

OPEN-O3 VIDEO: GROUNDED VIDEO REASONING WITH EXPLICIT SPATIO-TEMPORAL EVIDENCE

Anonymous authors

Paper under double-blind review

ABSTRACT

Most video reasoning models only generate textual reasoning traces without indicating when and where key evidence appears. Recent models such as OpenAI-o3 have sparked wide interest in evidence-centered reasoning for images, yet extending this ability to videos is more challenging, as it requires joint temporal tracking and spatial localization across dynamic scenes. We introduce **Open-o3 Video**, a non-agent framework that integrates explicit spatio-temporal evidence into video reasoning, and carefully collect training data and design training strategies to address the aforementioned challenges. The model highlights key timestamps, objects, and bounding boxes alongside its answers, allowing reasoning to be grounded in concrete visual observations. To enable this functionality, we first curate and build two high-quality datasets, **STGR-CoT-30k** for SFT and **STGR-RL-36k** for RL, with carefully constructed temporal and spatial annotations, since most existing datasets offer either temporal spans for videos or spatial boxes on images, lacking unified spatio-temporal supervision and reasoning traces. Then, we adopt a cold-start reinforcement learning strategy with multiple specially designed rewards that jointly encourage answer accuracy, temporal alignment, and spatial precision. On V-STAR benchmark, **Open-o3 Video** achieves state-of-the-art performance, raising mAM by 14.4% and mLGM by 24.2% on the Qwen2.5-VL baseline. Consistent improvements are also observed on a broad range of video understanding benchmarks, such as VideoMME, WorldSense, VideoMMU, **LongVideo-Reason-eval** and TVGBench. Beyond accuracy, the reasoning traces produced by Open-o3 Video also provide valuable signals for test-time scaling, enabling confidence-aware verification and improving answer reliability. The code and datasets will be made publicly available.

1 INTRODUCTION

Understanding complex video content is a long-standing goal for large multimodal models (Wang et al., 2025b; Team et al., 2025; Chen et al., 2024a; Zhang et al., 2024a; Ye et al., 2024; Zhang et al., 2023; 2024b; Wang et al., 2024), as videos encapsulate rich temporal dynamics and spatial interactions that far exceed the information in static images. While recent progress has advanced performance on tasks like action recognition and video question answering (Bai et al., 2025; Zhu et al., 2025; Zhang et al., 2024b; Li et al., 2024; Zhang et al., 2025a), building models that can perform reliable, fine-grained reasoning over long and cluttered scenes remains challenging.

Recent “thinking with images” attempts (OpenAI, 2025; Wang et al., 2025c;a; Zheng et al., 2025b) leverage explicit operations (such as cropping, zoom-in, and region selection) to interleave detailed *visual evidence* with language, achieving superior performance on fine-grained image comprehension. This success motivates extending a similar paradigm to the video domain.

However, this extension is difficult and non-trivial due to the requirement for *coherent localization across both time and space* precisely. The complexity of dynamic scenes, *e.g.*, replete with motion, occlusions, and camera changes, makes it incredibly challenging to pinpoint when and where events of interest occur. As a result, previous attempts to incorporate explicit reasoning in video have often been limited to *textual rationales* (Feng et al., 2025a; Wang et al., 2025d) or, coarse, *temporal-only* grounding (Li et al., 2025b; Wang et al., 2025f), failing to achieve the fine-grained spatio-temporal precision necessary for complex video reasoning. This gap is largely due to two interconnected

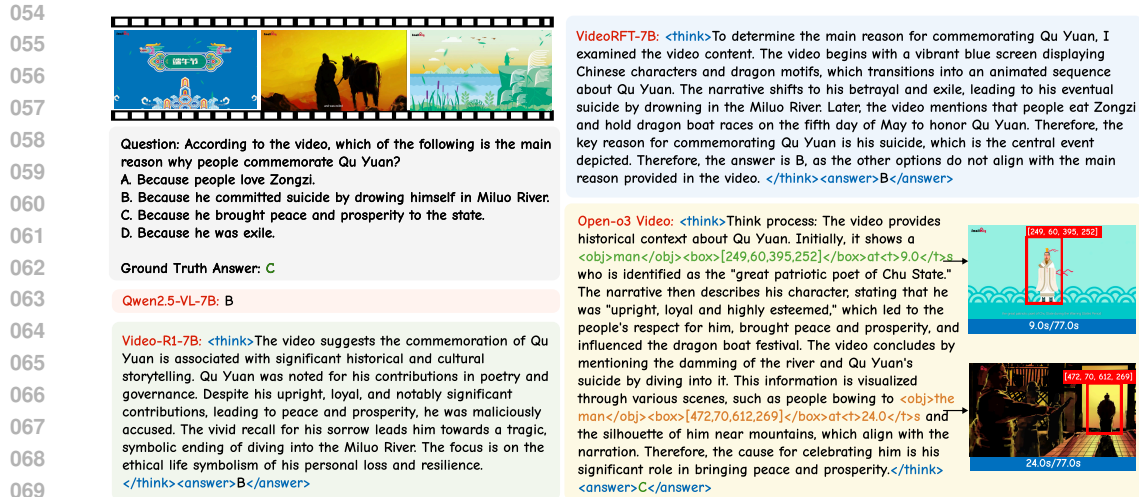


Figure 1: While prior video reasoning models (e.g., Video-R1 (Feng et al., 2025a), VideoRFT (Wang et al., 2025d)) only generate textual rationales, **Open-o3 Video** integrates explicit spatio-temporal grounding into the reasoning process. The model highlights key timestamps and object regions that directly support the answer, providing verifiable evidence for its prediction. More visualizations are provided in Appendix A.8.

obstacles: (1) the absence of *high-quality datasets* that provide joint spatio-temporal supervision for reasoning, and (2) the inherent difficulty of training a model to precisely localize objects in *time and space* simultaneously.

To address these challenges, we introduce **Open-o3 Video**, a framework that embeds *joint* spatio-temporal evidence directly into the reasoning process. Our first key contribution is the creation of a comprehensive training corpus designed to bridge this data gap. **We have curated two datasets, STGR-CoT-30k and STGR-RL-36k, for supervised fine-tuning and reinforcement learning, respectively.** These datasets integrate existing temporal-only and spatial-only grounding resources with *5.9k newly annotated high-quality spatio-temporal samples*. Each instance contains a question-answer pair, timestamped key frames, localized bounding boxes, and *a chain of thought that explicitly links the visual evidence to the reasoning steps*.

Building on this dataset, our second major contribution is a two-stage training strategy with **adaptive temporal proximity** and **temporal gating** to stably and efficiently optimize the model’s spatio-temporal reasoning capability. Although the model has acquired preliminary capabilities for generating structured, grounded chains of thought during the supervised fine-tuning stage, the subsequent reinforcement learning stage still cannot achieve stable training due to a critical *spatial collapse* issue. This is because spatial grounding rewards are usually conditioned on correctly identifying the timestamp. When temporal predictions are imprecise in the early stages, this leads to *near-zero spatial rewards*, stalling the learning process for localization. Therefore, we propose a novel *adaptive temporal proximity* technique, which relaxes the temporal requirement in early training to reduce reward sparsity, and gradually increases the precision demand over training time. This prevents premature saturation of the temporal reward and ensures that predicted timestamps keep approaching the ground truth, which is crucial for reliable spatial evaluation. In parallel, a complementary *temporal gating* mechanism computes spatial rewards only when temporal predictions are sufficiently accurate, preventing irrelevant objects from being rewarded and enforcing precise spatio-temporal alignment. Together, these mechanisms provide dense yet reliable feedback, forming a smoother learning curriculum that progressively strengthens both temporal accuracy and spatial grounding.

