

# 1 Appendices of Unleashing the Potential of Multimodal 2 LLMs for Zero-Shot Spatio-Temporal Video 3 Grounding

4		
5	<b>A Grounding Tokens</b>	<b>3</b>
6	A.1 Illustration of Input Tokens Components . . . . .	3
7	A.2 Quantitative Analysis of Special Tokens . . . . .	3
8	A.3 Qualitative Analysis of Special Tokens . . . . .	5
9	<b>B More Implementation Details</b>	<b>8</b>
10	B.1 Datasets . . . . .	8
11	B.2 Method Implementation . . . . .	8
12	B.3 Generation of Target-related Cues . . . . .	8
13	<b>C More Results and Ablation Analysis</b>	<b>10</b>
14	C.1 Comparison on VidSTG (Interrogative) and RefCOCO Benchmarks . . . . .	10
15	C.2 Temporal Consistency Analysis . . . . .	10
16	C.3 Effect of Optimization Times $N_{ep}$ and Learning Rate $lr$ . . . . .	11
17	C.4 Inference Efficiency . . . . .	11
18	C.5 Qualitative Spatio-Temporal Video Grounding Visualization . . . . .	12
19	<b>D Limitation and Future work</b>	<b>14</b>
20	<b>E Broader Impacts</b>	<b>14</b>

## 21 A Grounding Tokens

### 22 A.1 Illustration of Input Tokens Components

23 In Fig. 1, we give a clear illustration of input token components in MLLMs. In practice, besides  
 24 the visual and text prompt tokens (② and ③), there are also system tokens and some special tokens  
 25 (① and ④) equipped for language generation. In particular, the grounding ability of special tokens  
 26 subsequent to the text instruction prompt is significantly undervalued in previous works. Here we  
 27 focus on exploring the grounding ability of these special tokens. For different multi-modal large  
 28 language models, there often are different special tokens due to different instruction-tuning process.  
 29 For example, in LLaVA-1.5, the introduced special tokens include: {':', 'ANT', 'IST', 'SS',  
 30 '\_A'}. While for the Qwen2-VL, these special tokens include: {Ĉ, 'assistant', '<|im\_start|>',  
 31 '<|im\_end|>'}. Note that we also consider the '.' as one of the special tokens considering that it  
 32 is located in the end of the instruction prompt and is characteristic of sink tokens [11, 25].

33 In fact, special tokens are predefined symbols within a language model’s vocabulary. They focus on  
 34 guiding the model’s processing instead of representing real words and guide the model to generate  
 35 coherent and context-aware responses. In the scenario explored in our work (*i.e.*, dialogue systems),  
 36 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.

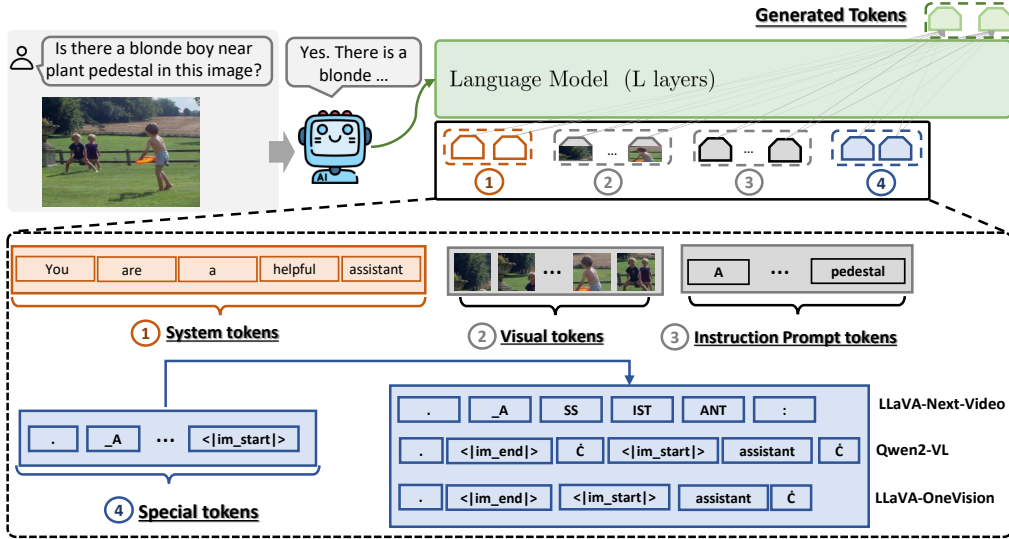


Figure 1: Illustration of the tokens input in MLLMs-based dialog system. Here, we take the image question answering as an example.

### 38 A.2 Quantitative Analysis of Special Tokens

39 In our work, we randomly selected a subset of 1000 image-text pairs from RefCOCOg [7] valida-  
 40 tion set for image MLLMs analysis, and a subset of 1000 video-text pairs from HC-STVG [22]  
 41 dataset for video MLLMs analysis. Particularly, we choose six typical MLLMs (*i.e.*, LLaVA-1.5,  
 42 Qwen-VL, Deepseek-VL, LLaVA-Next-Video, Qwen2-VL, LLaVA-OneVision) for pilot studies. In  
 43 the following, we will introduce our findings about special tokens in MLLMs.

