Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via Reinforcement Fine-Tuning

Xin Gu[1*] Haoji Zhang[2*] Qihang Fan[3] Jingxuan Niu[2] Zhipeng Zhang[4]
Libo Zhang[5] Guang Chen[1] Fan Chen[1] Longyin Wen[1] Sijie Zhu[1†]
[1]ByteDance Intelligent Creation [2]Tsinghua University
[3]Institute of Automation, Chinese Academy of Sciences [4] Shanghai Jiao Tong University
[5]Institute of Software, Chinese Academy of Sciences

Abstract

Spatio-temporal video grounding (STVG) requires localizing a target object in untrimmed videos both temporally and spatially from natural language descriptions. Despite their strong language understanding, multimodal large language models (MLLMs) underperform on STVG due to misaligned training objectives and weak fine-grained region-word alignment in standard visual encoders. To address this, we propose STVG-o1, the first framework that enables off-the-shelf MLLMs to achieve state-of-the-art STVG performance without any architectural modifications. Our method introduces a bounding-box chain-of-thought mechanism that explicitly reasons about spatio-temporal locations in an intermediate step before producing the final prediction. We further design a multi-dimensional reinforcement reward function consisting of format, consistency, temporal, spatial, and think rewards, which provides geometry-aware supervision through reinforcement fine-tuning. Evaluated on HCSTVG-v1/v2 and VidSTG, STVG-o1 sets new state-of-the-art results on HCSTVG, outperforming the best task-specific method by 7.3% m_tIoU on HCSTVG-v1, matching specialized models on VidSTG, and surpassing all existing MLLM-based approaches by large margins. It also demonstrates strong open-vocabulary generalization across datasets, establishing MLLMs as viable and powerful backbones for precise spatio-temporal grounding. Our code and models will be released.

1. Introduction

Spatio-Temporal Video Grounding (STVG) aims to localize a target object in an untrimmed video both temporally (*i.e.*, its start and end timestamps) and spatially (*i.e.*, bounding boxes in each frame) given a free-form natural language description [56]. As a core multimodal understanding task,

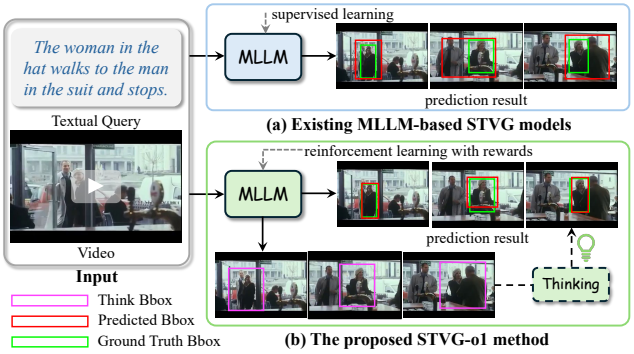

Figure 1. Comparison between existing MLLM-based STVG methods in (a) and the proposed STVG-o1 method in (b). *Best viewed in color for all figures.*

STVG requires fine-grained alignment between complex spatio-temporal video dynamics and textual semantics. Due to its fundamental role in video-language understanding and wide-ranging applications such as content-based video retrieval, intelligent surveillance, human-computer interaction, and service robotics, STVG has emerged as a major research focus in recent years.

Existing STVG methods [10, 11, 15, 45] primarily adopt transformer-based encoder-decoder architectures that model spatio-temporal alignment between video frames and textual descriptions in an end-to-end manner, achieving strong performance. These approaches typically build upon pre-trained image-level grounding models (*e.g.*, MDETR [16] or Grounding DINO [24]) and transfer their spatial reasoning capabilities to videos via fine-tuning. While effective, their performance is ultimately bounded by the limited language understanding and generalization capacity of the underlying grounding backbone. This limitation becomes especially pronounced with complex or abstract queries such as "*a person wearing red clothes jumps and then exits from the left*", where accurate semantic parsing and precise localization remain challenging. In contrast, Multimodal Large Language

*Equal contributions †Corresponding author

Models (MLLMs) [1, 2, 7, 20] trained on large-scale multimodal data offer significantly stronger language comprehension and cross-modal reasoning, suggesting a promising alternative for STVG. Recent works [2, 22, 35, 54] have indeed attempted to explore such MLLM-based solutions. Despite promising results, their performance still lags far behind task-specific methods. We argue that this under-performance stems not from a lack of inherent capability but from two key mismatches. ***First***, the training objective is ill-suited for grounding: most methods generate timestamps or bounding boxes as text tokens and optimize them with cross-entropy loss, which penalizes semantically correct predictions such as "1–9s" versus "0–8s" due to minor lexical differences. As a result, the training signal poorly reflects the actual localization quality, such as the Intersection over Union (IoU). ***Second***, current MLLM-based approaches simply inherit visual encoders pretrained for global image-text alignment (*e.g.*, CLIP-style [27]), which capture holistic semantics but lack fine-grained region-word correspondence—a capability essential for aligning phrases like "*person in red*" or "*jumping action*" with specific spatio-temporal regions.

To address these limitations, we propose leveraging reinforcement fine-tuning (RFT) for STVG, inspired by its success in guiding large models toward task-oriented reasoning through reward signals aligned with downstream objectives. Unlike maximum-likelihood training, RFT enables the direct optimization of metrics that reflect actual performance, such as IoU or temporal overlap, making it particularly suitable for grounding tasks. Prior work has successfully applied RFT to improve fine-grained understanding in visual question answering [43] and referring expression comprehension [3, 39, 44], suggesting its potential for STVG. Building on this insight, we present **STVG-o1**, a novel framework that unlocks the power of MLLMs for spatio-temporal grounding via reinforcement fine-tuning, as shown in Fig. 1. Particularly, drawing inspiration from the human cognitive process of "*thinking before deciding,*" STVG-o1 introduces a novel bounding-box chain-of-thought mechanism. Given a video and a textual query, the model first generates an intermediate reasoning step, a sequence of spatio-temporal bounding boxes denoted as `<think_bbox>`, which explicitly captures its initial hypothesis about the target object's location across key frames. It then refines this hypothesis into the final prediction `<pred_bbox>`. This two-stage generation enhances fine-grained region-word alignment and provides a structured intermediate signal that can be directly supervised by reward-based feedback. We design a multi-dimensional reward model consisting of five components, including *format reward*, *consistency reward*, *temporal reward*, *spatial reward*, and *think reward*. These rewards jointly encourage structurally valid outputs, coherent spatio-temporal trajectories, accurate timestamp prediction, precise bounding box localization, and meaningful intermediate reasoning in `<think_bbox>`. Using policy gradient methods, we fine-tune the MLLM end-to-end to maximize these rewards, effectively steering the model from semantic understanding toward precise spatio-temporal localization.

