.edu.hk`

Abstract

Spatio-temporal video grounding (STVG) aims at localizing the spatio-temporal tube of a video, as specified by the input text query. In this paper, we utilize multimodal large language models (MLLMs) to explore a zero-shot solution in STVG. We reveal two key insights about MLLMs: (1) MLLMs tend to dynamically assign special tokens, referred to as *grounding tokens*, for grounding the text query; and (2) MLLMs often suffer from suboptimal grounding due to the inability to fully integrate the cues in the text query (*e.g.*, attributes, actions) for inference. Based on these insights, we propose a MLLM-based zero-shot framework for STVG, which includes novel decomposed spatio-temporal highlighting (DSTH) and temporal-augmented assembling (TAS) strategies to unleash the reasoning ability of MLLMs. The DSTH strategy first decouples the original query into attribute and action sub-queries for inquiring the existence of the target both spatially and temporally. It then uses a novel logit-guided re-attention (LRA) module to learn latent variables as spatial and temporal prompts, by regularizing token predictions for each sub-query. These prompts highlight attribute and action cues, respectively, directing the model's attention to reliable spatial and temporal related visual regions. In addition, as the spatial grounding by the attribute sub-query should be temporally consistent, we introduce the TAS strategy to assemble the predictions using the original video frames and the temporal-augmented frames as inputs to help improve temporal consistency. We evaluate our method on various MLLMs, and show that it outperforms SOTA methods on three common STVG benchmarks. The code will be available at `https://github.com/zaiquanyang/LLaVA_Next_STVG`.

1 Introduction

Spatio-Temporal Video Grounding (STVG) aims to localize a target object in a video both spatially and temporally, given an input text query. This task is fundamental to many different applications (*e.g.*, video surveillance and autonomous driving [67]). However, it is also very challenging as it requires the model to be able to distinguish the target from distractors over time and identify the precise temporal boundary of the action. While existing methods handle the STVG task mainly in a fully-supervised setting [13; 59; 54], which often relies on costly frame-level annotations, several works [29; 3; 61; 21; 60; 25] attempt to introduce the weakly-supervised or zero-shot setting to alleviate the burden of dense annotations. For example, E3M [3] integrates CLIP [43] and an expectation maximization strategy to optimize spatio-temporal localization in a zero-shot manner. However, CLIP is known to be weak in localization [75; 14] as it simply aligns the global representation of image-text pairs.

[†] Joint corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

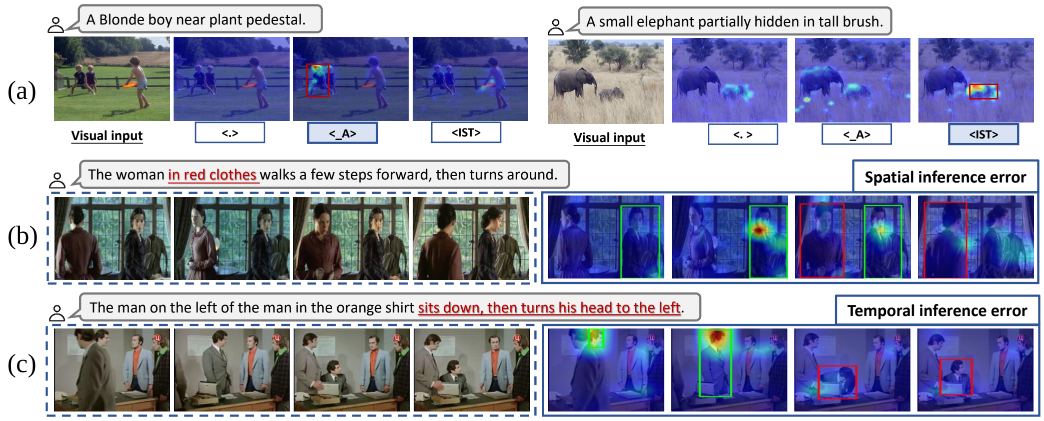


Figure 1: (a) The visual attention maps shows that some special tokens (marked as) can precisely attend to the target region of the input query. However, these special tokens, referred to as the *grounding tokens* in our work, underperform on complex STVG, where they often focus on part cues and ignore other cues (marked in red within the input text prompts). Examples (b) and (c) illustrate spatial/temporal grounding errors caused by ignoring the discriminative attribute/action cues. Red and green bounding boxes denote the ground truths and predictions, respectively.

Considering the strong capability of multi-modal large language models (MLLMs) [37; 27; 31; 8; 53; 36] in cross-modality alignment, several works explore the application of MLLMs in the visual grounding task. However, they typically require explicit fine-tuning [26; 44; 39] of MLLMs with additional grounding datasets and specific model modifications, which can be difficult to scale and generalize to novel visual data [4]. Although some recent works [4; 69] have investigated the attention maps in language models as done in previous ViT and CNN architectures [74; 52; 12], they only focus on the generated tokens while neglecting other token components input in the language model (*e.g.*, system tokens and special tokens). In particular, we observe that the special tokens play an important role in structuring the communication between the user input and the language model, and help guide the generation of coherent responses in dialogues.

With the above observation in mind, we delve deeper and find that the special tokens following the input instruction have outstanding grounding ability. In particular, a few special tokens are characterized by high visual activation and can attend to the region of interest well. Considering the left sample of Fig. 1(a) as an example, the special token '`_A`' provides tangible attention to the *'boy'* referred to by the given query, while in the right sample of Fig. 1(a), the token '`IST`' is assigned to ground the *'elephant'*. These special tokens are referred to as ***grounding tokens*** in our work. Despite the strong comprehension ability, the grounding tokens cannot optimally adapt to STVG as they tend to ignore some important cues (*e.g.*, attribute or action) in a complex video query. As shown in Fig. 1(b), the grounding token fails in the spatial grounding by ignoring the attributes. Similarly, in Fig. 1(c), it leads to the failure of temporal grounding by neglecting the action cues of the target.

In this work, we first conduct a systematic analysis to probe the special tokens of various MLLMs (Sec. 3.2), which demonstrates our aforementioned empirical findings and enables zero-shot spatio-temporal video grounding. To alleviate the problem of neglecting discriminative cues, we propose a novel decomposed spatio-temporal highlighting (DSTH) strategy (Sec. 3.3). Specifically, the language query is decomposed into attribute and action sub-queries, which are utilized as the text input of the MLLM for inquiring the target's existence spatially and temporally, respectively. We then propose a novel logit-guided re-attention (LRA) module to highlight the cues in attribute and action sub-queries. For each sub-query, LRA optimizes the learnable latent variable as the (spatial/temporal) prompts via enhancing the positive response generation while suppressing the negative response. As a result, the DSTH strategy can well adapt the model to mine faithful visual context and concentrate on spatially/temporally relevant regions. In addition, to further enhance the temporal consistency during spatial grounding by the attribute sub-query, we develop a temporal-augmented assembling (TAS) strategy (Sec. 3.4). The TAS strategy utilizes temporally perturbed frames as input to assemble different predictions into final spatial grounding result.

In summary, our contributions are as follows:

- We reveal that MLLMs dynamically assign the special tokens for precisely grounding text-related regions. We identify the special tokens characterized by the salient visual activation for deriving a novel zero-shot STVG framework.
- We propose a novel test-time tuning strategy named decomposed spatio-temporal highlighting (DSTH). It introduces an innovative logit-guided re-attention (LRA) module to adapt the grounding token for thorough spatio-temporal localization. We also develop a temporal-augmented assembling (TAS) strategy to further improve the robustness of spatial localization.
- We conduct extensive experiments on various MLLMs to demonstrate our empirical findings and validate the effectiveness of the proposed method. Our method outperforms existing methods by a remarkable margin on three STVG benchmarks.

2 Related work

Spatio-Temporal Video Grounding (STVG) aims to localize a spatio-temporal tube in a video corresponding to the text query. Unlike image-based visual grounding [10; 11; 63; 34; 62; 35], STVG [73; 51; 59] presents significant challenges, requiring models to distinguish targets from distractors both spatially and temporally. Fully-supervised STVG approaches [51; 54; 73; 49] have achieved promising results. However, these methods heavily depend on an extensive collection of labor-intensive annotations. Several recent works [29; 21; 25] tackle STVG in a more efficient manner, which only uses coarse video-level descriptions for training. E3M [3] leverages pre-trained vision-language models and proposes an expectation maximization framework to optimize spatio-temporal localization. However, contrastive objective based multimodal models are known to be weak in localization [75; 14] as they simply align the global representations of image-text pairs.

Multimodal Large Language Models (MLLMs) are capable of handling diverse language and vision tasks. To equip MLLMs with the grounding ability, current works [41; 26; 28; 66; 71] construct grounding-oriented supervision data for instruction tuning and propose novel architectural modifications. LLaVA-ST [28] achieves spatio-temporal understanding by introducing a progressive training strategy consisting of three sequential stages. However, the large amount of grounding data needed for instruction-tuning imposes high labeling costs. In addition, changing the focus of MLLMs to grounding tasks can degrade the original dialog capabilities due to catastrophic forgetting [68]. Recent works [4; 19; 69] reveal the inherent perception ability of MLLMs obtained by general instruction-tuning. Unlike these works that only focus on the generated tokens or in our work, we delve deeper into the more token components and reveal that the MLLMs always dynamically assign the special tokens following the instruction prompt for attending to the regions of interest.

Test-time Tuning (TTT) aims to optimize the inference of test samples online. With the progress of multimodal foundation models [43; 5]), TTT has attracted more attention [76; 40; 18; 48; 64] as it can learn effective prompts for test samples and well adapt the foundation model for zero-shot applications. TPT [48] pioneers the study on TTT by minimizing the prediction entropy between each test sample and its augmented views. HisTPT [70] explores memory learning for test-time prompt tuning by introducing the memory bank. However, the test prompt optimization for MLLMs is barely explored. The most relevant work to our work is ControlMLLM [56], which optimizes the attention map by taking the referring regions as supervision. With the lack of supervision, we propose to learn visual prompts by regularizing the token-level response to instruction input. Our work shows that we can rectify where text prompts attend to by altering the outputs of MLLMs.

3 Method

3.1 Preliminaries

Task Formulation. Given an untrimmed video $V = \{\mathrm{f}_t\}_{t=1}^{T_v}$ composed of T_v image frames and a sentence query Q, the goal of the STVG task is to localize the spatio-temporal tube of target $O = \{b_t\}_{t=t_s}^{t_e}$ described by Q. Here b_t represents the bounding box of the target in the t-th frame, t_s and t_e specify the starting and ending boundaries of the target tubelet, respectively. The STVG task can be solved through two sub-tasks: spatial grounding and temporal grounding. In our work, we first derive a zero-shot solution for the STVG task by unleashing the strong cross-modal comprehension ability of MLLMs.

The Setup of MLLMs. Current MLLMs typically consist of a visual encoder, a projector, and a large language model. Specifically, given an image-question pair (I, Q) as input, the image I is first projected into text-aligned visual tokens $\mathrm{T}_v = \{\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{M}\}$ by visual encoder and projector, and the question Q is converted into text tokens $\mathrm{T}_q = \{\mathrm{t}_1, \ldots, \mathrm{t}_{\mathrm{N_q}}\}$ by a text tokenizer and embedding layer. Here, M and $\mathrm{N_q}$ denote the numbers of visual tokens and question tokens, respectively. In practice, MLLMs also introduce system tokens $\mathrm{T_{sys}}$ and some special tokens $\mathrm{T_{role}}$ for instruction-following ability. In particular, the special tokens play an important role in structuring the well-organized conversation framework. In our work, the special tokens are positioned subsequent to the instruction prompts by the user. They help guide the generation of coherent responses. The number of special tokens $\mathrm{N_{role}}$ often differs among different MLLMs. As a result, the language model receives concatenated tokens, $\mathrm{T} = \{\mathrm{t}_1, \ldots, \mathrm{t}_{\mathrm{N_{sys}}}; \mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{M}; \mathrm{t}_1, \ldots, \mathrm{t}_{\mathrm{N_q}}; \mathrm{t}_1, \ldots, \mathrm{t}_{\mathrm{N_{role}}}\}$, as input. Appendix 6.1 provides visual illustration about the input tokens in MLLMs.