Through this combination of curated data and our training procedure, as shown in Figure 1, Open-o3 Video produces reasoning that is accurate, interpretable, and grounded in the visual evidence. We evaluate Open-o3 Video on the V-STAR benchmark and other video understanding tasks. On **V-STAR**, our model achieves state-of-the-art performance, surpassing GPT-4o and improving over Qwen2.5-VL by **+14.4%** mAM and **+24.2%** mLGM with a small amount of training data. Beyond V-STAR, Open-o3 Video also delivers consistent gains on VideoMME, WorldSense, VideoM-

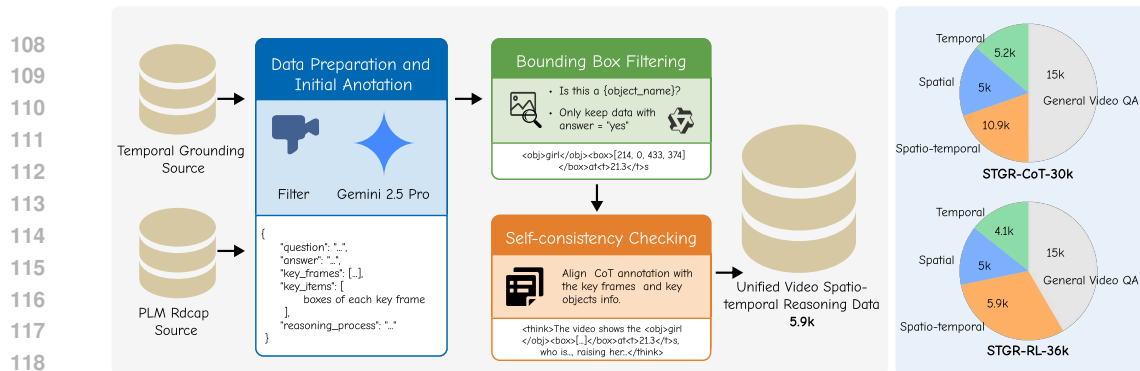


Figure 2: Overview of our data construction pipeline and dataset composition. **Left:** The annotation pipeline includes Gemini 2.5 Pro initial annotation, bounding box filtering, and self-consistency checking. **Right:** Distribution of data categories in STGR-CoT-30k (SFT) and STGR-RL-36k (RL), showing a balanced coverage across temporal, spatial, spatio-temporal, and general QA.

MMU, [LongVideo-Reason-eval](#), and TVGBench, demonstrating advantages in long-video reasoning, perception-oriented tasks, and fine-grained temporal localization. In addition, the explicit evidence traces support evidence-aware test-time scaling, where confidence-aware voting surpasses majority voting (e.g., +1.2% on WorldSense and 1.0% on VideoMMU), demonstrating that grounded evidence provides a reliable self-verification signal to improve inference accuracy.

2 RELATED WORKS

Video Reasoning. Recent advances in video reasoning ([Feng et al., 2025b](#); [Li et al., 2025b](#); [Wang et al., 2025g;d](#); [Zhang et al., 2025c](#); [Xie et al., 2025](#); [Chen et al., 2025b](#); [Zhang et al., 2025b](#); [Park et al., 2025](#); [Dang et al., 2025](#)) have largely been driven by reinforcement learning based post-training, which encourages models to move beyond direct question answering and exhibit step-by-step reasoning. Video-R1 ([Feng et al., 2025b](#)) shows that temporal-aware GRPO with curated reasoning data improves video understanding benchmarks, while VideoChat-R1 ([Li et al., 2025b](#)) extends to spatio-temporal perception tasks such as grounding and tracking without harming QA. Other variants, including Video-RTS ([Wang et al., 2025g](#)) and DeepVideo-R1 ([Park et al., 2025](#)), combine reinforcement learning with test-time scaling or difficulty-aware regularization to better exploit temporal information. These works demonstrate the potential of reinforcement-driven video reasoning, but still rely on text-only outputs without explicitly linking answers to visual evidence. In contrast, our approach generates spatio-temporal grounded evidence (timestamped frames and localized objects), enhancing perception, transparency, and verifiability.

Temporal and Spatial Grounding in Video. The problem of locating when and where relevant evidence appears in a video has attracted growing attention, leading to substantial progress in both temporal and spatial grounding ([Wang et al., 2025f;e](#); [Chen et al., 2025a](#); [Guo et al., 2024](#); [Ouyang et al., 2025](#); [Li et al., 2025c;b](#)). On the temporal side, Time-R1 ([Wang et al., 2025f](#)) introduces verifiable rewards for temporal grounding with strong generalization under limited supervision, while TVG-R1 ([Chen et al., 2025a](#)) improves robustness with curated cold-start and RL datasets. On the spatial side, SpaceR ([Ouyang et al., 2025](#)) leverages RL and a large corpus for object-centric grounding and geometric reasoning. Beyond these two sides, a number of approaches explore spatio-temporal localization. STCAT ([Jin et al., 2022](#)) and LRR ([Bhattacharyya et al., 2023](#)) improve the spatio-temporal grounding ability through model architectural optimization. EgoMask ([Liang et al., 2025](#)) enhances spatio-temporal localization for egocentric videos by fine-tuning models like Sa2Va ([Yuan et al., 2025](#)). LLaVA-ST ([Li et al., 2025a](#)) unify temporal and spatial grounding through positional embedding alignment and two-stream feature compression. However, these approaches do not combine such spatio-temporal localization ability with chain-of-thought reasoning. Aligning both timestamps and object regions within reasoning text, and further leveraging such grounded evidence to enhance video question answering, remain challenging. Our approach tackles both by explicitly linking boxes with temporal positions and integrating spatio-temporal evidence into reasoning, thereby strengthening perception and ensuring verifiability.

Thinking with Images. A growing line of research (OpenAI, 2025; Zheng et al., 2025b; Wang et al., 2025a;c; Fan et al., 2025) explores how multi-modal models improve reasoning by performing explicit visual operations such as cropping, zoom-in, and region selection, thereby producing intermediate evidence that is consumed within the reasoning chain. OpenAI-o3 (OpenAI, 2025) formalizes “thinking with images,” while DeepEyes (Zheng et al., 2025b) shows end-to-end RL can incentivize image–tool reasoning, and TreeBench (Wang et al., 2025a) provides methodology for traceable, box-level evidence. These advances demonstrate the promise of evidence-centric visual reasoning but are largely image-centric. Extending to videos adds challenges in temporal consistency, motion, and fine-grained event alignment. VITAL (Zhang et al., 2025b) adapts the paradigm via an agent-based, tool-augmented RL pipeline, yielding gains but relying on external orchestration. In contrast, our single-model framework “thinks with frames,” directly emitting timestamped crops and bounding boxes as evidence without complex tool pipelines.

3 STGR DATA CONSTRUCTION

3.1 DATA SOURCE AND STATISTICS

Building robust spatio-temporal reasoning models requires training signals that jointly supervise *when* and *where* evidence appears and how it is used in reasoning. Existing resources fall short in three ways: (i) temporal-only grounding datasets provide time spans but lack object regions; (ii) spatial or frame-level caption corpora offer boxes on isolated frames without timestamps; and (iii) most lack a chain of thought that *explicitly* ties objects and timestamps to the answer. These gaps make it impossible to learn coherent localization in dynamic scenes and to compute verifiable rewards for RL, since temporal and spatial supervision are not synchronized and reasoning traces are text-only.

To bridge this gap, we curate two complementary corpora: **STGR-CoT-30k** for supervised fine-tuning (SFT) and **STGR-RL-36k** for reinforcement learning (RL). Both combine existing temporal-only and spatial-only resources *with 5.9k newly annotated, high-quality spatio-temporal samples* produced by our pipeline (Sec. 3.2). Each new instance includes a question–answer pair, timestamped key frames, localized boxes, and a structured chain of thought that links visual evidence to reasoning steps. This design supplies synchronized temporal and spatial supervision for SFT to acquire grounded reasoning formats, and provides reliable, verifiable signals for RL to optimize alignment under complex video dynamics.

The SFT corpus consists of four components: (i) 4.1k temporal grounding CoT samples (TVG-Coldstart) (Chen et al., 2025a), (ii) 5k spatial grounding CoT samples (TreeVGR-SFT) (Wang et al., 2025a), (iii) 5.9k spatio-temporal samples curated by us, including 3.9k from temporal grounding datasets (video source: ActivityNet (Caba Heilbron et al., 2015), COIN (Tang et al., 2019), QueryD (Oncescu et al., 2021), QVHighlight (Lei et al., 2021), DiDeMo (Anne Hendricks et al., 2017)) and 2k from PLM-Rdcap (Cho et al., 2025), and (iv) 15k Video-R1-CoT samples (Feng et al., 2025b). The RL corpus further expands diversity: (i) 5.2k temporal grounding samples, including 2.3k from Time-R1 (Wang et al., 2025f) and 2.9k from TVG-RL (Chen et al., 2025a), (ii) 5k spatial grounding samples from VisCoT (Shao et al., 2024a), (iii) 10.9k spatio-temporal samples, comprising our 5.9k constructed data (via the pipeline) and an additional 5k filtered from VideoE-spresso Han et al. (2025) with consistency checks, and (iv) 15k Video-R1 samples (Feng et al., 2025b).

Overall, as shown in Figure 2 (right), the SFT set covers 13.7% temporal, 16.7% spatial, 19.7% spatio-temporal, and 50.0% general QA data, while the RL set includes 14.4% temporal, 13.9% spatial, 30.3% spatio-temporal, and 41.7% QA data. This design ensures that both phases expose the model to diverse supervisory signals while emphasizing spatio-temporal reasoning as the central capability. More details about the training data are provided in the Appendix A.2.