To the best of our knowledge, STVG-o1 is to date the first approach that unlocks the spatio-temporal grounding potential of off-the-shelf multimodal large language models (MLLMs) without any architectural modifications, relying solely on reinforcement-based fine-tuning with bounding-box chain-of-thought reasoning and a carefully designed multi-dimensional reward function. We evaluate our method on HCSTVG-v1/v2 and VidSTG. On HCSTVG, STVG-o1 achieves state-of-the-art performance, outperforming all prior methods, including all the task-specific methods, by a clear margin (*e.g.*, +7.3% m_tIoU over TA-STVG on HCSTVG-v1). On VidSTG, it matches the best task-specific models while surpassing all MLLM-based approaches. Across benchmarks, it yields dramatic gains over the MLLM base model [2] (*e.g.*, +34.7% m_tIoU and +25.0% m_vIoU on HCSTVG-v1). We also explore its open-vocabulary capability by training STVG-o1 on VidSTG and testing it on HCSTVG-v1, where it achieves strong cross-dataset performance, demonstrating robust generalization to unseen concepts and language.

In summary, the contributions of this work are as follows:

- We propose **STVG-o1**, the first framework that enables off-the-shelf multimodal large language models (MLLMs) to perform spatio-temporal video grounding without any architectural modifications, leveraging a novel bounding-box chain-of-thought reasoning mechanism.
- We design a multi-dimensional reinforcement reward function consisting of format reward, consistency reward, temporal reward, spatial reward, and think reward, which provides fine-grained supervision over both intermediate reasoning and final predictions and effectively compensates for the weak region-word alignment in standard MLLM visual encoders.
- We demonstrate state-of-the-art performance on HCSTVG-v1/v2, competitive results with task-specific methods on VidSTG, consistent superiority over all existing MLLM-based methods, and strong open-vocabulary generalization across datasets.

2. Related Work

Spatio-Temporal Video Grounding. STVG aims to localize a target of interest in an untrimmed video based on a natural language description, requiring both temporal and spatial localization. The development of STVG methods [10, 15, 23, 45, 55, 56] can be broadly divided into two phases. Early approaches [29, 55, 56] typically followed a two-stage pipeline: they first generated candidate proposals using a pre-trained object detector and then selected the correct proposal based on the textual query. More recent

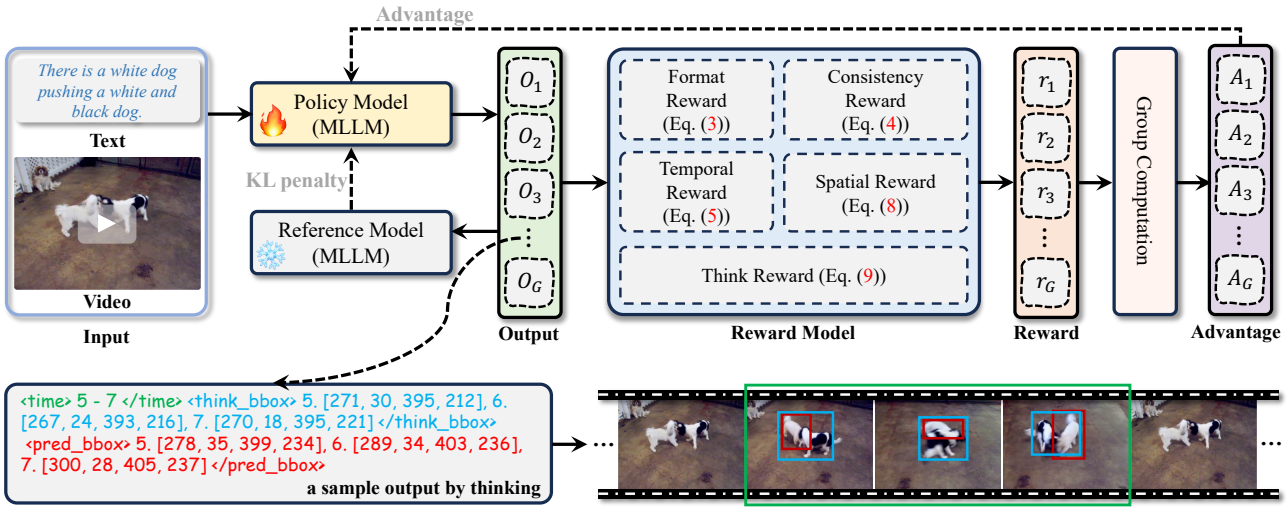

Figure 2. Overview of STVG-o1. Given a video and a natural language query, the base MLLM generates a chain-of-thought output sequence $O_1, \ldots, O_G$, where each step contains a temporal span, a sequence of thinking bounding boxes, and a sequence of final prediction bounding boxes. A reward model computes multi-dimensional rewards: format reward $\mathcal{R}_f$, consistency reward $\mathcal{R}_c$, temporal reward $\mathcal{R}_t$, spatial reward $\mathcal{R}_s$(combining GIoU and L1), and a think reward $\mathcal{R}_k$ that encourages refinement based on intermediate predictions. These rewards are aggregated to form a composite reward for reinforcement fine-tuning, enabling accurate spatio-temporal video grounding without architectural modifications. *Best viewed in color for all figures.*

methods [10, 15, 23, 46] have shifted toward a single-stage encoder-decoder framework to eliminate the heavy reliance on external detection models. In this paradigm, the encoder fuses multimodal features from video and text, while the decoder directly predicts the spatio-temporal location of the target without any external detector, achieving strong performance. However, these task-specific models still struggle to comprehend complex semantic descriptions or reason effectively in visually cluttered scenarios—areas where multimodal large language models (MLLMs) exhibit significantly stronger capabilities. **In contrast**, we propose to tackle STVG by activating the inherent spatio-temporal grounding potential of off-the-shelf MLLMs, without any architectural modifications.

Grounding in MLLMs. Recent multimodal large language models (MLLMs) [1, 2, 12, 13, 48, 57] have demonstrated promising capabilities in visual grounding tasks. Most existing efforts [2, 5, 12, 25, 40, 49, 57] focus on spatial grounding in static images, where the model is prompted to localize objects referred to in the text—typically by outputting bounding box coordinates or selecting from region proposals. For video understanding [4, 6, 14, 26, 32, 36, 50, 53], a few MLLMs extend grounding to the temporal dimension, localizing events or actions described in language along the time axis, *e.g.*, predicting start and end timestamps. However, current MLLM-based grounding models are limited to either spatial or temporal localization, but not both simultaneously. **In contrast**, we propose activating the joint spatio-temporal grounding capability of off-the-shelf MLLMs through a bounding-box chain-of-thought reasoning mechanism, enabling precise localization in both space and time within untrimmed videos.