44 **Some special tokens show outstanding grounding ability for text prompt input.** We compare  
 45 the grounding accuracy of G-DNIO [18] and the special tokens’ ones by probing the attention map.  
 46 Here we adopt the G-DNIO with Swin-T backbone and it is not finetuned on the refcoco series  
 47 datasets. Following previous visual grounding works [21, 6], we adopt the IoU@0.5 as the metric  
 48 of grounding accuracy. The results evaluated on refcoco series benchmarks are shown in the Fig. 2.  
 49 The yellow horizontal line denotes the result by G-DINO model. Though the G-DINO is trained on  
 50 image-text pairs in fully-supervised manner, some special tokens still achieve comparable and even

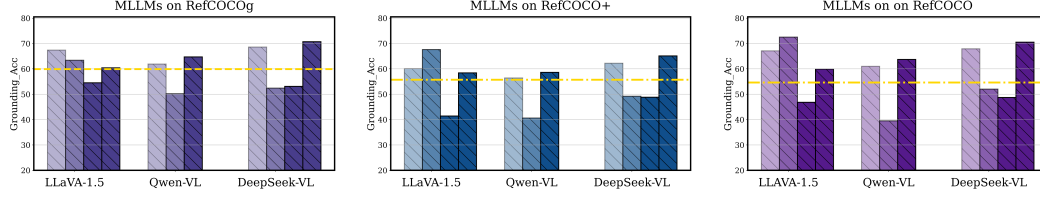


Figure 2: Comparison of grounding accuracy between G-DINO and special tokens on RefCOCO series datasets.

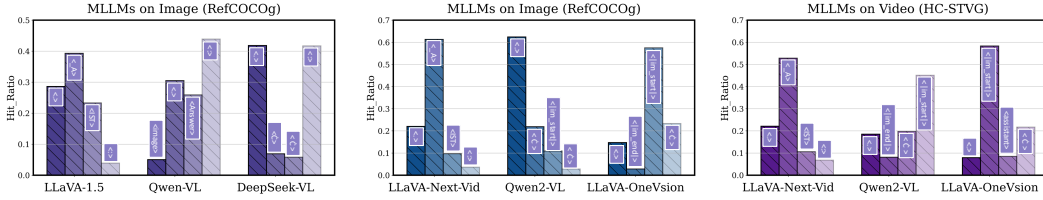


Figure 3: The hit ratio of different special tokens in image and video MLLMs.

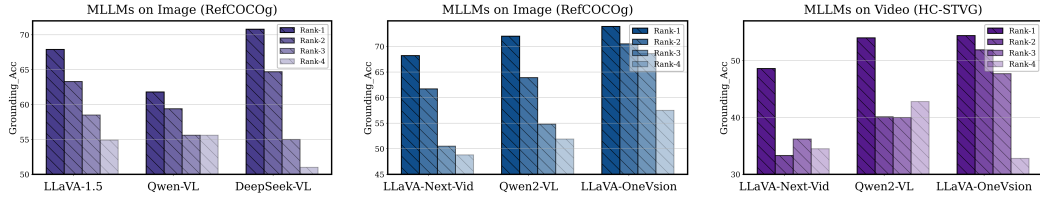


Figure 4: Grounding accuracy of special tokens ranked by the visual activation degree.

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

54 **MLLMs dynamically assign the special tokens to attend to the text-related regions.** For further  
 55 exploration on the grounding ability of special tokens, we define the *attention ratio* of each special  
 56 token as the ratio of maximum attention within the ground-truth bounding box  $b_{gt}$  to that outside it.  
 57 Here, a higher ratio indicates better target grounding ability. Given a test sample, we can identify the  
 58 token that yields the highest attention ratio as the superior token for grounding. Then a token’s *hit*  
 59 *ratio* is defined as the frequency of being the superior token for grounding across all test samples.  
 60 The Fig. 3(a) shows the hit ratio of special tokens in each MLLM. We choose four special tokens  
 61 for visualization. In addition to the findings that MLLMs dynamically assign the special tokens to  
 62 attend to the text-related regions, we also observe that the superior token for grounding in video  
 63 MLLMs shows a more concentrated trend than image MLLMs. For example, the highest hit ratio by  
 64 Qwen2-VL is more than 60%. Besides, for the same special token in the MLLM, the hit ratio in the  
 65 case of different datasets may be significantly different. For example, the token ‘.’ shows a high hit  
 66 ratio in RefCOCOg dataset, but its hit ratio is quite low in the HC-STVG dataset.

67 **The special token with higher visual activation tends to show superior grounding performance.**  
 68 In our work, with further analysis, we reveal that the superior token for grounding tends to show  
 69 higher visual activation. For each sample, we rank the special tokens according to the maximum  
 70 value of visual attention and evaluate their grounding accuracy by selecting the proposal with the  
 71 highest attention value as the prediction. In practice, we extract box proposals by the detector (*e.g.*,  
 72 G-DINO) and evaluate the grounding accuracy by the Acc@0.5 metric. Fig. 4 shows the results. We  
 73 can see that the grounding accuracy decreases as the rank of visual activation reduces (from left  
 74 to right). We also observe that models with better comprehension overall possess better grounding  
 75 ability.