3.2 DATA ANNOTATION PIPELINE

Spatio-temporal reasoning requires chain-of-thought data that include both temporal spans and spatial grounding. We construct 5.9k such samples by combining temporal grounding datasets with PLM-Rdcap sources (Figure 2, left). The pipeline follows three stages below.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

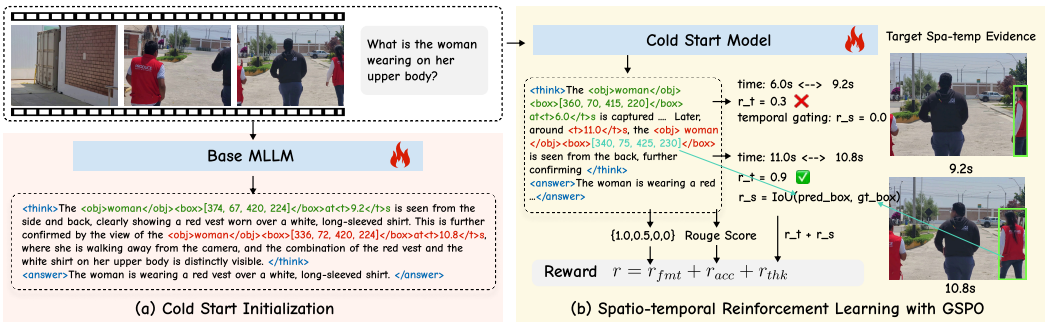


Figure 3: Overview of Open-o3 Video. We adopt a two-stage training paradigm: (a) cold-start initialization to learn structured, grounded outputs; (b) reinforcement learning with a composite reward that sharpens temporal alignment and spatial precision with adaptive temporal proximity and temporal gating.

Data Preparation and Initial Annotation. We begin by collecting two types of sources: temporal grounding datasets and PLM-Rdcap data that provide region-level dense captions. All videos are passed through Gemini 2.5 Pro (Comanici et al., 2025) API with carefully designed prompts (shown in the Appendix A.5) to generate structured annotations. Each annotation contains (i) a question-answer pair centered on a specific object or person, (ii) one to five key frames sampled from the annotated segment, (iii) bounding boxes for one to three salient objects in each key frame, and (iv) a reasoning process that must reference every object with explicit format: `<obj>object_name</obj><box>[x_min, y_min, x_max, y_max]</box>at<t>timestamp</t>s`.

Bounding Box Filtering. Initial annotations may contain noisy or incorrect boxes. We filter them with two rules: (i) boxes covering over 80% of the frame are removed as uninformative; (ii) each crop is verified by Qwen2.5-VL-7B (Bai et al., 2025) with the query “Is this a {object_name}?”. Only samples answered “yes” are kept, ensuring object mentions match validated boxes.

Self-consistency Checking and Quality Control. Our consistency checking aims to align timestamps, bounding boxes, and the spatio-temporal reasoning chain. Since each annotated sample contains timestamps, object names, and a reasoning chain with temporal and spatial references, we first ensure that all boxes and frames mentioned in the reasoning text appear in the “key_frames” and “key_items” annotations; missing elements are removed to keep the annotations complete. We then evaluate the relevance between each visual evidence and the corresponding sentence. We parse the reasoning text, crop the referenced region, and ask Qwen2.5-VL whether the cropped image matches the sentence. If it is judged unrelated, we remove the sample. These consistency checks improve the quality of the reasoning data and support cold-start training for grounded spatio-temporal reasoning.

4 OPEN-O3 VIDEO

As shown in Figure 3, our training recipe involves two stages: a cold-start initialization phase followed by reinforcement learning to enhance spatio-temporal reasoning under carefully designed rewards with adaptive temporal proximity and temporal gating mechanisms.

4.1 COLD START INITIALIZATION

We initialize our model from Qwen2.5-VL-7B (Bai et al., 2025), and further fine-tune it on the constructed STGR-CoT-30k corpus. This stage yields checkpoints that equip the model with basic capabilities in spatio-temporal grounding and structured reasoning output. Such a cold-start stage is essential, as found in the experiment. It reduces reward sparsity, stabilizes optimization, and allows subsequent reinforcement learning to focus on fine-grained temporal and spatial alignment instead of relearning basic reasoning skills.

4.2 REINFORCEMENT LEARNING WITH GSPO

We adopt Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025a) as our reinforcement learning algorithm. Compared with GRPO (Shao et al., 2024b), which operates at the token level, GSPO defines importance ratios and clipping at the sequence level, ensuring that optimization is aligned with sequence-level rewards. This eliminates high-variance token-wise corrections, stabilizes long-horizon training, and avoids collapse in chain-of-thought reasoning. Such stability is particularly important for video reasoning, where responses are longer, rewards combine accuracy, temporal, and spatial terms, and the training dynamics are more difficult to optimize. Our experiments further confirm that GSPO yields higher grounding accuracy and more stable training than GRPO (Section 5.2).

During training, given a video-question pair x , each generated response y is evaluated with a scalar reward $r(x, y)$ that reflects both correctness and reasoning quality. This reward serves as the optimization signal in GSPO, and more details of the GSPO algorithm are provided in Appendix A.6.

4.3 REWARD DESIGN

For each query-completion pair (x, y) , the scalar reward is defined as

$$r(x, y) = r_{\text{acc}}(x, y) + r_{\text{thk}}(x, y) + r_{\text{fmt}}(x, y), \quad (1)$$

which is group-normalized to obtain the advantage used by GSPO. Below we describe the three components.

Accuracy reward r_{acc} . Since the training data span multiple tasks, we design task-specific accuracy rewards. For multiple-choice questions we check exact correctness; for free-form QA we follow previous works and compute ROUGE score; for spatial grounding we use visual IoU; and for temporal grounding we use temporal IoU:

$$r_{\text{acc}}(x, y) = \begin{cases} 1 & \text{if task = MCQ and prediction matches ground truth,} \\ \text{ROUGE}(y^{\text{pred}}, y^{\text{gt}}) & \text{if task = Free-form QA,} \\ \text{vIoU}(\text{Box}^{\text{pred}}, \text{Box}^{\text{gt}}) & \text{if task = Spatial grounding,} \\ \text{tIoU}([s^{\text{pred}}, e^{\text{pred}}], [s^{\text{gt}}, e^{\text{gt}}]) & \text{if task = Temporal grounding.} \end{cases}$$

Thinking reward r_{thk} . We define the thinking reward as the sum of temporal and spatial terms:

$$r_{\text{thk}}(x, y) = r_{\text{t}}(x, y) + r_{\text{s}}(x, y). \quad (2)$$

Temporal term with adaptive temporal proximity. Let M be the number of timestamps $\{t_m\}_{m=1}^M$ parsed from $\langle \text{think} \rangle$. The temporal reward depends on the supervision type:

$$r_{\text{t}}(x, y) = \begin{cases} \frac{1}{M} \sum_{m=1}^M \mathbf{1}\{s^{\text{gt}} \leq t_m \leq e^{\text{gt}}\}, & \text{interval supervision } [s^{\text{gt}}, e^{\text{gt}}], \\ \frac{1}{M} \sum_{m=1}^M \exp\left(-\frac{\Delta t_m^2}{2\sigma^2}\right), \quad \Delta t_m = \min_j |t_m - t_j^{\text{gt}}|, & \text{point supervision } \{t_j^{\text{gt}}\}, \\ 0, & \text{no timestamp evidence.} \end{cases} \quad (3)$$

A key difficulty is that spatial rewards depend on accurate temporal predictions: IoU can only be computed reliably when the timestamp is close to the ground truth. If the temporal constraint is too strict (i.e., σ very small), the model receives little reward when its early temporal predictions are inaccurate, which slows down temporal learning and in turn prevents spatial grounding from being learned effectively. Conversely, if the constraint is always loose (i.e., σ large), temporal rewards quickly saturate and stop driving predicted timestamps closer to the ground truth, which again undermines spatial reward reliability. To resolve this trade-off, we propose **adaptive temporal proximity**: σ is large in early training to provide dense signals, and gradually decreases to enforce stricter alignment. This strategy ensures that the model first obtains stable gradients and later achieves precise timestamping, providing a solid foundation for spatial evaluation.