Reasoning-Enhanced MLLMs. Chain-of-thought (CoT) reasoning has been extended from language models to multimodal settings to improve the performance of MLLMs on complex vision-language tasks. Recent works [3, 8, 9, 19, 21, 38, 39, 44, 51, 52] enhance MLLMs with step-by-step visual or semantic rationales [34], such as object relations or scene descriptions [33], through in-context examples or intermediate token generation. Although effective, these approaches rarely support coordinate-level spatial or spatio-temporal reasoning. **In contrast**, we introduce a bounding-box chain-of-thought that enables MLLMs to explicitly reason about object locations over time, unlocking their potential for precise STVG.

3. Method

Overview. We present STVG-o1, a reinforcement fine-tuned framework that enables off-the-shelf multimodal large language models (MLLMs) to perform precise spatio-temporal video grounding without architectural modifications. As detailed in §3.1, our approach introduces a bounding box chain-of-thought mechanism: the model first generates intermediate spatio-temporal bounding box predictions as an explicit reasoning step and then produces refined final predictions. To provide geometry-aware supervision, we design a multi-dimensional reward function (§3.2) consisting of

format, consistency, temporal, spatial, and think rewards, which jointly evaluate both intermediate and final outputs. The entire system is trained via policy gradient-based optimization (§3.3) to maximize these rewards, directly aligning the MLLM's behavior with localization accuracy rather than token-level likelihood.

3.1. Framework of STVG-o1

STVG-o1 unlocks the latent spatio-temporal grounding ability of off-the-shelf multimodal large language models by structuring their output into a multi-stage bounding-box reasoning process. As illustrated in Fig. 2, the framework explicitly separates intermediate reasoning from final prediction, enabling coherent spatio-temporal localization without any architectural modifications. Specifically, given a video $\mathcal{V} = \{v_i\}_{i=1}^{N_v}$ and a natural language query $\mathcal{W} = \{w_i\}_{i=1}^{N_t}$, where N_v is the video length and N_t is the text length, the multimodal large language model (MLLM) first generates a raw output string:

$$O_{\text{str}} = \text{MLLM}(\mathcal{V}, \mathcal{W}) \tag{1}$$

This string is then parsed via regular expression matching to extract three key structured components:

$$\mathcal{T}^p,\ \mathcal{B}^t,\ \mathcal{B}^p = \text{RegexParse}(O_{\text{str}}) \tag{2}$$

These components consist of a temporal interval $\mathcal{T}^p = [t_s, t_e]$, wrapped within the `<time>` and `</time>` delimiters; an intermediate bounding box chain-of-thought $\mathcal{B}^t = \{b_i^t\}_{i=t_s}^{t_e}$, where each $b_i^t \in \mathbb{R}^4$ denotes the predicted bounding box in frame i, parameterized by its two opposite corner coordinates (x_1, y_1, x_2, y_2), and enclosed within `<think_bbox>` and `</think_bbox>`; and a final spatial prediction $\mathcal{B}^p = \{b_i^p\}_{i=t_s}^{t_e}$ with the same parameterization, enclosed in `<pred_bbox>` and `</pred_bbox>`. Both the reasoning chain and the final prediction consist of bounding boxes aligned with the predicted time span $[t_s, t_e]$. This design enables end-to-end spatio-temporal grounding while preserving the original MLLM architecture and leveraging its native reasoning capacity.

3.2. Multi-dimensional Reward

To enable precise spatio-temporal grounding, we introduce a multi-dimensional reward that jointly supervises intermediate and final predictions. It consists of five components: format, consistency, temporal, spatial, and think rewards, each targeting a distinct aspect of localization quality.

Format reward. To ensure the model outputs results in the desired format, we expect it to wrap temporal localization within `<time>...</time>`, enclose the intermediate bounding box within `<think_bbox>...</think_bbox>`, and place the final prediction within `<pred_bbox>...</pred_bbox>`. We use regular expression matching to determine whether the model's output conforms to the specified format:

$$\mathcal{R}_{\text{f}} = \begin{cases} 1, & \text{if output matches format,} \\ 0, & \text{if output doesn't match format.} \end{cases} \tag{3}$$

Consistency Reward. To ensure coherent spatio-temporal reasoning, we require both the intermediate bounding box $\mathcal{B}^t$ and the final prediction $\mathcal{B}^p$ to strictly align with the predicted temporal interval $\mathcal{T}^p$, meaning that both sequences must represent bounding boxes for frames $t_s, t_s + 1, \ldots, t_e$. A reward is given only if both $\mathcal{B}^t$ and $\mathcal{B}^p$ align with frames t_s through t_e:

$$\mathcal{R}_{\text{c}} = \begin{cases} 1, & \text{if the output is consistent,} \\ 0, & \text{If the output is not consistent.} \end{cases} \tag{4}$$

Temporal Reward. To improve temporal grounding, we design a temporal reward. Specifically, we use the Intersection over Union (IoU) between the predicted and ground truth temporal segments as the metric for this reward. This reward effectively guides the model to improve the precision of localizing the target event in time.

$$\mathcal{R}_{\text{t}} = \text{IoU}(\mathcal{T}^p,\ \mathcal{T}^{\text{gt}}) \tag{5}$$

where $\mathcal{T}^{\text{gt}}$ denotes the ground truth temporal segments.

Spatial Reward. To improve fine-grained spatial grounding, the spatial reward is computed only over the temporal intersection between the predicted and ground-truth event intervals. The intersection of the predicted time span $\mathcal{T}^p$ and the ground truth $\mathcal{T}^{\text{gt}}$ is defined as

$$\mathcal{T}^{\cap} = [\max(t_s, t_s^{\text{gt}}), \min(t_e, t_e^{\text{gt}})] = [t_s^{\cap}, t_e^{\cap}] \tag{6}$$

If $t_s^{\cap} > t_e^{\cap}$, the intersection is empty, and the spatial reward is set to zero. For a non-empty intersection, let $\mathcal{B} = \{b_i\}_{i=t_s^{\cap}}^{t_e^{\cap}}$ be the predicted bounding boxes (from either the reasoning chain or final prediction) and $\mathcal{B}^{\text{gt}} = \{b_i^{\text{gt}}\}_{i=t_s^{\cap}}^{t_e^{\cap}}$ the corresponding ground-truth boxes. The spatial reward is computed as:

$$\mathcal{R}_{\text{spa}} = \frac{1}{|\mathcal{T}^{\cap}|} \sum_{i=t_s^{\cap}}^{t_e^{\cap}} \left(\text{GIoU}(b_i, b_i^{\text{gt}}) - \|b_i - b_i^{\text{gt}}\|_1\right) \tag{7}$$

where $|\mathcal{T}^{\cap}| = t_e^{\cap} - t_s^{\cap} + 1$. The total spatial reward sums over both the intermediate reasoning and the final prediction:

$$\mathcal{R}_{\text{s}} = \mathcal{R}_{\text{spa}}^t + \mathcal{R}_{\text{spa}}^p \tag{8}$$

This design ensures that spatial feedback is provided only when temporal localization is reasonably accurate, thereby coupling spatial and temporal learning in principled manner.

