Annotations as Rollouts: Efficient and Scalable Reinforcement Learning for Video MLLMs

Yunheng Li[1], **Guohong Mu**[1], **Hao Li**[2], **Shengsheng Qian**[3], **Dingwen Zhang**[2], **Qibin Hou**[†1,4], **Ming-Ming Cheng**[1,4]

[1]VCIP, School of Computer Science, Nankai University
[2]Brain and Artificial Intelligence Lab, Northwestern Polytechnical University
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[4]NKIARI, Futian, Shenzhen
[†]Corresponding author.

Abstract: Multimodal large language models (MLLMs) have become a prevailing paradigm for unified video perception. However, post-training on large multi-task datasets remains challenging, as existing reinforcement learning methods sample on-policy groups with few high-quality rollouts even with costly chain-of-thought (CoT) generation. In this paper, we study the sample efficiency and scalability of RL post-training for video MLLMs and introduce OraRL. We identify an overlooked role for annotations: Beyond scoring rollouts, each can enter its on-policy group as an oracle rollout, a direct positive optimization target. Direct oracle integration, however, is nontrivial: a high-reward oracle raises the group baseline and inverts otherwise positive policy advantages, a failure we term advantage inversion. At the core of OraRL is a decoupled advantage estimator: policy rollouts determine an oracle-free baseline, while the oracle-policy gap modulates both a directional gain and a separate detached oracle advantage. Sign-balanced pruning improves efficiency: by retaining only the oracle and the strongest rollouts of each sign, OraRL requires just 2.2× the step time of SFT, less than half the 4.9× required by GRPO with CoT. OraRL scales with model size and data, surpassing its backbone from 0.8B to 9B and GRPO up to 100k prompts. Without chain-of-thought, Video-ORA-9B decodes in 130 ms instead of 4,780 ms. Compared with the respective prior best models, it raises temporal mIoU from 62.5 to 66.0, tracking AO from 73.0 to 78.2, segmentation from 64.3 to 70.4, and the three-benchmark spatial-intelligence macro average from 51.0 to 56.1; on VSI-Bench, it scores 73.1 against 55.0 for GPT-5 and 55.1 for Gemini-3-Pro.

Project: https://orarl.github.io/
Data: https://huggingface.co/datasets/OraRL/OraRL-Data
Model: https://huggingface.co/OraRL/models
Code: https://github.com/HVision-NKU/OraRL
Date: August 24, 2026

Keywords: Multimodal large language models, unified video perception, reinforcement learning.

HVision@Nankai

1 Introduction

Unified video perception is a key capability for physical intelligence [8], extending beyond coarse textual descriptions to precise temporal localization [58, 108], spatial grounding and segmentation [46, 103], object tracking [72], and spatial understanding [56, 96]. Recent generalist multimodal large language models (MLLMs), including LLaVA-OneVision-2 [1], Molmo2 [12], InternVideo3 [91], and VideoChat3 [44], increasingly integrate multiple such capabilities within a single model, yet still trail task-specific models in fine-grained perception. The largest proprietary models, including GPT-5 [67], Gemini-3-Pro [26], and Seed-2.0 [63], underperform open-source specialists with fewer than 10B parameters on temporal grounding [108, 118] and spatial intelligence [96, 113], suggesting that fine-grained precision is determined primarily by task-aligned post-training rather than by model scale. Scaling post-training data alone does not close this gap, because the central bottleneck is sample efficiency: under existing paradigms, additional training prompts yield only marginal, task-dependent gains and can even degrade performance.

This limitation is particularly pronounced in supervised fine-tuning (SFT) [60, 29, 46, 73, 38, 27], which uses each annotation as a maximum-likelihood target. This objective enforces the required output format but provides no task-level supervision to distinguish near-correct predictions from clearly incorrect ones. Recent video RL methods [20, 78, 43], often based on group relative policy optimization (GRPO) [65], address this limitation by comparing task rewards across multiple on-policy rollouts, but each annotation serves only as a scoring reference. Yet on-policy rollouts rarely recover the precise intervals, boxes, masks, or trajectories specified by these annotations, leaving many groups without a reliable positive anchor. Chain-of-thought (CoT) reasoning does not alleviate this scarcity; it lengthens every rollout, increasing training and inference costs without clear performance gains.

Instead of incorporating CoT, we use the annotation itself as a reliable positive target for policy optimization, a role overlooked by existing video RL methods. This principle, which we term annotation-as-rollout, is task-independent: each annotation is se-

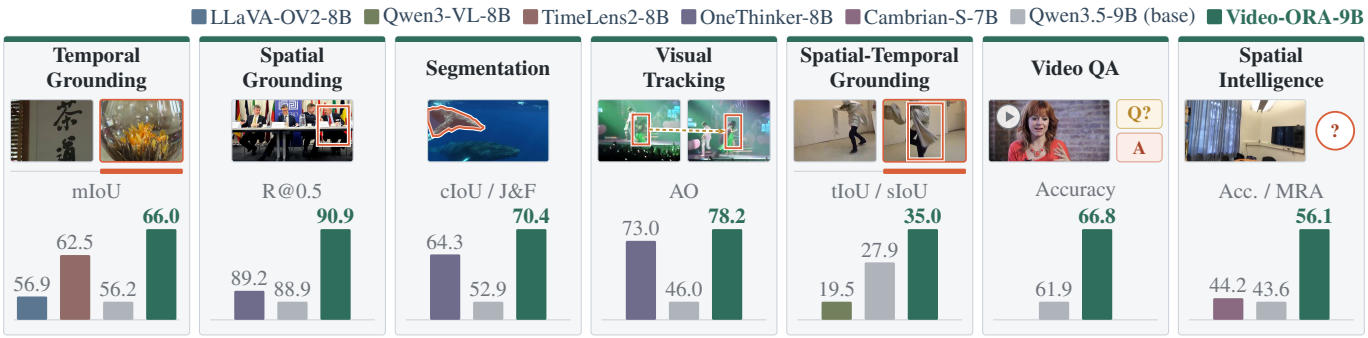


Figure 1 Video-ORA: one model, trained by a single OraRL recipe, for unified video perception. Each column presents a representative input and a family-level score computed over common benchmark coverage. Video-ORA-9B, evaluated without chain-of-thought decoding, outperforms every displayed baseline [71, 118, 21, 2, 96, 1] across seven task families. See Sec. 4 for detailed metrics.

rialized into the model's response format and appended to the on-policy group as an additional oracle rollout. However, standard normalization of this mixed-policy group includes the high-reward oracle in the advantage baseline, raising the threshold for a positive advantage. Consequently, on-policy rollouts with rewards above the on-policy mean may be assigned negative advantages, a phenomenon we term advantage inversion (Fig. 2(c)). These inverted advantages suppress the high-quality rollouts that should be reinforced, and in the worst case the entire on-policy group, collapsing learning toward oracle imitation while suppressing all exploratory rollouts and causing naive oracle augmentation to underperform standard GRPO, as analyzed in Sec. 4.10.

In this paper, we introduce a new RL paradigm, named OraRL, which excludes the oracle from the advantage baseline while retaining it as a detached optimization target, ensuring that advantage signs are determined solely by comparisons among on-policy rollouts (Fig. 2(d)). Because excluding the oracle removes the oracle-policy gap from policy-relative advantage estimation, OraRL encodes it in two update terms: a directional gain that increases with the gap to amplify above-average on-policy rollouts, and a weight on the detached oracle update that decays as the gap closes. To improve training efficiency, OraRL back-propagates through a sign-balanced subset that always includes the oracle, preserving the sign contrast that drives the update, whereas magnitude-only pruning [48] can retain rollouts of a single sign and push every retained rollout in the same direction. OraRL then re-centers and rescales the retained advantages to prevent selection from biasing or amplifying the policy update, while selective back-propagation yields a $1.48\times$ speedup over full-group optimization.

Our earlier Tempsamp-R1 [45] mitigated the same degradation with a hand-designed reward-shaping transform, which depended on task-specific reward semantics and left advantage inversion unexplained. Computing advantages from on-policy rewards and the oracle-policy gap removes this dependence, so one update rule handles any annotation expressed in the model's response format.

We train our new model Video-ORA with OraRL and evaluate it across the seven task families summarized in Fig. 1. Without CoT decoding, Video-ORA-9B achieves the best mIoU on all three TimeLens benchmarks (61.8, 63.6, and 72.5) and the best AO on GOT-10k (78.2). It also obtains leading segmentation scores on RefCOCO (79.4 cIoU) and MeViS (61.3 J&F), while ranking first on all eight RefCOCO comprehension splits and all four STVG metrics. Beyond these perception tasks, it ranks first among open-source models on five of seven Video QA benchmarks with a 66.8 macro average and on both MMSI-Bench and MindCube, while its VSI-Bench average of 73.1 is the best reported overall. OraRL improves over the corresponding backbone at all four scales from 0.8B to 9B and outperforms GRPO at every evaluated data budget up to 100k prompts. Answer-only generation also reduces median end-to-end latency on ten-minute videos from 29.0 to 24.3 seconds relative to the CoT-enabled backbone. Together, these results establish annotation-as-rollout as an efficient and scalable reinforcement learning principle for unified video perception. Our contributions are summarized as follows:

- We introduce annotation-as-rollout, a task-independent mechanism that converts each annotation into an oracle rollout, providing reliable positive supervision without requiring CoT or sacrificing on-policy exploration.
- We identify advantage inversion in naive oracle mixing and address it by separating policy advantages from oracle guidance. Sign-balanced pruning retains both advantage signs and, with post-selection moment correction, yields a $1.48\times$ speedup.
- We develop Video-ORA from 0.8B to 9B, achieving leading results across seven task families and consistent gains across data budgets and backbone families, together with improved training and inference efficiency.