### 76 A.3 Qualitative Analysis of Special Tokens

77 In Fig. 5, we visualize token-to-visual attention maps using some image-text pairs by LLaVA-1.5-7B  
78 model. In particular, for each image-text pair, we visualize the visual attendance of the semantic  
79 tokens from the text prompt and special tokens. The green bounding box denotes the ground truth of  
80 the target object. Notably, we notice that the semantic tokens in text prompts can provide tangible  
81 attention to related visual entities by the limited cues, while the special tokens are often assigned to  
82 integrate global instruction cues and attend to the exact target region. For example, given the text  
83 prompt '*a man getting ready to cut a cake*', the token 'man' and 'cake' both show reasonable visual  
84 response. The special token '\_A' can accurately attend to the target person. In addition, we see that  
85 the special token with high visual activation often shows better grounding performance. However,  
86 we also see that the special tokens may attend to the wrong instance even though the semantic tokens  
87 can properly capture their region of interest. It suggests that there is still room for improvement in  
88 inference in current multimodal language models.

89 In Fig. 6, we visualize attention maps of special tokens using some video-text pairs by  
90 ShareGPT4Video-8B model. The ground-truth spatio-temporal tube is denoted by red bound-  
91 ing boxes. Each row shows the attention maps of a specific token. We find that only a few special  
92 tokens can well capture the corresponding target regions, which necessitates the efficient selection of  
93 grounding tokens. Also, in some cases (*e.g.*, the below sample), there may be more than one special  
94 token that will pay attention to the target area. This shows that achieving localization by considering  
95 the integration of multiple effective special tokens can be a direction for improvement. We leave it to  
96 future research.