Spatial term with temporal gating. For each predicted timestamp t_m , let $j^*(m) = \arg \min_j |t_m - t_j^{\text{gt}}|$ be the nearest annotated time. Let \mathcal{B}_m be predicted boxes and $\mathcal{B}_{j^*(m)}^{\text{gt}}$ ground-truth boxes on that frame. The spatial reward is

$$r_s(x, y) = \frac{1}{M} \sum_{m=1}^M \mathbf{1}\{|t_m - t_{j^*(m)}^{\text{gt}}| \leq \tau\} \cdot \max_{b \in \mathcal{B}_m, b^{\text{gt}} \in \mathcal{B}_{j^*(m)}^{\text{gt}}} \text{IoU}(b, b^{\text{gt}}), \quad (4)$$

where τ is a temporal threshold. We further propose a **temporal gating** mechanism to guarantee the reliability of spatial supervision. Specifically, spatial rewards are only computed when temporal predictions are sufficiently close to the ground truth. This prevents rewarding salient but irrelevant objects at wrong timestamps, enforces spatio-temporal alignment, and ultimately improves both the interpretability and reliability of the reasoning process. Together, adaptive temporal proximity and temporal gating provide complementary solutions: the former supplies stable and progressive temporal supervision, while the latter ensures accurate and trustworthy spatial evaluation.

Format reward r_{fmt} . Strict usage of `<think>` and `<answer>` with correct `<obj>` `<box>` `<t>` gives 1.0. Having only `<think>` and `<answer>` yields 0.5. Otherwise, the reward is 0.0.

5 EXPERIMENTS

Implementation Details. We build upon the **Qwen2.5-VL-7B** model and train on 8 NVIDIA H100 GPUs. During training, we uniformly sample 16 frames from each video, where each frame has a resolution not exceeding $128 \times 28 \times 28$. If annotated key frames are available, they are inserted in addition to the uniformly sampled frames. To strengthen the model’s perception of temporal information, we prepend each frame with its absolute timestamp. More implementation details are provided in Appendix A.1.

Benchmarks. We adopt V-STAR (Cheng et al., 2025) as the main benchmark, since it is specifically designed to measure spatio-temporal grounding in videos. Unlike conventional video QA datasets, V-STAR requires models to not only answer questions but also localize *when* and *where* the supporting evidence occurs. It introduces two structured reasoning chains (“what–when–where” and “what–where–when”) and composite metrics combining accuracy with temporal and spatial IoU, thereby enabling comprehensive evaluation of spatio-temporal reasoning. We further evaluate on broader video understanding benchmarks. VideoMME (Fu et al., 2025) and VideoMMU (Hu et al., 2025) assess general video QA and multimodal comprehension across diverse domains, while WorldSense (Hong et al., 2025) emphasizes integrating multimodal signals with commonsense reasoning, and LongVideo-Reason-eval (Chen et al., 2025b) evaluates long-range reasoning on videos. In addition, TVGBench (Wang et al., 2025f) focuses on fine-grained temporal localization, STAR (Wu & Star, 2024) tests situated reasoning, and CameraBench (Lin et al., 2025) measures robustness under diverse camera motions.

5.1 MAIN RESULTS

Results on V-STAR. On the V-STAR benchmark, we compare our method with three groups of baselines: (i) closed-source commercial models such as GPT-4o (OpenAI, 2024) and Gemini-2-Flash (Team et al., 2024), which represent the current frontier of proprietary video LLMs. (ii) open-source general-purpose video understanding models, including Video-LLAMA3 (Zhang et al., 2025a), LLaVA-Video (Zhang et al., 2024b), VideoChat2 (Li et al., 2024), Oryx-1.5-7B (Liu et al., 2024), InternVL-2.5-8B (Chen et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). (iii) task-specialized approaches such as TRACE (Guo et al., 2024), designed for temporal video grounding, and Sa2VA (Yuan et al., 2025), optimized for fine-grained spatial grounding. As summarized in Table 1, our model consistently outperforms the baseline across different evaluation dimensions. In video question answering (*What*), our model achieves an accuracy of 61.03, representing a +27.6% point improvement over Qwen2.5-VL-7B. For temporal grounding (*When*), we report strong gains on both reasoning chains: Chain1 (*what–when–where*) improves by +9.1% points and Chain2 (*what–where–when*) by +10.2% points, showing robust performance regardless of the reasoning order. For spatial grounding (*Where*), our method surpasses the baseline by +8.4% points on Chain1 and +3.5% points on Chain2. Overall, compared with the Qwen2.5-VL baseline, our

Table 1: Performance on the **V-STAR** benchmark, which evaluates **spatio-temporal** reasoning across three dimensions. Chain1 denotes *what-when-where*, while Chain2 corresponds to *what-where-when*. mAM is the average of arithmetic mean, and mLGM is the average of modified logarithmic geometric mean, combining temporal and spatial alignment. * indicate we re-evaluate using the vLLM framework with 16 sampled frames. Bold numbers denote the best results, while underlined numbers indicate the second best.

Model	What	When (Temporal IoU)		Where (Visual IoU)		Overall	
	Acc	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
GPT-4o	<u>60.8</u>	16.7	12.8	6.5	3.0	26.8	<u>38.2</u>
Gemini-2-Flash	53.0	24.5	<u>23.8</u>	4.6	2.2	<u>26.9</u>	35.6
Video-LLaMA3	41.9	23.0	23.1	0.9	0.2	21.7	27.0
LLaVA-Video	49.5	10.5	12.2	1.9	1.3	20.8	27.3
VideoChat2	36.2	13.7	12.5	2.5	1.0	17.0	20.3
Oryx-1.5-7B	20.5	13.5	14.8	10.1	3.5	15.1	13.8
InternVL-2.5-8B	44.2	8.7	7.8	0.7	0.1	17.6	24.9
Qwen2.5-VL-7B* (base)	33.5	15.4	13.8	17.0	2.5	19.3	22.4
TRACE	17.6	19.1	17.1	0.0	0.0	12.0	13.3
Sa2VA-8B	16.4	0.1	0.0	32.3	37.5	17.1	20.3
Open-o3 Video (Ours)	61.0	24.5	24.0	<u>25.4</u>	<u>6.0</u>	33.7	46.6
Δ vs. Qwen2.5-VL-7B	\uparrow 27.5	\uparrow 9.1	\uparrow 10.2	\uparrow 8.4	\uparrow 3.5	\uparrow 14.4	\uparrow 24.2

Table 2: Performance across different video understanding and temporal grounding benchmarks. “LRR” refers to LongVideo-Reason-eval Benchmark. More evaluation results (on STAR and CameraBench) are provided in Appendix A.3.

Model	VideoMME		WorldSense		VideoMMMU		LRR	TVGBench	Avg
	Overall	Long	Overall	Recognition	Overall	Perception	Acc	mIoU	
GPT-4o	71.9	-	42.6	-	61.2	66.0	-	-	-
VideoLLaMA3-7B	60.6	48.7	37.3	<u>38.1</u>	46.5	59.7	59.8	22.2	45.3
InternVL-2.5-8B	62.3	51.2	39.6	38.5	42.4	57.0	62.0	6.3	42.5
Qwen2.5-VL-7B (Base)	<u>62.4</u>	<u>50.8</u>	36.1	33.7	51.2	64.7	59.3	16.3	45.1
VideoRFT-7B	59.8	50.7	<u>38.2</u>	36.6	51.1	<u>66.0</u>	69.4	14.3	<u>46.6</u>
VideoR1-7B	61.4	50.6	35.5	32.8	52.4	<u>65.3</u>	68.9	9.6	45.6
Open-o3 Video (Ours)	63.6	54.9	37.5	36.8	<u>52.3</u>	68.0	69.4	<u>20.8</u>	48.7
Δ vs. Qwen2.5-VL-7B	\uparrow 1.2	\uparrow 4.1	\uparrow 1.4	\uparrow 3.1	\uparrow 1.1	\uparrow 3.3	\uparrow 10.1	\uparrow 4.5	\uparrow 3.6

model improves performance by +14.4% mAM and +24.2% mLGM on V-STAR. It further surpasses proprietary models such as GPT-4o (OpenAI, 2024) and Gemini-2-Flash (Comanici et al., 2025) and achieves state-of-the-art performance. By extracting key frames and precise bounding boxes, Open-o3 Video brings o3-style, evidence-guided reasoning to videos, supplying more reliable and verifiable visual evidence during inference.

Results on General Video Understanding and Temporal Grounding Benchmarks. We further evaluate our method on a broad suite of video understanding benchmarks, comparing against three categories of baselines: (i) closed-source commercial models such as GPT-4o (OpenAI, 2024), (ii) open-source general-purpose video LLMs including Qwen2.5-VL-7B (Bai et al., 2025), Video-LLAMA3 (Zhang et al., 2025a), and InternVL-2.5-8B (Chen et al., 2024b), and (iii) recent reasoning-focused models such as VideoRFT-7B (Wang et al., 2025d) and VideoR1-7B (Feng et al., 2025b), which treat video understanding as text-only reasoning. In contrast, our method combines reasoning with explicit spatio-temporal grounding, enabling evidence-based inference. As shown in Table 2, Open-o3 Video achieves consistent improvements across all datasets. Across VideoMME, WorldSense, and VideoMMMU, our model shows consistent gains over Qwen2.5-VL-7B, with notable improvements on long videos (+4.1%) and perception-related tasks (+3.1% on WorldSense recognition and +3.3% on VideoMMMU perception), highlighting enhanced temporal reasoning and perceptual grounding. For long-range video reasoning, our model achieves 69.4% accuracy on the LongVideoReason-eval benchmark (LRR), and outperforms the baseline by +10.1%. Compared

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

Table 3: Ablation on Different training strategies.

