ByteDance

Open-o3 Video: Grounded Video Reasoning with Explicit Spatio-Temporal Evidence

Jiahao Meng[1,2], **Xiangtai Li**[2], **Haochen Wang**[2,3], **Yue Tan**[1], **Tao Zhang**[2,4], **Lingdong Kong**[2,5], **Yunhai Tong**[1†], **Anran Wang**[2], **Zhiyang Teng**[2], **Yujing Wang**[2], **Zhuochen Wang**[2]

Peking University[1]   ByteDance[2]   CASIA[3]   WHU[4]   NUS[5]
†Corresponding Author

Project Page: https://marinero4972.github.io/projects/Open-o3-Video/

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, including VideoMME, WorldSense, VideoMMMU, 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.

1 Introduction

Understanding complex video content is a long-standing goal for large multimodal models [5, 36, 37, 39, 46, 49, 51, 53], 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: According to the video, which of the following is the main reason why people commemorate Qu Yuan?
A. Because people love Zongzi.
B. Because he committed suicide by drowing himself in Miluo River.
C. Because he brought peace and prosperity to the state.
D. Because he was exile.

Ground Truth Answer: C

Qwen2.5-VL-7B: B

Video-R1-7B: <think>The video suggests the commemoration of Qu Yuan is associated with significant historical and cultural storytelling. Qu Yuan was noted for his contributions in poetry and governance. Despite his upright, loyal, and notably significant contributions, leading to peace and prosperity, he was maliciously accused. The vivid recall for his sorrow leads him towards a tragic, symbolic ending of diving into the Miluo River. The focus is on the ethical life symbolism of his personal loss and resilience. </think><answer>B</answer>

VideoRFT-7B: <think>To determine the main reason for commemorating Qu Yuan, I examined the video content. The video begins with a vibrant blue screen displaying Chinese characters and dragon motifs, which transitions into an animated sequence about Qu Yuan. The narrative shifts to his betrayal and exile, leading to his eventual suicide by drowning in the Miluo River. Later, the video mentions that people eat Zongzi and hold dragon boat races on the fifth day of May to honor Qu Yuan. Therefore, the key reason for commemorating Qu Yuan is his suicide, which is the central event depicted. Therefore, the answer is B, as the other options do not align with the main reason provided in the video. </think><answer>B</answer>

Open-o3 Video: <think>Think process: The video provides historical context about Qu Yuan. Initially, it shows a <obj>man</obj><box>[249,60,395,252]</box>at<t>9.0</t>s who is identified as the "great patriotic poet of Chu State." The narrative then describes his character, stating that he was "upright, loyal and highly esteemed," which led to the people's respect for him, brought peace and prosperity, and influenced the dragon boat festival. The video concludes by mentioning the damming of the river and Qu Yuan's suicide by diving into it. This information is visualized through various scenes, such as people bowing to <obj>the man</obj><box>[472,70,612,269]</box>at<t>24.0</t>s and the silhouette of him near mountains, which align with the narration. Therefore, the cause for celebrating him is his significant role in bringing peace and prosperity.</think><answer>C</answer>

Figure 1 While prior video reasoning models (e.g., Video-R1 [13], VideoRFT [41]) 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.6.

question answering [2, 22, 48, 53, 56], building models that can perform reliable, fine-grained reasoning over long and cluttered scenes remains challenging.

Recent "thinking with images" attempts [29, 38, 40, 55] 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 substantially 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* [13, 41] or, coarse, *temporal-only* grounding [23, 43], failing to achieve the fine-grained spatio-temporal precision necessary for complex video reasoning. This gap is largely due to two interconnected 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 meticulously curated two complementary 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 powerful combination of curated data and tailored training, as shown in Figure 1, Open-o3 Video moves beyond text-only responses to deliver reasoning that is accurate, interpretable, and tightly coupled with the visual content of the video. 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, VideoMMMU, 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 VideoMMMU), demonstrating that grounded evidence provides a reliable self-verification signal to improve inference accuracy.

2 Related works

Video Reasoning.