2 Related Work

Multimodal foundation models for video understanding. Multimodal large language models (MLLMs) typically connect a visual encoder to a language model through lightweight projection or cross-attention modules [87, 107, 37, 2]. Video MLLMs extend this architecture by representing sampled frames as visual token sequences, as exemplified by Video-ChatGPT [53], the Video-LLaMA series [106, 105], Chat-UniVi [31], the VideoChat series [40, 42, 44], and the LLaVA-OneVision and LLaVA-Video families [37, 112, 1]. Beyond architectural integration, large-scale video-text pretraining has produced dedicated video foundation models, from InternVideo [77] to InternVideo3 [91]. General open MLLMs, including Qwen3-VL [2], Qwen3.5 [71], InternVL3 [117], InternVL3.5 [76], Molmo2 [12], Keye-VL [70], MiniCPM-V [102], MiMo-VL [86], and Eagle2.5 [7], further

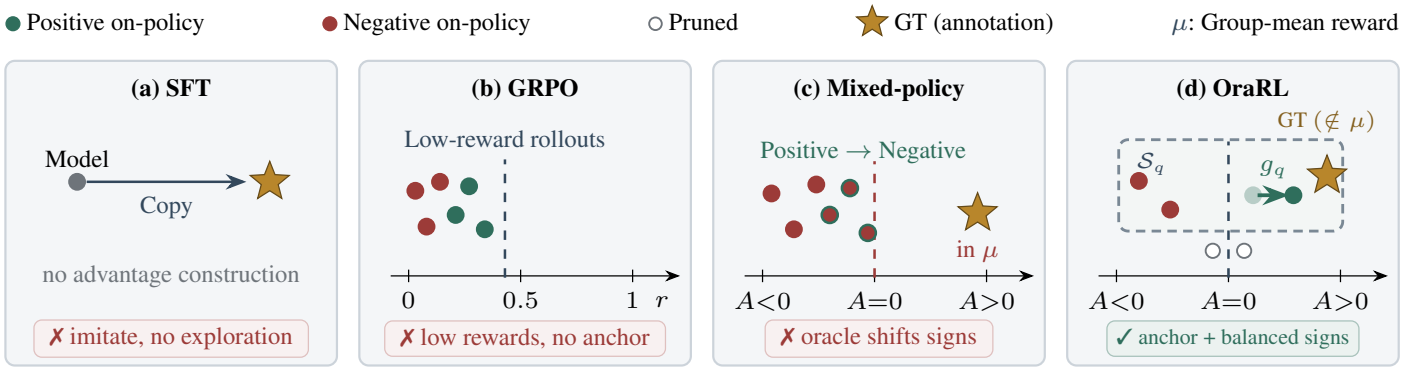


Figure 2 Four paradigms for using annotations in model adaptation. Panel (b) shows rewards, whereas panels (c) and (d) show advantages. Unlike mixed-policy normalization, OraRL excludes GT from the on-policy baseline while retaining it as an optimization target, preventing advantage inversion and enabling a sign-balanced update.

broaden image and video understanding within unified architectures. Long-video understanding has additionally motivated memory mechanisms and visual-token compression, as surveyed in [68]. Proprietary systems such as GPT-5 [67], Gemini [13, 26], Grok-4 [85], and Seed-2.0 [63] demonstrate strong capabilities in open-ended multimodal understanding. Collectively, these advances [44, 1, 91] have strengthened open-ended video understanding, but fine-grained temporal and spatial perception remains a distinct post-training challenge.

Fine-grained video perception. Temporal grounding and highlight detection localize a language query in time and have been studied extensively before the MLLM era [33, 24, 110, 36]. Transferring this ability to MLLMs has been dominated by supervised fine-tuning, from TimeChat [60], VTimeLLM [29], GroundingGPT [46], and VTG-LLM [27] to Grounded-VideoLLM [73], LLaVA-ST [38], and the TimeLens generalists [108, 118]. Fine-grained spatial perception has developed in parallel through reasoning segmentation and referring video object segmentation [18, 35, 61, 59, 89, 84, 3, 47, 50, 103, 16], as well as visual tracking, where transformer specialists [9, 14] remain the standard and MLLM-based trackers have appeared only recently [72]. More recently, research on spatial intelligence has focused on inferring metric and relational structure from visual observations [96, 56, 113, 81, 94, 109, 83, 8]. Although unified models such as OneThinker [21] cover many of these tasks, they rely on CoT supervision, whereas we train all seven families with answer-only rollouts under a single annotation interface and update rule.

Reinforcement learning for multimodal models. Reinforcement learning is widely used to adapt language and multimodal models beyond supervised imitation, with PPO [62] providing a standard policy-optimization foundation. GRPO [65] estimates advantages by normalizing rewards within a group of rollouts sampled for the same query, without requiring a learned critic. This formulation has been adopted for visual and video reasoning [20, 104], with task-specific variants for temporal grounding [78, 43], referring expression comprehension [100, 6], segmentation [99, 88, 25], and tracking [72]. Across these methods [78, 100, 99, 72], annotations serve only as reward references for sampled rollouts, so the learning signal remains bounded by the quality of the on-policy group. To strengthen weak groups, LUFFY [90] injects teacher traces with regularized importance sampling, and on-policy distillation has been applied to temporal grounding [39], whereas our oracle comes directly from the paired annotation and needs no teacher. Adding it to the group, however, shifts the baseline and inverts useful on-policy advantages (Sec. 4.10), which importance weighting cannot correct because all on-policy ratios equal one, so OraRL keeps the oracle as an optimization target while excluding it from the baseline. Group-based RL is also costly because each prompt needs several rollouts, especially with CoT [79], and while CPPO [48] prunes rollouts with low absolute advantages and reward shaping [28, 52] enriches a fixed budget, OraRL retains the oracle with equal numbers of positive and negative rollouts so the pruned group keeps its sign contrast.

3 Methodology

As shown in Fig. 3, OraRL appends each annotation as an oracle rollout to its corresponding on-policy group, providing a reliable positive target for every query. Standard group normalization, however, incorporates the high-reward oracle into the advantage baseline, raising the threshold for a positive advantage and causing above-average on-policy rollouts to be penalized rather than reinforced. OraRL avoids this by computing the baseline from on-policy rewards alone and encoding the oracle-policy gap in two terms: a directional gain for above-average rollouts and a separate, bounded oracle advantage. For efficiency, it back-propagates through a sign-balanced subset containing the oracle, then recenters and rescales the selected advantages.

3.1 Preliminaries

Let $q = (v, x)$ denote a multimodal query containing a video (or image) v and an instruction x. Given q, group relative policy optimization (GRPO) samples n rollouts

$$\mathcal{O}_{\text{op}} = \{o_i\}_{i=1}^{n}, \qquad o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid q), \tag{1}$$

and evaluates each rollout with a task reward $r_i = R(o_i, q)$. Without requiring a learned critic, vanilla GRPO estimates a group-relative advantage as

$$A_i^{\text{GRPO}} = \frac{r_i - \mu_{\text{grp}}}{\sigma_{\text{grp}} + \epsilon}, \qquad \mu_{\text{grp}} = \frac{1}{n}\sum_{j=1}^{n} r_j, \tag{2}$$

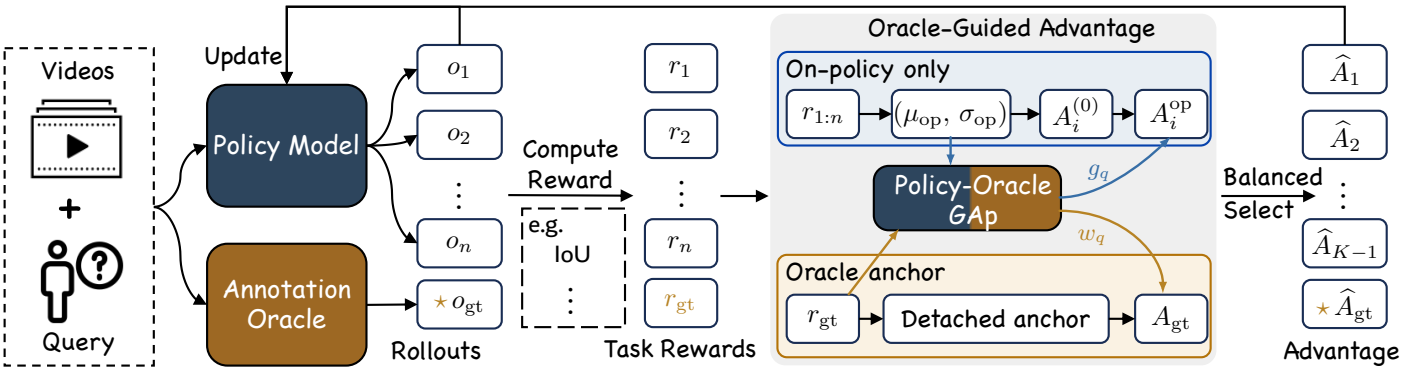


Figure 3 Overview of OraRL. The annotation is appended as an oracle rollout to n on-policy rollouts. On-policy rewards define the baseline, while the oracle separately strengthens promising rollouts and provides an adaptive anchor. Sign-balanced pruning retains informative positive and negative rollouts together with the oracle, yielding K rollouts for the policy update.

where σ_{grp} is the standard deviation of $\{r_j\}_{j=1}^{n}$ and ϵ is a small constant used throughout for numerical stability. For response token $o_{i,t}$, define the importance ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{3}$$

The clipped policy objective is

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(\rho_{i,t}A_i^{\text{GRPO}},\ \bar{\rho}_{i,t}A_i^{\text{GRPO}}\right), \tag{4}$$

where $\bar{\rho}_{i,t}$ denotes $\rho_{i,t}$ clipped to $[1-\epsilon_{\text{c}}, 1+\epsilon_{\text{c}}]$. OraRL retains this objective unchanged, modifying only rollout-group composition, advantage estimation, and rollout selection for policy updates.

3.2 OraRL

Annotation-as-rollout construction. Let y denote the annotation associated with q and T_{task} the transform that serializes it into the model's response format, so that the oracle rollout and the augmented group are

$$o_{\text{gt}} = T_{\text{task}}(y), \quad \mathcal{O}_{\text{aug}} = \mathcal{O}_{\text{op}} \cup \{o_{\text{gt}}\}, \quad |\mathcal{O}_{\text{aug}}| = n+1. \tag{5}$$

Appending the oracle rather than replacing an on-policy rollout preserves all n on-policy rollouts and the policy-relative comparisons among them, so the oracle adds supervision without reducing exploration. Depending on the task, o_{gt} encodes an answer choice, temporal interval, spatial box, box trajectory, interval with sampled boxes, or timestamped box-and-point segmentation prompt.

Advantage inversion under mixed-group normalization. Let $\mu_{\text{op}} = \frac{1}{n}\sum_{i=1}^{n} r_i$ be the on-policy reward mean and let r_{gt} be the oracle reward. If all $n+1$ rollouts are normalized together, the group mean becomes

$$\mu_{\text{aug}} = \frac{n\mu_{\text{op}} + r_{\text{gt}}}{n+1} = \mu_{\text{op}} + \frac{r_{\text{gt}} - \mu_{\text{op}}}{n+1}. \tag{6}$$

When $r_{\text{gt}} > \mu_{\text{op}}$, the oracle raises the baseline, and every on-policy rollout satisfying $\mu_{\text{op}} < r_i < \mu_{\text{aug}}$ outperforms the current policy on average but receives a negative advantage. Let A_i^{mix} denote the mixed-group advantage. For these rollouts, $A_i^{\text{GRPO}} > 0$ while $A_i^{\text{mix}} < 0$, which constitutes the advantage inversion quantified in Sec. 4.10. The inversion band has width $(r_{\text{gt}} - \mu_{\text{op}})/(n+1)$ and therefore grows linearly with the oracle-policy reward gap. Moreover, the oracle increases the mixed-group standard deviation, reducing the magnitudes of normalized on-policy advantages and their gradient contributions. OraRL avoids both effects by centering advantages on the on-policy mean without variance normalization and encoding the oracle-policy gap through separate scaling terms, as detailed below.