Setting	What	When (Temporal IoU)		Where (Visual IoU)		Overall	
	Acc	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Baseline	33.5	15.4	13.8	17.0	2.5	19.3	22.4
Pure SFT	53.0	19.6	17.2	23.3	4.6	28.5	37.1
Pure RL (GSPO)	56.4	21.6	20.7	23.7	3.7	30.4	40.7
SFT+RL (GRPO)	60.5	21.6	23.1	25.3	5.8	32.8	45.3
SFT+RL (GSPO)	61.0	24.5	24.0	25.4	6.0	33.7	46.6

Table 4: Impact of two reward designs.

Setting	mAM	mLGM
Open-o3 Video	33.7	46.6
w/o Ada.	33.0	45.2
w/o Gat.	32.3	44.9

Table 5: Impact of spatio-temporal training data.

Training data	mAM	mLGM
w/o spatio-temporal data	28.3	36.2
+ VideoEspresso	31.1	43.6
+ Our annotated data	33.7	46.6

with dedicated video reasoning methods, our model achieves comparable or even superior results, while providing more interpretable evidence in its reasoning process. On TVGBench, which directly measures temporal grounding, our model surpasses the baseline by a large margin (+4.5 mIoU), indicating significant gains in temporal localization. These results show that our approach **maintains the QA strength of general video LLMs** while enhancing the spatio-temporal grounding capability.

5.2 ABLATION AND ANALYSIS

Training strategy: RL provides larger gains than SFT, while their combination yields the best results, with GSPO offering the most stable improvements. As shown in Table 3, both SFT and RL substantially improve grounding over the base model. RL outperforms SFT (+2.1% mAM, +4.6% mLGM) by directly optimizing temporal and spatial alignment, while SFT ensures stable reasoning formats and basic grounding under supervision. Their combination is highly synergistic, reaching 33.7% mAM and 46.6% mLGM. Within this joint training, GSPO further surpasses GRPO (+0.9% mAM, +1.3% mLGM) by providing more stable rewards and better long-horizon temporal localization (+2.9% Chain1 tIoU).

Reward design: Both adaptive temporal proximity and temporal gating are effective. In the thinking reward, we introduce two mechanisms: adaptive temporal proximity (**Ada.**) and temporal gating (**Gat.**). To validate their effectiveness, we conduct ablation experiments on the V-STAR benchmark. Removing the proximity reward reduces performance by 0.7% mAM and 1.4% mLGM, showing that adaptive scaling helps the model better align predicted timestamps with annotated windows. Removing temporal gating causes larger drops of 1.4% mAM and 1.7% mLGM, confirming that gating is crucial for filtering irrelevant segments and preventing noisy spatial boxes. These results verify that our reward design effectively couples temporal and spatial grounding, leading to the strong performance.

Training data: High-quality spatio-temporal annotations significantly boost grounding. Without spatio-temporal (ST) supervision, the model exhibits substantially weaker performance, underscoring the necessity of Spatio-temporal annotations for effective grounding. Incorporating 9.6k filtered and rewritten *VideoEspresso* (Han et al., 2025) samples enables the model to perform basic spatio-temporal reasoning, leading to improvements of +2.8% mAM and +7.4% mLGM. Building upon this, we further construct 5.9k high-quality Spatio-temporal annotations through our dedicated pipeline (as illustrated in Figure 2), which bring a larger gain of +5.4% mAM and +10.4% mLGM. This shows the effectiveness of our pipeline and the critical role of high-quality spatio-temporal supervision.

Test-time scaling with grounded evidence: Confidence-aware voting with Open-o3 Video outperforms naive majority voting. Inspired by the scoring and adaptive voting mechanisms for video reasoning in CyberV (Meng et al., 2025), we introduce a confidence-aware voting scheme

486 that leverages grounded evidence to verify predictions at inference, as shown in Figure 7 in the ap-
487 pendix. Details, including scoring schemes, prompts, and results on WorldSense and VideoMMMU
488 are provided in Appendix A.7.

490 6 CONCLUSION

491
492 We introduced **Open-o3 Video**, a unified framework for grounded video reasoning that generates
493 explicit spatio-temporal evidence through timestamped frames and localized bounding boxes. With
494 carefully curated high-quality training data, a two-stage strategy combining supervised fine-tuning
495 and GSPO-based reinforcement learning, and novel thinking rewards incorporating adaptive tem-
496 poral proximity and temporal gating, our method substantially improves answer accuracy, temporal
497 alignment, and spatial grounding. Comprehensive experiments demonstrate that Open-o3 Video
498 achieves state-of-the-art performance on the V-STAR benchmark, surpassing strong baselines in-
499 cluding GPT-4o, while remaining broadly competitive across diverse video understanding tasks. In
500 future work, we aim to further align reasoning chains across text, time, space, and audio modalities,
501 and to extend our approach to more complex and longer video scenarios.

503 ETHICS STATEMENT

504
505 All datasets used for evaluation in this work are publicly available benchmarks for video under-
506 standing. In addition, we construct a new dataset based on open-source data sources, which will
507 be released to the community upon publication to ensure transparency and academic benefit. No
508 private or personally identifiable information is involved, and all data usage strictly follows the in-
509 tended research licenses. We also recognize potential risks such as biased annotations or unintended
510 harmful outputs, and we emphasize that our method is intended solely for academic research.

512 REPRODUCIBILITY STATEMENT

513
514 Comprehensive implementation details, including training procedures, hyperparameter configura-
515 tions, and evaluation protocols, are provided in the main paper (Section 5) and Appendix A.1. Fur-
516 thermore, upon acceptance of this paper, all source code, datasets, and trained model checkpoints
517 will be made publicly available.

519 LLM USAGE STATEMENT

520
521 Large language models (LLMs) are used solely to aid in polishing the writing of this paper, such as
522 improving grammar, clarity, and readability. No LLMs are used for research ideation, experimental
523 design, data analysis, or result generation. All technical contributions, experiments, and analyses
524 are conducted entirely by the authors.

526 REFERENCES

- 527
528 Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell.
529 Localizing moments in video with natural language. In *Proceedings of the IEEE international*
530 *conference on computer vision*, pp. 5803–5812, 2017.
- 531 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Siboz Song, Kai Dang, Peng Wang,
532 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
533 2025.
- 534 Apratim Bhattacharyya, Sunny Panchal, Mingu Lee, Reza Pourreza, Pulkit Madan, and Roland
535 Memisevic. Look, remember and reason: Grounded reasoning in videos with language models.
536 *arXiv preprint arXiv:2306.17778*, 2023.
- 537
538 Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet:
539 A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE*
conference on computer vision and pattern recognition, pp. 961–970, 2015.

- 540 Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng
541 Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video tempo-
542 ral grounding via reinforcement learning. *arXiv preprint arXiv:2507.18100*, 2025a.
- 543
544 Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian
545 Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for
546 long videos. *arXiv preprint arXiv:2408.10188*, 2024a.
- 547 Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu,
548 Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint*
549 *arXiv:2507.07966*, 2025b.
- 550 Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-
551 glong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source
552 multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*,
553 2024b.
- 554
555 Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking
556 video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.
- 557 Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan,
558 Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-
559 access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.
- 560
561 Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit
562 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
563 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
564 bilities. *arXiv preprint arXiv:2507.06261*, 2025.
- 565 Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng,
566 Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *arXiv*
567 *preprint arXiv:2505.24718*, 2025.
- 568 Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi
569 Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images.
570 *arXiv preprint arXiv:2505.15879*, 2025.
- 571
572 Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou
573 Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in mllms. *arXiv preprint*
574 *arXiv:2503.21776*, 2025a.
- 575 Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu,
576 Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-rl: Reinforcing video reasoning in
577 mllms. *arXiv preprint arXiv:2503.21776*, 2025b.
- 578
579 Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu
580 Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evalua-
581 tion benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.
- 582
583 Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal
584 grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.
- 585
586 Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi,
587 Yue Liao, and Si Liu. Videospresso: A large-scale chain-of-thought dataset for fine-grained
588 video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern*
589 *Recognition Conference*, pp. 26181–26191, 2025.
- 590
591 Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating
592 real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*,
593 2025.
- 594
595 Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei
596 Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos.
597 *arXiv preprint arXiv:2501.13826*, 2025.