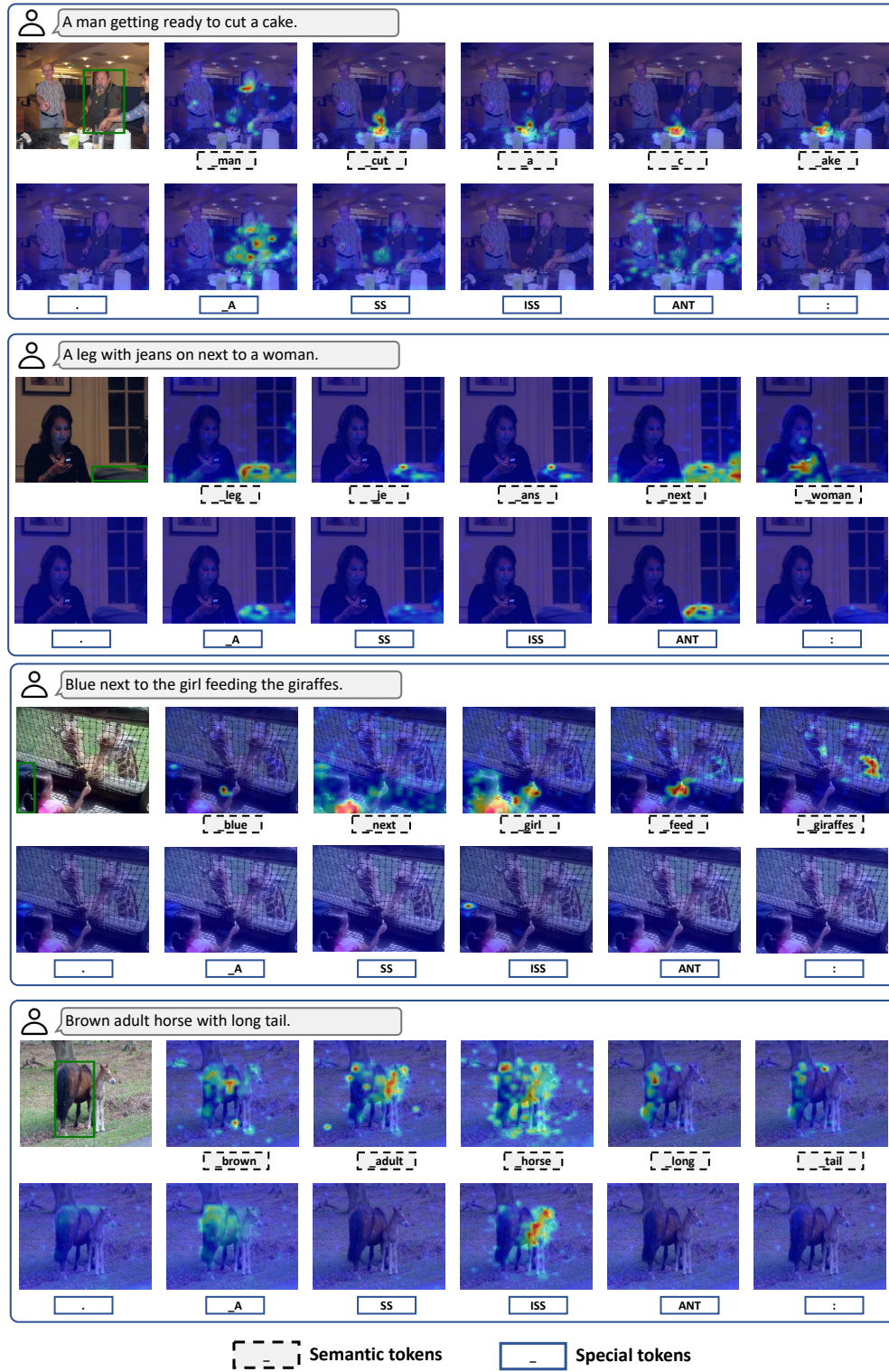


Figure 5: Visualization of tokens' attention maps on image-text pairs. The tokens with dotted box denote the semantic tokens in text prompts while the tokens with solid box denote the special tokens following the text prompts.

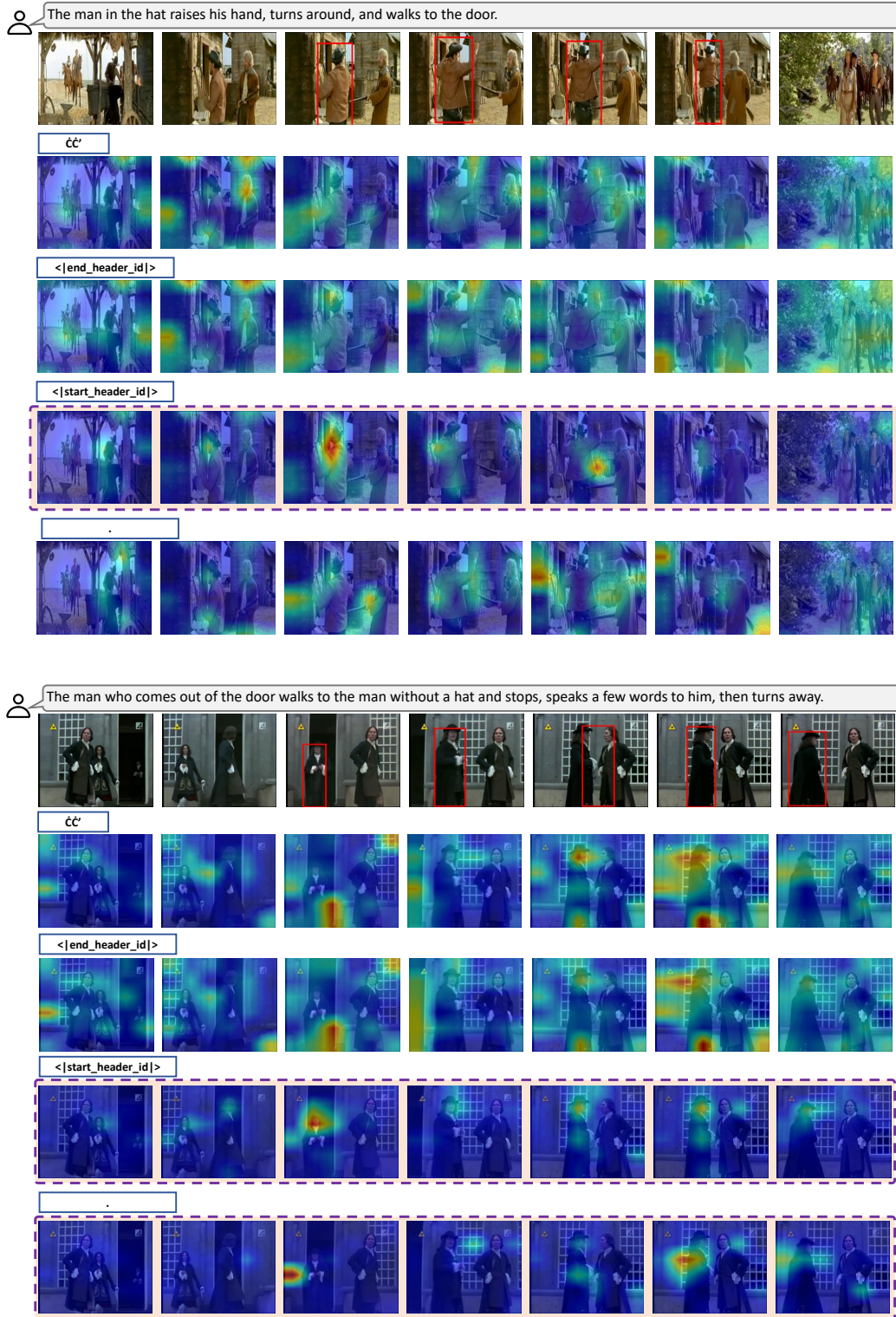


Figure 6: Visualization of tokens' attention maps on video-text pairs.

