EFFICIENT SPATIO-TEMPORAL GROUNDING WITH MULTIMODAL LARGE MODELS VIA SECOND-LEVEL TRACKING AND RL VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal video grounding (STVG) requires a model to identify when a queried event occurs and track the referred target throughout that interval. Applying multimodal large language models (MLLMs) frame by frame produces long visual contexts and autoregressive coordinate sequences. We formulate MLLM-based STVG around second-level trajectory control points and reconstruct dense tracks with deterministic inter-second smoothing. Training combines continued pre-training on broad grounding data, supervised fine-tuning with generated decision traces whose numerical labels are replaced by ground truth, and reinforcement learning with a verifier based on temporal and motion-aware spatial overlap. On 10-second videos, the second-level protocol reduces end-to-end latency by $29.3\times$ and prefill compute by $56.9\times$ relative to a 25-FPS frame-level alternative; on 60-second videos it remains executable when 5-FPS and 25-FPS inputs exceed the context window. A 9B task-aligned model improves over its SFT checkpoint on VidSTG and HC-STVG and also improves its base checkpoint on Video-MME and MMVU, showing that compact grounding preserves general video understanding.

1 INTRODUCTION

Spatio-temporal video grounding (STVG) asks a model to localize a natural-language query in both time and space (Su et al., 2021). Given a video and a sentence such as "the person in red picks up the bag", the model must identify the event interval and track the referred target throughout it. Combining temporal localization, referring expression comprehension, tracking, and video-language reasoning, STVG benchmarks language-controllable video understanding.

STVG provides a more complete grounding interface than temporal localization, frame-level referring localization, or tracking alone. Temporal-only methods cannot identify where the described target is, while tracking methods do not solve language-conditioned search by themselves. STVG combines these requirements into a single task, making it useful for video retrieval, interactive annotation, human activity understanding, robot perception, and video agents. Existing systems make progress by coupling proposal generation, temporal boundary prediction, detection, tracking, cross-modal fusion, and target-aware reasoning (Zhang et al., 2020a; Gu et al., 2025a), but many still rely on dense frame processing or specialized modules whose cost grows quickly with video length.

The difficulty becomes more pronounced in long videos. A short query may describe a brief action inside long background context, and the referred object may remain visible before and after the event with different relevance. Dense sampling preserves boundary cues but increases visual tokens and tracking states; sparse sampling is cheaper but may miss decisive transitions. Even within the correct segment, appearance change, occlusion, fast motion, and distractors can cause tracking errors.

Recent MLLMs offer stronger language understanding, broader visual knowledge, and structured intermediate reasoning. Complex STVG queries require parsing attributes, relations, actions, and event order to decide when and where the target appears. MLLMs can connect these linguistic constraints to visual evidence and express decisions in a flexible format.

However, there is still a clear gap between this promise and a robust STVG system. Direct frame-level MLLM grounding requires long visual contexts, fine temporal reasoning, and precise text-generated

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

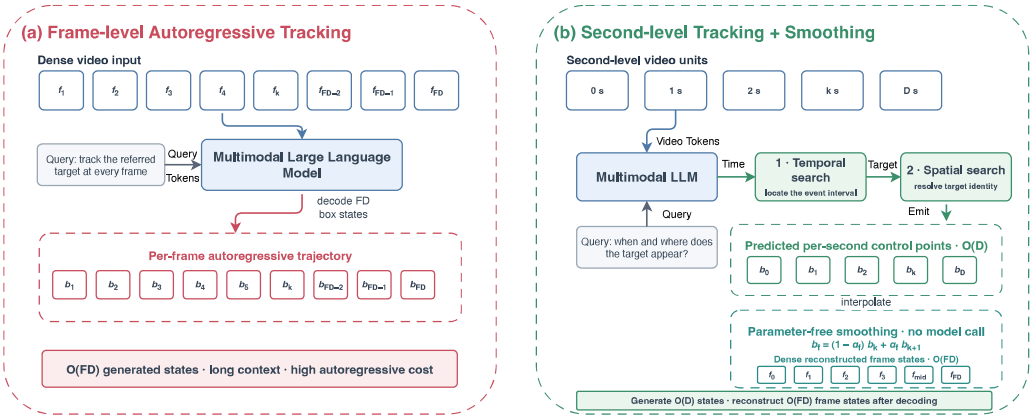

Figure 1: Comparison between our method and traditional VLMs. Left: existing VLMs performing per-frame tracking lead to token explosion. Right: our proposed per-second tracking mode with a post-smoothing strategy and the spatio-temporal reasoning paradigm.

trajectories. One workaround is to input frames iteratively and track step by step (Fu et al., 2025), but this turns long-video grounding into many sequential MLLM calls and makes the wall-clock cost too high. MLLM-based STVG therefore needs a formulation that preserves semantic reasoning while controlling sequence length and grounding noise.

Our work focuses on this practical missing piece: how should an MLLM-based STVG system handle long videos while producing usable temporal and spatial grounding? We argue that the basic reasoning unit should be the second rather than the frame. Many STVG queries describe actions and interactions that persist across short temporal neighborhoods; second-level units therefore shorten the sequence, preserve semantic event structure, and yield compact trajectories. A lightweight inter-second smoothing step restores continuity and reduces jitter.