Think Reward. To encourage the model to refine its spatial predictions through reasoning, we introduce a think reward that measures the improvement from the intermediate bounding box to the final prediction. Specifically, let R^t_{spa} and R^p_{spa} denote the spatial rewards (computed over the temporal intersection as in Eq. (8)) for the reasoning sequence (<think_bbox>) and the final output (<pred_bbox>), respectively. The reward is computed as:

$$\mathcal{R}_{\text{k}} = \max(\mathcal{R}^p_{\text{spa}} - \mathcal{R}^t_{\text{spa}}, 0) \tag{9}$$

A positive r_k indicates that the model successfully enhances its localization accuracy through deliberative reasoning, thereby reinforcing productive chain-of-thought behavior. The total reward is formulated as:

$$\mathcal{R} = \mathcal{R}_f + \mathcal{R}_c + \mathcal{R}_t + \mathcal{R}_s + \lambda_k * \mathcal{R}_k \tag{10}$$

where λ_k is a hyperparameter that effectively balances the reward contributions.

3.3. Optimization

We optimize the policy model π_θ using GRPO [13]. For each training sample consisting of a video $\mathcal{V}$ and a query $\mathcal{W}$, we generate $n = 8$ responses from the current policy $\pi_{\theta_{\text{old}}}$. Each response o_i receives a scalar reward $\mathcal{R}(o_i)$ as defined in Equation (10). The advantage for each response is computed by normalizing rewards within the group:

$$A_i = \frac{\mathcal{R}(o_i) - \mu}{\sigma + \delta} \tag{11}$$

where μ and σ are the mean and standard deviation of the group rewards, and $\delta = 10^{-6}$ ensure numerical stability. The policy update maximizes the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(\tilde{\mathcal{V}},q)\sim\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1-\epsilon, 1+\epsilon)A_i) - \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid q) \,\|\, \pi_{\text{ref}}(\cdot \mid q)\right)\Big)\right] \tag{12}$$

where $r_i(\theta) = \pi_\theta(o_i \mid q)/\pi_{\theta_{\text{old}}}(o_i \mid q)$, ϵ limit the step size via clipping, and β controls the strength of KL regularization against a frozen reference policy π_{ref}.

4. Experiments

Implementation. We implement STVG-o1 in Python using PyTorch, with Qwen2.5-VL-7B as the base model. Our training framework is based on verl [28] and integrates components from vLLM [18] to support efficient multi-modal sequence generation and policy optimization. The model is trained using the AdamW [17] optimizer with a learning rate of $1e-6$ and a weight decay of $1e-2$. We adopt Group Relative Policy Optimization [13] with 8 rollouts per iteration and a global batch size of 128, achieved via gradient accumulation when necessary. Input videos are uniformly sampled at 2 frames per second. Each frame is resized so that its longer side is at most 336 pixels, preserving the original aspect ratio, resulting in a maximum resolution of 336×336. The think reward weight parameter λ_k is set to 0.5.

Datasets. We evaluate our method on two standard spatio-temporal video grounding benchmarks: HC-STVG [31] and VidSTG [56]. HC-STVG focuses on multi-person scenes, where each untrimmed video is paired with a textual description of human attributes and actions. The original version, HCSTVG-v1, contains 5,660 videos (4,500 for training and 1,160 for testing). The extended HCSTVG-v2 includes 10,131 training, 2,000 validation, and 4,413 test samples. Since the test annotations for v2 are not public, we report results on the validation set, following prior work [10, 11, 15]. VidSTG consists of 6,924 untrimmed videos with 99,943 sentences (declarative and interrogative) referring to 80 object categories, forming 44,808 video–triplet instances. Following the standard split [56], it uses 5,563 / 618 / 743 videos for training, validation, and testing, associated with 80,684 / 8,956 / 10,303 sentences, respectively.

Metrics. Following prior work [10, 11, 15, 45], we adopt three standard metrics for spatio-temporal video grounding: m_tIoU, m_vIoU, and vIoU@R. m_tIoU evaluates temporal grounding accuracy by averaging the Intersection-over-Union (tIoU) between predicted and ground-truth time intervals over all test samples. m_vIoU assesses spatial grounding quality by computing the average 3D IoU across space and time between predicted and annotated spatio-temporal tubes. vIoU@R measures the percentage of test samples whose vIoU exceeds a threshold R (*e.g.*, $R = 0.3, 0.5$), reflecting performance under stricter localization criteria. For detailed metrics, please kindly refer to [45].

4.1. State-of-the-Art Comparison

HC-STVG Datasets. We evaluate STVG-o1 on HCSTVG-v1 and HCSTVG-v2, two challenging benchmarks for spatio-temporal video grounding. As shown in Tab. 1 and Tab. 2, our method achieves state-of-the-art results, outperforming both specialized models and existing MLLM-based approaches. On HCSTVG-v1, STVG-o1 obtains 60.3% m_tIoU and 44.1% m_vIoU, surpassing the previous best method, TA-STVG, by +7.3% and +5.0%, respectively. Compared to the base model Qwen2.5-VL (25.6% m_tIoU), we achieve a +34.7% absolute gain, demonstrating that reinforcement fine-tuning effectively unlocks the latent grounding capability of MLLMs.

On the larger HCSTVG-v2 benchmark, STVG-o1 further improves to 63.8% m_tIoU and 41.2% m_vIoU, establishing a new state of the art. Notably, our method achieves this without any architectural modifications, unlike prior

Table 1. Comparison on HCSTVG-v1 (%).

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Task-specific Methods				
STVGBert [ICCV2021] [29]	-	20.4	29.4	11.3
TubeDETR [CVPR22] [45]	43.7	32.4	49.8	23.5
STCAT [NeurIPS22] [15]	49.4	35.1	57.7	30.1
SGFDN [ACMMM23] [37]	46.9	35.8	56.3	37.1
STVGFormer [CVPR23] [23]	-	36.9	62.2	34.8
VG-DINO [CVPR24] [42]	-	38.3	62.5	36.1
CG-STVG [CVPR24] [10]	52.8	38.4	61.5	36.3
TA-STVG [ICLR25] [11]	53.0	39.1	63.1	36.8
MLLM-based Methods				
Gemini-2.5-Pro [Arxiv25] [7]	55.1	25.9	39.1	9.9
GPT-4o [Arxiv23] [1]	27.5	7.9	4.0	0.3
GroundingGPT [ACL24] [22]	22.2	16.7	15.0	4.9
Qwen2.5-VL [Arxiv25] [2]	25.6	19.1	20.2	12.6
LLaVA-Video-SFT [TMLR25] [54]	52.8	27.7	43.1	21.3
Qwen2.5-VL-SFT [Arxiv25] [2]	53.5	28.6	45.2	21.9
SpaceVLLM [AAAI26] [35]	56.9	39.3	66.6	36.9
STVG-o1 (ours)	**60.3**	**44.1**	**73.3**	**43.5**

Table 2. Comparison on HCSTVG-v2 (%).