## 97 B More Implementation Details

### 98 B.1 Datasets

99 We evaluate on three video benchmark datasets: HCSTVG-v1, HCSTVG-v2 [22], and VidSTG [29].  
 100 HCSTVG-v1 has 4500 training and 1160 testing videos with sentence descriptions of human at-  
 101 tributes/actions. HCSTVG-v2 extends v1 to 16,544 videos, including 10,131 training, 2,000 vali-  
 102 dation, and 4,413 testing videos. Since the test set is unavailable, we evaluate on the validation set  
 103 following prior works [26, 13]. VidSTG includes 99,943 video-sentence pairs (44,808 declarative,  
 104 55,135 interrogative), covering 10,303 videos and 80 object categories. Training, validation, and test  
 105 sets have 80,684, 8,956, and 10,303 pairs, respectively.

### 106 B.2 Method Implementation

107 In this work, we adopt G-DINO [18] with 0.4 for both phrase and box thresholds to detect object  
 108 proposals, and then utilize SAM2 [19] as tracker for tubelet proposals generation. Considering the  
 109 information redundancy, we run the detector every 10 frames for efficiency. We utilize GPT-4o to  
 110 decompose the original query sentence into spatial and temporal sub-queries. To demonstrate the  
 111 efficiency of our method, we consider four LLaVA-like video MLLMs: LLaVA-Next-Video-7B [14],  
 112 Qwen2-VL-7B [23], ShareGPT4Video-8B [3], LLaVA-OneVision-7B [14]. In practice, with the  
 113 limited tokens context length and computing efficiency, we sample 20 frames by default as the visual  
 114 input of MLLMs for each video. When adopting the test-time tuning strategy DSTH, we set the  
 115 learning rate  $lr$  and iteration times  $N_{ef}$  as 8.0 and 1 according to the ablation analysis in Tab. 3. We  
 116 conduct all experiments on an A100 GPU with 80G VRAM based on Pytorch framework. To better  
 117 capture text-to-visual attention, we introduce ‘Describe this video in details.’ as general query prompt  
 118  $Q_{cap}$  and implement the relative attention strategy [28] to reduce the effect of visual registers [4, 24].  
 119 In Algo. 1, we outline the implementation of the proposed decomposed spatio-temporal highlighting  
 120 strategy during test-time tuning.

### 121 B.3 Generation of Target-related Cues

122 In this work, we extract the attributes and actions descriptions from the original query  $Q$  as textual  
 123 cues to enhance the spatial and temporal comprehension, respectively. To obtain multiple target-  
 124 related cues, we leverage the strong in-context capability of the Large Language Model (LLM). In  
 125 particular, we construct the in-context instruction to prompt llm for completion. The whole prompt  
 126 used in this work includes: general instruction (in brown), output constraints (in blue), in-context task  
 127 examples (in green) and input question (in yellow), shown in Fig. 7.

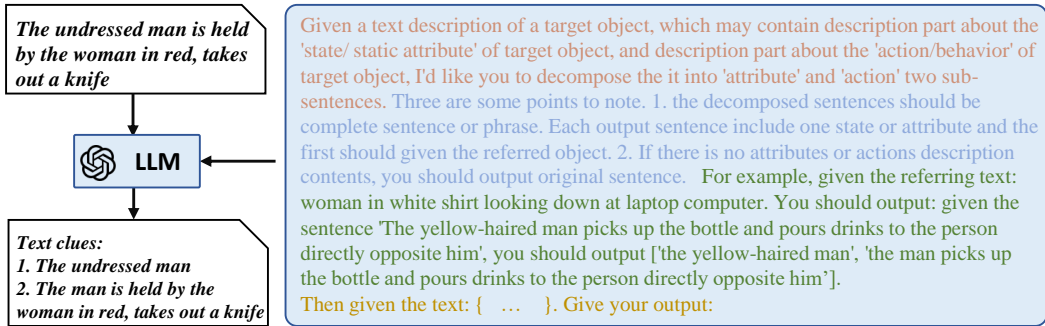


Figure 7: Flow of LLM-based generation of target-related cues.

128 In Fig. 8, we present some examples of LLM generation. After decomposing the original text  
 129 description into attributes and actions-related descriptions, we will further transform the descriptions  
 130 into interrogative queries by a fixed template to inquire about the existence of the target. In our work,  
 131 we adopt the template ‘Is there a \_\_ in this video?’.



<ul style="list-style-type: none"> <li>• <u>Original Text</u>: The man in purple clothes takes a step forward and sits on the bench.</li> <li>• <u>Attributes Text</u> : the man in purple clothes.</li> <li>• <u>Actions Text</u> : The man takes a step forward and sits on the bench.</li> </ul>	10_phVLLTMzmKk.mp4
<ul style="list-style-type: none"> <li>• <u>Original Text</u>: The yellow-haired man takes the opposite woman's hand and kisses it, then puts it down.</li> <li>• <u>Attributes Text</u> : the yellow-haired man.</li> <li>• <u>Actions Text</u> : The man takes the opposite woman's hand and kisses it, then puts it down.</li> </ul>	159_vsMgg4snZzM.mkv
<ul style="list-style-type: none"> <li>• <u>Original Text</u>: The white-bearded man points at the man on his right and then points at himself again.</li> <li>• <u>Attributes Text</u> : the white-bearded man.</li> <li>• <u>Actions Text</u> : The man points at the man on his right and then points at himself again.</li> </ul>	111_pSdPmmJ3-ng.mp4

Figure 8: Examples of LLM-based cues generation.