- 594 Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-
595 temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–
596 29204, 2022.
- 597
- 598 Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural
599 language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.
- 600
- 601 Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and
602 Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal under-
603 standing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8592–
604 8603, 2025a.
- 605
- 606 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,
607 Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In
608 *CVPR*, pp. 22195–22206, 2024.
- 609
- 610 Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao,
611 Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforce-
612 ment fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025b.
- 613
- 614 Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal
615 video temporal grounding with generative multi-modal large language models. *arXiv preprint*
616 *arXiv:2506.18883*, 2025c.
- 617
- 618 Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, and Liwei Wang. Fine-grained spatiotemporal
619 grounding on egocentric videos. In *Proceedings of the IEEE/CVF International Conference on*
620 *Computer Vision*, pp. 9385–9395, 2025.
- 621
- 622 Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling,
623 Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any
624 video. *arXiv preprint arXiv:2504.15376*, 2025.
- 625
- 626 Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-
627 demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*,
628 2024.
- 629
- 630 Jiahao Meng, Shuyang Sun, Yue Tan, Lu Qi, Yunhai Tong, Xiangtai Li, and Longyin Wen. Cyberv:
631 Cybernetics for test-time scaling in video understanding. *arXiv preprint arXiv:2506.07971*, 2025.
- 632
- 633 Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie.
634 Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE*
635 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2265–2269.
636 IEEE, 2021.
- 637
- 638 OpenAI. Hello gpt4-o. <https://openai.com/index/hello-gpt-4o/>, 2024.
- 639
- 640 OpenAI. Openai-o3. [https://openai.com/index/
641 introducing-o3-and-o4-mini/](https://openai.com/index/introducing-o3-and-o4-mini/), 2025.
- 642
- 643 Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun.
644 Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.
- 645
- 646 Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement
647 fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025.
- 648
- 649 Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hong-
650 sheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and
651 benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*,
652 37:8612–8642, 2024a.
- 653
- 654 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
655 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
656 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

- 648 Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sam-
649 pling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recogni-
650 tion Conference*, pp. 29118–29128, 2025.
- 651 Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie
652 Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings
653 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.
- 654 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,
655 Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal under-
656 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- 657 Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling
658 Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint
659 arXiv:2507.01949*, 2025.
- 660 Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxi-
661 ang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024.
- 662 Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng,
663 Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning:
664 Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.
- 665 Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang
666 Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *arXiv preprint
667 arXiv:2504.01901*, 2025b.
- 668 Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang,
669 Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint
670 arXiv:2506.11991*, 2025c.
- 671 Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning
672 capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025d.
- 673 Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei
674 Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal
675 grounding. *arXiv preprint arXiv:2507.13353*, 2025e.
- 676 Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju,
677 Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal
678 video grounding. *arXiv preprint arXiv:2503.13377*, 2025f.
- 679 Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit
680 Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and en-
681 hanced video reasoning. *arXiv preprint arXiv:2507.06485*, 2025g.
- 682 Bo Wu and Shoubin Yu Star. A benchmark for situated reasoning in real-world videos. *Advances in
683 Neural Information Processing Systems (NeurIPS)*, 3, 2024.
- 684 Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning
685 for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.
- 686 Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and
687 Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large
688 language models. *arXiv preprint arXiv:2408.04840*, 2024.
- 689 Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi,
690 Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded under-
691 standing of images and videos. *arXiv*, 2025.
- 692 Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong
693 Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation
694 models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

702 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language
703 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
704

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

708 Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue
709 Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision.
710 *arXiv preprint arXiv:2406.16852*, 2024a.

711 Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller
712 lms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025c.
713

714 Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video
715 instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.
716

717 Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang,
718 Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint*
719 *arXiv:2507.18071*, 2025a.

720 Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and
721 Xing Yu. Deepeyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv*
722 *preprint arXiv:2505.14362*, 2025b.

723 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao
724 Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for
725 open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

A APPENDIX

Overview. This appendix provides additional details and analyses to complement the main paper. Section A.1 gives more implementation details. Section A.2 describes training dataset preparation and ablations on the ratio of general VideoQA data. Section A.3 provides more experimental results on STAR benchmark and CameraBench. Section A.4 provides analysis of inference frame rate. Section A.5 presents the prompts used for data annotation with Gemini. Section A.6 provides the full mathematical formulation of GSPO algorithm. Section A.7 details the confidence-aware test-time scaling procedure and reports additional results. Section A.8 provides further qualitative visualizations of spatio-temporal reasoning. Finally, Section A.9 discusses limitations of our current framework and directions for future work.

A.1 MORE IMPLEMENTATION DETAILS.

The training process of Open-o3 Video consists of two stages. In the cold-start stage, we train on the STGR-CoT-30k dataset for one epoch with a learning rate of 1×10^{-6} . In the GSPO stage, we further train on the STGR-RL-36k dataset for one epoch, also with a learning rate of 1×10^{-6} . For the thinking reward, the standard deviation parameter σ is annealed from 4 to 1 and then kept constant. The gating mechanism employs a temporal threshold τ of 3s. At test time, we employ the vLLM framework, requiring the model to first produce a spatio-temporal grounded reasoning process, followed by the final answer.

A.2 MORE DETAILS AND ABLATION ON TRAINING DATA.

Data Preparation. Beyond reporting corpus sizes, we describe here the sampling and filtering strategy applied to each source. For temporal grounding data, we adopt strict constraints to ensure annotation quality and manageable reasoning length. Specifically, for TVG-Coldstart, we retain only samples with chain-of-thought length under 6,000 characters and with ground-truth spans covering less than 70% of the total video duration. The same filtering is applied to Time-R1, resulting in 2.3k samples. For additional temporal grounding video sources (ActivityNet, COIN, QueryD, QVHighlight, and DiDeMo), we keep videos of duration between 10 seconds and 3 minutes, further discarding those where the annotated action lasts more than 50% of the video; TVG-RL is filtered with the same rules, and 2.9k samples are randomly selected. For spatial grounding data, we randomly sample 5k instances from both TreeVGR-SFT and VisCoT. For general video QA data, 15k Video-R1 samples are randomly drawn without additional filtering. For PLM-based video dense captioning data (PLM-Rdcap), we initially sample 3k videos for annotation, from which 2k remain after filtering for quality and consistency. This careful selection yields a high-quality dataset that balances temporal, spatial, and general reasoning tasks. The resulting dataset provides diverse yet clean supervision signals, making it particularly suitable for training and evaluating spatio-temporal reasoning models.

Ablation on Different Ratios of General VideoQA Data. To enhance the model’s grounding ability, we emphasize temporal and spatial grounding data during training. However, excessive focus on grounding may weaken the model’s original strength in general VideoQA. Thus, an important design choice is how much general VideoQA data to include in the STGR dataset. We compare different ratios and evaluate performance on both grounding-oriented (VSTAR) and QA-oriented (VideoMME) benchmarks. As shown in Table 6, adding 15k VideoQA samples significantly improves QA accuracy without harming grounding performance. In contrast, adding 30k yields no further QA gain while slightly reducing grounding accuracy. Therefore, we adopt 15k VideoQA samples as a balanced choice, offering strong QA capability while preserving grounding ability, and maintaining training efficiency.

A.3 MORE EVALUATION RESULTS

As shown in Table 7, Open-o3 Video performs better than the base model on both STAR and CameraBench. On STAR, it improves accuracy by +3.2%, showing that Open-o3 Video can better handle situated reasoning tasks when involving spatio-temporal cues. We also evaluate our model on the CameraBench VQA task and compare it with the baseline model as well as models trained without adaptive temporal proximity or without temporal gating. We find that our model performs better

Table 6: Impact of different amounts of general VideoQA data. 15k achieves the best balance between grounding and general QA performance.

VideoQA Data	VSTAR (mAM)	VideoMME (Acc)
w/o Video-R1 data	33.4	60.7
+5k	33.0	63.2
+15k	33.7	63.6
+30k	31.7	63.6

Table 7: Performance on STAR and CameraBench.

Models	STAR		CameraBench VQA		
	Overall	Overall	Confusable Motion	Motion and Steadiness	Motion Speed
Qwen2.5-VL-7B	67.3	57.5	49.3	56.7	69.0
Open-o3 Video	70.5	58.8	51.3	57.6	69.3
w/o Ada.	70.1	58.5	50.0	56.7	68.7
w/o Gat.	69.6	57.8	50.3	55.9	67.0

than the baseline and shows gains in challenging motion settings, such as confusable motion, motion and steadiness, and different motion speeds. It also outperforms the variants without Ada. or without Gat. These results indicate that both the model and the training techniques remain stable under camera motions that differ from the training data distribution.