Text-to-visual Attention. The LLM in MLLMs typically processes the input tokens through L transformer blocks [37] with the multi-head attention (MHA) for interactions of different tokens. Particularly, the text-to-visual attention represents the relationships between the visual and the textual tokens. We can derive the text-to-visual attention matrix $A \in \mathbb{R}^{\mathrm{L \times H \times N \times N}}$, where $\mathrm{N} = \mathrm{N_{sys}} + \mathrm{M} + \mathrm{N_q} + \mathrm{N_{role}}$. L and H denote the numbers of layers and heads in the transformer. Our empirical observation from Fig. 1 is that the special tokens $\mathrm{T_{role}}$ show outstanding grounding ability with a global comprehension. For the simplicity of the following analysis, we omit the affect of different layers and heads by the mean operation. We can then obtain the text-to-visual attention matrix, $A_{\mathrm{role}} = [\mathrm{N_{sys}} + \mathrm{M} + \mathrm{N_q} : \mathrm{N}, \mathrm{N_{sys}} : \mathrm{N_{sys}} + \mathrm{M}]$.

3.2 Grounding Token Identification

In this section, we conduct a pilot study to quantitatively analyze the grounding ability of the special tokens. Without loss of generality, we randomly select a subset of `1,000` image-text pairs from the RefCOCOg [17] validation set for image MLLMs analysis, and a subset of `1,000` video-text pairs from the HC-STVGv2 [51] dataset for video MLLMs analysis. Particularly, we choose three typical MLLMs (*i.e.*, `LLaVA-1.5`, `Qwen-VL`, `Deepseek-VL`) for the study on image input and three MLLMs (*i.e.*, `LLaVA-Next-Video`, `Qwen2-VL`, `LLaVA-OneVision`) for the study on video input. We have two key findings from our studies.

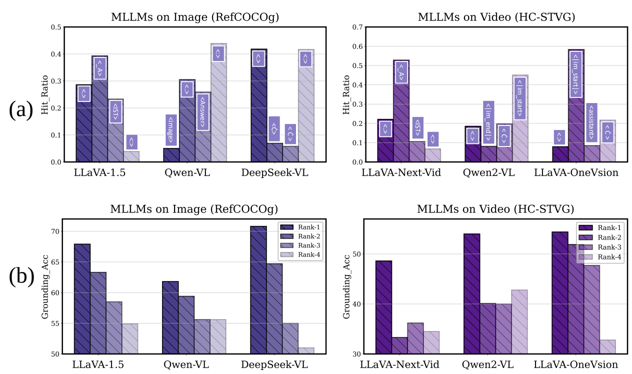


Figure 2: In (a), the results show the frequency with which different special tokens (such as **'_A'**, **'.'**) have superior grounding ability than other tokens, *i.e.*, hit_ratio. In (b), the results represent the grounding accuracy of tokens ranked by their visual activation degrees. For each MLLM, we select four tokens for visualization.

MLLMs dynamically assign the special tokens to attend to the text-related regions. To demonstrate the finding, we first define the ***attention ratio*** of each special token as the ratio of maximum attention within the ground-truth bounding box b_{gt} to that outside it. For example, the attention ratio of the special token `'_A'` can be computed as:

$$\mathrm{R}_{\mathrm{att}}^{\texttt{\_A}} = \frac{\max(A_{\mathrm{role}}^{\texttt{\_A}} \odot f_{\mathrm{B2M}}\left(b_{gt}\right))}{\max\left(A_{\mathrm{role}}^{\texttt{\_A}} \odot \left(1 - f_{\mathrm{B2M}}\left(b_{gt}\right)\right)\right)}, \qquad A_{\mathrm{role}}^{\texttt{\_A}} \in \mathbb{R}^{1 \times \mathrm{M}}, \tag{1}$$

where f_{B2M} denotes the function that transforms a bounding box into its corresponding binary mask. Here, a higher ratio indicates a better target grounding ability. Given a test sample, we can identify the token yielding the highest attention ratio as the superior token for grounding. The ***hit ratio*** of a token is then defined as the frequency of being the superior token for grounding across all test samples. Fig. 2(a) shows the hit ratio of four special tokens in each MLLM. Notably, the fixed token does not consistently exhibit the best grounding ability for different samples. For example, the highest hit ratio achieved by token `'.'` in `LLaVA-1.5` is not more than 50%. In addition, for different MLLMs, the token at a fixed position does not always yield the best localization performance. For example, the

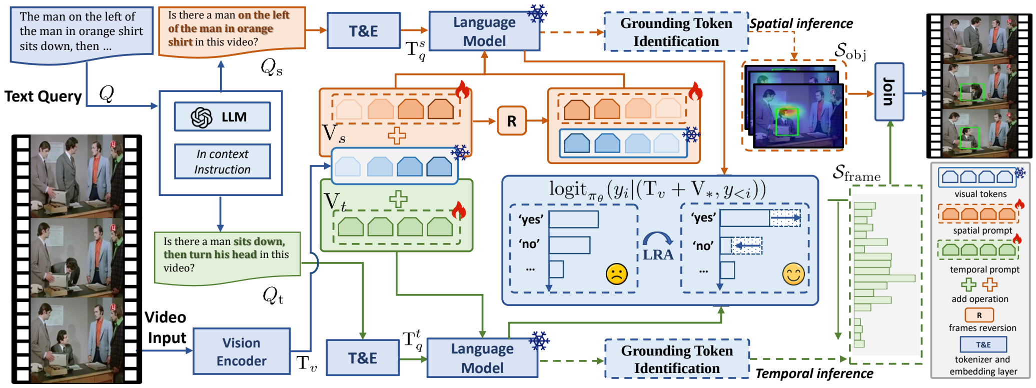


Figure 3: Overview of the proposed approach for zero-shot STVG. Given a video-text pair, we first decompose the text Q into spatially and temporally related sub-queries, Q_s and Q_t. The text prompt tokens converted from Q_s and Q_t are then concatenated with visual tokens T_v for spatial and temporal inferences, respectively. In addition, we introduce learnable variables as visual prompts and optimize them by the logit-guided re-attention (LRA) module. For spatial grounding, we also develop a temporal-augmented assembling (TAS) strategy by reversing the frames to enhance temporal consistency. After optimization, we obtain the object track score $\mathcal{S}_\mathrm{obj}$ and frame score $\mathcal{S}_\mathrm{frame}$ based on the grounding token identification. The final prediction is derived by joining $\mathcal{S}_\mathrm{obj}$ and $\mathcal{S}_\mathrm{frame}$.

last special token `':'`, which is adopted in a previous work [69], obtains a high hit ratio in `Qwen-VL`, but its hit ratio is quite low in `LLaVA-1.5`. This observation holds for both image and video inputs.

The special token with higher visual activation tends to show superior grounding performance. Since the superior grounding tokens vary across different samples and MLLMs, identifying them for grounding is a problem that needs to be resolved when ground truth is unavailable as a prior. With further analysis, we reveal that the superior token for grounding tends to show higher visual activation. For each sample, we rank the special tokens according to the maximum value of visual attention and then evaluate their grounding accuracy by selecting the proposal with the highest attention value. Following the paradigm in previous works [50; 16], we extract box proposals using a detector and evaluate the grounding accuracy by the Acc@0.5 metric. Fig. 2(b) shows the results. We can see that the grounding accuracy decreases as the rank of visual activation reduces (from left to right). This supports our hypothesis that the special token with higher visual activation tends to show superior grounding ability. We have also observed that models with better comprehension possess better grounding ability overall. For example, the special tokens in `LLaVA-OneVision` achieve better grounding performance than those of the other MLLMs.

A straightforward solution for zero-shot STVG. Inspired by the grounding ability of special tokens subsequent to the text prompt, we refer to them as ***grounding tokens*** in our work. By identifying the special tokens characterized by high visual activation, we derive a strong training-free framework for STVG. Specifically, given the video-text pair (V, Q), we first extract the object track proposals $\mathcal{O}_{pro} = \{O_1, \ldots, O_P\}$ from the video V, as done in previous works [3; 25], where P denotes the number of proposals, and $O_p = \{b'_t\}_{t=1}^{t=T_v}$ is the set of bounding boxes of the p-th proposal in T_v frames. Based on the foregoing findings, we then select the token with the highest attention value for locating the target object, and denote the text-to-image attention matrix of the selected token as $A_\mathrm{g} \in \mathbb{R}^{1\times M}$. As a result, we obtain each object track score by computing the maximum attention value inside each object track as:

$$\mathcal{S}_\mathrm{obj} = \{s_1^o, s_2^o, \cdots, s_P^o\}, \text{ with } s_p^o = \max(A_\mathrm{g} \odot f_\mathrm{B2M}(O_p)), \tag{2}$$

where $\odot$ denotes element-wise multiplication. We choose the track with the highest score as spatial prediction O'_pred. In a similar manner, we can compute the frame score $\mathcal{S}_\mathrm{frame} = \{s_1^t, s_2^t, \cdots, s_{\mathrm{T}_v}^t\}$. By selecting the top-K frames with the highest scores from the temporal prediction $\mathcal{S}_\mathrm{frame}$, we obtain the final spatio-temporal prediction $O'_\mathrm{pred} = \{b'_t\}_{t=t_{s'}}^{t_{e'}}$, where $t_{s'}$ and $t_{e'}$ denote the starting and ending boundary predictions. Though simple, this solution has achieved comparable or even superior performance than current zero-shot methods. For example, based on `LLaVA-OneVision` model, this solution achieves 23.3% on the m_vIoU metric on HC-STVGv1 dataset, which outperforms previous SOTA result (19.1%) by E3M [3].

3.3 Decomposed Spatio-Temporal Highlighting

Despite a strong performance by the above solution, these grounding tokens often neglect some important cues during spatio-temporal localization especially when processing complex video queries. As shown in Fig. 1(b), the attribute cue *'in red clothes'* is overlooked during inference, which causes the incorrect spatial localization. We observe that the language query often contains attribute and action descriptions of the target object, which are beneficial for spatial and temporal localization, respectively. To this end, we propose a novel decomposed spatio-temporal highlighting (DSTH) strategy in our framework, which aims at highlighting the attributes/ actions cues in language query and enhances the spatial/temporal reasoning, respectively.

Generation of target-related cues. For comprehensive spatial and temporal reasoning, it is essential to extract the attribute and action descriptions from the original query Q as textual cues of the target. Here, we leverage the strong in-context capability of the LLM [1] to extract attribute and action descriptions. Following previous works [62; 16], we construct a prompt with general instructions and in-context task examples. As shown in Fig. 3, we feed the prompt into LLM, and then these related descriptions Q_{s} and Q_{t} are generated through the LLM completion. More implementation details can be found in Appendix 7.3.