On-policy advantage estimation. To preserve policy-relative comparisons, OraRL defines $A_i^{(0)} = r_i - \mu_{\text{op}}$ for $i = 1, \ldots, n$, without variance normalization. No rollout that outperforms the on-policy mean can therefore receive a negative advantage, so this inversion interval is empty by construction. To retain information about the oracle-policy gap, we compute the on-policy and augmented reward dispersions as $\sigma_{\text{op}} = \text{Std}(\{r_i\}_{i=1}^{n})$ and $\sigma_{\text{aug}} = \text{Std}(\{r_i\}_{i=1}^{n} \cup \{r_{\text{gt}}\})$, respectively. These quantities enter only the directional scaling introduced next.

Oracle-gap directional gain. The base advantage $A_i^{(0)}$ preserves policy-relative comparisons but not the oracle–policy discrepancy. We estimate this discrepancy from the change in reward dispersion with the bounded gain:

$$g_q = \text{clip}\left[\left(\frac{\sigma_{\text{aug}}}{\sigma_{\text{op}} + \epsilon}\right)^{1/4}, 1, 4\right]. \tag{7}$$

Since σ_{aug} includes the oracle, g_q grows as its reward deviates from the on-policy distribution. Clipping preserves the base scale and prevents excessive amplification when σ_{op} is small. We apply g_q only to above-mean rollouts:

$$U_i = \begin{cases} g_q A_i^{(0)}, & A_i^{(0)} > 0, \\ A_i^{(0)}, & A_i^{(0)} \le 0. \end{cases} \tag{8}$$

This transform increases the advantages of rollouts above the on-policy mean without amplifying those below it. Because the asymmetric scaling generally shifts the group mean, we remove this shift by setting $A_i^{\text{op}} = U_i - n^{-1}\sum_{j=1}^{n} U_j$. By construction, $\sum_i A_i^{\text{op}} = 0$, while the increased separation between rollouts above and below the on-policy mean is preserved.

Detached oracle advantage. Rather than using the raw difference $r_{\mathrm{gt}} - \mu_{\mathrm{op}}$ directly, OraRL derives a detached oracle advantage whose scale is calibrated to both the remaining oracle-policy gap and the strongest useful on-policy signal. We first define the normalized reward-gap weight

$$w_q = \left[\mathrm{clip}\left(\frac{r_{\mathrm{gt}} - \mu_{\mathrm{op}}}{r_{\mathrm{gt}} + \epsilon}, 0,\, 1\right)\right]^2 . \tag{9}$$

The weight becomes zero when the on-policy mean reaches the oracle reward and approaches one as the mean decreases relative to that reward. It therefore measures the residual supervision provided by the annotation, while the exponent sharpens its decay as the policy improves. We then calibrate this gap-dependent scale against the strongest positive on-policy advantage, $A^{+}_{\max} = \max\big(0, \max_i A_i^{\mathrm{op}}\big)$:

$$A_{\mathrm{gt}} = \min\big(2w_q,\ \mathrm{clip}\big(1.2\, A^{+}_{\max},\, 0.05,\, 1\big)\big) . \tag{10}$$

The first term sets the nominal oracle scale, whereas the second caps it relative to the strongest useful on-policy signal, preventing the oracle from dominating the group. If no positive on-policy rollout exists, the cap is 0.05, yielding a small bootstrap signal when $w_q > 0$ and vanishing when $w_q = 0$. Finally, g_q is not applied to A_{gt} because both terms respond to the oracle-policy discrepancy and would double-count the same correction.

Sign-balanced advantage pruning. To reduce update cost, OraRL computes all on-policy and oracle advantages before pruning, then performs the policy forward and backward passes only on the retained subset. Pruning therefore leaves the pre-correction advantages unchanged. We retain $K = \lfloor n(1-\kappa) \rfloor$ rollouts, where $\kappa \in [0, 1)$ is chosen such that $K \geq 1$ and the budget is defined relative to the original n on-policy rollouts. Let s_i denote the sequence-level advantage of rollout o_i, computed as the masked mean over its valid response tokens. We define the positive and negative candidate sets as $\mathcal{O}^{+} = \{o_i : s_i > 0\}$ and $\mathcal{O}^{-} = \{o_i : s_i < 0\}$. The retained set is

$$\mathcal{S}_q = \{o_{\mathrm{gt}}\} \cup \mathrm{Top}_{K_+}\big(\mathcal{O}^{+},\, |s_i|\big) \cup \mathrm{Top}_{K_-}\big(\mathcal{O}^{-},\, |s_i|\big) , \tag{11}$$

where $1 + K_+ + K_- = K$. The oracle is always retained and treated as a positive anchor when allocating the quota. The remaining $K - 1$ positions are divided as evenly as possible between positive and negative rollouts, with candidates ranked by $|s_i|$ within each sign. If one sign has insufficient candidates, its unused positions are reassigned to the other sign, and zero-advantage rollouts are selected only when necessary. For $n = 8$ and $\kappa = 0.5$, the retained set contains the oracle, one positive on-policy rollout, and two negative rollouts. Unlike magnitude-only pruning in CPPO [48], which may select only one sign, the sign quota preserves both reinforcing and suppressive signals when available.

Post-selection moment correction. Sign-balanced pruning changes the mean and scale of the retained advantages because it always keeps the positive oracle and favors large-magnitude rollouts. Without correction, the reduced group can therefore produce a nonzero mean and a disproportionate update scale. Let a_i denote the pre-correction advantages of the K rollouts in $\mathcal{S}_q$. We first restore a zero mean by setting $z_i = a_i - K^{-1} \sum_{\ell \in \mathcal{S}_q} a_\ell$. If $z_{\mathrm{gt}} < 0$, centering would penalize the known-correct oracle. We therefore project the centered vector onto the constraints $\sum_i \widetilde{z}_i = 0$ and $\widetilde{z}_{\mathrm{gt}} \geq 0$:

$$\widetilde{z}_i = \begin{cases} 0, & i = \mathrm{gt}, \\ z_i + \dfrac{z_{\mathrm{gt}}}{K-1}, & i \neq \mathrm{gt}. \end{cases} \tag{12}$$

This projection sets the oracle advantage to zero and distributes its offset uniformly across the remaining rollouts while preserving the zero sum. When $z_{\mathrm{gt}} \geq 0$, we simply set $\widetilde{z}_i = z_i$. We next control the retained scale using the pre-pruning on-policy RMS, $\mathrm{RMS}_{\mathrm{op}} = \sqrt{n^{-1} \sum_{i=1}^{n} (A_i^{\mathrm{op}})^2}$, and the selected RMS, $\mathrm{RMS}_{\mathcal{S}} = \sqrt{K^{-1} \sum_i \widetilde{z}_i^2}$. The corrected advantages are $\widehat{A}_i = \lambda_q \widetilde{z}_i$, where

$$\lambda_q = \mathrm{clip}\left(\frac{\mathrm{RMS}_{\mathrm{op}}}{\mathrm{RMS}_{\mathcal{S}} + \epsilon},\, 0.25,\, 1\right) . \tag{13}$$

The ratio matches the pre-pruning RMS whenever it lies within the clipping range. The upper bound prevents amplification, while the lower bound limits attenuation to a factor of four. We skip scale matching when σ_{op} is numerically zero because the on-policy RMS then provides no meaningful reference. Overall, the correction restores a zero mean, keeps the oracle nonnegative, and controls the update scale relative to the full on-policy group.

Unified policy update. We evaluate each rollout in the augmented group, including the serialized oracle, under the frozen pre-update policy to obtain its old token log-probabilities. Let $N_q = \sum_{i \in \mathcal{S}_q} |o_i|$ be the retained response-token count. Broadcasting $\widehat{A}_i$ to these tokens gives

$$\mathcal{J}_{\mathrm{ours}}(\theta) = \frac{1}{N_q} \sum_{i \in \mathcal{S}_q} \sum_{t=1}^{|o_i|} \min\Big(\rho_{i,t} \widehat{A}_i,\, \bar{\rho}_{i,t} \widehat{A}_i\Big) . \tag{14}$$

The objective increases the likelihood of the oracle and positive policy rollouts while reducing that of negative ones.

3.3 Oracle Serialization and Task Rewards

Each task adapter serializes annotation y as the oracle rollout o_{gt} and evaluates rollout o with the scalar score $R_k(o, y)$. Let $F_k(o) \in \{0, 1\}$ indicate output validity. Thus, $R(o, q) = R_k(o, y) F_k(o)$, with malformed outputs assigned zero. We instantiate adapters for temporal and highlight grounding, spatial and spatial-temporal grounding, visual tracking, segmentation, video question answering, and spatial intelligence. Video question answering and spatial intelligence use textual answers rather than geometric predictions, showing that annotation-as-rollout also applies beyond localization. During reinforcement learning, each rollout group uses one adapter, so normalization never mixes task reward scales. Appendix C specifies all oracle serializations and task scores.

4 Experiments

4.1 Experimental Setup

Training data. Supervised fine-tuning and reinforcement learning use 284,779 and 100,032 distinct prompts, respectively, from the same seven task families. Answer-only tasks comprise 82% of the supervised mixture, whereas structured tasks comprise 63% of the reinforcement learning mixture.

Table 1 Temporal grounding results. We report R1@0.3, R1@0.5, R1@0.7, and mIoU on three TimeLens benchmarks.