A.4 ABLATION STUDIES ON INFERENCE FRAME RATE

We analyze the effect of inference frame rate on long-video understanding using the LongVideo-Reason-eval benchmark, as shown in Table 8. For the comparison between high and low frame rates, we find that higher frame rates (64 frames) give some improvement, but even with only 16 frames, our model performs well and surpasses the baseline and other reasoning models. And the gain of increasing more frames is small. For variable frame rates, we follow AKS (Tang et al., 2025) and apply the key frames selection strategy. This strategy reaches 70.1% accuracy, showing that key frames sampling can offer a small improvement over uniform sampling when involving spatio-temporal reasoning.

A.5 PROMPT FOR DATA ANNOTATION.

To obtain high-quality spatio-temporal annotations, we design structured prompts for the Gemini 2.5 Pro API, separately tailored to the two data sources described in Section 3: PLM-Rdcap data and temporal grounding datasets. The goal of these prompts is to guide the model to produce question-answer pairs, key frame selection, bounding boxes, and reasoning chains in a consistent JSON format.

For PLM-Rdcap, as shown in Figure 4, the input is the dense video captions and total frame count, and the output is a JSON with *question*, *answer*, *key_frames*, and *reasoning_process*. Since only frame indices are given, we post-process them into timestamps and align reasoning mentions with annotated object names and boxes.

For temporal grounding datasets, as shown in Figure 5, the input includes the annotated segment, video duration, and segment descriptions, and the output JSON contains the *question*, *answer*, *key_frames* with timestamps, objects and boxes, and the spatio-temporal grounded *reasoning_process*.

We further apply strict filtering and consistency checks, retaining only annotations with validated boxes, aligned timestamps, and coherent reasoning. This ensures a high-quality dataset with reliable spatio-temporal evidence, essential for robust training and evaluation.

Table 8: Ablation on inference frame rate on LongVideo-Reason-eval.

Models	Qwen2.5-VL	Video-R1	VideoRFT	Open-o3 Video
Number of Frames	64	64	16	64 (+AKS) 16
LongVideo-Reason-eval	59.3	68.9	67.3	69.4 68.0 69.4 70.1 69.2

Prompt for Gemini 2.5 Pro (PLM-Rdcap)

The video contains a total of {item['total_frames']} frames, with the following dense captions information: {str(dcap)}

Please complete the following tasks based on the video and caption information:

1. Generate a question-answer pair. Since the dense caption is centered on a specific object or person, the question should also focus on this object or person. You can consider aspects such as its color, clothing, actions, and so on.
2. Output key_frames, which should be the critical frames needed to answer the question. The key_frames must be a list of integer values and fall within the frame range mentioned in the dense caption. (at least one and at most five).
3. Generate a reasoning process:
 - Reasoning must use visual evidence grounded in the video.
 - When referencing the target object or person, you MUST use the following strict format: <obj>object_name</obj>at<t>Frame frame_num</t>
 - The reasoning must not exceed 200 words.
 - The frame number must be in key_frames. The mentioned frame numbers and the visual content of those frames must match consistently.
 - All object names must be identical.
 - Every time you mention the object name (<obj>), you must use the format '<obj>object_name</obj>at<t>Frame frame_num</t>' to specify the corresponding frame.
 - In the reasoning process, except for the text between <t> </t>, the words "frames", "frame" and similar terms MUST not appear.

You must strictly follow the following JSON format (with no additional text outside the JSON):

```

{{
  "question": "...",
  "answer": "...",
  "key_frames": [...],
  "reasoning_process": "..."
}}
```

Figure 4: Annotation Prompt for PLM-Video-Human Region Dense Temporal Captioning Data source.

A.6 DETAILS OF GSPO TRAINING

For completeness, we provide the full formulation of Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025a), which is used in our reinforcement learning stage.

Given a query x , the model generates a group of G candidate responses $\{y_i\}_{i=1}^G$ sampled from the old policy $\pi_{\theta_{\text{old}}}(\cdot|x)$. Each response is scored by a reward function $r(x, y_i)$, and its normalized advantage is computed as

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}(\{r(x, y_j)\}_{j=1}^G)}{\text{std}(\{r(x, y_j)\}_{j=1}^G)}. \quad (5)$$

The importance ratio is defined at the sequence level as

$$s_i(\theta) = \left(\frac{\pi_{\theta}(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)} \right)^{\frac{1}{|y_i|}} = \exp \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \right), \quad (6)$$

where $|y_i|$ denotes the response length.

The GSPO objective is then

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x, \{y_i\} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i \right) \right], \quad (7)$$

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

Prompt for Gemini 2.5 Pro (Temporal Grounding)

The video has a duration of {item['duration']} seconds. The temporal grounding annotation for the video is as follows:
Description: {item['conversations']}
Annotated time segment: {str(item['gt_segment'])}

Based on this annotation, please complete the following tasks:

1. Construct a question-answer pair in open-ended Q&A format.
 - The question should be adapted from the temporal grounding description.
 - The question should focus on a specific object or person, rather than their action.
 - The answer should be concise and not exceed 30 words.
 - Do NOT mention timestamps or annotated time segments in the question.
2. Select at least ONE and at most FIVE keyframes.
 - Each timestamp MUST be within the annotated time segment and be written as a float rounded to exactly one decimal place.
3. For each keyframe, include at least ONE and at most THREE detected objects.
 - Each bounding box coordinates are normalized floats (rounded to exactly two decimal place) in the format [x_min, y_min, x_max, y_max].
4. Generate the reasoning process for answering the question:
 - Reasoning must use visual evidence grounded in the video.
 - When referencing any object, person, or visual element, you MUST use the following strict format: <obj>object_name</obj><box>[x_min, y_min, x_max, y_max]</box>at<t>time_in_seconds</t>s
 - Both the 'time_in_seconds' and the box coordinates MUST be consistent with the info in the key frames.
 - The reasoning must not exceed 200 words.

You must strictly follow the following JSON format (with no additional text outside the JSON):

```

{{
  "question": "...",
  "answer": "...",
  "key_frames": [
    {{
      "timestamp": time in second,
      "items": {{
        "object1_name": box1,
        "object2_name": box2,
      }}
    }},
    {{
      "timestamp": time in second,
      ...
    }}
  ],
  "reasoning_process": "..."
}}

```

Figure 5: Annotation Prompt for Temporal Grounding Data Source.

with ϵ controlling the clipping range.

Unlike GRPO, which clips per-token updates, GSPO clips entire responses, thereby aligning reward assignment with optimization granularity. In practice, this leads to more stable gradients and better performance on long chain-of-thought reasoning tasks.

A.7 MORE DETAILS ABOUT TEST TIME SCALING.

To further enhance robustness at inference, we adopt a **confidence-aware test-time scaling** procedure, as shown in Figure 7. Given a video question, the model first generates N independent responses in parallel (In our experiments, $N = 8$, with temperature set to 1.0). Each response contains spatio-temporal grounding annotations in the format

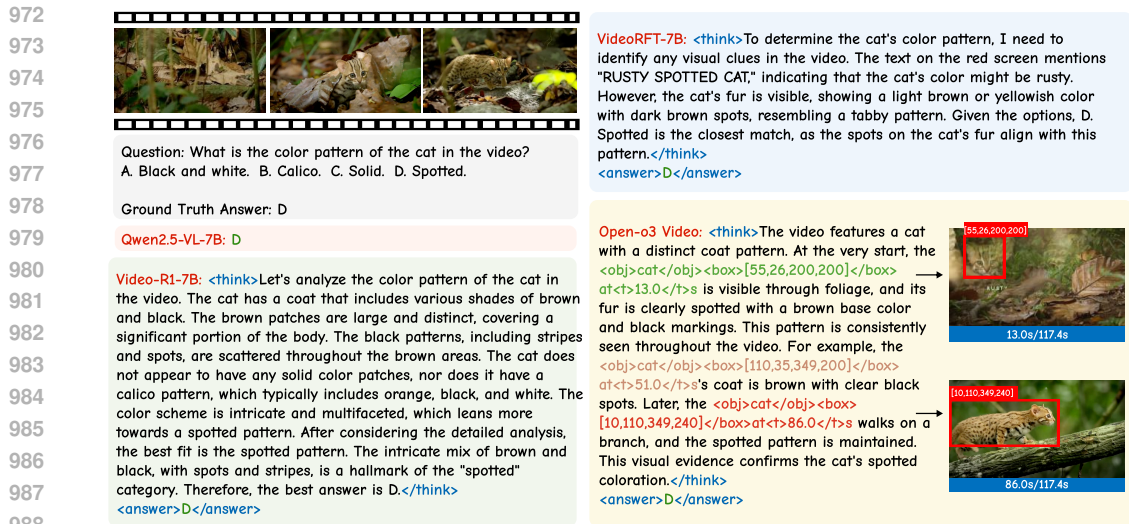


Figure 6: **Visualization.** On simple appearance perception tasks, both our model and related base-lines can provide correct answers; however, our approach additionally offers explicit spatio-temporal evidence.