Spatio-temporal prompt learning. With the decomposed attribute and action descriptions as cues for spatial and temporal reasoning, the next question is how to efficiently direct the model to focus on the corresponding visual regions. Reliable responses in visual question answering (VQA) necessitate careful attention to the relevant visual context as pointed out by previous studies [58; 42]. Inspired by this, we innovatively propose regularizing the response to the questions constructed from the attribute and action descriptions for adjusting the visual attention of MLLMs. Specifically, we first transform the descriptions into interrogative queries by a fixed template to inquire the existence of the target. As shown in Fig. 3, from the original text input Q, we can obtain the attribute description '*a man on the left of the man in the orange shirt*', which is further transformed into interrogative sub-query Q_{s}: '*Is there a man on the left of the man in the orange shirt in this video?*' for spatial inquiry. In a similar way, we can obtain the interrogative sub-query Q_{t} for temporal inquiry. These interrogative sub-queries will be taken as input instructions of MLLMs for response generation.

Logit-guided re-attention. With extracted sub-queries above, we then propose a novel logit-guided re-attention (LRA) module to regularize the token prediction during response generation. Taking the spatial sub-query as an example, we first initialize a learnable variable V_s with the same shape as the visual tokens T_v, and then add it to T_v as the visual input of the language model. Sub-query Q_{s} is converted as text prompt tokens T_q^s by the text tokenizer and embedding layer. Given the input token sequence, the next token probability prediction over the vocabulary set $\mathcal{V}$ is formulated as:

$$\mathrm{p}_y = \exp\left(\mathrm{logit}_{\pi_\theta}(y_i|(\mathrm{T}_v + \mathrm{V}_s, \mathrm{T}_q^s, y_{<i}))\right), \tag{3}$$

where π_θ denotes the parameter of the language model and is frozen in our work. $\mathrm{logit}_{\pi_\theta}$ is the log probability of the generated token at time step i. $y_{<t}$ denotes the text tokens sequence prior to prediction time step i. We define the optimization objective by contrasting the probabilities of positive token *'yes'* and negative token *'no'*:

$$\mathcal{L}_s = 1 - \exp\left(\mathrm{logit}_{\pi_\theta}(y_i^{yes}|(\mathrm{T}_v + \mathrm{V}_s, \mathrm{T}_q^s, y_{<i})) - \mathrm{logit}_{\pi_\theta}(y_i^{no}|(\mathrm{T}_v + \mathrm{V}_s, \mathrm{T}_q^s, y_{<i}))\right). \tag{4}$$

During the inference process, we conduct backpropagation to optimize the learnable variable V_s as the spatial prompt. The process is iterated N_{ep} times by test-time tuning paradigm. By enhancing the positive response towards the sub-query Q_{s}, we can prompt the MLLM to effectively mine target-related contextual information during VQA, which in turn highlights the attribute cues in the original text. Similarly, we can obtain the temporal prompt V_t by optimizing the temporal inference.

Joint inference. Based on the spatial and temporal visual prompts, we derive the attention maps A_{g}^S and $A_{\mathrm{g}}^T \in \mathbb{R}^{\mathrm{T}_v \times \mathrm{h} \times \mathrm{w}}$ of the special token with high visual activation for spatial and temporal predictions, where h and w denote the token numbers of the height and width. From Eq. 2, we obtain the object track score $\mathcal{S}_{\text{obj}}$ and the temporal score $\mathcal{S}_{\mathrm{frame}}$ based on A_{g}^S and A^T, respectively. Finally, we integrate the predictions $O'_{\text{pred}} = \{b'_t\}_{t=t_{s'}}^{t_{e'}}$ as the spatio-temporal grounding result.

3.4 Temporal-augmented Assembling

The attribute sub-query, which provides a static state description, exhibits temporal independence for spatial grounding. In other words, the spatial grounding by the attribute sub-query should be

Table 1: Quantitative comparison on HCSTVG (v1&v2) and VidSTG (Declarative) benchmarks.

Sup	Method	HCSTVG-v1			HCSTVG-v2			VidSTG (Declarative)		
		m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
Full	TubeDETR [59] [CVPR2022]	32.4	49.8	23.5	36.4	58.8	30.6	30.4	42.5	28.2
	STCAT [20] [NeurIPS2022]	35.0	57.7	30.0	–	–	–	33.1	46.2	32.6
	CSDVL [33] [CVPR2023]	36.9	62.2	34.8	38.7	65.5	33.8	33.7	47.2	32.8
	CG-STVG [15] [CVPR2024]	38.4	61.5	36.3	39.5	64.5	36.3	34.0	47.7	33.1
Weak	WINNER [29] [CVPR2023]	14.2	17.2	6.1	–	–	–	11.6	14.1	7.4
	VEM [21] [ECCV2024]	14.6	18.6	5.8	–	–	–	14.5	18.6	8.8
	CoSPaL [25] [ICLR2025]	22.1	31.8	19.6	22.2	31.4	18.9	16.0	20.1	13.1
	STPro [13] [CVPR2025]	17.6	27.0	12.9	20.0	31.1	14.6	15.5	19.4	12.7
ZS	RedCircle [47] [CVPR2023]	9.2	7.8	1.6	–	–	–	8.6	7.6	0.9
	ReCLIP [50] [ACL2022]	14.4	18.3	4.9	–	–	–	14.2	17.5	7.9
	E3M [3] [ECCV2024]	19.1	29.4	10.6	–	–	–	16.2	20.5	11.9
	Ours LLaVA-Next-Video-7B	20.4	33.6	12.4	23.6	36.8	15.5	16.6	26.8	11.1
	Ours ShareGPT4Video-8B	20.0	32.2	10.9	24.4	38.9	15.4	17.1	27.8	11.6
	Ours Qwen2-VL-7B	23.6	39.0	14.4	25.6	40.5	17.1	17.0	27.4	11.4
	Ours LLaVA-OneVision-7B	24.8	41.5	16.3	27.7	44.7	19.5	18.0	29.8	12.2

temporally consistent for temporal augmentation (*e.g.*, reversing the order of video frames). However, there exists temporal inconsistency when introducing temporal augmentation in current MLLMs. Here, we propose a metric to measure the temporal consistency. By denoting the spatial attention maps before and after reversing the order of input frames as A_{g}^{S} and $\tilde{A}_{\mathrm{g}}^{S}$, the temporal consistency can be measured as:

$$S_{\text{cons}} = \max\{s_1, \cdots, s_P\}, \quad s_p = \max\left((A_{\mathrm{g}}^{S} \odot f_{\text{B2M}}(O_p)) \odot (\tilde{A}_{\mathrm{g}}^{S} \odot f_{\text{B2M}}(O_p)) \right), \tag{5}$$

where a higher S_{cons} indicates better temporal consistency. We then divide the testing samples into ten groups in descending order of temporal consistency. Fig. 4 shows the average grounding accuracy of each group samples. The results demonstrate a pronounced association between temporal consistency and spatial grounding performance. The temporal inconsistency tends to cause worse spatial localization.

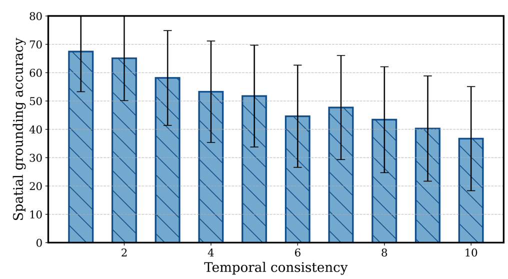


Figure 4: Spatial grounding accuracy of different groups of samples on the HC-STVGv1 dataset. These groups are ranked by descending temporal consistency.

To this end, we integrate a temporal-augmented assembling (TAS) strategy in our framework. As shown in Fig. 3, we perform a frame-level reversion operation on the visual tokens and the spatial prompt simultaneously, and then optimize the spatial prompts for the input of the original frames and the temporal-augmented input, respectively. During inference, we derive the spatial prediction by assembling the attention maps of temporal-augmented input frames. The proposed TAS strategy alleviates the effect of temporal inconsistency and improves the robustness of spatial grounding.

4 Experiments

4.1 Settings

Datasets. We evaluate on three video benchmark datasets: HCSTVG-v1, HCSTVG-v2 [51], and VidSTG [73]. We provide detailed introduction about them in Appendx 7.1.

Implementation Details. We adopt G-DINO [38] and SAM2 [45] for detection and tracking tubelet generation, and use GPT-4o to decompose the original query sentence into spatial and temporal sub-queries. We consider four widely-used video MLLMs: `LlaVA-Next-Video-7B` [27], `Qwen2-VL-7B` [53], `ShareGPT4Video-8B` [7], `LlaVa-OneVision-7B` [27] for demonstrating the efficiency of our method. Refer to Appendix 7.2 for more details.

Evaluation Metrics. We follow the standard evaluation protocol [59; 25] and use $\mathrm{m\_vIoU}$, and $\mathrm{vIoU@R}$ to assess the performance of spatio-temporal grounding. Specifically, let S_i, S_u denote the intersection and union between the predicted and ground-truth frames. The vIoU is computed by $\frac{1}{\mathrm{S}_u}\sum_{t\in\mathrm{S}_i}\mathbf{IoU}(b'_t, b_t)$, where b'_t and b_t denote the detected and ground-truth bounding box at frame t, respectively. The $\mathrm{m\_vIoU}$ represents the vIoU averaged over all testing samples, and $\mathrm{vIoU@R}$ denotes the proportion of data samples in the testing subset with vIoU greater than the threshold $\mathrm{R} \in \{0.3, 0.5\}$.

4.2 Performance Comparison

Quantitative Comparison. Tab. 1 presents a comparison of our method against 11 methods from three categories, including zero-shot, weakly-supervised, and fully-supervised methods. Specifically, we compare our approach with existing zero-shot SOTA approaches (*e.g.*, E3M [3], ReCLIP [50], RedCircle [47]). Our method consistently outperforms these methods by a remarkable margin on all benchmarks. Based on `LLaVA-Next-Video-7B`, our method outperforms E3M by 4.2% on $\mathrm{vIoU@0.3}$ and 1.8% on $\mathrm{vIoU@0.5}$. When integrated into the better `LLaVA-OneVision-7B` model, the corresponding improvements reach 12.1% and 5.7%. This shows that our framework can adapt to various MLLMs well, and better performances can be achieved by using better MLLMs. Even on the VidSTG dataset, which contains fewer action cues related to the target and thus makes temporal grounding particularly challenging, our framework still outperforms the previous SOTA methods overall. This demonstrates the strong generalization capability of our method. Besides, our method even surpasses current SOTA weakly-supervised methods on most metrics. For example, on the HCTSVG-v2 benchmark, our method outperforms CoSPaL [25] by a margin of 5.5% on the $\mathrm{m\_vIoU}$ metric. Furthermore, compared to the fully-supervised methods, our method can still achieve comparable results, which further validates the superiority of our approach. We provide quantitative comparison on the VidSTG (Interrogative) and image benchmarks in Appendix 8.1

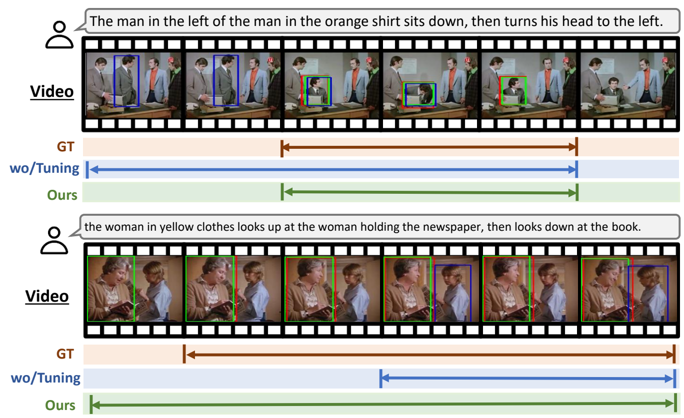


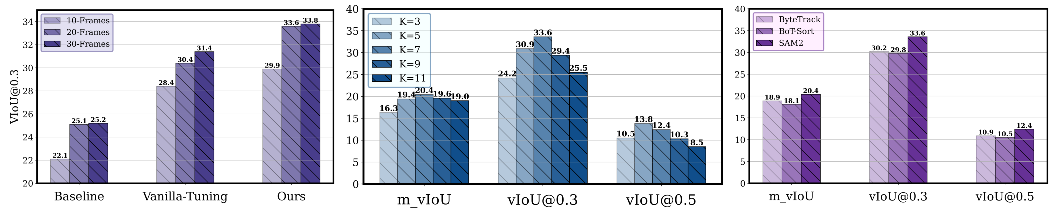
Figure 5: Qualitative results on the HC-STVGv1 test set. Better spatio-temporal grounding results (green) are obtained when the DSTH strategy is being used for optimization.