Model	Charades-TimeLens [24, 108]				ActivityNet-TimeLens [33, 108]				QVHighlights-TimeLens [36, 108]			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Proprietary Models												
GPT-5 [67]	59.3	42.0	22.0	40.5	57.4	44.9	30.4	42.9	72.4	60.4	46.4	56.8
Gemini-2.5-Pro [13]	74.1	61.1	34.0	52.8	72.3	64.2	47.1	58.1	84.1	75.9	61.1	70.4
Open-Source Models												
Qwen3.5-35B-A3B [71]	69.5	50.4	27.3	48.2	59.4	58.6	40.1	52.6	81.6	72.6	56.5	66.0
Qwen3-VL-30B-A3B [2]	70.3	46.5	25.1	48.1	61.7	51.1	36.5	49.4	79.3	67.6	52.6	63.2
Keye-VL-2.0-30B-A3B [70]	-	-	-	58.4	-	-	-	58.5	-	-	-	70.1
VideoChat-Flash-7B [42]	60.2	37.9	17.8	39.7	35.5	21.8	10.5	24.8	45.2	30.6	16.7	32.7
Time-R1-7B [78]	57.9	32.0	16.9	36.6	44.8	31.0	19.0	33.1	65.8	51.5	36.1	49.2
Tempsamp-R1-7B [45]	59.1	37.3	16.2	39.8	48.3	33.2	18.6	34.4	75.6	60.6	41.1	55.1
MiMo-VL-7B [86]	57.9	42.6	20.5	39.6	49.3	38.7	22.4	35.5	57.1	42.6	28.4	41.5
Video-O3-7B [104]	58.7	34.2	16.2	38.6	53.2	41.0	24.4	38.9	62.7	49.5	34.4	47.4
Molmo2-4B [12]	44.4	31.1	18.5	34.7	50.8	40.3	29.9	40.9	73.7	62.6	51.8	60.8
VideoChat3-4B [44]	78.4	64.9	35.9	56.1	70.1	60.5	42.2	54.6	81.0	72.5	56.8	67.0
Qwen3-VL-8B [2]	69.2	53.4	27.5	48.3	62.1	51.2	34.4	46.8	74.2	64.6	49.3	59.4
InternVideo3-8B [91]	75.8	61.7	32.8	53.2	60.0	51.4	33.4	46.3	72.2	62.6	48.9	59.2
Video-OPD-8B [39]	73.1	45.8	32.4	-	60.5	45.6	35.8	-	73.8	60.3	50.4	-
LLaVA-OneVision-2-8B [1]	73.1	59.6	34.0	52.6	65.6	57.9	40.9	52.4	78.2	70.0	57.0	65.7
TimeLens2-8B [118]	80.9	68.0	38.9	58.6	74.0	65.6	46.6	58.6	84.4	76.1	61.2	70.2
Qwen3.5-4B [71]	73.0	53.5	26.7	49.4	68.9	57.9	38.8	51.9	80.7	71.7	55.3	64.5
Qwen3.5-9B [71]	73.0	57.3	30.2	50.6	69.8	59.6	38.7	52.2	80.1	72.7	56.7	65.8
Video-ORA-4B	77.1	64.9	37.9	56.8	69.3	60.1	43.8	57.3	81.4	72.6	58.6	67.5
Video-ORA-9B	**84.5**	**71.6**	**42.4**	**61.8**	**77.4**	**69.5**	**53.3**	**63.6**	**84.9**	**77.6**	**64.0**	**72.5**

Training protocol and system optimizations. All controlled variants share the same initialization, training data, and optimization budget, isolating algorithmic differences. We implement OraRL in veRL [66] with vLLM [34] for rollout generation and FSDP [114] for training. Task queues are interleaved to balance exposure, and each rollout group uses one adapter and reward function. Our pipeline reduces preprocessing overhead by decoding each video once and sharing its frames and temporal metadata between the rollout and actor engines. With $n = 8$ and $\kappa = 0.5$, it compacts each nine-rollout group to four retained rollouts and balances sequence lengths across data-parallel ranks, reducing backward computation. Rewards are evaluated asynchronously; the single-update protocol reuses detached old log-probabilities from the pre-update forward pass, avoiding a second old-policy evaluation.

Benchmarks and metrics. Controlled analyses cover three tasks: temporal grounding on the TimeLens [108] splits of Charades-STA [24], ActivityNet Captions [33], and QVHighlights [36], measured by mIoU; visual tracking on GOT-10k [30], measured by AO; and mask-aware video segmentation on MeViS [17] and ReasonVOS [3], measured by J&F. The unified evaluation adds spatial grounding and image segmentation on the RefCOCO series [101, 54], reported by R@0.5 and cIoU; spatial-temporal grounding on STVG [38], reported by tIoU@0.5, mean tIoU, sIoU@0.5, and mean sIoU; seven Video QA benchmarks [22, 23, 41, 115, 11, 82, 116], reported by accuracy; and spatial intelligence on VSI-Bench [93], MindCube [75], and MMSI-Bench [95], reported by exact-match accuracy and, for numerical VSI questions, mean relative accuracy. Unless stated otherwise, Video-ORA generates answers directly, without chain-of-thought decoding. Boldface marks the best result, underlining the second best, and "–" an unreported result.

4.2 Main Results

Temporal grounding. Tab. 1 reports the three TimeLens splits, where Video-ORA-9B leads every metric, improving mIoU over the temporal specialist TimeLens2-8B by 2.3 to 5.0 points and exceeding Gemini-2.5-Pro on all three. The margin over TimeLens2-8B is not uniform across thresholds: on ActivityNet and QVHighlights it widens as the required overlap tightens, reaching 6.7 points at R1@0.7 on ActivityNet. This is the pattern that annotation-as-rollout predicts, since coarse retrieval of the right region is already within reach of the backbone, whereas an exactly correct boundary is what on-policy sampling rarely produces and what the annotation supplies in every group. The gains also hold at 4B: Video-ORA-4B ranks second in mIoU among open models below 10B, behind only TimeLens2-8B.

Referring grounding. Video-ORA-9B ranks first in R@0.5 on all eight validation and test splits of RefCOCO, RefCOCO+, and RefCOCOg in Tab. 2. It exceeds the strongest baseline by 0.5 to 2.7 points, with the largest margin on RefCOCO+ validation. Because RefCOCO+ excludes absolute location words, the larger gains are consistent with stronger grounding from visual appearance and object relations rather than coarse positional information.

Video question answering. For multiple-choice Video QA, the annotated option serves as the oracle rollout and is rewarded by exact match. Among the open-source models reported in Tab. 3, Video-ORA-9B ranks first on five of seven benchmarks and improves the macro average over its Qwen3.5-9B backbone from

Table 2 Spatial grounding results on the RefCOCO family. All values are R@0.5 on RefCOCO, RefCOCO+, and RefCOCOg.

Models	RefCOCO [101]			RefCOCO+ [101]			RefCOCOg [54]	
	testA	testB	val	testA	testB	val	test	val
Perception-R1 [100]	91.4	84.5	89.1	86.8	74.3	81.7	85.4	85.7
Ground-R1-7B [6]	93.9	88.0	92.9	90.8	78.8	86.5	90.2	90.1
Game-R1-7B [57]	92.3	84.7	90.1	89.6	77.3	84.3	86.6	88.1
OneThinker-8B [21]	93.7	88.9	92.0	91.4	82.7	87.0	88.8	89.2
Qwen3.5-4B [71]	91.1	87.2	89.6	86.2	77.5	82.9	89.0	87.3
Qwen3.5-9B [71]	93.6	89.2	92.1	90.4	82.1	86.7	87.3	89.9
Video-ORA-4B	93.0	89.2	91.9	90.2	81.7	87.0	90.6	89.1
Video-ORA-9B	**94.6**	**90.6**	**93.4**	**92.7**	**84.4**	**89.7**	**91.0**	**90.8**

Table 3 Video question answering results. We evaluate VideoMME [22], VideoMME-v2 [23], MV-Bench [41], MMVU [115], VideoHolmes [11], LongVideoBench [82], and MLVU [116]. LongVideoBench and MLVU-MC use validation and development M-Avg., respectively.

Models	VideoMME	VideoMME-v2	MV-Bench	MMVU(mc)	VideoHolmes	LongVideoBench	MLVU-MC
InternVL3.5-8B [76]	66.0	26.0	72.1	–	–	62.1	70.2
Keye-VL-1.5-8B [92]	73.0	23.8	56.9	–	–	66.0	75.0
Eagle2.5-8B [7]	72.4	24.9	74.8	–	–	66.4	77.6
MiniCPM-V-4.5-8B [102]	67.9	26.5	60.5	–	–	63.9	70.2
InternVideo3-8B [91]	73.8	27.6	**75.0**	–	–	66.8	77.3
LLaVA-OneVision-2-8B [1]	71.9	19.9	66.2	–	–	**66.9**	76.6
Qwen3.5-9B [71]	71.1	29.9	68.2	72.5	50.3	64.6	76.7
Video-ORA-4B	69.8	26.1	67.2	65.4	61.9	60.2	73.2
Video-ORA-9B	**76.7**	**32.0**	72.5	**77.0**	**65.5**	66.2	**77.7**

Table 4 Tracking results on GOT-10k. We report average overlap (AO) and recall at IoU thresholds 0.3, 0.5, and 0.7.

Models	Frame	GOT-10k [30]			
		AO	R@0.3	R@0.5	R@0.7
Qwen3-VL-8B [2]	32	33.7	51.1	28.9	10.6
Qwen3.5-4B [71]	32	45.3	65.5	46.4	25.2
Qwen3.5-9B [71]	32	46.0	66.9	46.9	25.8
OneThinker-8B [21]	32	73.0	93.9	84.4	68.8
Video-ORA-4B	32	67.5	87.2	76.8	57.8
Video-ORA-9B	32	**78.2**	**94.1**	**87.8**	**75.3**

Table 5 Spatial-temporal grounding results on STVG. We report tIoU@0.5, mean tIoU, sIoU@0.5, and mean sIoU.

Models	Frame	STVG [38]			
		tIoU@0.5	tIoU	sIoU@0.5	sIoU
Grounded-VideoLLM [73]	-	30.0	33.0	-	-
Qwen3-VL-8B [2]	128	24.4	25.4	11.6	13.6
Qwen3.5-4B [71]	128	35.8	37.7	6.3	11.2
Qwen3.5-9B [71]	128	37.0	39.3	11.9	16.5
Video-ORA-4B	128	39.0	40.5	22.2	24.2
Video-ORA-9B	128	**40.0**	**41.7**	**27.4**	**28.4**

61.9 to 66.8. Gains reach 15.2 points on VideoHolmes, 5.6 on VideoMME, and 4.5 on MMVU, but only 1.6 on LongVideoBench.

Visual tracking. On GOT-10k, which evaluates complete trajectories, Video-ORA-9B leads AO and recall at every threshold in Tab. 4, surpassing OneThinker-8B by 5.2 AO. Its recall margin grows from 0.2 points at R@0.3 to 6.5 at R@0.7. Video-ORA-4B also reaches 67.5 AO, compared with 45.3 and 46.0 for the Qwen3.5-4B and 9B backbones.