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Task-specific Methods				
PCC [arxiv22] [47]	-	30.0	-	-
2D-Tan [arxiv22] [30]	-	30.4	50.4	18.8
MMN [AAAI22] [41]	-	30.3	49.0	25.6
TubeDETR [CVPR22] [45]	53.9	36.4	58.8	30.6
STVGFormer [CVPR23] [23]	58.1	38.7	65.5	33.8
VG-DINO [CVPR24] [42]	-	39.9	67.1	34.5
CG-STVG [CVPR24] [10]	60.0	39.5	64.5	36.3
TA-STVG [ICLR25] [11]	60.4	40.2	65.8	36.7
MLLM-based Methods				
Gemini-2.5-Pro [Arxiv25] [7]	60.4	24.5	34.6	9.6
GPT-4o [Arxiv23] [1]	32.7	9.1	5.7	0.0
GroundingGPT [ACL24] [22]	19.6	14.7	16.6	3.1
Qwen2.5-VL [TMLR25] [54]	22.9	13.0	15.6	6.4
LLaVA-Video-SFT [TMLR25] [54]	54.2	24.8	40.1	15.5
Qwen2.5-VL-SFT [Arxiv25] [2]	55.3	26.5	38.6	20.2
SpaceVLLM [AAAI26] [35]	58.0	34.0	56.9	24.7
STVG-o1 (ours)	**63.8**	**41.2**	**68.5**	**39.6**

Table 3. Comparison with existing state-of-the-art methods on VidSTG (%).

Methods	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Task-specific Methods								
STGRN [CVPR20] [56]	48.5	19.8	25.8	14.6	47.0	18.3	21.1	12.8
STVGBert [ICCV21] [29]	-	24.0	30.9	18.4	-	22.5	26.0	16.0
TubeDETR [CVPR22] [45]	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT [NeurIPS22] [15]	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
SGFDN [ACMMM23] [37]	45.1	28.3	41.7	29.1	44.8	25.8	36.9	23.9
STVGFormer [CVPR23] [23]	-	33.7	47.2	32.8	-	28.5	39.9	26.2
CG-STVG [CVPR24] [10]	51.4	34.0	47.7	33.1	49.9	29.0	40.5	27.5
VG-DINO [CVPR24] [42]	52.0	**34.7**	48.1	**34.0**	**50.8**	**29.9**	41.0	27.6
TA-STVG [ICLR25] [11]	51.7	34.4	48.2	33.5	50.2	29.5	**41.5**	**28.0**
MLLM-based Methods								
Gemini-2.5-Pro [Arxiv25] [7]	49.9	22.5	33.6	12.0	45.4	13.7	17.3	8.1
GPT-4o [Arxiv23] [1]	38.3	9.2	7.1	1.6	39.8	6.1	3.5	0.6
GroundingGPT [ACL24] [22]	15.5	12.3	13.2	4.1	11.9	8.7	9.6	2.9
Qwen2.5-VL [Arxiv25] [2]	16.8	10.9	14.3	5.4	13.8	8.5	11.3	4.4
LLaVA-Video-SFT [TMLR25] [54]	39.5	19.2	20.3	13.8	38.6	15.7	15.3	10.4
LLaVA-ST [CVPR25] [20]	45.5	24.8	36.0	22.9	43.2	20.0	28.1	17.5
Qwen2.5-VL-SFT [Arxiv25] [2]	41.6	20.3	26.1	15.4	40.9	17.1	17.6	13.9
SpaceVLLM [AAAI26] [35]	47.7	27.4	39.1	26.2	48.5	25.4	35.9	22.2
STVG-o1 (ours)	**52.1**	33.5	**48.4**	32.0	50.5	27.9	39.8	26.0

Table 4. Performance comparisons of the state-of-the-art on HCSTVG-v1 in open-vocabulary setting.

Method	Pre-training	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
TubeDETR [45]	VidSTG	-	16.8	22.3	9.2
STCAT [15]	VidSTG	-	22.6	32.1	20.8
CG-STVG [10]	VidSTG	31.2	21.7	31.1	17.8
VG-DINO [42]	VidSTG	-	27.5	40.1	**29.9**
TA-STVG [11]	VidSTG	30.1	20.9	30.8	11.5
STVG-o1 (ours)	VidSTG	**45.9**	**32.2**	**50.8**	25.5

approaches [20, 35] that rely on additional detection heads or external modules. This underscores the effectiveness of our "thinking with bounding boxes" mechanism combined with task-aligned reinforcement learning rewards. The results show that off-the-shelf MLLMs, when optimized with localization-aware objectives, can match or surpass specialized STVG models.

VidSTG Dataset. We also evaluate STVG-o1 on the VidSTG benchmark, which includes both declarative and interrogative language queries to test grounding robustness under diverse linguistic forms. As shown in Tab. 3, our method achieves competitive performance against specialized task-specific models while consistently outperforming all existing MLLM-based approaches. On declarative sentences, STVG-o1 significantly surpasses the best prior MLLM-based method, SpaceVLLM (47.7% m_tIoU, 27.4% m_vIoU), by 4.4% and 6.1%, respectively. For interrogative sentences, STVG-o1 achieves 50.5% m_tIoU and 27.9% m_vIoU, again outperforming all MLLM-based methods and remaining competitive with task-specific architectures. Notably, compared to the Qwen2.5-VL-SFT, which has been supervised fine-tuned on the STVG dataset, our method yields

absolute gains of 10.5% in m_tIoU and 13.2% in m_vIoU, highlighting the effectiveness of our STVG-o1.

4.2. Open-Vocabulary Comparison

To assess open-vocabulary generalization, we train STVG-o1 on VidSTG and evaluate it on the HCSTVG-v1 test set. As shown in Tab. 4, our method achieves 45.9% m_tIoU and 32.2% m_vIoU, substantially outperforming the TA-STVG by +15.8% in m_tIoU and +11.3% in m_vIoU. This strong cross-dataset performance highlights a key advantage of our approach: by fine-tuning a general-purpose MLLM with reinforcement learning guided by spatio-temporal rewards, STVG-o1 learns to interpret arbitrary object descriptions through language without relying on fixed detection vocabularies or task-specific modules. In contrast, most prior methods suffer from domain shift or limited linguistic coverage when evaluated outside their training distribution. The high vIoU@0.3 (50.8%) further confirms that our predictions maintain accurate spatial alignment even under open-vocabulary conditions.