Qualitative Comparison. We present qualitative results in Fig. 5. In the example below, before introducing the DSTH optimization strategy, the model neglects the attribute cues in the language query and suffers from the spatial grounding error. By highlighting the attribute cues of the target, our method can direct the MLLMs toward reliable visual context and improve spatial localization.

4.3 Ablation Study

In this section, based on HC-STVGv1 dataset, we analyse the effect of different proposed components when integrated into `LlaVa-Next-Video` and `LlaVa-OneVision`. We also conduct extensive ablation experiments with the hyper-parameters based on `LlaVa-Next-Video`.

Component analysis. In Tab. 2, we first average the attention maps of all special tokens (***1st*** row) as the baseline. This solution has achieved outstanding performance. Next, in the ***2nd*** row, we integrate the selection of the superior token introduced in grounding token identification (GTI). It brings consistent improvements on different MLLMs. The results show that identifying the superior special token can effectively unleash the powerful comprehension ability of MLLMs. We them validate the efficiency of the decomposed spatio-temporal highlighting strategy (***3rd*** and ***4th*** rows). $\mathbf{S}_{\text{tune}}$ and $\mathbf{T}_{\text{tune}}$ denote adopting the prompt learning in spatial and temporal inferences, respectively. It is demonstrated that the MLLMs can be directed to focus on the spatial/temporal related regions better



(a) (b) (c)

Figure 6: (a) Comparison on the number of frames N_f as input, using vIoU@0.3. (b) Ablation on the number of selected frames K during temporal prediction. (c) Ablation on different trackers.

when highlighting the attribute/action cues with proper prompts. When utilizing prompt learning on the two sub-tasks, the final performance can be further refined (***5th*** row). In addition, we also find that the DSTH strategy can improve the grounding performance especially for the less efficient MLLMs (*e.g.*, `LLaVA-Next-Video`). Even for the MLLMs (*e.g.*, `LLaVA-OneVision`) with strong temporal comprehension ability, performance improvement can still be achieved by the test-time optimization. Finally, we introduce the temporal-augmented assembling (TAS) strategy (***6th*** row). The overall performance can be further improved by refining the spatial localization.

Table 2: Component ablation on `LLaVA-Next-Video` and `LLaVA-OneVision`.

GTI	$\mathrm{S_p}$	$\mathrm{T_p}$	TAS	LLaVA-Next-Video			LLaVA-OneVision		
				m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
✗	✗	✗	✗	15.2	25.1	8.5	21.3	36.1	12.6
✓	✗	✗	✗	16.3	26.6	9.3	23.3	38.8	15.1
✓	✓	✗	✗	18.0	30.1	10.1	24.1	40.3	16.0
✓	✗	✓	✗	18.4	29.5	10.1	23.8	39.7	15.5
✓	✓	✓	✗	19.9	32.1	11.9	24.3	40.7	16.0
✓	✓	✓	✓	20.4	33.6	12.4	24.8	41.5	16.3

Why does spatial and temporal prompt learning by contrasting logits of *'yes'* and *'no'* work. Though contrasting logits of *'yes'* and *'no'* tokens is heuristic, it is reasonable. We empirically find that given the question prompt, the model often gives ambiguous or even wrong response predictions even though the objects referred to by the text do indeed exist in the video. This indicates that the model cannot understand the video content well in certain cases. By optimizing the logit of the *'yes'* token toward positive prediction instead of the *'no'* token, we encourage the model to carefully attend to the target-related visual information and achieve better visual localization.

To further validate the hypothesis, we measure the distribution of attention peaks before and after test-time optimization. Based on the evaluation on the HC-STVG dataset, we find that the average attention peak before test-time optimization is about 6.7 (without normalization), and after optimization, it increases to 23.9 (without normalization). This change in attention values indicates that our proposed optimization strategy can encourage the model to perceive more visual modal information for generating visually-grounded responses, thereby justifying our previous hypothesis.

Comparison of different learning strategies. We propose logit-guided re-attention by contrasting logits of *'yes'* and *'no'* tokens. Here, we also consider entropy minimization (i.e., increasing the predicted probability of the *'yes'* token across the entire vocabulary) during test-time tuning. The results obtained by `LLaVa-Next-Video-7B` on the HC-STVG-v2 dataset are shown in the table below. The ***1st*** row denotes the results of introducing grounding token identification (GTI). The ***2nd*** and ***3rd*** rows denote the test-time tuning results achieved by entropy minimization and contrasting the logits of *'yes'* and *'no'*, respectively. We find that entropy minimization does not bring noticeable improvement. We believe this is because entropy minimization may be achieved by reducing the token prediction of abundant irrelevant tokens in the vocabulary while ignoring the differentiation between the *'yes'* and the *'no'* token. As a result, it is less efficient than explicitly contrasting the logits in promoting the model to attend to abundant visual information.

Table 3: Comparison of different learning strategies.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
GTI	19.2	28.5	12.4
GTI + entropy minimization	19.5	28.5	12.5
GTI + contrasting the logits	22.1	32.9	14.0

Generality analysis of our method. We conduct the scalability verification on `InternVL3` and `Qwen2.5-VL` series models. With limited computational resources, we evaluate our method on `InternVL3-4B`, `InternVL3-8B`, `Qwen2.5-VL-3B` and `Qwen2.5-VL-7B` models in the table below. We average the attention maps of all special tokens as the baseline. Benefiting from our grounding token identification and test-time optimization, we can obviously achieve better grounding performance not only in the smaller but also the larger models, which well demonstrates both the generality and scalability of our method. Moreover, we find that the model size significantly impacts the inherent grounding ability of special tokens. The performance of small models (*e.g.*, `InternVL3-4B`) is inferior to the larger models (*e.g.*, `InternVL3-8B`). It is because fewer parameters in language models perform poorly in multimodal alignment.

Table 4: Evaluation on different scales of models.

Methods	vIoU@0.3	m_vIoU	t_iou	gt_viou
`InternVL3-4B` + Baseline	24.5	17.0	38.4	38.2
`InternVL3-4B` + Ours	41.1	25.4	45.5	49.5
`InternVL3-8B` + Baseline	36.5	22.9	42.9	46.1
`InternVL3-8B` + Ours	45.3	27.5	47.2	53.2
`Qwen2.5-VL-3B` + Baseline	28.5	17.3	37.1	39.1
`Qwen2.5-VL-3B` + Ours	41.7	25.4	45.9	49.3
`Qwen2.5-VL-7B` + Baseline	39.7	23.6	44.1	47.2
`Qwen2.5-VL-7B` + Ours	45.4	27.7	46.9	51.6

Effect of input frames N_f. Fig. 6(a) analyzes the effect of using different numbers of video frames N_f as input. We can see that more input frames provide richer visual context for model inference and lead to a better overall performance. Besides, the performance becomes slightly peaked as the number of input frames increases. To balance between performance and efficiency during inference, we uniformly sample 20 frames as the visual input of MLLMs. In Fig. 6, we also evaluate `Vanilla-Tuning`, a solution that directly applies test-time optimization to the entire text query Q instead of decomposing it into attribute/action sub-queries. We observe that the performance of `Vanilla-Tuning` is inferior to our method. It demonstrates that decomposing the text query into attribute/action sub-queries can help facilitate the intensive spatio-temporal comprehension of MLLMs.

Ablation on the number K of predicted frames. Fig. 6(b) analyzes the effect of predicted frame numbers K for temporal grounding based on the HCSTVG-v1 dataset. Here, we consider selecting the optimal frame number from the set $\{3, 5, 7, 9, 11\}$. In general, when K is set as 7, the optimal result can be obtained. Thus, we select top-7 frames as the prediction during temporal grounding.

Trackers ablation. Fig. 6(c) analyzes the effect of different trackers. Besides SAM2 [45], a foundation model for tracking, we also consider two other tracking models (*i.e.*, ByteTrack [72] and BoTSort [2]) for analysis. It is reasonable that when a stronger tracking model (*e.g.*, SAM2) is used for generating spatio-temporal tubelets, we can obtain better STVG performances. Besides, when suboptimal tracking models are used, our method can still achieve comparable or better performances than the current SOTA methods, which shows the generalization of our approach.

5 Conclusion

In this paper, we have presented a novel MLLM-based zero-shot framework for the spatio-temporal video grounding (STVG) task. Our approach is initiated by identifying the grounding capability of special tokens in widely used MLLMs during response generation. To leverage and unleash the comprehension ability of MLLMs, we have proposed the decomposed spatio-temporal highlighting (DSTH) strategy. It first decomposes the text query into attributes and actions sub-queries. It then employs a logit-guided re-attention (LRA) module to sharpen the spatial/temporal visual context comprehension. We have also proposed the temporal-augmented assembling (TAS) strategy to alleviate the effect of temporal inconsistency. Extensive experiments conducted on three STVG benchmarks demonstrate the effectiveness of our proposed framework.