Spatial-temporal grounding. STVG requires both a temporal interval and framewise boxes, jointly evaluating event localization and spatial grounding. Video-ORA-9B ranks first on all four metrics, with Video-ORA-4B ranking second throughout (Tab. 5). Relative to Qwen3.5-9B, it gains 3.0 and 2.4 points on tIoU@0.5 and mean tIoU, but 15.5 and 11.9 points on sIoU@0.5 and mean sIoU, showing that the larger improvement comes from frame-level spatial localization. It also exceeds Grounded-VideoLLM by 10.0 and 8.7 points on its two reported temporal metrics, while using the same serialized output interface as the other tasks.

Segmentation. Tab. 7 shows that Video-ORA-9B leads every RefCOCO-family cIoU and the MeViS J&F, while Video-ORA-4B is best on ReasonVOS. Recent specialists are tightly clustered on image cIoU, where Video-ORA-9B adds at most 1.0 point over the best baseline, whereas the video benchmarks move much further, from 52.7 to 61.3 on MeViS and from 59.9 to 63.8 on ReasonVOS. Video segmentation is also the task with the largest absolute change, improving over the Qwen3.5-9B backbone by 29.2 J&F on MeViS and 42.2 on ReasonVOS, since the backbone cannot produce usable mask prompts on its own. On ReasonVOS the 4B and 9B models finish within 0.1 J&F of each other, so the residual error on that benchmark is not capacity-limited. Appendix D reports the remaining per-metric scores.

Spatial intelligence. Video-ORA-9B achieves the highest re-

Table 6 Complete results on VSI-Bench [93]. Numerical and multiple-choice tasks use MRA and accuracy; Avg. is their macro average.

Models	Numerical Question				Multiple-Choice Question				Avg.
	Obj. Count	Abs. Dist	Obj. Size	Room Size	Rel. Dis	Rel. Dir	Route Plan	Appr. Order	
Proprietary Models									
Seed-2.0 [63]	49.4	25.3	69.5	25.8	61.8	44.9	44.3	71.0	49.0
Grok-4 [85]	37.1	32.9	60.8	45.4	53.1	39.6	47.4	66.8	47.9
Gemini-2.5-Pro [13]	46.0	37.3	68.7	54.3	61.9	43.9	47.4	68.7	53.5
Gemini-3-Pro [26]	49.0	42.8	71.5	41.8	56.6	57.5	**61.9**	60.0	55.1
Kimi-K2.5 [69]	57.2	34.9	69.3	54.4	59.6	41.3	52.1	67.0	54.5
GPT-5 [67]	53.3	34.4	73.3	47.5	63.7	48.6	50.2	68.9	55.0
Open-source General Models									
LLaVA-OneVision-72B [37]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
LLaVA-Video-72B [112]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
InternVL3-8B [117]	66.0	34.8	43.6	47.5	48.0	39.3	26.2	31.3	42.1
Qwen3-VL-8B [2]	67.5	47.0	76.3	61.9	58.0	50.9	35.0	66.3	57.9
LLaVA-OneVision-2-8B [1]	–	–	–	–	–	–	–	–	70.9
Qwen3.5-4B [71]	56.5	40.4	68.9	59.3	65.5	75.7	36.1	74.9	59.7
Qwen3.5-9B [71]	62.0	44.0	73.6	61.9	69.7	79.9	45.4	26.7	57.9
Open-source Spatial Intelligence Models									
SpaceR-7B [56]	44.5	24.7	53.5	37.3	41.9	46.1	29.3	54.8	41.5
ViLaSR-7B [83]	58.1	33.8	61.4	28.8	45.0	46.5	29.9	53.2	44.6
VST-7B [94]	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2	61.1
Cambrian-S-7B [96]	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1	67.5
Spatial-MLLM-4B [81]	65.3	34.8	63.1	45.1	41.3	46.9	33.5	46.3	47.0
SpatialStack-5B [109]	71.0	55.6	69.1	68.2	67.3	84.1	41.2	83.5	67.5
SpaceMind [113]	73.3	**61.4**	77.4	**74.2**	67.2	**88.4**	44.3	70.6	69.6
Video-ORA-4B	72.2	51.2	75.8	64.7	67.5	81.9	41.2	75.4	66.2
Video-ORA-9B	**76.1**	58.3	**78.2**	72.1	**75.2**	86.9	47.4	**90.6**	**73.1**

Table 7 Segmentation results. RefCOCO/+/g [101, 54] use cIoU, whereas MeViS [17] and ReasonVOS [3] use J&F.

Model	Image (cIoU)			Video (J&F)	
	RefCOCO	+	g	MeViS	ReasonVOS
PixelLM-7B [61]	73.0	66.3	69.3	–	–
LISA-7B [35]	74.1	62.4	66.4	37.2	31.1
VISA-13B [89]	72.4	59.8	65.5	44.5	–
Seg-R1-7B [99]	74.3	62.6	71.0	–	–
Sa2VA-4B [103]	78.9	71.7	74.1	46.2	–
MomentSeg-7B [16]	77.8	68.2	70.8	–	–
VideoSeg-R1-7B [88]	78.2	71.8	73.1	–	–
ReferFormer [84]	–	–	–	31.0	32.9
VideoLISA-3.8B [3]	–	–	–	44.4	47.5
Veason-R1-7B [25]	–	–	–	52.2	59.9
Qwen3-VL-8B [2]	73.8	65.1	70.1	22.9	19.6
OneThinker-8B [21]	75.8	67.1	70.8	52.7	54.9
Qwen3.5-4B [71]	74.7	64.0	69.0	27.7	20.7
Qwen3.5-9B [71]	75.2	65.6	70.0	32.1	21.5
Video-ORA-4B	76.7	68.7	73.0	57.5	**63.8**
Video-ORA-9B	**79.4**	**72.6**	**75.1**	**61.3**	63.7

Table 8 Results on MMSI-Bench [95] and MindCube-Tiny [75]. Both report accuracy; Avg. is their mean.

Model	MMSI	MindCube	Avg.
Proprietary Models			
Gemini-2.5-Pro [13]	38.0	57.6	47.8
Grok-4 [85]	37.8	**63.6**	**50.7**
GPT-5 [67]	**41.8**	56.3	49.1
Open-source General Models			
InternVL3-8B [117]	28.0	41.5	34.8
Qwen3-VL-8B [2]	31.1	29.4	30.3
Qwen3.5-4B [71]	31.6	46.1	38.9
Qwen3.5-9B [71]	31.7	41.2	36.5
Open-source Spatial Intelligence Models			
Spatial-MLLM-4B [81]	26.1	33.5	29.8
SpaceR-7B [56]	27.4	38.0	32.7
ViLaSR-7B [83]	30.2	35.1	32.7
VST-7B [94]	32.5	39.7	36.1
Cambrian-S-7B [96]	27.1	37.9	32.5
Video-ORA-4B	31.9	48.6	40.3
Video-ORA-9B	37.9	57.4	47.7

ported VSI-Bench average of 73.1 in Tab. 6, exceeding LLaVA-OneVision-2-8B by 2.2 points and the strongest proprietary model by 18.0. Within this comparison, it ranks first in object count, object size, relative distance, and appearance order. The largest margin occurs on appearance order, where it scores 90.6 versus 83.5, while route planning remains behind Gemini-3-Pro, Kimi-K2.5, and GPT-5. Among the open-source models in Tab. 8, it scores 37.9 on MMSI-Bench and 57.4 on MindCube-Tiny. Their average of 47.7 remains below Grok-4 at 50.7, GPT-5 at 49.1, and Gemini-2.5-Pro at 47.8. Task-level results in Appendix E

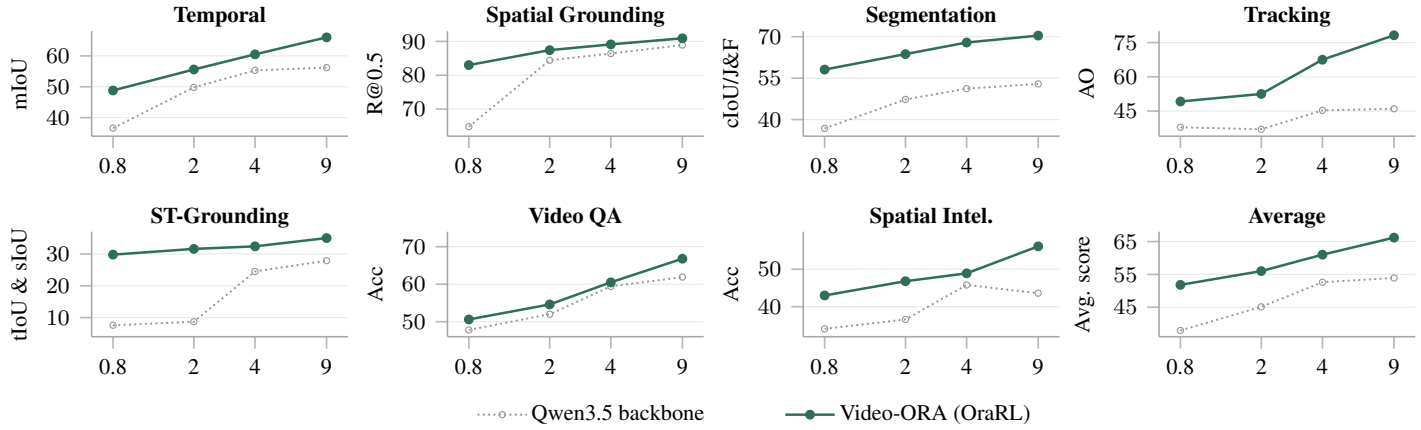


Figure 4 Model scaling from 0.8B to 9B. Under the same OraRL recipe, Video-ORA improves across all task families with model size (x-axis, in billions of parameters, log scale) and outperforms its Qwen3.5 backbone at every scale. The final panel reports the macro-average; Spatial Grounding averages RefCOCO-family R@0.5, while ST-Grounding averages tIoU and sIoU.