Table 9: Test-time scaling results on WorldSense and VideoMMM, showing that the confidence-aware voting (N=8) with grounded evidence consistently outperforms base model (N=1) and naive majority voting (N=8).

Setting	WorldSense	VideoMMM
Base	37.5	52.3
Majority Voting	37.3	53.1
Confidence-aware Voting	38.5	54.1

<obj>...</obj><box>...</box>at<t>...</t>s, from which we extract the predicted bounding boxes. The corresponding regions are then cropped from the original video frames and paired with the question to form a new input. This input is passed back into the model to obtain a confidence score $s \in \{0, 1, 2\}$, where:

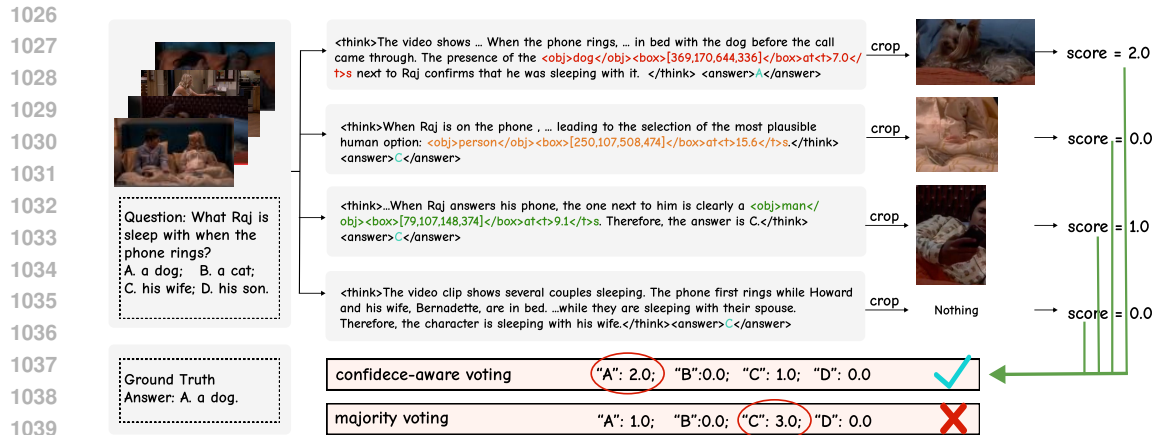
- $s = 2$: the cropped evidence is highly supportive for answering the question,
- $s = 1$: the evidence may be partially useful,
- $s = 0$: the evidence is irrelevant.

Each initial response is assigned a confidence-weighted score by averaging its evidence scores across all mentioned objects. The final prediction is selected via weighted voting over the N responses. This process effectively filters out hallucinated reasoning traces and highlights consistent evidence across responses.

As reported in Table 9, confidence-aware voting consistently improves over *naive majority voting*, achieving +1.0 on WorldSense and +1.0 on VideoMMM. This demonstrates that our o3-style spatio-temporal evidence not only enhances grounding, but also provides a natural mechanism for scalable inference and self-correction at test time.

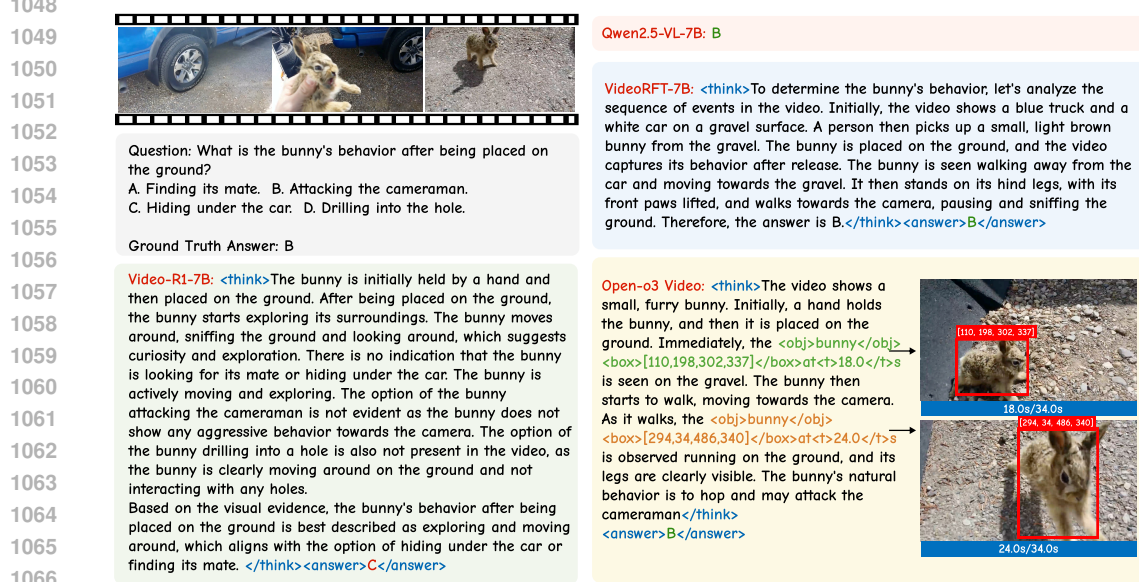
A.8 MORE VISUALIZATIONS.

As shown in Figure 6,8,9, we provide additional qualitative examples to illustrate the spatio-temporal reasoning ability of Open-o3 Video. These visualizations demonstrate that our model can obtain spatio-temporal evidence and achieve better results.



1042
1043
1044
1045
1046
1047

Figure 7: Illustration of our **confidence-aware test-time scaling**. The model generates multiple responses with spatio-temporal traces, from which visual regions are cropped and scored for evidence relevance ($s \in \{0, 1, 2\}$). Final predictions are obtained by confidence-weighted voting. Unlike naive majority voting that is misled by spurious patterns (predicting “C”), our method highlights consistent supportive evidence and correctly predicts “A”, improving robustness at inference.



1069
1070
1071
1072


Figure 8: **Visualization**. For action recognition, our model precisely localizes both the time and location of the action, achieving superior performance compared to Video-R1.

1073 A.9 LIMITATIONS AND FUTURE WORK.

1074
1075
1076
1077
1078
1079

While our framework demonstrates strong performance, several limitations remain. First, handling longer videos with complex scenes and smaller objects is still challenging, as high-quality spatio-temporal data for such cases is still relatively scarce. Second, reasoning-intensive queries that require multi-step inference beyond direct grounding remain difficult to fully address. Finally, our current design does not integrate audio or speech information, which often carries crucial cues for understanding video content. Future work will focus on extending the approach to longer and more

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133



Qwen2.5-VL-7B: C

VideoRFT-7B: <think>To determine the weather during takeoff, I first recall the scene. The video initially shows a character in a metallic suit on a platform with a helicopter labeled 'INDUSTRIES' in the background. The scene transitions to an exterior view of the helicopter as it takes off, indicating it is cloudy, with dark blue skies and patches of light breaking through. The helicopter exits the hangar, and the sky remains cloudy throughout the takeoff. There are no visible signs of rain, such as water droplets on the helicopter's surface or a wet ground. The visual evidence clearly shows the helicopter taking off in a cloudy sky, with no significant presence of rain or snow. The sky is cloudy, and the takeoff sequence is smooth, indicating a cloudy sky.</think>
<answer>A</answer>

Video-R1-7B: <think>As an AI reading the provided video description, it is difficult to tell the weather during the helicopter takeoff accurately due to the absence of visual evidence of weather phenomena such as rain, snow, sunlight, or shadow. The scene involves a black helicopter flying over a body of water that looks dark, reflecting light, which might suggest certain weather conditions. Using the genre of the identified overview video, we can posit that bystanders might perceive the weather as part of the dramatic effect, perhaps with scattered clouds or a suggestive atmosphere. </think>
<answer>A</answer>

Open-o3 Video: <think>The video shows a <obj>helicopter</obj> taking off from a helipad. There is visible rain or water droplets on the helicopter's surface and the surrounding area, indicating it is raining. The rain continues as the helicopter gains altitude, with the dark, cloudy sky in the background. Therefore, the weather is rainy.</think>
<answer>C</answer>

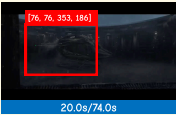


Figure 9: **Visualization.** In weather reasoning tasks, our model identifies more effective supporting evidence, whereas related video reasoning models perform poorly.

complex videos, enriching supervision for fine-grained object grounding, and unifying multimodal signals including speech to further enhance logical reasoning.