Our approach does have limitations. For example, it may struggle to process long videos well due to high computational consumption caused by MLLMs. As a future work, we would like to consider incorporating token pruning and key frame selection techniques into the model design.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)

[2] Aharon, N., Orfaig, R., Bobrovsky, B.Z.: Bot-sort: Robust associations multi-pedestrian tracking. arXiv:2206.14651 (2022)

[3] Bao, P., Shao, Z., Yang, W., Ng, B.P., Kot, A.C.: E3m: zero-shot spatio-temporal video grounding with expectation-maximization multimodal modulation. In: ECCV. pp. 227–243. Springer (2024)

[4] Cao, S., Gui, L.Y., Wang, Y.X.: Emerging pixel grounding in large multimodal models without grounding supervision. arXiv:2410.08209 (2024)

[5] Chen, D., Wu, Z., Liu, F., Yang, Z., Zheng, S., Tan, Y., Zhou, E.: Protoclip: Prototypical contrastive language image pretraining. TNNLS (2023)

[6] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv:2306.15195 (2023)

[7] Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Lin, B., Tang, Z., et al.: Sharegpt4video: Improving video understanding and generation with better captions. In: NeurIPS (2024)

[8] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)

[9] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv:2309.16588 (2023)

[10] Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: ICCV. pp. 1769–1779 (2021)

[11] Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV. pp. 16321–16330 (2021)

[12] Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: ICCV. pp. 2886–2895 (2021)

[13] Garg, A., Kumar, A., Rawat, Y.S.: Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding. arXiv:2502.20678 (2025)

[14] Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV. pp. 540–557. Springer (2022)

[15] Gu, X., Fan, H., Huang, Y., Luo, T., Zhang, L.: Context-guided spatio-temporal video grounding. In: CVPR. pp. 18330–18339 (2024)

[16] Han, Z., Zhu, F., Lao, Q., Jiang, H.: Zero-shot referring expression comprehension via structural similarity between images and captions. In: CVPR. pp. 14364–14374 (2024)

[17] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV. pp. 108–124. Springer (2016)

[18] Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727. Springer (2022)

[19] Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., Yang, X.: Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. arXiv:2411.16724 (2024)

[20] Jin, Y., Li, Y., Yuan, Z., Mu, Y.: Embracing consistency: a one-stage approach for spatio-temporal video grounding. In: NeurIPS. pp. 29192–29204 (2022)

[21] Jin, Y., Mu, Y.: Weakly-supervised spatio-temporal video grounding with variational cross-modal alignment. In: ECCV. pp. 412–429. Springer (2024)

[22] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: CVPR. pp. 1780–1790 (2021)

[23] Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. arXiv:2503.03321 (2025)

[24] Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. arXiv:2503.06287 (2025)

[25] Kumar, A., Kira, Z., Rawat, Y.S.: Contextual self-paced learning for weakly supervised spatio-temporal video grounding. arXiv:2501.17053 (2025)

[26] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)

[27] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv:2408.03326 (2024)

[28] Li, H., Chen, J., Wei, Z., Huang, S., Hui, T., Gao, J., Wei, X., Liu, S.: Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. arXiv:2501.08282 (2025)

[29] Li, M., Wang, H., Zhang, W., Miao, J., Zhao, Z., Zhang, S., Ji, W., Wu, F.: Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In: CVPR. pp. 23090–23099 (2023)

[30] Liang, Y., Cai, Z., Xu, J., Huang, G., Wang, Y., Liang, X., Liu, J., Li, Z., Wang, J., Huang, S.L.: Unleashing region understanding in intermediate layers for mllm-based referring expression generation. NeurIPS **37**, 120578–120601 (2024)

[31] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: EMNLP. pp. 5971–5984 (2024)

[32] Lin, F., Yuan, J., Wu, S., Wang, F., Wang, Z.: Uninext: Exploring a unified architecture for vision recognition. In: ACMMM. pp. 3200–3208 (2023)

[33] Lin, Z., Tan, C., Hu, J.F., Jin, Z., Ye, T., Zheng, W.S.: Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In: CVPR. pp. 23100–23109 (2023)

[34] Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)

[35] Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., Hancke, G., Lau, R.: Referring image segmentation using text supervision. In: ICCV. pp. 22124–22134 (2023)

[36] Liu, F., Liu, Y., Xu, K., Ye, S., Hancke, G.P., Lau, R.W.: Language-guided salient object ranking. In: CVPR. pp. 29803–29813 (2025)

[37] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[38] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024)

[39] Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: ECCV. pp. 417–435. Springer (2024)

[40] Ma, J., Li, H., Xiang, X.: Decoupled entropy minimization. In: NeurIPS (2025)

[41] Munasinghe, S., Thushara, R., Maaz, M., Rasheed, H.A., Khan, S., Shah, M., Khan, F.: Pg-video-llava: Pixel grounding large video-language models. arXiv:2311.13435 (2023)

[42] Prasad, A., Stengel-Eskin, E., Bansal, M.: Rephrase, augment, reason: Visual grounding of questions for vision-language models. arXiv:2310.05861 (2023)

[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)

[44] Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: CVPR. pp. 13009–13018 (2024)

[45] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv:2408.00714 (2024)

[46] Shen, H., Zhao, T., Zhu, M., Yin, J.: Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In: AAAI. vol. 38, pp. 4766–4775 (2024)

[47] Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. In: ICCV. pp. 11987–11997 (2023)

[48] Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: NeurIPS (2022)

[49] Su, R., Yu, Q., Xu, D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: ICCV. pp. 1533–1542 (2021)

[50] Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. In: ACL. pp. 5198–5215 (2022)

[51] Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. TCSVT **32**(12), 8238–8249 (2021)

[52] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357. PMLR (2021)

[53] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191 (2024)

[54] Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In: CVPR. pp. 18909–18918 (2024)

[55] Woo, S., Kim, D., Jang, J., Choi, Y., Kim, C.: Don't miss the forest for the trees: Attentional vision calibration for large vision language models. arXiv:2405.17820 (2024)

[56] Wu, M., Cai, X., Ji, J., Li, J., Huang, O., Luo, G., Fei, H., JIANG, G., Sun, X., Ji, R.: Controlmllm: Training-free visual prompt learning for multimodal large language models. In: NeurIPS (2024)

[57] Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)

[58] Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: CVPR. pp. 13204–13214 (2024)

[59] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: CVPR. pp. 16442–16453 (2022)

[60] Yang, Z., Ke, Z., Hancke, G.P., Lau, R.W.: Cross-domain semantic decoupling for weakly-supervised semantic segmentation. In: BMVC (2023)

[61] Yang, Z., Liu, Y., Xu, W., Huang, C., Zhou, L., Tong, C.: Learning prototype via placeholder for zero-shot recognition. In: IJCAI. pp. 1559–1565 (2022)

[62] Yang, Z., Liu, Y., Lin, J., Hancke, G., Lau, R.: Boosting weakly supervised referring image segmentation via progressive comprehension. Advances in Neural Information Processing Systems **37**, 93213–93239 (2024)

[63] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022)

[64] Yoon, H.S., Yoon, E., Tee, J.T.J., Hasegawa-Johnson, M.A., Li, Y., Yoo, C.D.: C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In: ICLR (2024)

[65] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: ICLR (2024)

[66] Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: CVPR. pp. 28202–28211 (2024)

[67] Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv:2505.17685 (2025)

[68] Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language models. arXiv:2309.10313 (2023)

[69] Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. arXiv:2502.17422 (2025)

[70] Zhang, J., Huang, J., Zhang, X., Shao, L., Lu, S.: Historical test-time prompt tuning for vision foundation models. In: NeurIPS (2024)

[71] Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C.C., Shuicheng, Y.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In: NeurIPS

[72] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV. pp. 1–21. Springer (2022)

[73] Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: CVPR. pp. 10668–10677 (2020)

[74] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)

[75] Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV. pp. 696–712. Springer (2022)

[76] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)

[77] Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: ECCV. pp. 598–615. Springer (2022)

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: In Sec.1, we make the claims to reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA 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 NA 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: In Appendix 9, we discuss the limitations of the work.

   Guidelines:

   - The answer NA 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: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA 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 provide implementation details in Sec. 4.1 and the Appendix for reproducibility.

   Guidelines:

   - The answer NA 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: We will consider releasing the data and code once the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details about the experimental setting (*e.g.*, hyper-parameters) in Sec. 4.1 and the Appendix for reproducibility.

   Guidelines:

   - The answer NA 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: [No]

   Justification: We do not report error bars due to limited computing resources.

   Guidelines:

   - The answer NA 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: We provide the information on the computer resources in Appendix 7.2.

   Guidelines:

   - The answer NA 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 conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: The paper discusses societal impacts of the work in the Appendix 10.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA 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: The paper properly cites the existing assets.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendices of Unleashing the Potential of Multimodal LLMs for Zero-Shot Spatio-Temporal Video Grounding

6 Grounding Tokens

6.1 Illustration of Input Tokens Components

In Fig. 7, we give a clear illustration of input token components in MLLMs. In practice, besides the visual and text prompt tokens (② and ③), there are also system tokens and some special tokens (① and ④) equipped for language generation. In particular, the grounding ability of special tokens subsequent to the text instruction prompt is significantly undervalued in previous works. Here we focus on exploring the grounding ability of these special tokens. For different multi-modal large language models, there often are different special tokens due to different instruction-tuning process. For example, in `Llava-1.5`, the introduced special tokens include: {‘`:`’, ‘`ANT`’, ‘`IST`’, ‘`SS`’, ‘`_A`’}. While for the `Qwen2-VL`, these special tokens include: { Ċ, ‘`assistant`’, ‘`<|im_start|>`’, ‘`<|im_end|>`’}. Note that we also consider the ‘`.`’ as one of the special tokens considering that it is located in the end of the instruction prompt and is characteristic of sink tokens [23; 57].

In fact, special tokens are predefined symbols within a language model’s vocabulary. They focus on guiding the model’s processing instead of representing real words and guide the model to generate coherent and context-aware responses. In the scenario explored in our work (*i.e.*, dialogue systems), special tokens can differentiate between a user’s question and the assistant’s answer. By using special tokens, models become better at understanding structure and context.

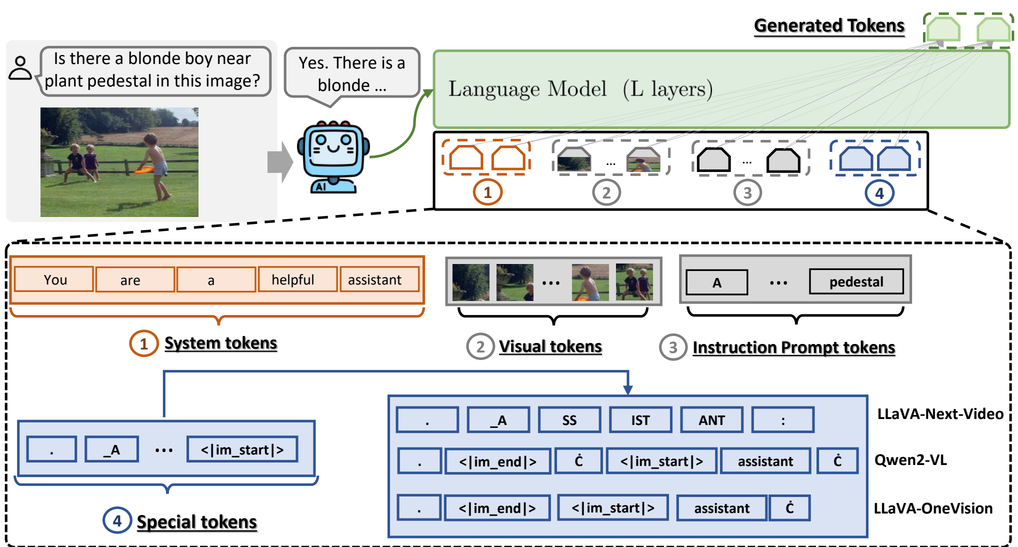


Figure 7: Illustration of the tokens input in MLLMs-based dialog system. Here, we take the image question answering as an example.

6.2 Quantitative Analysis of Special Tokens

In our work, we randomly selected a subset of 1000 image-text pairs from RefCOCOg [17] validation set for image MLLMs analysis, and a subset of 1000 video-text pairs from HC-STVG [51] dataset for video MLLMs analysis. Particularly, we choose six typical MLLMs (*i.e.*, `LLaVA-1.5`, `Qwen-VL`, `Deepseek-VL`, `LLaVA-Next-Video`, `Qwen2-VL`, `LlavA-OneVision`) for pilot studies. In the following, we will introduce our findings about special tokens in MLLMs.