4.3. Ablation Study

Impact of thinking with bounding boxes. We conduct ablation studies to evaluate the effectiveness of our bounding box chain-of-thought mechanism. As shown in Tab. 5, the base model, without any training, achieves poor performance, highlighting the need for task-specific optimization. Supervised fine-tuning (SFT) significantly improves results, but reinforcement fine-tuning (RFT), guided by our multi-dimensional reward function, further boosts performance across all metrics, demonstrating the benefit of aligning training with downstream grounding objectives through geometry-aware rewards. Most notably, adding the `<bbox_think>` reasoning step under RFT yields substantial gains of +1.8% m_tIoU and +2.3% m_vIoU, indicating that explicit intermediate reasoning enhances both temporal and spatial localization accuracy. This confirms that the thinking-with-bounding-boxes mechanism, when supervised by our multi-dimensional rewards, effectively guides the model toward more precise spatio-temporal predictions.

Table 5. Ablations of thinking with bounding boxes.

Training	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
-	25.6	19.1	20.2	12.6
SFT	53.5	28.6	45.2	21.9
RFT	58.5	41.8	68.7	38.9
RFT + Thinking	**60.3**	**44.1**	**73.3**	**43.5**

Impact of the think reward. We ablate the think reward, which encourages the model to refine bounding boxes based on its intermediate reasoning. As shown in Tab. 6, removing this reward leads to a drop in performance: m_tIoU decreases from 60.3% to 59.3%, and m_vIoU drops from 44.1% to 42.2%. This confirms that the think reward plays a key role in guiding the MLLM to iteratively improve spatial localization accuracy during reinforcement learning, enabling more precise grounding through self-corrective reasoning.

Table 6. Ablations of think reward.

Think Reward	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
-	59.3	42.2	69.1	39.9
✓	**60.3**	**44.1**	**73.3**	**43.5**

Impact of spatial reward. We analyze the spatial reward, which consists of GIoU and L1 distance reward components. As shown in Tab. 7, using only GIoU or L1 leads to inferior performance (*e.g.*, m_vIoU drops to 40.6% and 39.9%, respectively). In contrast, combining both achieves 60.3% m_tIoU and 44.1% m_vIoU, significantly improving spatial grounding accuracy and demonstrating the effectiveness of joint optimization.

Table 7. Ablations of spatial reward.

Spatial Reward	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
$\mathcal{G}$	58.6	40.6	67.6	36.6
$\mathcal{L}_1$	58.5	39.9	67.0	34.6
$\mathcal{G} + \mathcal{L}_1$	**60.3**	**44.1**	**73.3**	**43.5**

Impact of think reward weight. We study the effect of the think reward weight λ_k on performance. As shown in Tab. 8, we can see that the model performs best by setting λ_k to 0.5.

Table 8. Ablations of think reward weight.

λ_k	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
0.2	59.5	42.5	70.9	40.6
0.5	60.3	**44.1**	**73.3**	**43.5**
1.0	**60.5**	43.8	72.4	42.9

4.4. Grounding Performance Across Object Scales

To understand how STVG-o1 performs spatial grounding across object scales, we analyze the average vIoU of intermediate `<think_bbox>` and final `<pred_bbox>` predictions, grouped by bounding box area. As shown in Fig. 4, the model achieves a lower vIoU in the early reasoning stage (blue curve), indicating initial coarse localization. However, the final output (red curve) consistently outperforms the intermediate prediction across all size ranges, with a maximum improvement of +11% in the 0.04–0.09 area ratio bin, which typically corresponds to medium-sized objects. This suggests that our reinforcement-based refinement process is particularly effective for objects of moderate scale, where both visual context and language cues are rich enough to guide accurate adjustments. Notably, even for small or large objects, the `<pred_bbox>` maintains higher accuracy than `<think_bbox>`, confirming that the model reliably refines its spatial predictions, regardless of object size. These results validate that our method robustly enhances grounding precision across diverse object scales.

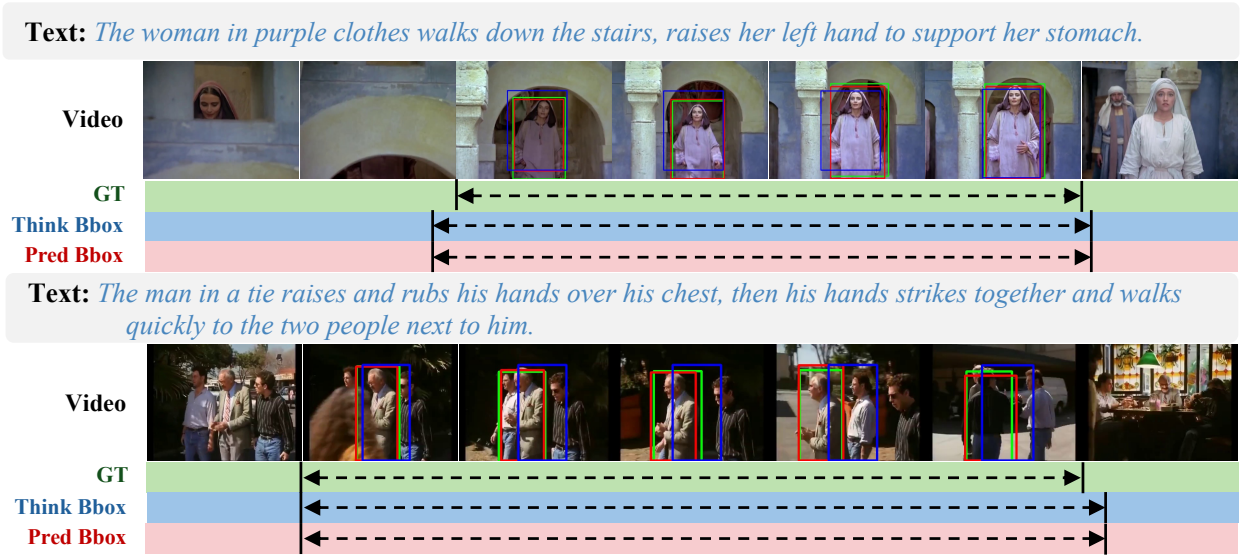

Figure 3. Qualitative results of our STVG-o1. Green boxes denote ground truth, blue boxes represent intermediate `<think_bbox>` predictions during reasoning, and red boxes indicate final `<pred_bbox>` outputs. *Best viewed in color for all figures.*

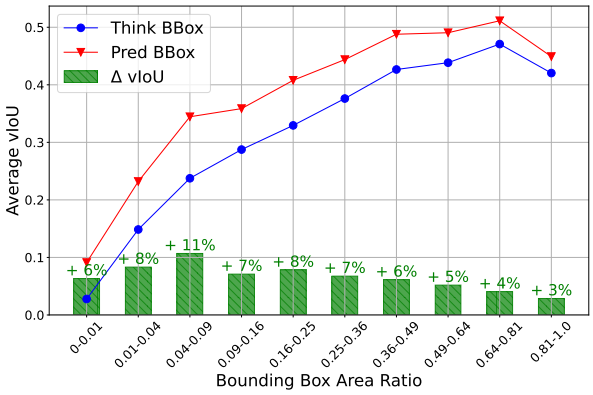

Figure 4. Analysis of average vIoU across different bounding box areas. Blue curve shows performance of intermediate `<think_bbox>` predictions, red curve shows final `<pred_bbox>` outputs, and green bars indicate the relative improvement (Δ vIoU) from think bbox to predicted bbox.