Recent advances in video reasoning [6, 11, 14, 23, 31, 41, 44, 45, 50, 52] 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 [14] shows that temporal-aware GRPO with curated reasoning data improves video understanding benchmarks, while VideoChat-R1 [23] extends to spatio-temporal perception tasks such as grounding and tracking without harming QA. Other variants, including Video-RTS [44] and DeepVideo-R1 [31], 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 [4, 16, 23, 24, 30, 42, 43]. On the temporal side, Time-R1 [43] introduces verifiable rewards for temporal grounding with strong generalization under limited supervision, while TVG-R1 [4] improves robustness with curated cold-start and RL datasets. On the spatial side, SpaceR [30] leverages RL and a large corpus for object-centric grounding and geometric reasoning. Moreover, LLaVA-ST [21] bridges the two sides by employing positional embedding alignment and two-stream feature compression, achieving spatio-temporal localization. However, 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 [12, 29, 38, 40, 55] 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 [29] formalizes "thinking with images," while DeepEyes [55] shows end-to-end RL can incentivize image–tool reasoning, and TreeBench [38] 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 [50] 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
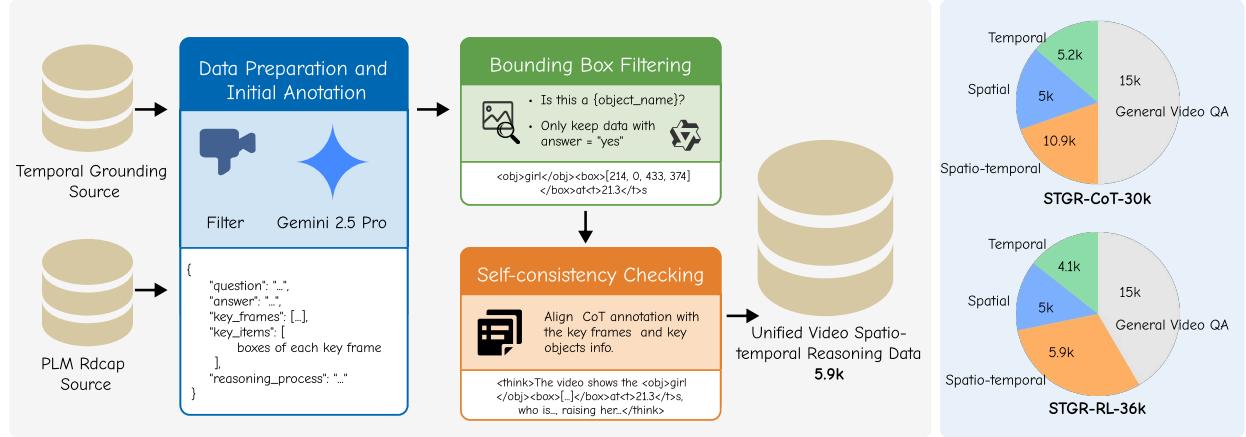
Figure 2 Overview of our data construction pipeline and dataset composition. **Left**: The annotation pipeline includes Gemini 2.5 Pro initial annotation, bounding box filtering, and self-consistency checking. **Right**: Distribution of data categories in STGR-CoT-30k (SFT) and STGR-RL-36k (RL), showing a balanced coverage across temporal, spatial, spatio-temporal, and general QA.

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) [4], (ii) 5k spatial grounding CoT samples (TreeVGR-SFT) [38], (iii) 5.9k spatio-temporal samples curated by us, including 3.9k from temporal grounding datasets (video source: ActivityNet [3], COIN [34], QueryD [27], QVHighlight [20], DiDeMo [1]) and 2k from PLM-Rdcap [9], and (iv) 15k Video-R1-CoT samples [14]. The RL corpus further expands diversity: (i) 5.2k temporal grounding samples, including 2.3k from Time-R1 [43] and 2.9k from TVG-RL [4], (ii) 5k spatial grounding samples from VisCoT [32], (iii) 10.9k spatio-temporal samples, comprising our 5.9k constructed data (via the pipeline) and an additional 5k filtered from VideoEspresso [17] with consistency checks, and (iv) 15k Video-R1 samples [14].

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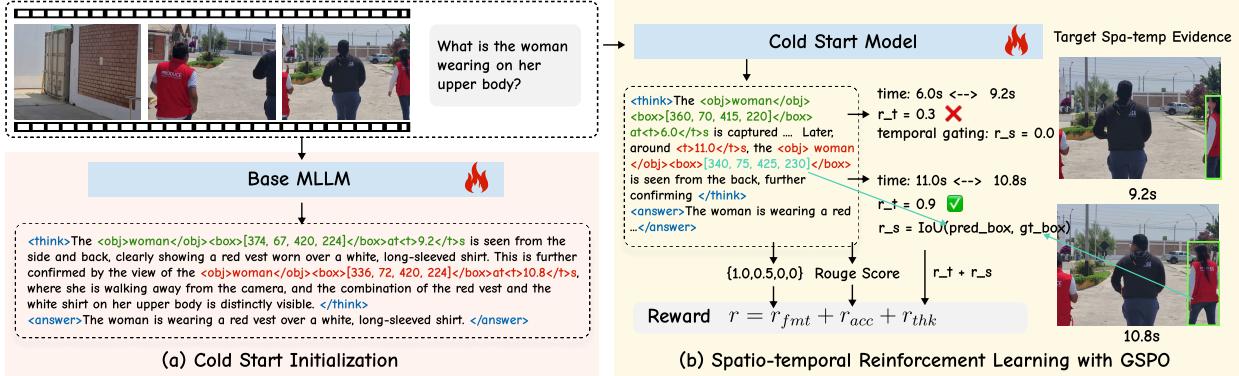
Figure 3 Overview of Open-o3 Video. We adopt a two-stage training paradigm: (a) cold-start initialization to learn structured, grounded outputs; (b) reinforcement learning with a composite reward that sharpens temporal alignment and spatial precision with adaptive temporal proximity and temporal gating.