The second-level formulation also changes the training problem. We use three stages: CPT for broad spatio-temporal grounding, supervised fine-tuning for structured target and event reasoning, and reinforcement learning for metric-aligned localization. During SFT, generated traces expose temporal and identity decisions, but final coordinates and segments are replaced with annotations; RL then scores complete predictions with temporal and motion-aware spatial overlap, aligning optimization with evaluation more directly than token likelihood alone.

Taken together, our formulation couples second-level tracking, ground-truth-corrected reasoning supervision, and verifier-driven RL. These components address long visual inputs, drift-prone traces, text-token coordinates, and metric-misaligned supervision. We evaluate on VidSTG and HC-STVG, measure latency, KV-cache size, tokens, and prefill compute against frame-level alternatives, and test whether grounding alignment preserves general video understanding.

Our contributions are summarized as follows:

- We formulate MLLM-based STVG around second-level tracking, reducing long-video sequence cost while maintaining track continuity through inter-second smoothing.
- We develop a three-stage training pipeline that combines CPT, corrected reasoning SFT, and verifier-driven RL for spatio-temporal localization.
- We construct reasoning supervision with generated decision traces and ground-truth coordinate replacement, separating useful semantic reasoning from unreliable generated geometry.
- We provide direct efficiency measurements across video durations, controlled output-granularity and reward ablations, and general-video evaluation on Video-MME and MMVU.

2 Related Work

Video grounding and spatio-temporal localization. Spatio-temporal video grounding (STVG) localizes both the temporal interval and frame-level trajectory of a target described by language. Early non-MLLM studies approached this through language-guided video retrieval, object referring, and sentence-guided video segmentation Yamaguchi et al. (2017); Vasudevan et al. (2018); Gavrilyuk et al. (2018). Later conventional methods rely on task-specific graph reasoning, two-stage localization, visual-linguistic Transformers, DETR-style tube prediction, consistency modeling, context-guided decoding, and open-vocabulary grounding Tan et al. (2022); Yang et al. (2022); Jin et al. (2022); Lin et al. (2022); Gu et al. (2024); Wasim et al. (2024). These systems provide strong task-specific baselines, but they remain specialized cross-modal matching or tube regression pipelines with limited high-level reasoning and open-world knowledge.

Grounding in Multimodal Large Language Models. Recent MLLMs show promising grounding ability, but most work studies either spatial grounding in images or temporal grounding in videos. Spatial grounding models such as Qwen-VL (Bai et al., 2023) connect language to regions through location tokens, natural-language coordinates, image-caption-box alignment, referring mechanisms, or masks, while video MLLMs such as TimeChat (Ren et al., 2024) and query-enrichment methods (Pramanick et al., 2025) localize relevant moments from language. Later MLLMs (Team et al., 2026c; Bai et al., 2025; Team et al., 2026b) increasingly treat spatial and temporal grounding as core capabilities, yet they are still often evaluated separately (Cheng et al., 2025; Meng et al., 2026b); methods that jointly predict temporal boundaries and spatial trajectories remain less common (Team et al., 2026d; Li et al., 2025b; Clark et al., 2026). Recent systems such as LLaVA-ST (Li et al., 2025a) and SpaceVLLM (Wang et al., 2025) explore fine-grained spatial-temporal outputs, and other work studies zero-shot STVG or reinforcement fine-tuning (Fu et al., 2025; Meng et al., 2026a; Yang et al., 2025b; Gu et al., 2025b). However, many approaches depend on architecture changes (Chamiti et al., 2025; Yang et al., 2026; Wang et al., 2024) or external-tool pipelines (Gao et al., 2025; Wu et al., 2026; Munasinghe et al., 2025; Wu et al., 2026), limiting generality (Yang et al., 2025a). We study joint spatial-temporal localization and RL with the existing architecture and inference pipeline.

3 Method

3.1 Problem Definition

We study spatio-temporal video grounding (STVG) under a language-conditioned setting. Given an input video $V = \{f_1, \dots, f_T\}$ with frame rate F and a natural-language query q, the model needs to output both the temporal interval in which the described event occurs and the spatial location of the referred target throughout that interval. The ground-truth annotation can be written as

$$y^* = (t_s^*, t_e^*, \mathcal{B}^*), \qquad \mathcal{B}^* = \{b_k^*\}_{k=t_s^*}^{t_e^*}, \tag{1}$$

where $[t_s^*, t_e^*]$ denotes the target temporal span and b_k^* is the target box at time index k. In frame-level STVG, k usually corresponds to a frame. This is precise, but the sequence length grows linearly with the number of frames and quickly becomes impractical for MLLMs on long videos.