Some special tokens show outstanding grounding ability for text prompt input. We compare the grounding accuracy of G-DNIO [38] and the special tokens’ ones by probing the attention map. Here we adopt the G-DNIO with Swin-T backbone and it is not finetuned on the refcoco series datasets. Following previous visual grounding works [50; 16], we adopt the IoU@0.5 as the metric of grounding accuracy. The results evaluated on refcoco series benchmarks are shown in the Fig. 8. The yellow horizontal line denotes the result by G-DINO model. Though the G-DINO is trained on image-text pairs in fully-supervised manner, some special tokens still achieve comparable and even

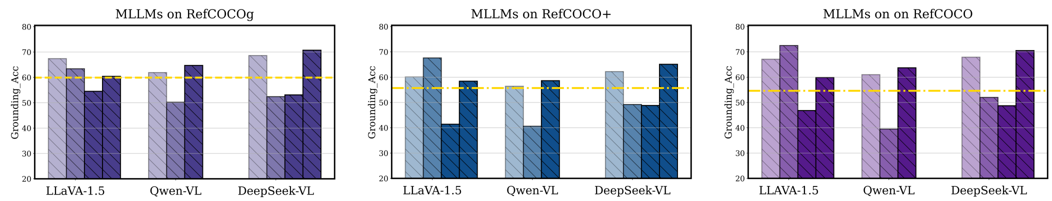


Figure 8: Comparison of grounding accuracy between G-DINO and special tokens on RefCOCO series datasets.

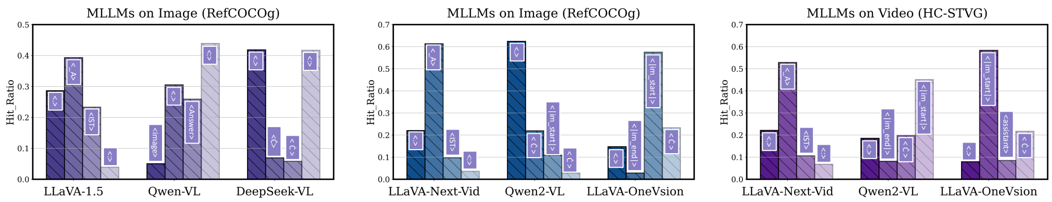


Figure 9: The hit ratio of different special tokens in image and video MLLMs.

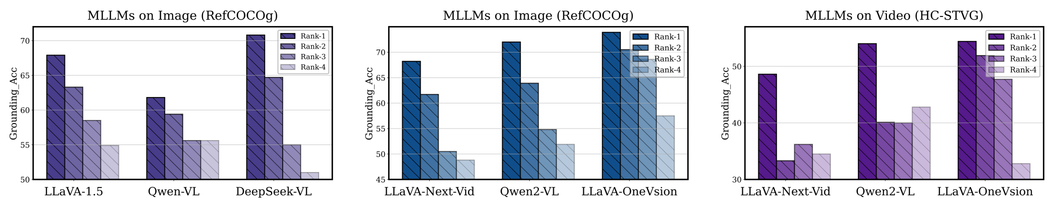


Figure 10: Grounding accuracy of special tokens ranked by the visual activation degree.

better performance. The results show that the special token can effectively ground the text prompt input by integrating the textual cues. Besides, we also notice that there exists a pronounced difference for grounding ability of different special tokens.

MLLMs dynamically assign the special tokens to attend to the text-related regions. For further exploration on the grounding ability of special tokens, we define the ***attention ratio*** of each special token as the ratio of maximum attention within the ground-truth bounding box b_{gt} to that outside it. Here, a higher ratio indicates better target grounding ability. Given a test sample, we can identify the token that yields the highest attention ratio as the superior token for grounding. Then a token's ***hit ratio*** is defined as the frequency of being the superior token for grounding across all test samples. The Fig. 9(a) shows the hit ratio of special tokens in each MLLM. We choose four special tokens for visualization. In addition to the findings that MLLMs dynamically assign the special tokens to attend to the text-related regions, we also observe that the superior token for grounding in video MLLMs shows a more concentrated trend than image MLLMs. For example, the highest hit ratio by `Qwen2-VL` is more than 60%. Besides, for the same special token in the MLLM, the hit ratio in the case of different datasets may be significantly different. For example, the token '`.`' shows a high hit ratio in RefCOCOg dataset, but its hit ratio is quite low in the HC-STVG dataset.

The special token with higher visual activation tends to show superior grounding performance. In our work, with further analysis, we reveal that the superior token for grounding tends to show higher visual activation. For each sample, we rank the special tokens according to the maximum value of visual attention and evaluate their grounding accuracy by selecting the proposal with the highest attention value as the prediction. In practice, we extract box proposals by the detector (*e.g.*, G-DINO) and evaluate the grounding accuracy by the Acc@0.5 metric. Fig. 10 shows the results. We can see that the grounding accuracy decreases as the rank of visual activation reduces (from left to right). We also observe that models with better comprehension overall possess better grounding ability.

6.3 Qualitative Analysis of Special Tokens

In Fig. 11, we visualize token-to-visual attention maps using some image-text pairs by `LLaVA-1.5-7B` model. In particular, for each image-text pair, we visualize the visual attendance of the semantic tokens from the text prompt and special tokens. The green bounding box denotes the ground truth of the target object. Notably, we notice that the semantic tokens in text prompts can provide tangible attention to related visual entities by the limited cues, while the special tokens are often assigned to integrate global instruction cues and attend to the exact target region. For example, given the text prompt *'a man getting ready to cut a cake'*, the token `'man'` and `'cake'` both show reasonable visual response. The special token `'_A'` can accurately attend to the target person. In addition, we see that the special token with hight visual activation often shows better grounding performance. However, we also see that the special tokens may attend to the wrong instance even though the semantic tokens can properly capture their region of interest. It suggests that there is still room for improvement in inference in current multimodal language models.

In Fig. 12, we visualize attention maps of special tokens using some video-text pairs by `ShareGPT4Video-8B` model. The ground-truth spatio-temporal tube is denoted by red bounding boxes. Each row shows the attention maps of a specific token. We find that only a few special tokens can well capture the corresponding target regions, which necessitates the efficient selection of grounding tokens. Also, in some cases (*e.g.*, the below sample), there may be more than one special token that will pay attention to the target area. This shows that achieving localization by considering the integration of multiple effective special tokens can be a direction for improvement. We leave it to future research.

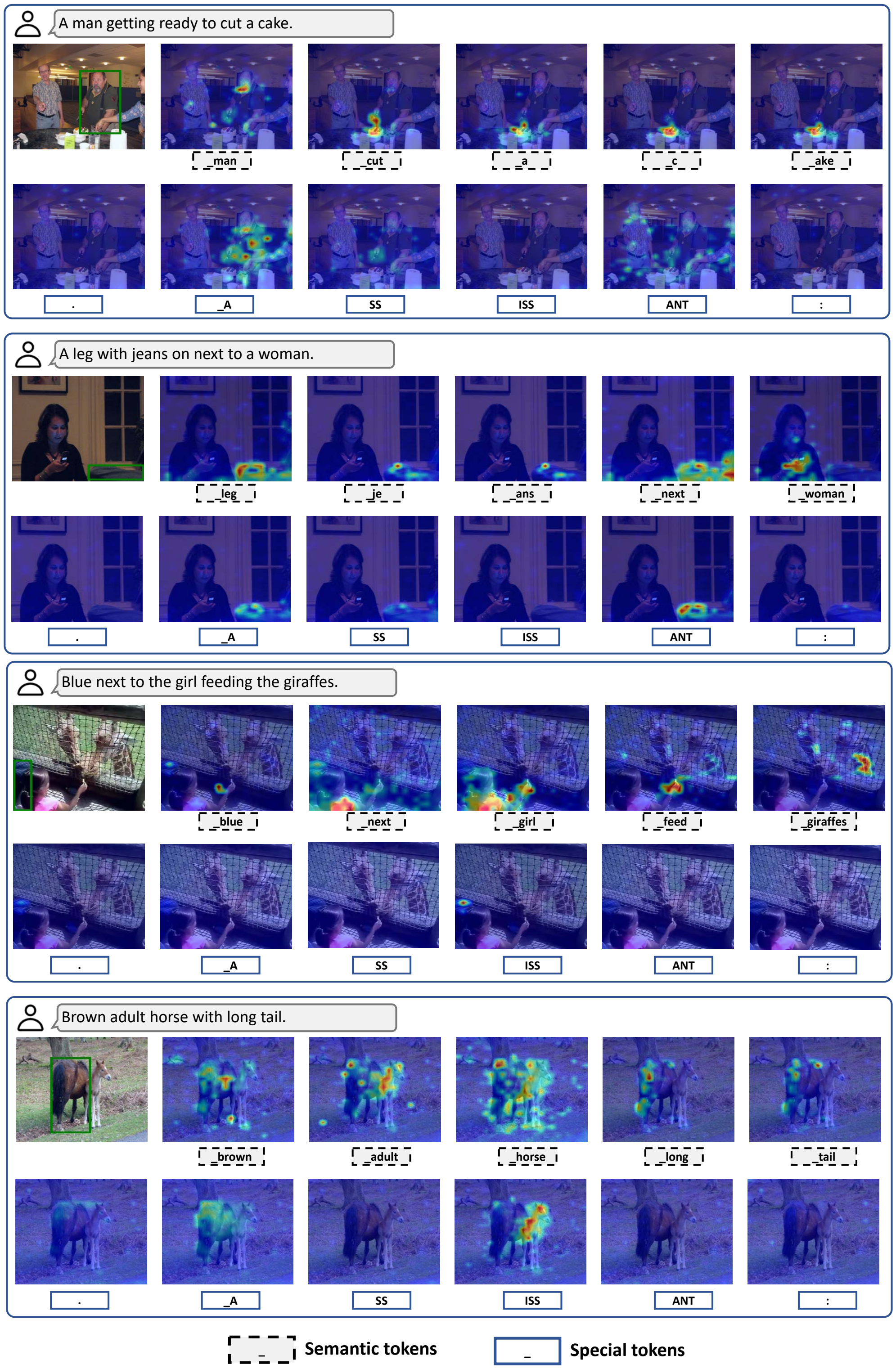


Figure 11: Visualization of tokens' attention maps on image-text pairs. The tokens with dotted box denote the semantic tokens in text prompts while the tokens with solid box denote the special tokens following the text prompts.