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.

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 [10] API with carefully designed prompts (shown in the Appendix A.3) 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 [2] 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. We then enforce consistency across annotations, boxes, and reasoning chains. Each mentioned entity must have a valid box and timestamp; redundant or unmatched references are removed, and mismatched samples discarded. This ensures every instance contains coherent spatio-temporal evidence, with answers, timestamps, and boxes aligned to visual content.

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 framework from Qwen2.5-VL-7B [2], 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.

5

4.2 Reinforcement Learning with GSPO

We adopt Group Sequence Policy Optimization (GSPO) [54] as our reinforcement learning algorithm. Compared with GRPO [33], 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.4.

4.3 Reward Design

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

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

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:

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

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

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

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

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

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 $j^\star(m) = \arg\min_j |t_m - t_j^{\text{gt}}|$ be the nearest annotated time. Let $\mathcal{B}_m$ be predicted boxes and $\mathcal{B}_{j^\star(m)}^{\text{gt}}$ ground-truth boxes on that frame. The spatial reward is

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

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 [8] 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 [15] and VideoMMMU [19] assess *general* video QA and multimodal comprehension across diverse domains, while WorldSense [18] emphasizes integrating multimodal signals with commonsense reasoning. TVG-Bench [43] focuses on fine-grained temporal localization.

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 [28] and Gemini-2-Flash [35], which represent the current frontier of proprietary video LLMs. (ii) open-source general-purpose video understanding models, including Video-LLAMA3 [48], LLaVA-Video [53], VideoChat2 [22], Oryx-1.5-7B [25], InternVL-2.5-8B [7], and Qwen2.5-VL-7B [2]. (iii) task-specialized approaches such as TRACE [16], designed for temporal video grounding, and Sa2VA [47], 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, our model sets a new **state-of-the-art** with **mAM improved by +14.4% and mLGM by +24.2%**, surpassing GPT-4o [28] and Gemini-2-Flash [10]. These results demonstrate that our approach **brings significant advances in temporal and spatial grounding.** 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:

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	↑ 27.5	↑ 9.1	↑ 10.2	↑ 8.4	↑ 3.5	↑ 14.4	↑ 24.2

(i) closed-source commercial models such as GPT-4o [28], (ii) open-source general-purpose video LLMs including Qwen2.5-VL-7B [2], and (iii) recent reasoning-focused models such as VideoRFT-7B [41] and VideoR1-7B [14], 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. Compared 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

Table 2 Performance across different video understanding and temporal grounding benchmarks. Open-o3 Video achieves comparable or even superior results to other video reasoning models, while providing more intuitive spatio-temporal evidence.

Model	VideoMME		WorldSense		VideoMMMU		TVGBench
	Overall	Long	**Overall**	Recognition	**Overall**	Perception	mIoU
GPT-4o	71.9	-	42.6	-	61.2	66.0	-
Qwen2.5-VL-7B (base)	<u>62.4</u>	<u>50.8</u>	36.1	33.7	51.2	64.7	<u>16.3</u>
VideoRFT-7B	59.8	50.7	**38.2**	<u>36.6</u>	51.1	<u>66.0</u>	14.3
VideoR1-7B	61.4	50.6	35.5	32.8	**52.4**	65.3	9.6
Open-o3 Video (Ours)	**63.6**	**54.9**	<u>37.5</u>	**36.8**	<u>52.3</u>	**68.0**	**20.8**
Δ vs. Qwen2.5-VL-7B	↑ 1.2	↑ 4.1	↑ 1.4	↑ 3.1	↑ 1.1	↑ 3.3	↑ 4.5

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**

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* [17] 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 [26], we introduce a confidence-aware voting scheme that leverages grounded evidence to verify predictions at inference, as shown in Figure 6 in the appendix. Details, including scoring schemes, prompts, and results on WorldSense and VideoMMMU are provided in Appendix A.5.