---

**Algorithm 1:** Decomposed Spatio-Temporal Highlighting

---

**Input:** MLLM  $\pi_\theta$ , video query input  $Q$ , video  $X$ , track proposals  $O_{\text{pro}} = \{o_1, \dots, o_P\}$ ,  
test-time tuning epoches  $N_{ep}$

**Output:** Generated tube  $O'_{\text{pred}} = \{b_t\}_{t=t_{s'}}^{t_{e'}}$ , 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**  $i_{ep} < N_{ep}$  **do**

// spatial prompt Learning

Compute positive and negative logit prediction for attribute subquery  $Q_s$ :

$p^{yes} = \text{logit}_{\pi_\theta}(y_i^{yes} | (T_v + V_s, T_q^s, y_{<i}))$ ,  $p^{no} = \text{logit}_{\pi_\theta}(y_i^{no} | (T_v + V_s, T_q^s, y_{<i}))$

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

// temporal prompt Learning

Compute positive and negative logit prediction for action query  $Q_t$ :

$p^{yes} = \text{logit}_{\pi_\theta}(y_i^{yes} | (T_v + V_t, T_q^t, y_{<i}))$ ,  $p^{no} = \text{logit}_{\pi_\theta}(y_i^{no} | (T_v + V_t, T_q^t, y_{<i}))$

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

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

**end**

**end**

prepare general query prompt  $Q_{\text{cap}}$ .

**while inference do**

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

// inference with  $T_v + V_s$ .

cache the visual attention map  $A_g^S$  by query  $Q$  with  $T_v + V_s$  as visual tokens.

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

obtain the object track prediction.

// inference with  $T_v + V_t$ .

cache the visual attention map  $A_g^T$  by query  $Q$  with  $T_v + V_t$  as visual tokens.

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

obtain the target frames prediction.

Combine the object track prediction and target frames prediction.

**end**

---

## C More Results and Ablation Analysis

### C.1 Comparison on VidSTG (Interrogative) and RefCOCO Benchmarks

In Tab. 1, 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. 2, 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 indicated by the number in lower right corner. Note that our insight about grounding tokens is orthogonal to other works [28, 16, 12]. We believe that it can be integrated with other works to achieve better results.

Table 1: Comparison on VidSTG (Interrogative) benchmark.

Sup	Method	VidSTG (Interrogative)		
		m_vIoU	vIoU@0.3	vIoU@0.5
<b>Full</b>	TubeDETR [26] [CVPR2022]	25.7	35.7	23.2
	CSDVL [8] [CVPR2023]	28.5	39.9	26.2
	CG-STVG [5] [CVPR2024]	29.0	40.5	27.5
<b>Weak</b>	WINNER [15] [CVPR2023]	10.2	12.0	5.5
	VEM [9] [ECCV2024]	13.3	16.7	7.7
	CoSPaL [13] [ICLR2025]	13.5	16.4	10.2
<b>ZS</b>	ReCLIP [21] [ACL2022]	8.4	8.0	2.3
	E3M [1] [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 2: Comparison of different methods on RefCOCO series datasets.

Sup	Method	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
<b>Full</b>	MDETR [10] [CVPR2021]	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9
	SeqTR [30] [ECCV2022]	87.0	90.2	83.6	78.7	84.5	71.9	82.7	83.4
	UNINEXT [17] [ACMMM2023]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
	Shikra-7B [2] [arXiv]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	Ferret-7B [27] [ICLR2024]	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8
<b>Zero</b>	ReCLIP [21] [ACL2022]	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
	ZS_REC [6] [CVPR2024]	49.4	47.8	51.7	48.9	50.0	46.9	61.0	60.0
	GroundVLP [20] [AAAI2024]	65.0	73.5	55.0	68.8	78.1	57.3	74.7	75.0
	Ours LLaVA-OneVision-7B	68.7 <sub>75.4%</sub>	77.2 <sub>80.8%</sub>	62.0 <sub>72.9%</sub>	67.1 <sub>74.4%</sub>	76.7 <sub>80.2%</sub>	57.5 <sub>67.7%</sub>	72.3 <sub>81.3%</sub>	73.7 <sub>83.2%</sub>
	Oracle	91.1	95.6	85.0	91.2	95.6	84.9	88.9	88.6

### C.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. 9, we show the relation between spatial grounding accuracy and temporal