We instead convert the video into second-level units $\mathcal{S} = \{S_1, \dots, S_K\}$, where $K = \lceil T/F \rceil$ and S_k contains the frames falling into the k-th second. The model predicts

$$\hat{y} = (\hat{t}_s, \hat{t}_e, \hat{\mathcal{B}}), \qquad \hat{\mathcal{B}} = \{\hat{b}_k\}_{k=\hat{t}_s}^{\hat{t}_e}. \tag{2}$$

Here $[\hat{t}_s, \hat{t}_e]$ is a second-level temporal span and $\hat{b}_k$ is the predicted target state for second k. When original annotations are available at frame granularity, we convert them to second-level supervision by aggregating visible target boxes within each second. This representation keeps the event structure needed by STVG while reducing the number of visual units from T frames to roughly T/F seconds. It also matches the fact that many language queries describe actions, interactions, and state changes that persist across short temporal neighborhoods rather than isolated frames.

3.2 Why Direct MLLM Grounding is Difficult

MLLMs are attractive for STVG because they can parse complex referring expressions and reason over event context, but direct STVG asks them to solve continuous localization through long discrete

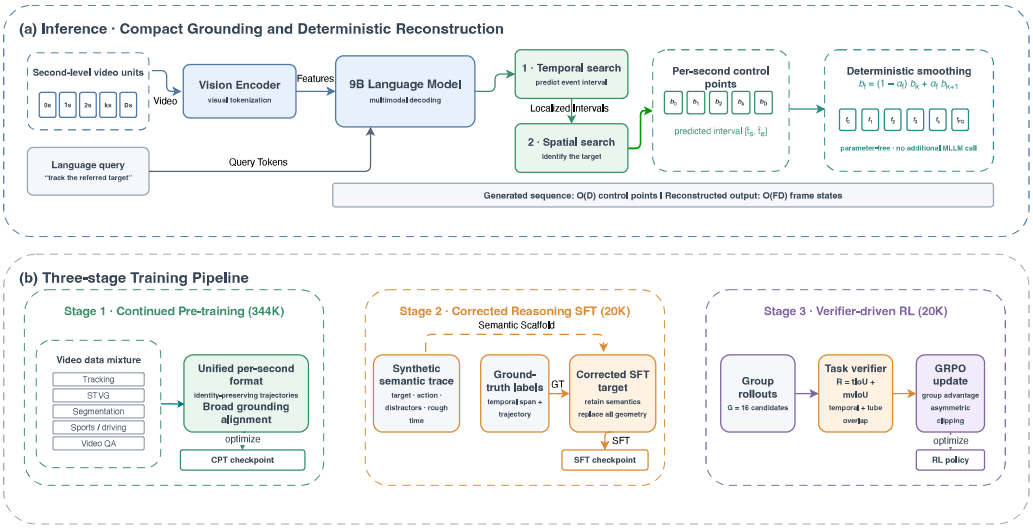

Figure 2: Model architecture and three-stage training. We train a GLM-V-style multimodal architecture to output per-second tracking results. The reasoning adopts a “time-then-space" logical CoT. Training includes CPT, logical CoT-focused SFT, and RL for spatio-temporal grounding.

generation. A straightforward design would feed many frames to the MLLM and ask it to generate a temporal span together with dense coordinates. This creates several mismatches between what STVG requires and what MLLMs are naturally optimized to produce.

Frame-level visual input quickly creates an expensive context problem: dense sampling preserves boundary cues but increases visual-token cost, while sparse sampling may remove the evidence that distinguishes the queried event from surrounding context. Direct grounding also lengthens the output, so generated reasoning and coordinate sequences can drift across similar frames, visible distractors, nearby temporal units, or language priors that are insufficiently grounded in visual evidence (Chen et al., 2024). Finally, coordinates are a brittle text-generation interface for dense geometry: small numerical errors can produce jitter, and independently generated boxes do not guarantee stable identity. SFT optimizes token likelihood, so we use compact second-level outputs, corrected geometry, and RL rewards for grounding overlap and trajectory consistency.

3.3 Three-Stage Training Pipeline

Our three-stage pipeline addresses these failure modes using a GLM-V-style architecture. CPT connects language, temporal units, and object locations through broad grounding supervision. SFT teaches reasoning about the queried target and approximate temporal window. RL optimizes numerical grounding quality with task-level rewards, aligning the final output with STVG metrics.

3.4 CPT: Broad Spatio-Temporal Grounding

Data construction. The CPT stage uses approximately 344K video samples spanning general single- and multi-object tracking, sports, traffic and driving, aerial and drone footage, animal tracking, human activity, spatio-temporal grounding, video segmentation, and general video understanding. The corpus combines TAO, GOT-10k, LaSOT, TC128, OTB-2015, MOSEv2, BFT, TLP, TNL2K, DanceTrack, SportsMOT, TeamTrack, SoccerNet, ReferKITTI-v2, BDD, UA-DETRAC, VisDrone, UAV123, SeaDroneSee, AnimalTrack, HiEve, VidSTG, HC-STVG, YouTube-VOS, VideoEspresso, and VideoChat2. All tracking annotations are converted to our unified per-second format. The mixture exposes the model to instance discrimination, identity preservation, distractors, diverse motion patterns, language-conditioned grounding, and general video reasoning.