4.5. Inference Complexity Analysis

We analyze the inference complexity of STVG-o1 and compare it with existing methods on a single GPU. As shown in Tab. 9, the two task-specific models, CG-STVG [10] and TA-STVG [11], achieve fast inference due to their lightweight architectures and non-autoregressive design. In contrast, all MLLM-based approaches, such as Qwen2.5-VL-SFT [2], LLaVA-ST [20], and our STVG-o1, require autoregressive generation of bounding boxes, resulting in significantly higher latency (approximately 15–17 seconds). Please note that since SpaceVLLM [35] has not been released, we do not include it in this comparison. Despite the shared AR generation paradigm, STVG-o1 achieves faster inference than both Qwen2.5-VL-SFT (15.49 s *vs.* 16.37 s) and LLaVA-ST (15.49 s *vs.* 17.63 s), thanks to its more efficient and compact output format for bounding boxes. It also uses less GPU memory (24.2 GB *vs.* 38.7 GB) than LLaVA-ST. These results demonstrate that while MLLMs introduce higher inference costs compared to task-specific models, they can still achieve competitive speed and better efficiency when optimized with a streamlined output design.

Table 9. Comparison on model complexity. The 'AR' refers to autoregressive generation of bounding boxes.

Methods	AR	Params	Time	GPU Mem
CG-STVG [10]		231 M	0.61 s	29.7 G
TA-STVG [11]		234 M	0.57 s	28.4 G
Qwen2.5-VL-SFT [2]	✓	8.3 B	16.37 s	26.6 G
LLaVA-ST [20]	✓	8.3 B	17.63 s	38.7 G
STVG-o1 (ours)	✓	8.3 B	15.49 s	24.2 G

4.6. Qualitative Analysis

To further qualitatively validate the effectiveness of our proposed method, Fig. 3 shows the qualitative results of STVG-o1 on HCSTVG [31]. Green boxes denote ground truth, blue boxes represent intermediate `<think_bbox>` predictions during reasoning, and red boxes indicate final `<pred_bbox>` outputs. The model first generates a coarse localization (blue) and then refines it to better align with the target (red), illustrating how the bounding-box chain-of-thought enables progressive spatial grounding.

Due to limited space, please refer to supplementary material for more visualizations, analyzes, and experimental details.

5. Conclusion

We present STVG-o1, the first framework that enables off-the-shelf multimodal large language models to perform spatio-temporal video grounding without architectural changes. By introducing a bounding-box chain-of-thought mechanism and a multi-dimensional reward function consisting of format reward, consistency reward, temporal reward, spatial reward, and think reward, we provide fine-grained, geometry-aware supervision through reinforcement fine-tuning. This design effectively overcomes the misaligned training objective and the weak region-word alignment inherent in standard MLLMs. STVG-o1 achieves state-of-the-art results on HCSTVG-v1/v2, matches specialized models on VidSTG, and outperforms all existing MLLM-based approaches. It also demonstrates strong open-vocabulary generalization across datasets. Our work shows that, with proper task-oriented rewards, MLLMs can be efficiently adapted to complex grounding tasks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 6, 8

[3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025. 2, 3

[4] Wayner Barrios, Mattia Soldan, Fabian Caba Heilbron, Alberto Mario Ceballos-Arroyo, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In *ICCV*, 2023. 3

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3

[6] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. In *NeurIPS*, 2021. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 6

[8] Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19597–19607, 2025. 3

[9] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18941–18951, 2023. 3

[10] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *CVPR*, 2024. 1, 2, 3, 5, 6, 8

[11] Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. *arXiv preprint arXiv:2502.11168*, 2025. 1, 5, 6, 8

[12] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 3

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3, 5

[14] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *ECCV*, 2022. 3

[15] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, 2022. 1, 2, 3, 5, 6

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023. 5

[19] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462*, 2025. 3

[20] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8592–8603, 2025. 2, 6, 8

[21] Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *arXiv preprint arXiv:2509.18056*, 2025. 3

[22] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 6657–6678, 2024. 2, 6

[23] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, 2023. 2, 3, 6

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55, 2024. 1

[25] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 3

[26] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[28] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 5

[29] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021. 2, 6

[30] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Xiang Li, and Wei-Shi Zheng. Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. *arXiv*, 2021. 6

[31] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE TCSVT*, 32(12):8238–8249, 2021. 5, 8

[32] Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025. 3

[33] Jiacong Wang, Bohong Wu, Haiyong Jiang, Xun Zhou, Xin Xiao, Haoyuan Guo, and Jun Xiao. World to code: Multi-modal data generation via self-instructed compositional captioning and filtering. *arXiv preprint arXiv:2409.20424*, 2024. 3

[34] Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025. 3

[35] Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. *arXiv preprint arXiv:2503.13983*, 2025. 2, 6, 8

[36] Lan Wang, Gaurav Mittal, Sandra Sajeev, Ye Yu, Matthew Hall, Vishnu Naresh Boddeti, and Mei Chen. Protege: Untrimmed pretraining for video temporal grounding by video temporal grounding. In *CVPR*, 2023. 3

[37] Weikang Wang, Jing Liu, Yuting Su, and Weizhi Nie. Efficient spatio-temporal video grounding with semantic-guided feature decomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4867–4876, 2023. 6

[38] Yiqin Wang, Haoji Zhang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Hierarchical memory for long video qa. *arXiv preprint arXiv:2407.00603*, 2024. 3

[39] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding, 2025b. *arXiv preprint arXiv:2503.13377*, 2025. 2, 3

[40] Yiqin Wang, Haoji Zhang, Jingqi Tian, and Yansong Tang. Ponder & press: Advancing visual gui agent towards general computer control. In *Findings of the*

Association for Computational Linguistics: ACL 2025, pages 1461–1473, 2025. 3

[41] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, 2022. 6

[42] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024. 6

[43] Zhucun Xue, Jiangning Zhang, Xurong Xie, Yong Liu, Xiangtai Li, Dacheng Tao, et al. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[44] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 2, 3

[45] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 1, 2, 5, 6

[46] Jiali Yao, Xinran Deng, Xin Gu, Mengrui Dai, Bing Fan, Zhipeng Zhang, Yan Huang, Heng Fan, and Libo Zhang. Omnistvg: Toward spatio-temporal omni-object video grounding. *arXiv preprint arXiv:2503.10500*, 2025. 3