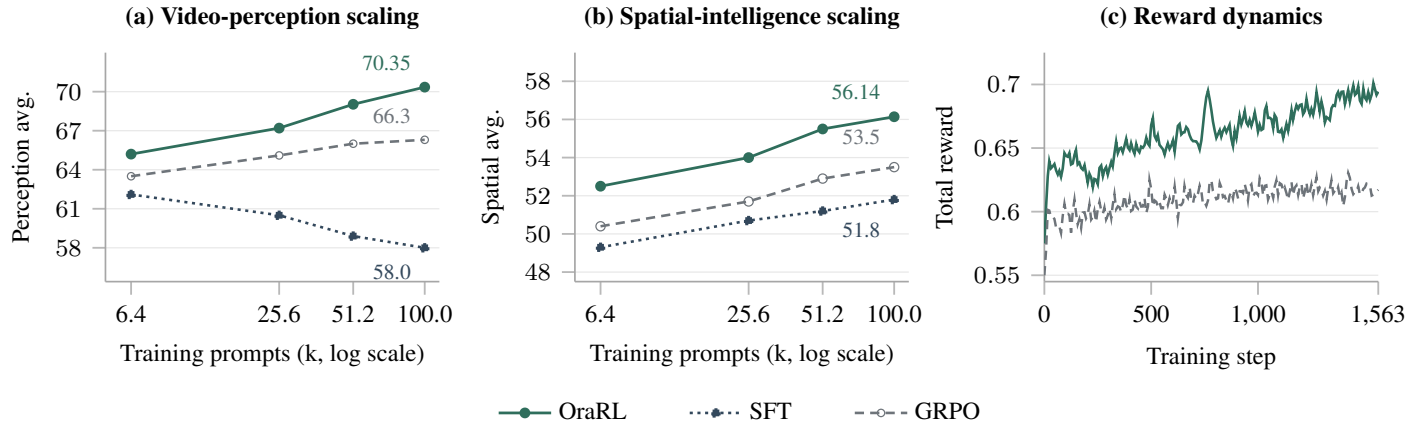


Figure 5 Task-group data scaling and reward dynamics under the controlled 9B protocol. (a) The video-perception aggregate is the macro average over temporal grounding, tracking, segmentation, and video question answering. **(b)** The spatial-intelligence aggregate averages VSI-Bench, MMSI-Bench, and MindCube. **(c)** Total training reward over one epoch for OraRL and GRPO.

show strengths in MMSI camera-object relations and MindCube "among" and "around," but weak planning and rotation. Appendix F repeats the VSI-Bench comparison on the re-annotated ReVSI benchmark, where Video-ORA-9B again attains the highest open-source score.

4.3 Model Scaling

Fig. 4 compares Video-ORA at 0.8B, 2B, 4B, and 9B parameters, each trained under the same OraRL configuration and evaluated against its corresponding Qwen3.5 backbone. At every scale, Video-ORA improves all seven task families and exceeds its backbone by over 8 points in macro average. Its macro average rises monotonically from 51.8 at 0.8B to 66.2 at 9B, showing that performance scales consistently with model size under the same training configuration. Tracking benefits most from scale, with the margin widening from 11.3 at 0.8B to 32.2 at 9B and becoming the largest of any family at 4B and above, suggesting that larger models use framewise box supervision more effectively. Below 4B the widest margins appear on spatial-temporal grounding: the backbone scores under 9, while OraRL adds over 22 points.

4.4 Data Scaling and Reward Dynamics

We compare SFT, GRPO, and OraRL at matched checkpoints spanning 6.4k to 100k processed prompts. Video perception aggregates temporal grounding, tracking, segmentation, and Video QA, while spatial intelligence aggregates VSI-Bench, MMSI-Bench, and MindCube-Tiny. Across this range, OraRL gains 5.2 and 3.6 points on the two aggregates, compared with 2.8 and 3.1 for GRPO. Their respective margins therefore widen from 1.7 and 2.1 points to 4.1 and 2.6 in the first two panels of Fig. 5. This scaling advantage is consistent with annotation-as-rollout supplying a reliable positive target for every added prompt, whereas GRPO relies solely on reward variation among its sampled rollouts. SFT instead loses 4.1 points on video perception while gaining 2.5 on spatial intelligence, indicating that continued imitation does not transfer uniformly across task groups.

4.5 Does Video Data Improve Spatial Intelligence?

To assess whether video perception supervision transfers to spatial intelligence, we compare two RL runs initialized from the same SFT checkpoint. One uses 9,000 spatial-intelligence prompts,

Table 9 Effect of video perception data on spatial intelligence. Avg. is the mean across the three benchmarks.

Training data	RL prompts		VSI-Bench [93]	MMSI [95]	MindCube [75]	Avg.
	Video	Spatial				
Qwen3.5-9B, no training [71]	–	–	57.9	31.7	41.2	43.6
Spatial intelligence only	–	9k	68.1	33.5	49.2	50.3
+ video perception	9k	9k	70.0	34.2	50.3	51.5
Video-ORA-9B, seven tasks	100k		**73.1**	**37.9**	**57.4**	**56.1**

Table 10 Backbone generalization. Each model family is evaluated after SFT, GRPO, and OraRL under the same protocol.

Backbone	Stage	Temporal	Tracking	Video Seg.	Avg.
Qwen3-VL-8B	Backbone	51.5	33.7	21.3	35.5
	+SFT	53.2	67.7	55.0	58.6
	+GRPO	58.0	68.5	57.8	61.4
	+OraRL	60.8	71.3	59.9	64.0
Qwen3.5-9B	Backbone	56.2	46.0	26.8	43.0
	+SFT	58.7	68.0	57.2	61.3
	+GRPO	60.2	73.6	60.0	64.6
	+OraRL	62.3	75.0	60.6	66.0

Table 11 Chain-of-thought ablation. Temporal-grounding scores are averaged across Charades, ActivityNet, and QVHighlights; training cost is reported in seconds per step.

Method	Char.	ANet.	QV.	Avg.	Train s/step
CoT + answer					
Backbone	50.3	45.0	58.9	51.4	N/A
GRPO	54.8	53.4	67.2	58.5	135.6
Answer only					
Backbone	49.4	51.9	64.5	55.3	N/A
GRPO	54.0	54.3	67.7	58.7	93.9
OraRL	**58.0**	**57.5**	**68.7**	**61.4**	**62.4**

Table 12 Training paradigm comparison. All post-training runs use the same Qwen3.5-4B SFT initialization, data, and optimization budget.

Method	Temporal	Tracking	Video Seg.	Avg.
Supervised baselines				
Backbone	55.3	45.3	24.2	41.6
SFT init.	57.6	60.2	55.6	57.8
Continued SFT	55.2	63.1	57.7	58.7
On-policy group RL				
GRPO [65]	58.1	63.9	58.9	60.3
Dr. GRPO [51]	58.3	65.1	58.6	60.7
GDPO [49]	58.0	64.4	58.9	60.4
CPPO [48]	58.5	64.2	58.8	60.5
Group RL with off-policy rollouts				
LUFFY-style [90]	57.0	50.0	57.2	54.7
OraRL	**60.5**	**67.1**	**60.6**	**62.7**

while the other augments them with 9,000 video prompts drawn from temporal grounding, tracking, mask-aware video segmentation, spatial-temporal grounding, and Video QA. The pretrained backbone and full seven-task model are included as reference points. As shown in Tab. 9, the augmented run improves VSI-Bench, MMSI-Bench, and MindCube-Tiny by 1.9, 0.7, and 1.1 points, respectively, raising their mean by 1.2 points. Although the augmented run uses twice as many RL updates, its consistent gains across all three benchmarks support the practical value of adding video perception data under an expanded training budget.

4.6 Backbone Generalization

We apply the same three-task protocol to Qwen3-VL-8B and Qwen3.5-9B without separate retuning. For both families, OraRL outperforms GRPO, SFT, and the pretrained backbone on every task in Tab. 10. OraRL improves the GRPO average by 2.6 points on Qwen3-VL-8B and 1.4 points on Qwen3.5-9B. The larger gain on the weaker Qwen3-VL SFT checkpoint may arise from stronger oracle guidance when the policy remains farther from the oracle.

4.7 Is Chain-of-Thought Necessary?

Tab. 11 compares CoT and answer-only training on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens. Before RL, CoT lowers the backbone average by 3.9 points, showing that the deficit precedes policy optimization. Under GRPO, CoT reduces the average from 58.7 to 58.5 while increasing step time by 44.4%, from 93.9 to 135.6 s. By contrast, OraRL improves all three datasets over answer-only GRPO, raising the average by 2.7 points while reducing step time by 33.5%. For structured temporal grounding, these results favor direct oracle supervision over longer reasoning traces in both accuracy and efficiency.

4.8 Comparison of Training Paradigms

All post-training runs in Tab. 12 start from the same answer-only SFT checkpoint and use the same three-task data and optimization budget. Continued SFT improves the average by only 0.9 points, whereas the on-policy group methods gain between 2.5 and 2.9, indicating that comparative supervision is more effective than further imitation once the output format has been learned. Those methods, however, differ from one another by at most 0.4 points, with Dr. GRPO reaching 60.7 against 60.5 for CPPO, 60.4 for GDPO, and 60.3 for GRPO, so removing the normalization biases or pruning by advantage magnitude leaves the outcome essentially unchanged. By contrast, OraRL reaches 62.7 and leads all three tasks, exceeding the strongest variant by 2.0 points, so its gain follows from the added oracle rather than from a different advantage estimator or rollout budget. LUFFY-style optimization averages only 54.7 and lowers tracking to 50.0, consistent with a mismatch

Table 13 Annotation injection under the shared three-task protocol. Shaping follows Tempsamp-R1.

Variant	Temporal	Tracking	Video Seg.	Avg.
GRPO	58.1	63.9	58.9	60.3
Naive GT injection	57.2	52.0	57.0	55.4
+ Reward shaping	60.4	64.0	59.2	61.2
OraRL	**60.5**	**67.1**	**60.6**	**62.7**

Table 14 Advantage inversion across tasks. Flip rates measure GRPO-positive rollouts assigned negative advantages, so lower is better.

Variant	Temporal	Tracking	Video Seg.	Overall
Rollout-level flip rate				
Naive oracle mix	18.8	38.7	4.5	22.4
+ reward shaping [45]	9.7	21.7	1.3	11.9
OraRL, before pruning	**1.0**	3.5	0.3	1.9
OraRL, after pruning	1.1	**0.2**	**0.0**	**0.3**
Group-level incidence, naive oracle mix				
Any inverted rollout	25.9	73.4	14.4	42.5
No positive rollout left	17.5	9.2	1.4	8.3

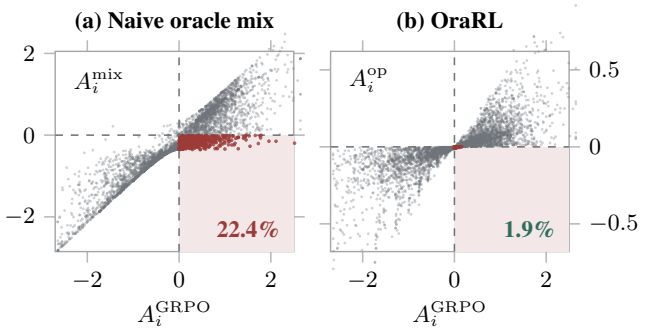


Figure 6 Advantage inversion. Shading marks inverted rollouts; labels give full-data rates among GRPO-positive samples.

between policy and oracle signals weakening its weighted update.