Most videos in CPT are sampled at 1 FPS, which matches our second-level formulation and keeps the visual context affordable. We also mix in a small portion of higher-FPS videos to preserve distributional coverage and ensure that the model still has sufficient processing ability when evaluation videos contain denser temporal evidence. This mixture limits specialization to 1 FPS.

All grounding-related data sources are converted into a unified second-level format. For tracking data, each training instance is formed by pairing a target description or template-derived prompt with a second-level trajectory. For segmentation data, prompts are constructed around object categories or referred regions, and the target mask or box is assigned to the corresponding second-level unit. For STVG data, the temporal span and trajectory are already aligned with language queries and are converted directly. Mixing these sources is important because STVG datasets alone are usually not large enough to teach robust tracking, object identity preservation, and spatial precision across diverse scenes at the CPT stage. We further add general video captioning and video question-answering data so that CPT does not collapse into a narrow coordinate-prediction task. These samples preserve conversational and video understanding abilities for grounding open-ended queries.

3.5 SFT: Reasoning Without Coordinate Drift

Training data. The SFT stage uses 20K examples sampled from the VidSTG and HC-STVG training sets. For each example, we use Gemini to generate a chain-of-thought style decision trace. The trace focuses on two questions: what object or person should be tracked, and what rough temporal window contains the described event. Before training, we manually spot-check generated traces for correct target identification, relevant temporal evidence, and logical consistency. This supervision is intended to teach the model to make its grounding decision through explicit visual-language reasoning rather than by directly emitting coordinates from shallow pattern matching.

Reasoning format. We deliberately do not ask the model to output spatial coordinates inside the reasoning chain. This design keeps the trace focused on semantic decisions rather than dense geometry: the model describes the target, discriminative attributes, relevant actions, and approximate temporal evidence, while the final structured answer compactly provides the temporal span and trajectory supervised by ground truth. We also encourage the trace to spend more reasoning on locally ambiguous moments, such as adjacent temporal units with distractor objects or similar actions, instead of repeatedly verbalizing routine frame-level observations.

Coordinate correction. Although Gemini-generated traces are useful for semantic reasoning, we do not trust generated geometry as supervision. For any final structured output, temporal spans and boxes are replaced with ground-truth annotations. This produces a corrected sequence τ that preserves high-level reasoning while ensuring that all numerical labels are geometrically valid. The SFT loss is the standard next-token objective:

$$\mathcal{L}_{\text{sft}} = -\mathbb{E}_{(\mathcal{S},q,\tau)} \sum_{u} \log \pi_\theta(\tau_u \mid \tau_{<u}, \mathcal{S}, q). \tag{3}$$

This stage teaches the expected response structure and reasoning behavior for STVG, while leaving metric-level localization quality to the RL stage.

3.6 RL: Direct Optimization of Grounding Quality

Training data. The grounding RL mixture contains approximately 20K video-object-tracking samples. Following the rollout-based filtering used for data construction, we select challenging CPT examples for which the model does not solve every sampled rollout, thereby retaining prompts that provide informative reward variation during RL.

Reward design. For each prompt, the policy samples a group of candidate outputs. We parse each output into a temporal span and second-level trajectory, then score it with a verifier:

$$R = t_{\text{IoU}} + m v_{\text{IoU}}. \tag{4}$$

The temporal reward t_{IoU} measures overlap between the predicted and ground-truth time windows:

$$t_{\text{IoU}} = \frac{|[\hat{t}_s, \hat{t}_e] \cap [t_s^*, t_e^*]|}{|[\hat{t}_s, \hat{t}_e] \cup [t_s^*, t_e^*]|}. \tag{5}$$

This term directly supervises event localization before spatial trajectory evaluation.

The motion-aware spatial reward mv_{IoU} evaluates whether the predicted trajectory follows the correct target during the relevant time span. We compute spatial overlap over valid second-level states and aggregate it across the matched temporal region:

$$mv_{\mathrm{IoU}} = \frac{1}{|\Omega|} \sum_{k \in \Omega} \mathrm{IoU}(\hat{b}_k, b_k^*), \qquad \Omega = [\hat{t}_s, \hat{t}_e] \cap [t_s^*, t_e^*]. \tag{6}$$

If the temporal intersection is empty, the spatial reward is set to zero. This term encourages accurate localization, stable identity tracking, and motion-aware alignment after the model has found the correct event region. Together, t_{IoU} and mv_{IoU} score the two coupled parts of STVG without collapsing temporal and spatial errors into a single signal.

GRPO-style optimization. We initialize RL from the SFT model. For each prompt x, we sample a group of responses $\{y_i\}_{i=1}^{G}$ and compute normalized group advantages $\hat{A}_i$ from their verifier rewards. Let

$$r_i(\theta) = \frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}. \tag{7}$$

We optimize the clipped objective

$$\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}\left[\min\Big(r_i(\theta)\hat{A}_i, \mathrm{clip}_{[1-\epsilon_{\text{low}},\, 1+\epsilon_{\text{high}}]}(r_i(\theta))\,\hat{A}_i\Big)\right]. \tag{8}$$

We use asymmetric clipping with $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$. The larger upper range allows beneficial updates to move slightly more when the verifier strongly prefers an output, while the lower range keeps harmful probability decreases conservative. We do not add an explicit KL penalty.