[47] Yi Yu, Xinying Wang, Wei Hu, Xun Luo, and Cheng Li. 2rd place solutions in the hc-stvg track of person in context challenge 2021. *arXiv*, 2021. 6

[48] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025. 3

[49] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3

[50] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025. 3

[51] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. *arXiv preprint arXiv:2506.23825*, 2025. 3

[52] Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24203–24213, 2025. 3

[53] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. Text-visual prompting for efficient 2d temporal video grounding. In *CVPR*, 2023. 3

[54] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2, 6

[55] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *IJCAI*, 2020. 2

[56] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020. 1, 2, 5, 6

[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3

Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via Reinforcement Fine-Tuning

Supplementary Material

For a better understanding of this work, we offer additional details, analysis, and results as follows:

- **A** ***Analysis of Qualitative Results***
  In this section, we show qualitative results of our method and a comparison to the TA-STVG method.
- **B** ***Analysis of Failure Cases***
  In this section, we discuss the failure cases of our proposed method.
- **C** ***Prompt Details for Closed-Source Model***
  In this section, we present the prompt used to perform STVG with closed-source models.

A. Analysis of Qualitative Results

A.1 Thinking with Bounding Boxes

We present qualitative results in Fig. 5 to illustrate how STVG-o1 refines its spatio-temporal predictions through a structured chain-of-thought process. In each example, the intermediate `<think_bbox>` (blue) represents the model's initial reasoning output, while the final `<pred_bbox>` (red) reflects the refined prediction after iterative refinement. As shown, the model consistently generates plausible but slightly imprecise bounding boxes during reasoning, often exhibiting minor localization errors or temporal drift. However, the final prediction demonstrates significant improvement; it aligns more closely with the ground truth (green), both spatially and temporally. For instance, in the first example, the initial blue box fails to fully capture the man's movement trajectory, but the red box correctly adjusts to track his motion across frames. Similarly, in the third example, the think bbox incorrectly includes part of the background, whereas the pred bbox tightens the region to focus on the target. These observations highlight that STVG-o1 does not merely generate a single-shot prediction; instead, it engages in a deliberative reasoning process in which intermediate outputs serve as stepping stones toward accurate grounding. The progressive refinement from `<think_bbox>` to `<pred_bbox>` underscores the effectiveness of our reward-driven optimization in encouraging the model to improve upon its own reasoning, leading to robust and precise spatio-temporal grounding.

A.2 Comparison with TA-STVG

We also present qualitative comparisons between our STVG-o1 and the task-specific method TA-STVG in Fig. 6. Green boxes denote ground truth, blue boxes represent TA-STVG's predictions, and red boxes indicate our STVG-o1 outputs. The results demonstrate that STVG-o1 achieves more accurate and semantically consistent spatio-temporal grounding. For example, in the first instance, TA-STVG activates its prediction prematurely, starting before the target man in the suit is visible. This leads to an incorrect temporal onset that includes irrelevant motion prior to the actual event. In contrast, STVG-o1 waits until the subject becomes visually distinct, specifically when he steps fully into view, before initiating the grounding process. These results show the effectiveness of STVG-o1.

B. Analysis of Failure Cases

Despite the strong performance of STVG-o1, it still faces challenges in certain scenarios, as illustrated in Fig. 7. We analyze three typical failure modes: *(i)* **Missing small objects.** Due to the limited resolution of input video frames, our model struggles to localize small or distant targets, especially during their initial appearance. As shown in the top example of Fig. 7, the girl is not localized until she moves closer, which indicates a limitation in early-stage grounding. *(ii)* **Extremely short events.** With a fixed frame sampling rate, brief actions may span fewer than two sampled frames, causing critical transitions to be missed entirely. As illustrated in the middle example, the model fails to pinpoint the exact moment when the man stands up, resulting in an over-extended duration prediction. *(iii)* **Occlusion-induced confusion.** When the target object is occluded by another object, the model may shift its attention to a visually similar but incorrect entity. As seen in the bottom example (Fig. 7), after the standing man is partially blocked by a woman, the model incorrectly localizes the woman instead, highlighting the challenge of maintaining consistent grounding under occlusion. To address these limitations, we plan to explore adaptive frame sampling, dynamic input resolution, and enhanced chain-of-thought reasoning to improve robustness for spatio-temporal grounding.

C. Prompt Details for Closed-Source Model

To investigate the STVG performance of closed-source models, we also experiment with the APIs of closed-source models, including GPT-4o and Gemini-2.5 Pro. To ensure these models generate outputs in the required format, we designed the prompt shown in Fig. 8.

Figure 5. Qualitative results of our STVG-o1. Green boxes denote ground truth, blue boxes represent intermediate `<think_bbox>` predictions during reasoning, and red boxes indicate final `<pred_bbox>` outputs. *Best viewed in color for all figures*.

Figure 6. Qualitative results comparing our STVG-o1 with the task-specific TA-STVG method. Green boxes denote ground truth, blue boxes represent TA-STVG's predictions, and red boxes indicate our STVG-o1 outputs. *Best viewed in color for all figures*.

Figure 7. Failure cases of STVG-o1. Green boxes denote ground-truth bounding boxes, and red boxes indicate predictions from our method. *Best viewed in color for all figures*.

Task:
Given the following video and query, please perform spatial-temporal video grounding.

Query:
"The man in brown clothes pours the contents of the bag into his hand, and then takes out a piece of paper from the bag and opens it."

Instructions:
1. **First**, predict the temporal range in which the described target appears in the video.
   - Output the temporal range in the format: <start_time>-<end_time>
   - start_time and end_time are absolute seconds (integer, starting from 0)
   - Only output one temporal range, corresponding to the entire action described in the query.

2. **Next**, output a CSV list with the following columns: **timestamp, x1, y1, x2, y2**
   - timestamp: absolute second (integer, starting from 0)
   - x1, y1, x2, y2: bounding box coordinates in relative format (values between 0 and 1)
   - Only output for seconds when the described target is present in the video
   - Output at most one bounding box per second (for each second, output at most one line if the target is present)
   - Maximum 30 bounding boxes in total. If the target is present for more than 30 seconds, only output the first 30.
   - If the target is not present in a given second, do not output that second.

3. **Output format**:
   - First line: temporal range, <start_time>-<end_time>, e.g. 12-24
   - Then the CSV list, starting with the header: timestamp, x1, y1, x2, y2

4. **Example Output**:
   12-24
   timestamp, x1, y1, x2, y2
   12, 0.42, 0.02, 0.879, 1.0
   13, 0.43, 0.01, 0.88, 1.0
   ...

Video Duration: {duration} seconds, frames are sampled at fixed FPS.
Query: "{question}"

Please analyze the video and return the temporal range and CSV list as specified. Do not output anything else.

Figure 8. Prompt design for STVG with closed-source models, specifying temporal and spatial output formats.