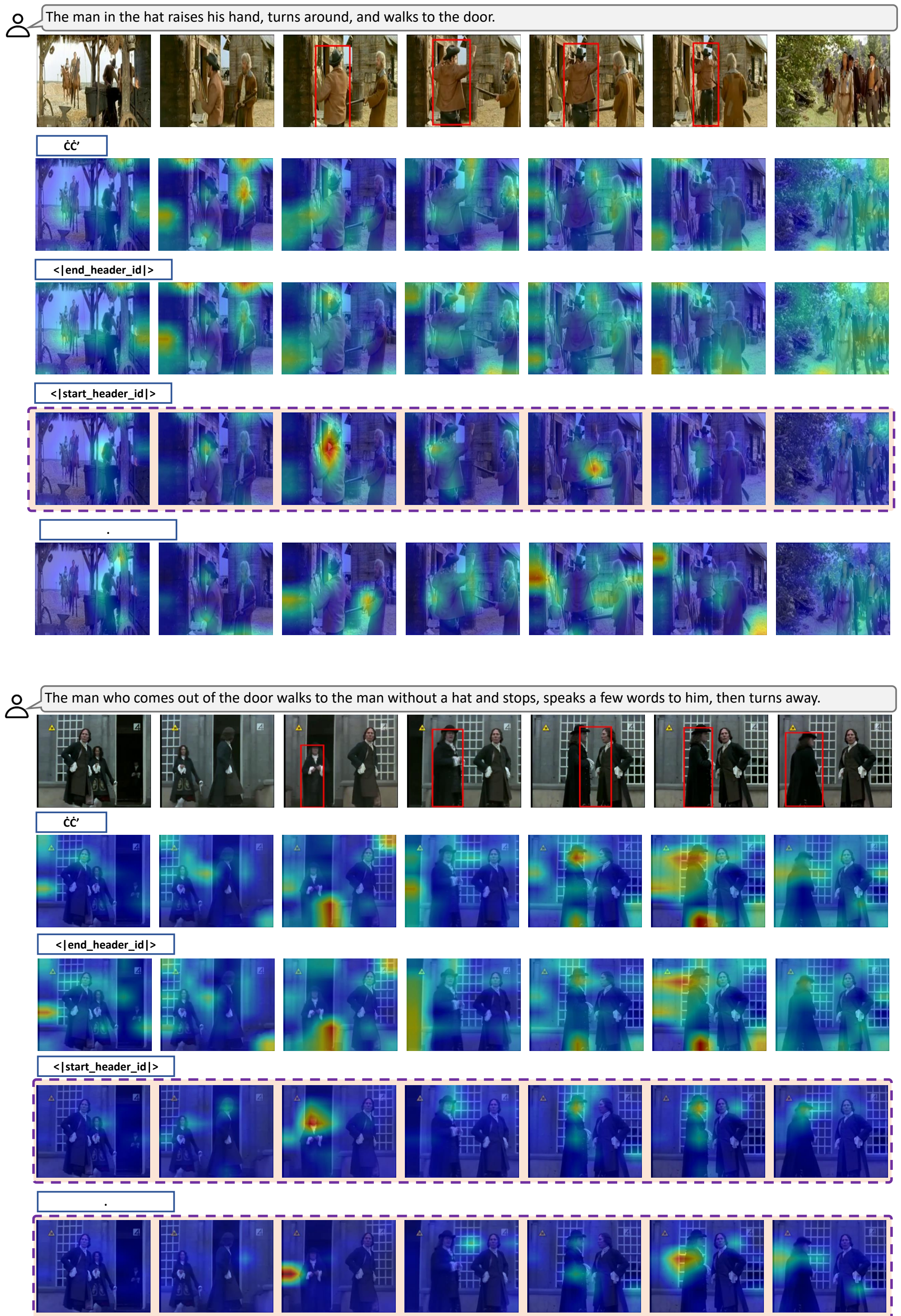


Figure 12: Visualization of tokens' attention maps on video-text pairs.

7 More Implementation Details

7.1 Datasets

We evaluate on three video benchmark datasets: HCSTVG-v1, HCSTVG-v2 [51], and VidSTG [73]. HCSTVG-v1 has 4500 training and 1160 testing videos with sentence descriptions of human attributes/actions. HCSTVG-v2 extends v1 to 16,544 videos, including 10,131 training, 2,000 validation, and 4,413 testing videos. Since the test set is unavailable, we evaluate on the validation set following prior works [59; 25]. VidSTG includes 99,943 video-sentence pairs (44,808 declarative, 55,135 interrogative), covering 10,303 videos and 80 object categories. Training, validation, and test sets have 80,684, 8,956, and 10,303 pairs, respectively.

7.2 Method Implementation

In this work, we adopt G-DINO [38] with 0.4 for both phrase and box thresholds to detect object proposals, and then utilize SAM2 [45] as tracker for tubelet proposals generation. Considering the information redundancy, we run the detector every 10 frames for efficiency. We utilize GPT-4o to decompose the original query sentence into spatial and temporal sub-queries. To demonstrate the efficiency of our method, we consider four LLaVA-like video MLLMs: `LlaVA-Next-Video-7B` [27], `Qwen2-VL-7B` [53], `ShareGPT4Video-8B` [7], `LlaVa-OneVision-7B` [27]. In practice, with the limited tokens context length and computing efficiency, we sample 20 frames by default as the visual input of MLLMs for each video. When adopting the test-time tuning strategy DSTH, we set the learning rate lr and iteration times N_{ef} as 8.0 and 1 according to the ablation analysis in Tab. 7. We conduct all experiments on an A100 GPU with 80G VRAM based on Pytorch framework. To better capture text-to-visual attention, we introduce *'Describe this video in details.'* as general query prompt Q_{cap} and implement the relative attention strategy [69] to reduce the effect of visual registers [9; 55]. In Algo. 1, we outline the implementation of the proposed decomposed spatio-temporal highlighting strategy during test-time tuning.

7.3 Generation of Target-related Cues

In this work, we extract the attributes and actions descriptions from the original query Q as textual cues to enhance the spatial and temporal comprehension, respectively. To obtain multiple target-related cues, we leverage the strong in-context capability of the Large Language Model (LLM). In particular, we construct the in-context instruction to prompt llm for completion. The whole prompt used in this work includes:general instruction (in brown), output constraints (in blue), in-context task examples (in green) and input question (in yellow), shown in Fig. 13.

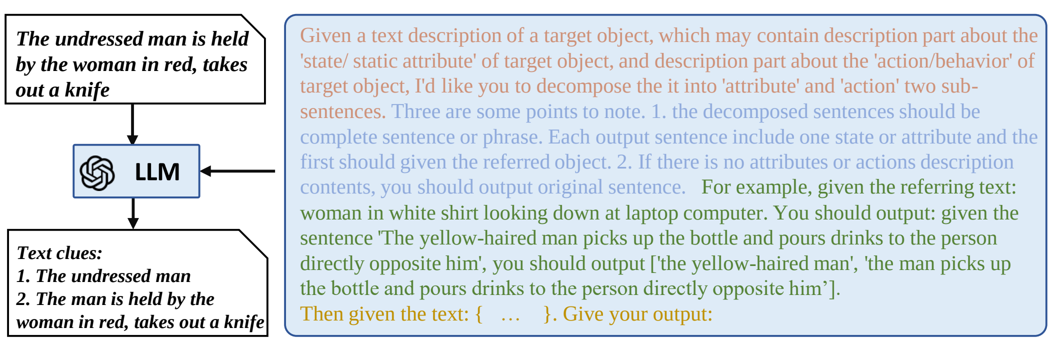


Figure 13: Flow of LLM-based generation of target-related cues.

In Fig. 14, we present some examples of LLM generation. After decomposing the original text description into attributes and actions-related descriptions, we will further transform the descriptions into interrogative queries by a fixed template to inquire about the existence of the target. In our work, we adopt the template *'Is there a __ in this video?'*.

- Original Text: The man in purple clothes takes a step forward and sits on the bench.
- Attributes Text : the man in purple clothes.
- Actions Text : The man takes a step forward and sits on the bench.

10_phVLLTMzmKk.mp4

- Original Text: The yellow-haired man takes the opposite woman's hand and kisses it, then puts it down.
- Attributes Text : the yellow-haired man.
- Actions Text : The man takes the opposite woman's hand and kisses it, then puts it down.

159_vsMgg4snZzM.mkv

- Original Text: The white-bearded man points at the man on his right and then points at himself again.
- Attributes Text : the white-bearded man.
- Actions Text : The man points at the man on his right and then points at himself again.

111_pSdPmmJ3-ng.mp4

Figure 14: Examples of LLM-based cues generation.

Algorithm 1: Decomposed Spatio-Temporal Highlighting

Input: MLLM π_θ, video query input Q, video X, track proposals $O_{\text{pro}} = \{\mathrm{o}_1, \ldots, \mathrm{o}_P\}$, test-time tuning epoches N_{ep}

Output: Generated tube $O'_{\text{pred}} = \{\mathrm{b}_t\}_{t=t_{s'}}^{t_{e'}}$ based on query input Q and video X

Decompose the Q into attribute/action sub-queries Q_{s} and Q_{t} for spatial/temporal inquiry.

Initialize spatial prompt V_s and temporal prompt V_t.

Initialize $i_{ep} = 0$.

Embed the video X into visual tokens T_v. Embed the Q_{s} and Q_{t} into text tokens T_q^s and T_q^t.

while *training* **do**

  while $i_{ep} < \mathrm{N}_{ep}$ **do**

    `// spatial prompt Learning`

    Compute positive and negative logit prediction for attribute subquery Q_{s}:

    $\mathrm{p}^{yes} = \mathrm{logit}_{\pi_\theta}(y_i^{yes} | (\mathrm{T}_v + \mathrm{V}_s, \mathrm{T}_q^s, y_{<i}))$, $\mathrm{p}^{no} = \mathrm{logit}_{\pi_\theta}(y_i^{no} | (\mathrm{T}_v + \mathrm{V}_s, \mathrm{T}_q^s, y_{<i}))$

    Update V_s by minimizing $\mathcal{L}_{\text{s}} = -\exp(\mathrm{p}^{yes} - \mathrm{p}^{no})$.

    `// temporal prompt Learning`

    Compute positive and negative logit prediction for action query Q_{t}:

    $\mathrm{p}^{yes} = \mathrm{logit}_{\pi_\theta}(y_i^{yes} | (\mathrm{T}_v + \mathrm{V}_t, \mathrm{T}_q^t, y_{<i}))$, $\mathrm{p}^{no} = \mathrm{logit}_{\pi_\theta}(y_i^{no} | (\mathrm{T}_v + \mathrm{V}_t, \mathrm{T}_q^t, y_{<i}))$

    Update V_t by minimizing $\mathcal{L}_{\text{t}} = -\exp(\mathrm{p}^{yes} - \mathrm{p}^{no})$.

    $i_{ep} = i_{ep} + 1$.

  end

end

prepare general query prompt Q_{cap}.

while *inference* **do**

  cache the visual attention map A_{cap} by query Q_{cap} as input.

  `// inference with` $\mathrm{T}_v + \mathrm{V}_s$.

  cache the visual attention map A_{g}^S by query Q with $\mathrm{T}_v + \mathrm{V}_s$ as visual tokens.

  compute spatial-related visual attention map $A_{\text{g}}^S = A_{\text{g}}^S / A_{\text{cap}}$,

  obtain the object track prediction.

  `// inference with` $\mathrm{T}_v + \mathrm{V}_t$.

  cache the visual attention map A_{g}^T by query Q with $\mathrm{T}_v + \mathrm{V}_t$ as visual tokens.

  compute temporal-related visual attention map $A_{\text{g}}^T = A_{\text{g}}^T / A_{\text{cap}}$,

  obtain the target frames prediction.

  Combine the object track prediction and target frames prediction.

end

8 More Results and Ablation Analysis

8.1 Comparison on VidSTG (Interrogative) and RefCOCO Benchmarks

In Tab. 5, we compare our approach with other SOTA methods on the VidSTG (Interrogative) benchmarks. Even given the interrogative sentence as query, our method still show superior performance with a 4.0% improvement on the vIoU@0.5 metric when integrated with `LLaVA-Next-Video-7B` model, which show the strong generalization capability of our framework. Undoubtedly, our method can also achieve superior localization performance in the image grounding tasks. In Tab. 6, we compare our framework with other SOTA visual grounding methods on RefCOCO series benchmarks. Here we just introduce the selection of superior token for grounding and has achieved comparable and even better performance than current zero-shot SOTA methods. The The 'Oracle' denotes the result of choosing the most exact one of candidate boxes (extracted by G-DINO) with knowledge of the ground truth. We also show the proportion of our method's performance relative to oracle's, which is indicatd by the number in lower right corner. Note that our insight about grounding tokens is orthogonal to other works [69; 30; 24]. We believe that it can be integrated with other works to achieve better results.