3.7 Inference with Second-Level Tracking and Smoothing

At inference time, the model receives the language query and second-level video units, then outputs a temporal span and a compact trajectory. Rather than treating this trajectory as a set of unrelated detections, we interpret the predicted boxes as *sparse control points* of a continuous target path. The MLLM performs semantic grounding once per second; deterministic trajectory reconstruction requires no additional model calls or learned parameters.

Let the video frame rate be F, let c_k denote the frame index at the beginning of second k, and let $\hat{b}_k = (x_{1,k}, y_{1,k}, x_{2,k}, y_{2,k}) \in [0,1]^4$ be the corresponding valid predicted box. For a frame $f \in [c_k, c_{k+1}]$ between two consecutive valid control points, we set

$$\alpha_f = \frac{f - c_k}{c_{k+1} - c_k}, \qquad \tilde{b}_f = (1-\alpha_f)\hat{b}_k + \alpha_f \hat{b}_{k+1}, \tag{9}$$

where interpolation is applied independently to the four normalized box coordinates. We then clip the coordinates to $[0,1]$ and enforce $x_1 < x_2$ and $y_1 < y_2$. At a trajectory boundary, or when only one neighboring control point is valid, we use the nearest valid box without extrapolating beyond the model-predicted temporal interval. Missing internal control points are bridged only when valid boxes with the same object identity exist on both sides; otherwise the corresponding frames remain invalid. Smoothing thus preserves predicted temporal boundaries and target identity.

For a video of duration D seconds, this representation generates $O(D)$ box states, compared with $O(FD)$ states for a frame-level textual trajectory. The reconstruction itself costs $O(FD)$ elementary coordinate operations but no autoregressive decoding, so the expensive generated sequence remains proportional to video duration in seconds. Section 4.3 compares latency, memory, and token usage against a matched frame-level alternative.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on two standard STVG leaderboards. **VidSTG** (Zhang et al., 2020b) is built from the VidOR (Shang et al., 2019) video relation dataset and asks models to ground multi-form language queries in untrimmed videos. It contains both declarative and interrogative sentence

Table 1: Deterministic inter-second smoothing. The procedure converts second-level control points into a frame-level trajectory when dense output is required. It has no learned parameters.

Step	Operation
1	Parse the predicted interval $[\hat{t}_s, \hat{t}_e]$, object identity, and second-level boxes $\{\hat{b}_k\}$. Remove malformed boxes and boxes outside the interval.
2	For every pair of consecutive valid, identity-consistent control points, reconstruct intermediate frames using Eq. equation 9.
3	Use the nearest valid control point at an interval boundary; do not extrapolate outside $[\hat{t}_s, \hat{t}_e]$.
4	Clip coordinates to the image domain, restore valid corner ordering, and return either the compact control points or the reconstructed frame-level tube required by the evaluator.

forms, making it a useful benchmark for testing whether a model can handle different linguistic expressions of the same grounding problem. **HC-STVG** focuses on human-centric spatio-temporal grounding in complex multi-person scenes. Compared with generic object grounding, HC-STVG stresses target disambiguation, interaction understanding, and identity preservation under crowded motion. We additionally evaluate general video understanding on **Video-MME-v2** (Fu et al., 2026), **Video-MME** (Fu et al., 2024), and **MMVU** (Zhao et al., 2025).

Metrics. Following standard STVG evaluation, we report temporal localization and tube localization metrics. $m_{t\mathrm{IoU}}$ measures the mean temporal IoU between predicted and ground-truth time windows. $m_{v\mathrm{IoU}}$ measures the mean spatio-temporal tube IoU and is the primary indicator of full grounding quality. We also report $v\mathrm{IoU}@R$, which measures the percentage of examples whose tube IoU exceeds threshold R (e.g., 0.3 or 0.5).

Compared models. We compare against five general MLLM baselines: Qwen3.5, Seed2.0-pro, Kimi-k2.5 (Team et al., 2026a), Gemini3-Flash, and Gemini3-Pro. These baselines test grounding within the same broad MLLM interface, which also supports open-ended video question answering and general video understanding. General MLLMs are not necessarily task-naive: for example, Qwen3-VL is trained with large-scale spatio-temporal video-grounding annotations at the object, action, and person levels (Bai et al., 2025). We report both our SFT checkpoint and the final RL checkpoint; a controlled reward ablation additionally compares mv_{IoU} alone with $t_{\mathrm{IoU}} + mv_{\mathrm{IoU}}$.

4.2 Main Results

Table 2 compares general-purpose MLLMs under our evaluation protocol. Our final 9B task-aligned RL model achieves the strongest grounding performance on VidSTG and HC-STVG, outperforming several substantially larger baselines.