4.9 From Naive Oracle Injection to OraRL

Under the matched three-task protocol, naive oracle injection lowers the GRPO average from 60.3 to 55.4 and tracking by 11.9 points in Tab. 13. This shows that oracle correctness alone is insufficient. Including its high reward in the normalization statistics shifts the baseline and suppresses useful policy rollouts, especially when the oracle-policy gap is large. Reward shaping as in Tempsamp-R1 [45] recovers the average to 61.2 but requires a manually designed transform for each task reward. Using raw rewards, OraRL reaches 62.7 and outperforms shaping on every task, with the largest gain of 3.1 points on tracking.

4.10 Advantage Inversion Analysis

Fig. 6 visualizes 4,000 uniformly sampled rollouts, while Tab. 14 reports pooled rates over all 92,024 rollouts from 11,503 groups. Naive oracle mixing inverts 22.4% of rollouts that GRPO assigns positive advantages. At the group level, 42.5% contain at least one inversion and 8.3% lose every positive rollout. The task variation

Table 15 Component ablation. All variants use the same initialization, data, rollout budget, and optimization schedule.

Variant	Temp.	Track.	Seg.	Avg.
On-policy scaling				
w/o directional gain	59.0	64.4	60.1	61.2
Oracle advantage				
w/o detached oracle	58.0	65.5	60.3	61.3
w/o reward-gap weight	58.9	66.1	59.8	61.6
Rollout pruning				
w/o sign balance	58.7	67.5	60.0	62.1
w/o pruning ($\kappa = 0$)	60.4	**68.0**	**60.8**	**63.1**
w/o moment correction	60.0	66.0	60.1	62.0
OraRL	**60.5**	67.1	60.6	62.7

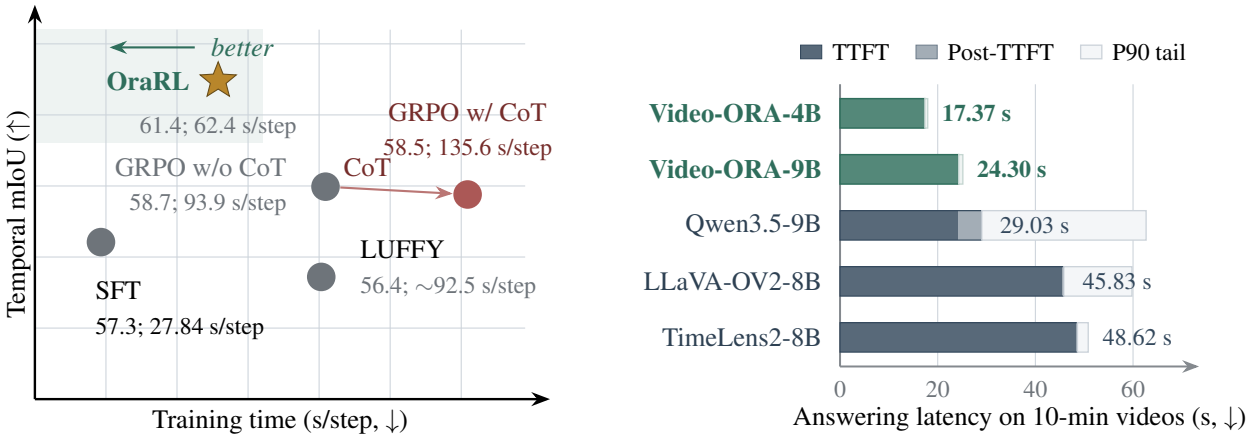
Table 16 Accuracy and efficiency trade-off under sign-balanced pruning. K is the number of retained rollouts from each nine-rollout group, memory is peak per-GPU allocation, and trade-off is time saved per point of average-score loss. OraRL uses $\kappa = 0.5$.

κ	K	s/step	Mem. (GB)	Avg.	Trade-off (s/pt)
0	8	92.5	62.4	**63.1**	–
0.25	6	75.0	60.0	62.8	58.3
0.50	4	62.4	50.9	62.7	**75.3**
0.75	2	45.0	47.5	60.2	16.4

follows the gap dependence predicted by Eqn. (6). Tracking has the highest flip rate at 38.7%, whereas segmentation has the lowest at 4.5%. Temporal grounding is especially vulnerable at the group level, with sparse rewards leaving 17.5% of groups without a positive rollout. Task-specific reward shaping reduces the pooled flip rate to 11.9% but cannot remove the baseline shift because the oracle remains in the normalization statistics. OraRL removes the oracle-induced baseline shift by estimating policy advantages from on-policy rewards alone, leaving a residual flip rate of 1.9%. The residual comes from re-centering rather than from the baseline: the directional gain amplifies only rollouts above the on-policy mean, so the advantages acquire a small positive mean whose subtraction pushes the weakest of them below zero. These rollouts carry negligible magnitude and are almost all pruned, leaving 0.3% inverted among the rollouts that receive gradients.

4.11 Component Ablations

Tab. 15 evaluates individual component removals under the shared three-task protocol. Removing the directional gain, detached oracle advantage, and reward-gap weight lowers the average by 1.5, 1.4, and 1.1 points, respectively. Their task effects are complementary. Removing the directional gain costs 2.7 points on tracking, whereas removing the detached oracle advantage costs 2.5 on temporal grounding. Magnitude-only selection raises tracking by 0.4 points but lowers temporal grounding by 1.8 and the average by 0.6, indicating that sign balance yields more consistent performance across tasks. Relative to no pruning, pruning without moment correction loses 1.1 average points. Moment correction restores 0.7, limiting the loss to 0.4 points at a 1.48× speedup.



(a) Performance vs. training cost.

(b) Single-request inference latency.

Figure 7 Training and inference efficiency. (a) Accuracy against training cost, where OraRL gives the strongest trade-off and CoT substantially raises cost. (b) End-to-end single-request latency on one H20 in BF16. All models see the same 2-fps input of about 120K tokens per video. Solid bars give median latency split at the first token, pale tails run to P90, and the numbers report medians.

4.12 Training and Inference Efficiency

Pruning ratio. In Tab. 16, retaining four rollouts at $\kappa = 0.5$ cuts step time from 92.5 to 62.4 s while lowering the average by 0.4 points. Peak memory also drops from 62.4 to 50.9 GB. At $\kappa = 0.75$, retaining only the oracle and one policy rollout removes sign contrast and lowers the average to 60.2.

Training cost. In the left panel of Fig. 7, OraRL raises GRPO without CoT from 58.7 to 61.4 mIoU and cuts step time from 93.9 to 62.4 s. SFT is faster but scores 4.1 points lower; GRPO with CoT and LUFFY trail OraRL in both speed and accuracy.

Inference latency. The right panel of Fig. 7 reports latency for individual requests on ten videos, each 10 minutes long and sampled at 2 fps. Video-ORA-9B and its Qwen3.5-9B backbone have nearly identical TTFTs of 24.17 and 24.25 s, but generate medians of 13.5 answer tokens and 808.5 reasoning tokens. Consequently, latency after TTFT rises from 0.13 to 4.78 s and total median latency from 24.30 to 29.03 s. At P90, variable rationale length raises the backbone latency to 62.67 s, while Video-ORA-9B remains at 25.15 s.

5 Conclusions

This paper presents OraRL, an efficient and scalable reinforcement learning framework for unified video MLLMs. OraRL serializes each annotation as an oracle rollout, providing a reliable positive target while preserving on-policy exploration. We characterize advantage inversion caused by direct oracle mixing and remove the resulting baseline shift by estimating policy advantages from on-policy rewards alone and calibrating oracle guidance to the current policy gap. Sign-balanced pruning with moment correction reduces step time from 92.5 to 62.4 s with an average loss of 0.4 points. Matched evaluations show that OraRL outperforms SFT and GRPO across all controlled tasks. Consistent improvements are observed across model scales from 0.8B to 9B, two backbone families, and data budgets up to 100k prompts. These results demonstrate the effectiveness and scalability of annotation-as-rollout for reinforcement learning in unified video perception.

Limitations and future work. The current formulation assumes that each annotation can be serialized as a valid oracle rollout and evaluated by a scalar task reward. Its behavior under ambiguous, partial, or noisy supervision and learned oracles has not been evaluated. Experiments cover seven task families and two backbone families, although spatial tasks requiring complex reasoning still underperform some proprietary models. Future work should explore broader supervision, model families, and reasoning.

References

[1] Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026. 1, 2, 3, 6, 7, 8

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 6, 7, 8, 19, 20

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Adv. Neural Inform. Process. Syst.*, volume 37, pages 6833–6859, 2024. 3, 6, 8, 19

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Adv. Neural Inform. Process. Syst.*, 2021. 17

[5] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7879–7890, 2026. 17

[6] Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025. 3, 7

[7] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *Adv. Neural Inform. Process. Syst.*, pages 91077–91100, 2026. 2, 7

[8] Jieneng Chen, Wenxin Ma, Ruisheng Yuan, Yunzhi Zhang, Jiajun Wu, and Alan Yuille. Thinking with spatial code for physical-world video reasoning. *arXiv preprint arXiv:2603.05591*, 2026. 1, 3

[9] Xin Chen, Bin Yan, Jiawen Zhu, Huchuan Lu, Xiang Ruan, and Dong Wang. High-performance transformer tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8507–8523, 2023. 3

[10] 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. In *Int. Conf. Learn. Represent.*, pages 18227–18246, 2025. 17

[11] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 6, 7, 17

[12] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 28652–28668, 2026. 1, 2, 6

[13] 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. 3, 6, 8, 19

[14] Yutao Cui, Cheng Jiang, Gangshan Wu, and Limin Wang. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4129–4146, 2024. 3

[15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 17

[16] Ming Dai, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Momentseg: Moment-centric sampling for enhanced video pixel understanding. *arXiv preprint arXiv:2510.09274*, 2025. 3, 8, 19

[17] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Int. Conf. Comput. Vis.*, pages 2694–2703. IEEE, 2023. 6, 8, 17, 19

[18] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7900–7916, 2022. 3

[19] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 31054–31065, 2026. 20

[20] 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. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 99114–99137, 2025. 1, 3

[21] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5432–5443, 2026. 2, 3, 7, 8, 17, 19

[22] Chaoyou Fu, Yuhan Dai, Yongdong 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24108–24118, 2025. 6, 7

[23] Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026. 6, 7

[24] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Int. Conf. Comput. Vis.*, Oct 2017. 3, 6