6 Conclusion

We introduced **Open-o3 Video**, a unified framework for grounded video reasoning that generates explicit spatio-temporal evidence through timestamped frames and localized bounding boxes. With carefully curated high-quality training data, a two-stage strategy combining supervised fine-tuning and GSPO-based reinforcement learning, and novel thinking rewards incorporating adaptive temporal proximity and temporal gating, our method substantially improves answer accuracy, temporal alignment, and spatial grounding. Comprehensive experiments demonstrate that Open-o3 Video achieves state-of-the-art performance on the V-STAR benchmark, surpassing strong baselines including GPT-4o, while remaining broadly competitive across diverse video

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**

understanding tasks. In future work, we aim to further align reasoning chains across text, time, space, and audio modalities, and to extend our approach to more complex and longer video scenarios.

Ethics Statement

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

Reproducibility Statement

Comprehensive implementation details, including training procedures, hyperparameter configurations, and evaluation protocols, are provided in the main paper (Section 5) and Appendix A.1. Furthermore, upon acceptance of this paper, all source code, datasets, and trained model checkpoints will be made publicly available.

LLM Usage Statement

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

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In Proceedings of the IEEE international conference on computer vision, pages 5803–5812, 2017.

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

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the ieee conference on computer vision and pattern recognition, pages 961–970, 2015.

[4] Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. arXiv preprint arXiv:2507.18100, 2025.

[5] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188, 2024.

[6] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. arXiv preprint arXiv:2507.07966, 2025.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[8] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495, 2025.

[9] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. arXiv preprint arXiv:2504.13180, 2025.

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

[11] Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. arXiv preprint arXiv:2505.24718, 2025.

[12] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. arXiv preprint arXiv:2505.15879, 2025.

[13] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776, 2025.

[14] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776, 2025.

[15] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In CVPR, 2025.

[16] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. arXiv preprint arXiv:2410.05643, 2024.

[17] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 26181–26191, 2025.

[18] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. arXiv preprint arXiv:2502.04326, 2025.

[19] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826, 2025.

[20] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems, 34:11846–11858, 2021.

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

[22] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In CVPR, pages 22195–22206, 2024.

[23] 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. arXiv preprint arXiv:2504.06958, 2025.

[24] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. arXiv preprint arXiv:2506.18883, 2025.

[25] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. arXiv preprint arXiv:2409.12961, 2024.

[26] Jiahao Meng, Shuyang Sun, Yue Tan, Lu Qi, Yunhai Tong, Xiangtai Li, and Longyin Wen. Cyberv: Cybernetics for test-time scaling in video understanding. arXiv preprint arXiv:2506.07971, 2025.

[27] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2265–2269. IEEE, 2021.

[28] OpenAI. Hello gpt4-o. https://openai.com/index/hello-gpt-4o/, 2024.

[29] OpenAI. Openai-o3. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[30] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805, 2025.

[31] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. arXiv preprint arXiv:2506.07464, 2025.

[32] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642, 2024.

[33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[34] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1207–1216, 2019.

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

[36] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. arXiv preprint arXiv:2507.01949, 2025.

[37] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. arXiv preprint arXiv:2410.09575, 2024.

[38] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. arXiv preprint arXiv:2507.07999, 2025.

[39] Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. arXiv preprint arXiv:2504.01901, 2025.

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

[41] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. arXiv preprint arXiv:2505.12434, 2025.

[42] Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. arXiv preprint arXiv:2507.13353, 2025.

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

[44] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. arXiv preprint arXiv:2507.06485, 2025.

[45] Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. arXiv preprint arXiv:2508.20478, 2025.

[46] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840, 2024.

[47] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv, 2025.

[48] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106, 2025.

[49] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

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

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

[52] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641, 2025.

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

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

[55] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362, 2025.

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

Appendix

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 presents the prompts used for data annotation with Gemini. Section A.4 provides the full mathematical formulation of GSPO algorithm. Section A.5 details the confidence-aware test-time scaling procedure and reports additional results. Section A.6 provides further qualitative visualizations of spatio-temporal reasoning. Finally, Section A.7 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 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.

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**

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.

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.

Table 7 Test-time scaling results on WorldSense and VideoMMMU, 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	VideoMMMU
Base	37.5	52.3
Majority Voting	37.3	53.1
Confidence-aware Voting	**38.5**	**54.1**

A.4 Details of GSPO Training

For completeness, we provide the full formulation of Group Sequence Policy Optimization (GSPO) [54], 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

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

The importance ratio is defined at the sequence level as

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

where $|y_i|$ denotes the response length.

The GSPO objective is then

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

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.5 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 6. 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 `<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 7, confidence-aware voting consistently improves over *naive majority voting*, achieving +1.0 on WorldSense and +1.0 on VideoMMMU. 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.6 More Visualizations.

As shown in Figure 7,8,9, 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.

A.7 Limitations and Future Work.

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

Figure 5 Annotation Prompt for Temporal Grounding Data Source.

Figure 6 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.

Figure 7 Visualization. On simple appearance perception tasks, both our model and related baselines can provide correct answers; however, our approach additionally offers explicit spatio-temporal evidence.

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.

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