The progression from Ours-SFT to RL isolates the effect of metric alignment. SFT adapts the model to the required output format and reasoning behavior, while RL improves the final grounding metrics by optimizing completed predictions with task-level rewards. The combined $t_{\mathrm{IoU}} + mv_{\mathrm{IoU}}$ verifier improves both temporal and motion-aware tube metrics over mv_{IoU} alone because the temporal component provides a denser signal before the model predicts accurate trajectories. For the final RL model, we report the mean over three evaluation runs on VidSTG and HC-STVG.

4.3 Efficiency and Component Ablations

Because efficiency is a central motivation, we compare second-level prediction with frame-level alternatives under the same model and measurement environment. The frame-level variants encode densely sampled frames, whereas our protocol uses one frame and emits one trajectory control point per second, followed by the deterministic reconstruction in Table 1. We report the number of encoded frames, total input tokens, time to first token (TTFT), end-to-end latency, KV-cache size, and prefill compute. Metric definitions are specified in Appendix A.

For the 10-second video, per-second processing reduces input tokens by $24.5\times$, TTFT by $37.5\times$, end-to-end latency by $29.3\times$, KV-cache size by $24.7\times$, and prefill compute by $56.9\times$ relative to 25-FPS frame-level processing. At 40 seconds, it reduces these quantities by $5.0\times$, $6.6\times$, $2.5\times$, $4.8\times$,

Table 2: VidSTG and HC-STVG results. Ours-RL: three-run mean; "–": unavailable.

Benchmark	Model	Params	$m_{t\mathrm{IoU}}\uparrow$	$m_{v\mathrm{IoU}}\uparrow$	$v\mathrm{IoU}@0.5\uparrow$
VidSTG	Qwen3.5	397B	27.88	15.32	15.50
	Seed2.0-pro	undisclosed	26.28	16.66	9.78
	Kimi-k2.5	1T	23.99	16.05	10.46
	Gemini3-Flash	undisclosed	18.76	8.90	4.11
	Gemini3-Pro	undisclosed	20.07	11.57	4.98
	Ours-SFT	9B	27.66	19.47	11.86
	Ours-RL	9B	**31.35**	**23.79**	**19.49**
HC-STVG	Qwen3.5	397B	50.99	28.20	16.18
	Seed2.0-pro	undisclosed	42.17	26.68	9.37
	Kimi-k2.5	1T	43.9	10.99	9.04
	Gemini3-Flash	undisclosed	–	16.64	2.84
	Gemini3-Pro	undisclosed	43.81	20.24	4.82
	Ours-SFT	9B	51.23	33.75	26.18
	Ours-RL	9B	**60.04**	**40.99**	**34.12**

Table 3: Efficiency scaling with video duration and temporal sampling protocol. "Overflow" indicates that the input exceeds the supported context window. TTFT denotes time to first token.

Duration	Input protocol	Frames	Input tokens
10s	Per-second	10	5,227
10s	Per-frame, 25 FPS	250	128,106
40s	Per-second	40	20,587
40s	Per-frame, 5 FPS	200	102,506
60s	Per-second	60	30,827
60s	Per-frame, 5/25 FPS	300/1,500	Overflow

Duration	Input protocol	TTFT (s)↓	End-to-end (s)↓	KV cache (GiB)↓	Prefill (TFLOPs)↓
10s	Per-second	0.74	1.32	0.20	159
10s	Per-frame, 25 FPS	27.74	38.73	4.93	9,049
40s	Per-second	3.07	12.41	0.82	729
40s	Per-frame, 5 FPS	20.29	31.26	3.95	6,381
60s	Per-second	4.53	14.10	1.22	1,195
60s	Per-frame, 5/25 FPS	Overflow			

and $8.8\times$, respectively, relative to the 5-FPS frame-level protocol. At 60 seconds, the per-second protocol remains executable with 30,827 input tokens and 14.10s end-to-end latency, while both 5-FPS and 25-FPS frame-level inputs exceed the context window.

We next isolate output granularity while holding the model and visual evidence fixed. Both rows in Table 4 use the same model, 1-FPS video input, and a matched 55,790 video-token budget on DanceTrack. Per-second JSON output reaches 39.08 HOTA, compared with 39.02 for per-frame JSON output. Thus the compact output protocol preserves tracking quality in this controlled setting while avoiding a dense autoregressive trajectory. The SFT-to-RL comparison in Table 2 and the reward comparison in Figure 4 provide the remaining confirmed stage and reward ablations.

Beyond STVG, Figure 3 evaluates whether grounding-heavy alignment preserves general video understanding. On Video-MME-v2, our model improves over GLM-4.1V-Thinking from 9.4 to 12.6 on non-linear reasoning and from 23.7 to 28.0 on simple accuracy. It also improves the base checkpoint from 66.07 to 68.80 on Video-MME and from 59.4 to 64.3 on MMVU. These consistent gains show that the training pipeline does not narrow the model into a localization-only system.

Figure 4 in Appendix D consolidates the reward ablation and training-FPS sensitivity. On DanceTrack, 1 FPS achieves 39.08 HOTA, compared with 31.4 at 0.5 FPS and 38.5 under dynamic-FPS sampling.

Table 4: Output-granularity ablation on DanceTrack with the same model, 1-FPS visual input, and matched train/inference video-token budget.