Table 5: Comparison on VidSTG (Interrogative) benchmark.

Sup	Method	VidSTG (Interrogative) m_vIoU	vIoU@0.3	vIoU@0.5
Full	TubeDETR [59] [CVPR2022]	25.7	35.7	23.2
	CSDVL [20] [CVPR2023]	28.5	39.9	26.2
	CG-STVG [15] [CVPR2024]	29.0	40.5	27.5
Weak	WINNER [29] [CVPR2023]	10.2	12.0	5.5
	VEM [21] [ECCV2024]	13.3	16.7	7.7
	CoSPaL [25] [ICLR2025]	13.5	16.4	10.2
ZS	ReCLIP [50] [ACL2022]	8.4	8.0	2.3
	E3M [3] [ECCV2024]	10.6	12.2	5.4
	Ours LLaVA-Next-Video-7B	14.6	23.1	9.4
	Ours Qwen2-VL-7B	13.0	19.5	7.8
	Ours ShareGPT4Video-8B	13.1	21.0	9.1
	Ours LLaVA-OneVision-7B	13.5	21.4	8.2

Table 6: Comparison of different methods on RefCOCO series datasets.

Sup	Method	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
Full	MDETR [22] [CVPR2021]	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9
	SeqTR [77] [ECCV2022]	87.0	90.2	83.6	78.7	84.5	71.9	82.7	83.4
	UNINEXT [32] [ACMMM2023]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
	Shikra-7B [6] [arXiv]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	Ferret-7B [65] [ICLR2024]	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8
Zero	ReCLIP [50] [ACL2022]	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
	ZS_REC [16] [CVPR2024]	49.4	47.8	51.7	48.9	50.0	46.9	61.0	60.0
	GroundVLP [46] [AAAI2024]	65.0	73.5	55.0	68.8	78.1	57.3	74.7	75.0
	Ours LLaVA-OneVision-7B	68.7$_{75.4\%}$	77.2$_{80.8\%}$	62.0$_{72.9\%}$	67.1$_{74.4\%}$	76.7$_{80.2\%}$	57.5$_{67.7\%}$	72.3$_{81.3\%}$	73.7$_{83.2\%}$
	Oracle	91.1	95.6	85.0	91.2	95.6	84.9	88.9	88.6

8.2 Temporal Consistency Analysis

In our work, we obtain the attributes and actions related sub-queries. Specially, the decomposed attribute sub-query, which provides static state description, should be temporally consistent for spatial grounding. Interestingly, there exists evident inconsistency when introducing temporal augmentation in current MLLMs. We develop a temporal inconsistency metric by introducing reversing the order of input frames. In Fig. 15, we show the relation between spatial grounding accuracy and temporal

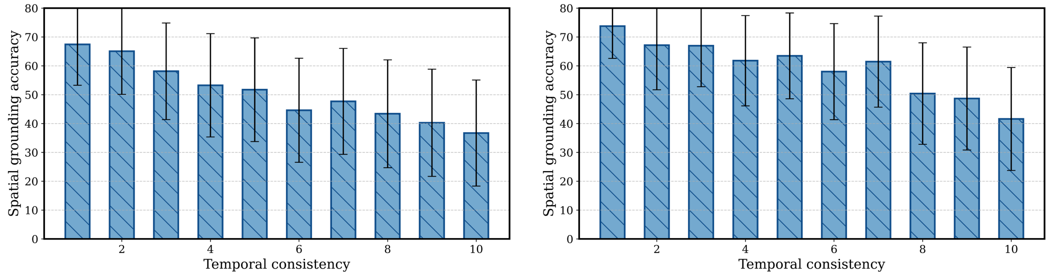


Figure 15: Spatial grounding accuracy of different groups of samples on the HC-STVGv1 dataset. These groups are ranked by descending temporal consistency.

Table 7: Iteration times N_{ep} and learning rate lr ablation.

N_{ep}	lr	m_vIoU	vIoU@0.3	vIoU@0.5
0	0	15.2	25.1	8.5
1	4	19.9	32.6	11.6
1	**8**	**20.4**	**33.6**	**12.4**
1	16	20.2	32.3	12.0
2	4	20.5	34.0	12.1
2	8	20.3	32.6	12.1
2	16	20.3	33.2	12.2

consistency based on `LLaVA-Next-Video-7B` model (left of Fig. 15) and `LLaVA-OneVision-7B` model (right of Fig. 15). We can find that although the `LLaVA-OneVision-7B` model achieves better localization performance, there is still the pronounced temporal inconsistency caused by temporal augmentation. Our findings provide guidance and insight for subsequent research in video MLLMs.

8.3 Effect of Optimization Times N_{ep} and Learning Rate lr

The test-time tuning strategy DSTH is iterated N_{ep} times for optimization with learning rate lr. Here we analyse effect of the hyper-parameters. As shown in Tab. 7, better results can be achieved as the optimization progresses and the our optimization strategy is relatively robust to these hyper-parameters. In particular, when setting $\mathrm{N}_{ep} = 1$ and $\mathrm{lr} = 8.0$, optimal performance is achieved in general.

8.4 Inference Efficiency

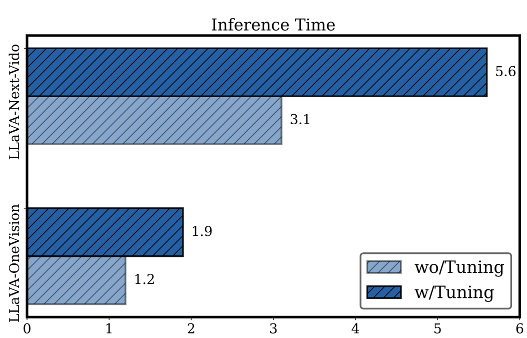


Figure 16: Comparison of inference time before/after introducing test-time tuning.

In Fig. 16, we show the inference efficiency of our proposed zero-shot framework. We test the inference speed on all video samples from HC-STVGv1 dataset on a single A100 GPU and then compute the average value. Specifically, before introducing the test-time tuning strategy DSTH for inference, our framework costs about 3.1 seconds for each video based on `LLaVA-Next-Video` model. After adopting the test-time tuning strategy, the cost is still acceptable though with some additional resource consumption. We believe that it would be more efficient to apply the resource-friendly inference schemes in the future.

8.5 Qualitative Spatio-Temporal Video Grounding Visualization

In Fig. 17, we present some video grounding examples for qualitative analysis. Here, we compare the results before introducing the test-time tuning (denoted with yellow boxes) with the results (denoted with green boxes) after introducing the optimization. The ground truth boxes are denoted with red boxes. By highlighting the attribute/action cues of the target by test-time tuning, our method can direct the MLLMs toward reliable visual context and improve spatial/temporal localization. We also give some failure cases (*e.g.*, the case (d)). Despite optimization during testing, the current model still pinpoints the wrong object instance. We attribute the less efficient comprehension to the poor visual conditions and suboptimal spatial inference in current MLLMs.

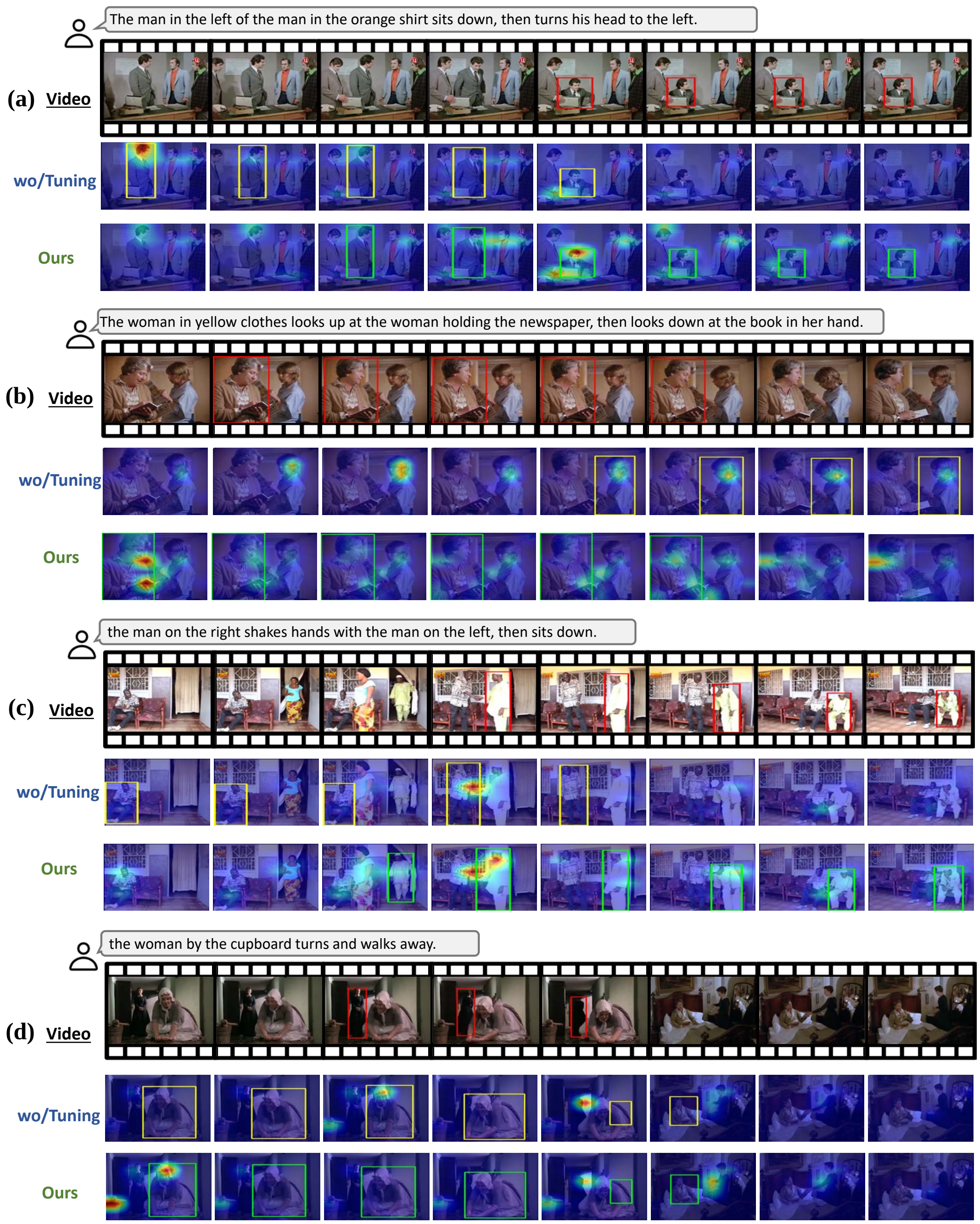


Figure 17: Comparison of attention maps on the HC-STVG dataset.

9 Limitation and Future work

Our work is simple yet effective, and also provides insights for related fields (*e.g.*, hallucination detection and attention-guided MLLMs pruning). Our work does have limitations. For example, based on MLLMs our framework can only receive a limited number of video frames as visual input due to the limit of computing resource. Besides, our framework needs to obtain attention for spatial and temporal inference, which is not compatible with the flash-attention mechanism adopted in current MLLMs. In the future, we will consider introducing a more efficient solution for the comprehension of long videos by incorporating token pruning and key frame selection techniques.

10 Broader Impacts

While we do not foresee our method causing any direct negative societal impact, it may potentially be leveraged by malicious parties to create applications that could misuse the grounding capabilities for unethical or illegal purposes. We urge the readers to limit the usage of this work to legal use cases.