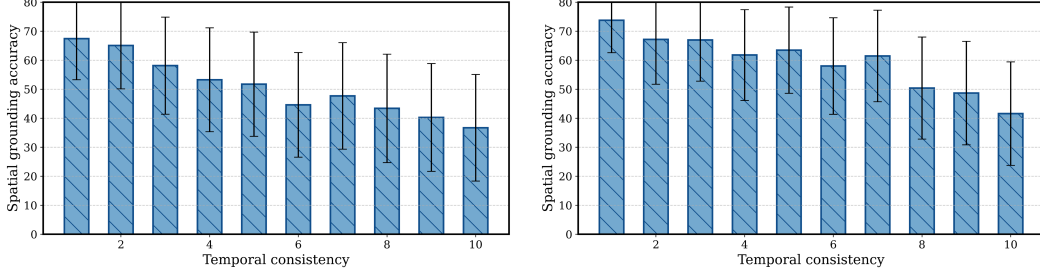


Figure 9: Spatial grounding accuracy of different groups of samples on the HC-STVGv1 dataset. These groups are ranked by descending temporal consistency.

Table 3: 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
<b>1</b>	<b>8</b>	<b>20.4</b>	<b>33.6</b>	<b>12.4</b>
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. 9) and LLaVA-OneVision-7B model (right of Fig. 9). 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.

### C.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. 3, 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  $N_{ep} = 1$  and  $lr = 8.0$ , optimal performance is achieved in general.

### C.4 Inference Efficiency

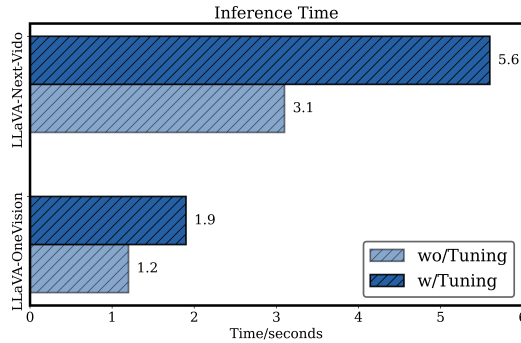


Figure 10: Comparison of inference time before/after introducing test-time tuning.

164 In Fig. 10, we show the inference efficiency of our proposed zero-shot framework. We test the  
165 inference speed on all video samples from HC-STVGv1 dataset on a single A100 GPU and then  
166 compute the average value. Specifically, before introducing the test-time tuning strategy DSTH  
167 for inference, our framework costs about 3.1 seconds for each video based on LLaVA-Next-Video  
168 model. After adopting the test-time tuning strategy, the cost is still acceptable though with some  
169 additional resource consumption. We believe that it would be more efficient to apply the resource-  
170 friendly inference schemes in the future.

## 171 C.5 Qualitative Spatio-Temporal Video Grounding Visualization

172 In Fig. 11, we present some video grounding examples for qualitative analysis. Here, we compare the  
173 results before introducing the test-time tuning (denoted with yellow boxes) with the results (denoted  
174 with green boxes) after introducing the optimization. The ground truth boxes are denoted with red  
175 boxes. By highlighting the attribute/action cues of the target by test-time tuning, our method can  
176 direct the MLLMs toward reliable visual context and improve spatial/temporal localization. We also  
177 give some failure cases (*e.g.*, the case (d)). Despite optimization during testing, the current model  
178 still pinpoints the wrong object instance. We attribute the less efficient comprehension to the poor  
179 visual conditions and suboptimal spatial inference in current MLLMs.