Output protocol	Train video tokens	Inference video tokens	HOTA↑
JSON per-frame	55,790	55,790	39.02
JSON per-second	55,790	55,790	**39.08**

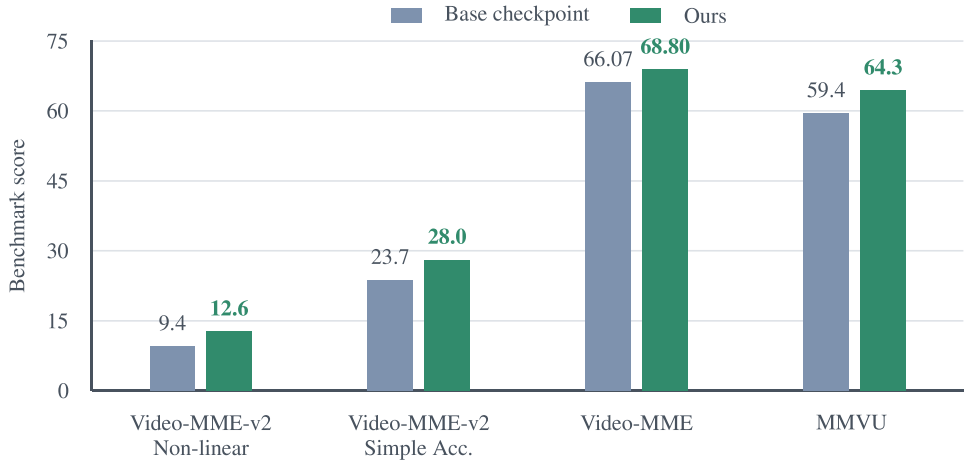

Figure 3: General video understanding on Video-MME-v2, Video-MME, and MMVU. Each group reports its benchmark's official score, which cannot be compared directly across benchmarks.

5 Conclusion

We present a practical MLLM framework for STVG. The method uses second-level tracking to reduce long context, corrected reasoning SFT to teach target and temporal reasoning, and verifier-driven RL to directly optimize numerical grounding quality. On VidSTG and HC-STVG, the 9B task-aligned model outperforms the evaluated general MLLM baselines, while direct efficiency measurements show substantial reductions in token count, latency, KV-cache size, and prefill compute. Improvements over the base checkpoint on Video-MME and MMVU further show that localization-oriented alignment can preserve and strengthen general video understanding.

6 Limitations

Second-level control points assume that target motion can be reconstructed from sparse observations. Linear interpolation may be inaccurate for abrupt motion, heavy occlusion, shot changes, or sub-second events. The method trades fine temporal precision for shorter autoregressive output. Our efficiency and output-granularity comparisons quantify this trade-off but do not establish that interpolation is optimal in every motion regime.

The three-stage pipeline, especially group-based RL sampling with a 64K context, adds substantial training compute. RL uses one node with eight H100 GPUs; Appendix A reports optimization and parallelism settings to distinguish training cost from inference latency and memory. Generated SFT traces may inherit a proprietary model's stylistic and domain biases. Manual inspection and ground-truth coordinate replacement reduce quality and geometry errors but do not establish invariance to the teacher or reasoning style.

VidSTG and HC-STVG do not cover all open-world conditions. Video-MME and MMVU extend evaluation to general video understanding; future work should address real-time streams, very long videos, rare object categories, and rapidly changing multi-object identities. These settings require further evaluation of temporal precision and long-term identity consistency.

AI USE STATEMENT

We used AI tools to assist with writing, language editing, and LaTeX formatting, including figure labels and table presentation. Gemini also generated the SFT reasoning traces described in Section 3.5; temporal boundaries and coordinates were replaced with ground-truth annotations. The authors take responsibility for the final manuscript, including its scientific claims and reported results.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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

Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. Refergpt: Towards zero-shot referring multi-object tracking, 2025. URL https://arxiv.org/abs/2504.09195.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models, 2024. URL https://arxiv.org/abs/2411.15268.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning, 2025. URL https://arxiv.org/abs/2503.11495.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Teng Fu, Mengyang Zhao, Ke Niu, Kaixin Peng, and Bin Li. Omnipt: Unleashing the potential of large vision language models for pedestrian tracking and understanding, 2025. URL https://arxiv.org/abs/2511.17053.

Shida Gao, Feng Xue, Xiangfeng Wang, Anlong Ming, Teng Long, Yihua Shao, Haozhe Wang, Zhaowen Lin, Wei Wang, and Nicu Sebe. 1 + 1 > 2: Detector-empowered video large language model for spatio-temporal grounding and reasoning, 2025. URL https://arxiv.org/abs/2512.06673.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence, 2018. URL https://arxiv.org/abs/1803.07485.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding, 2024. URL https://arxiv.org/abs/2401.01578.

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, 2025a. URL https://arxiv.org/abs/2502.11168.

Xin Gu, Haoji Zhang, Qihang Fan, Jingxuan Niu, Zhipeng Zhang, Libo Zhang, Guang Chen, Fan Chen, Longyin Wen, and Sijie Zhu. Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning, 2025b. URL https://arxiv.org/abs/2511.21375.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding, 2022. URL https://arxiv.org/abs/2209.13306.