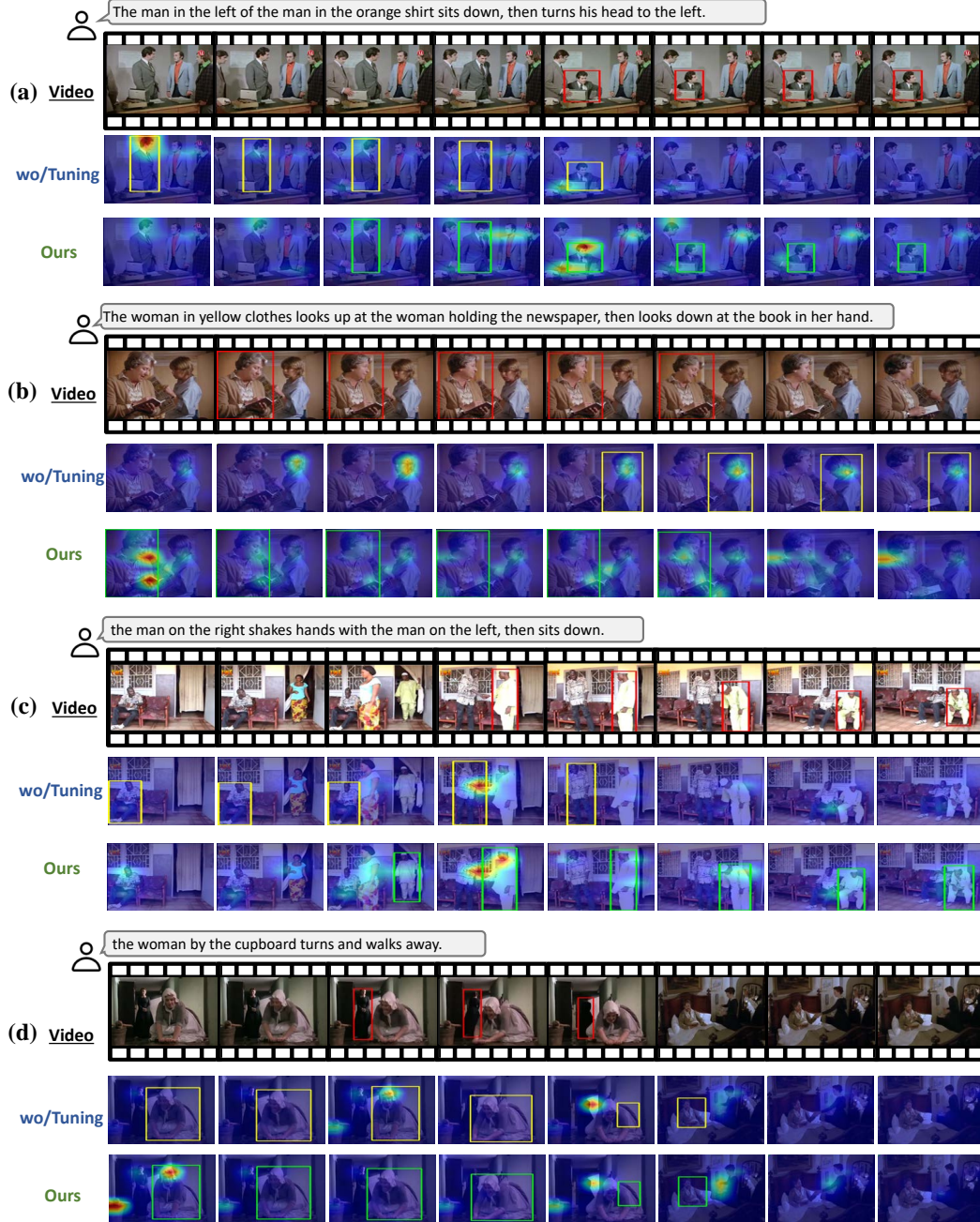


Figure 11: Comparison of attention maps on the HC-STVG dataset.



## 180 **D Limitation and Future work**

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

## 188 **E Broader Impacts**

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

## References

- [1] 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)
- [2] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
- [3] 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: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track
- [4] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
- [5] Gu, X., Fan, H., Huang, Y., Luo, T., Zhang, L.: Context-guided spatio-temporal video grounding. In: CVPR. pp. 18330–18339 (2024)
- [6] 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)
- [7] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 108–124. Springer (2016)
- [8] Jin, Y., Li, Y., Yuan, Z., Mu, Y.: Embracing consistency: a one-stage approach for spatio-temporal video grounding. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 29192–29204 (2022)
- [9] Jin, Y., Mu, Y.: Weakly-supervised spatio-temporal video grounding with variational cross-modal alignment. In: ECCV. pp. 412–429. Springer (2024)
- [10] 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)
- [11] Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. arXiv preprint arXiv:2503.03321 (2025)
- [12] Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. arXiv preprint arXiv:2503.06287 (2025)
- [13] Kumar, A., Kira, Z., Rawat, Y.S.: Contextual self-paced learning for weakly supervised spatio-temporal video grounding. arXiv preprint arXiv:2501.17053 (2025)
- [14] 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 preprint arXiv:2408.03326 (2024)
- [15] 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)
- [16] 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. *Advances in Neural Information Processing Systems* **37**, 120578–120601 (2024)
- [17] Lin, F., Yuan, J., Wu, S., Wang, F., Wang, Z.: Uninext: Exploring a unified architecture for vision recognition. In: ACM-MM. pp. 3200–3208 (2023)
- [18] 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)
- [19] 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 preprint arXiv:2408.00714 (2024)
- [20] 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: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4766–4775 (2024)

- 243 [21] Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A  
244 strong zero-shot baseline for referring expression comprehension. In: ACL. pp. 5198–5215  
245 (2022)
- 246 [22] Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-  
247 temporal video grounding with visual transformers. TCSVT **32**(12), 8238–8249 (2021)
- 248 [23] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.:  
249 Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv  
250 preprint arXiv:2409.12191 (2024)
- 251 [24] Woo, S., Kim, D., Jang, J., Choi, Y., Kim, C.: Don’t miss the forest for the trees: Attentional  
252 vision calibration for large vision language models. arXiv preprint arXiv:2405.17820 (2024)
- 253 [25] Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with  
254 attention sinks. In: ICLR
- 255 [26] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video  
256 grounding with transformers. In: CVPR. pp. 16442–16453 (2022)
- 257 [27] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.:  
258 Ferret: Refer and ground anything anywhere at any granularity. In: The Twelfth International  
259 Conference on Learning Representations
- 260 [28] Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free  
261 perception of small visual details with multimodal llms. arXiv preprint arXiv:2502.17422 (2025)
- 262 [29] Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal  
263 video grounding for multi-form sentences. In: CVPR. pp. 10668–10677 (2020)
- 264 [30] Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr:  
265 A simple yet universal network for visual grounding. In: ECCV. pp. 598–615. Springer (2022)