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, 2025a. URL https://arxiv.org/abs/2501.08282.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning, 2025b. URL https://arxiv.org/abs/2504.06958.

Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Stvgformer: Spatio-temporal video grounding with static-dynamic cross-modal understanding, 2022. URL https://arxiv.org/abs/2207.02756.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence, 2026a. URL https://arxiv.org/abs/2510.20579.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification, 2026b. URL https://arxiv.org/abs/2604.01569.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos, 2025. URL https://arxiv.org/abs/2411.04923.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms, 2025. URL https://arxiv.org/abs/2510.17023.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024. URL https://arxiv.org/abs/2312.02051.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 279–287. ACM, 2019.

Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1533–1542, October 2021.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion, 2022. URL https://arxiv.org/abs/2111.14690.

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, 2022. URL https://arxiv.org/abs/2106.10634.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihan Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, Lihang Pan, Ke Ning, Junhui Ji, Jinjiang Wang, Jing Chen, Jiazheng Xu, Jiale Zhu, Jiale Cheng, Ji Qi, Guobing Gan, Guo Wang, Cong Yao, Zijun Dou, Zihao Zhou, Zihan Wang, Zhiqi Ge, Zhijie Li, Zhenyu Hou, Zhao Xue, Zehui Wang, Zehan Qi, Zehai He, Yutao Zhang, Yusen Liu, Yukuo Cen, Yuchen Li, Yuan Wang, Yu Yang, Yongbin Liu, Yijian Lu, Yifan Xu, Yanzi Wang, Yanxiao Zhao, Yanfeng Wang, Yadong Xue, Yabo Xu, Xinyu Zhang, Xinyu Liu, Xiao Liu, Wenyi Zhao, Wenkai Li, Tianyu Tong, Tianshu Zhang, Shudan Zhang, Shengdong Yan, Qinkai Zheng, Mingde Xu, Licheng Bao, lat Long long, Jiaxing Xu, Jiaxin Fan, Jiawen Qian, Jiali Chen, Jiahui Lin, Jiadai Sun, Haozhi Zheng, Haoran Wang, Haochen Li, Hanyu Liu, Han Xu, Fan Yang, Dan Zhang, Da Yin, Chuangxin Zhao, Chengcheng Wu, Boyan Shi, Bowen Lv, Bowei Jia, Bo Li, Bin Chen, Baoxu Wang, Peng Zhang, Debing Liu, Bin Xu,

Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-5v-turbo: Toward a native foundation model for multimodal agents, 2026b. URL https://arxiv.org/abs/2604.26752.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026c. URL https://arxiv.org/abs/2507.01006.

Vidi Team, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Fanding Lei, Feng Gao, Guang Chen, Haoji Zhang, Haojun Zhao, Jin Liu, Jingjing Zhuge, Lili Fang, Lingxi Zhang, Longyin Wen, Lu Guo, Lu Xu, Lusha Li, Qihang Fan, Rachel Deng, Shaobo Fang, Shu Zhang, Sijie Zhu, Stuart Siew, Weiyan Tao, Wen Zhong, Xiaohui Shen, Xin Gu, Ye Yuan, Yicheng He, Yiming Cui, Zhenfang Chen, Zhihua Wu, and Zuhua Lin. Vidi2.5: Large multimodal models for video understanding and creation, 2026d. URL https://arxiv.org/abs/2511.19529.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze, 2018. URL https://arxiv.org/abs/1801.01582.

Han Wang, Yanjie Wang, Yongjie Ye, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm, 2024. URL https://arxiv.org/abs/2403.16558.

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, 2025. URL https://arxiv.org/abs/2503.13983.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Video-groundingdino: Towards open-vocabulary spatio-temporal video grounding, 2024. URL https://arxiv.org/abs/2401.00901.

Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. Mass: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models, 2026. URL https://arxiv.org/abs/2511.18373.

Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries, 2017. URL https://arxiv.org/abs/1704.07945.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers, 2022. URL https://arxiv.org/abs/2203.16434.

Fan Yang, Shurong Zheng, Hongyin Zhao, Yufei Zhan, Xin Li, Yousong Zhu, Chaoyang Zhao Ming Tang, and Jinqiao Wang. Tracevision: Trajectory-aware vision-language model for human-like spatial understanding, 2026. URL https://arxiv.org/abs/2602.19768.

Yi Yang, Yiming Xu, Timo Kaiser, Hao Cheng, Bodo Rosenhahn, and Michael Ying Yang. Multi-object tracking retrieval with llava-video: A training-free solution to mot25-stag challenge, 2025a. URL https://arxiv.org/abs/2511.03332.

Zaiquan Yang, Yuhao Liu, Gerhard Hancke, and Rynson W. H. Lau. Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding, 2025b. URL https://arxiv.org/abs/2509.15178.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Nicholas Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding, 2020a. URL https://arxiv.org/abs/2008.06941.

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, 2020b. URL https://arxiv.org/abs/2001.06891.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380.