[25] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Jiazuo Yu, Pingping Zhang, Xu Jia, and Huchuan Lu. Reinforcing video object segmentation to think before it segments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3835–3844, 2026. 3, 8, 19

[26] Google DeepMind. Gemini 3 pro model card, November 2025. 1, 3, 8, 20

[27] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *AAAI*, volume 39, pages 3302–3310, 2025. 1, 3

[28] Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. In *Adv. Neural Inform. Process. Syst.*, volume 35, pages 15281–15295, 2022. 3

[29] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14271–14280, 2024. 1, 3

[30] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1562–1577, 2019. 6, 7, 17

[31] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13700–13710, June 2024. 2

[32] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conf. Comput. Vis.*, pages 123–141, 2018. 17

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, Oct 2017. 3, 6

[34] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM Symp. Operating Syst. Princ.*, pages 611–626, 2023. 6

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9579–9589, 2024. 3, 8, 19

[36] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Adv. Neural Inform. Process. Syst.*, volume 34, pages 11846–11858, 2021. 3, 6

[37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and

Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025. 2, 8

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8592–8603, 2025. 1, 3, 6, 7, 17

[39] Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. In *Inter. Conf. Mach. Learning*, 2026. 3, 6

[40] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Sci. China Inf. Sci.*, 68(10):200102, 2025. 2

[41] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22195–22206, 2024. 6, 7

[42] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 2, 6

[43] 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. 1, 3

[44] Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026. 1, 2, 3, 6

[45] Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-R1: Effective temporal sampling with reinforcement fine-tuning for video LLMs. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 40692–40716, 2025. 2, 6, 11, 18

[46] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Annu. Meet. Assoc. Comput. Linguist.*, pages 6657–6678, 2024. 1, 3

[47] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8658–8667, 2025. 3

[48] Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. CPPO: Accelerating the training of group relative policy optimization-based reasoning models. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 61043–61068, 2025. 2, 3, 5, 10

[49] Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In *Inter. Conf. Mach. Learning*, 2026. 10

[50] Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 126078–126108, 2025. 3

[51] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conf. Lang. Model.*, 2025. 10

[52] Haozhe Ma, Zhengding Luo, Thanh Vinh Vo, Kuankuan Sima, and Tze-Yun Leong. Highly efficient self-adaptive reward shaping for reinforcement learning. In *Int. Conf. Learn. Represent.*, 2025. 3

[53] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annu. Meet. Assoc. Comput. Linguist.*, 2024. 2

[54] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 6, 7, 8, 17, 19

[55] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Eur. Conf. Comput. Vis.*, pages 300–317, 2018. 17

[56] 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. 1, 3, 8, 19, 20

[57] Zhangyang Qi, Jinsong Li, Hongjian Wu, Jiaqi Wang, and Hengshuang Zhao. Game ground bench: Probing the limits of lvlms in complex semantic grounding across game universes. In *AAAI*, volume 40, pages 8493–8501, 2026. 7

[58] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: advancing video large language model with fine-grained temporal reasoning. In *Inter. Conf. Mach. Learning*, pages 41340–41356, 2024. 1

[59] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13009–13018, 2024. 3

[60] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14313–14323, June 2024. 1, 3

[61] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 26374–26383, 2024. 3, 8, 19

[62] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[63] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026. 1, 3, 8

[64] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Eur. Conf. Comput. Vis.*, pages 208–223, 2020. 17

[65] 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. 1, 3, 10

[66] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *ACM Eur. Conf. Comput. Syst.*, pages 1279–1297, 2025. 6

[67] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 1, 3, 6, 8, 19, 20

[68] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al.

Video understanding with large language models: A survey. *IEEE Trans. Circuit Syst. Video Technol.*, 2025. 3

[69] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. 8

[70] Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, et al. Kwai keye-vl-2.0 technical report. *arXiv preprint arXiv:2606.10651*, 2026. 2, 6

[71] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. 2, 6, 7, 8, 10, 19, 20

[72] Biao Wang, Wenwen Li, and Jiawei Ge. R1-track: Direct application of mllms to visual object tracking via reinforcement learning. *arXiv preprint arXiv:2506.21980*, 2025. 1, 3

[73] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. In *Conf. Empir. Methods Nat. Lang. Process.*, 2025. 1, 3, 7

[74] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *Eur. Conf. Comput. Vis.*, pages 166–184, 2024. 17

[75] Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Int. Conf. Learn. Represent.*, 2026. 6, 8, 10, 17, 19

[76] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 7, 20

[77] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[78] 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. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 83330–83364, 2025. 1, 3, 6

[79] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst.*, volume 35, pages 24824–24837, 2022. 3

[80] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Adv. Neural Inform. Process. Syst.*, 2021. 17

[81] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 13569–13597, 2025. 3, 8, 20

[82] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Adv. Neural Inform. Process. Syst.*, pages 28828–28857, 2024. 6, 7

[83] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 143297–143330, 2025. 3, 8, 19

[84] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4974–4984, 2022. 3, 8, 19

[85] xAI. Grok 4 model card, August 2025. 3, 8, 19

[86] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. 2, 6

[87] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12113–12132, 2023. 2

[88] Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. Videoseg-r1: reasoning video object segmentation via reinforcement learning. In *AAAI*, volume 40, pages 11496–11504, 2026. 3, 8, 19

[89] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *Eur. Conf. Comput. Vis.*, pages 98–115. Springer, 2024. 3, 8, 17, 19

[90] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 117157–117186, 2025. 3, 10

[91] Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. Internvideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026. 1, 2, 3, 6, 7

[92] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 7

[93] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10632–10643, 2025. 6, 8, 10

[94] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 3, 8, 19, 20

[95] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *Int. Conf. Learn. Represent.*, pages 157051–157088, 2026. 6, 8, 10, 17, 19

[96] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *Int. Conf. Learn. Represent.*, 2025. 1, 2, 3, 8, 17, 18, 19, 20

[97] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Int. Conf. Comput. Vis.*, pages 12–22, 2023. 17

[98] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *Int. Conf. Learn. Represent.*, 2020. 17

[99] Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025. 3, 8, 19

[100] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, et al. Perception-r1: Pioneering perception policy with reinforcement learning. In *Adv. Neural Inform. Process. Syst.*, volume 38, pages 94827–94853, 2025. 3, 7

[101] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85, 2016. 6, 7, 8, 17, 19

[102] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11704–11715, 2026. 2, 7

[103] Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with mllm for dense grounded understanding of images and videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2026. 1, 3, 8, 17, 19

[104] Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026. 3, 6

[105] 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. 2

[106] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conf. Empir. Methods Nat. Lang. Process.*, pages 543–553, 2023. 2

[107] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5625–5644, 2024. 2

[108] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10419–10429, 2026. 1, 3, 6, 17

[109] Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 38678–38688, 2026. 3, 8

[110] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9073–9087, 2022. 3

[111] Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. In *Inter. Conf. Mach. Learning*, 2026. 18, 20

[112] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025. 2, 8, 17, 20

[113] Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025. 1, 3, 8

[114] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, 2023. 6

[115] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8475–8489, 2025. 6, 7

[116] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13691–13701, 2025. 6, 7

[117] 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. 2, 8, 19

[118] Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. Timelens2: Generalist video temporal grounding with multimodal llms, 2026. 1, 2, 3, 6

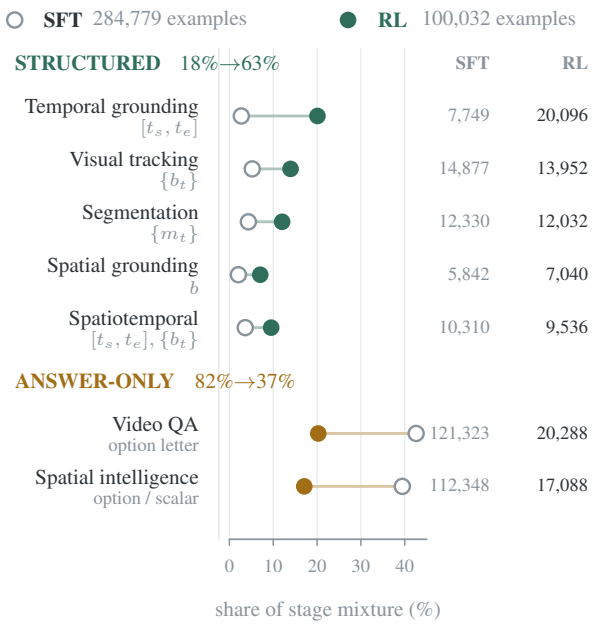


Figure 8 Composition of the SFT and RL training data. Open and filled markers indicate each task's mixture proportion during SFT and RL, respectively. Tasks are grouped by output format, and the rightmost columns report exact example counts.

Table 17 Training hyperparameters. The RL batch size counts prompts before rollout generation.

Setting	SFT	RL
Epochs	1	1
Framework	ms-swift	veRL + vLLM
Distributed strategy	ZeRO-2	FSDP full shard
Precision	bfloat16	bfloat16
Trainable module	language model	language model
Global prompt batch	64	64
Update micro-batch / GPU	1	1
Optimizer	AdamW	AdamW
Learning rate	1×10^{-5}	2×10^{-6}
Weight decay	1×10^{-6}	0
LR schedule	cosine	constant
Warmup ratio	0.02	0
Gradient clipping	1.0	1.0
Maximum sequence length	12,288	–
Prompt length cap	–	24,576
Response length cap	–	2,048
Policy rollouts / prompt	–	8
Sampling (T, p)	–	$(1.0, 0.85)$
Actor epochs / PPO clip	–	1 / 0.2
KL coefficient	–	0
Pruning ratio κ	–	0.5

A Training Data Construction

SFT and RL cover the same seven task families but use different sampling distributions (Fig. 8). SFT contains 284,779 prompts (18% structured and 82% answer-only), whereas RL contains 100,032 prompts (63% structured and 37% answer-only). The SFT mixture assigns lower sampling proportions to structured perception tasks, particularly temporal grounding, to mitigate task imbalance and preserve general video understanding during supervised fine-tuning. RL instead prioritizes structured perception and temporal grounding, whose interval annotations provide precise oracle rollouts, while retaining video QA and spatial intelligence at lower proportions. All examples are drawn from public training splits. We identify videos shared between the training and evaluation sets and remove the overlapping training examples.

Sources. Temporal grounding is taken from TimeLens-100K [108]. Tracking uses GOT-10k [30], TrackingNet [55], and ElysiumTrack [74]. Image and video segmentation use the One-Thinker [21] training pack, covering ReVOS [89], Ref-YouTube-VOS [64], Ref-SAV [103], MeViS [17], DAVIS-17 [32], and COCO referring expressions [101, 54]. Spatial grounding uses RefCOCO/+/g [101, 54] on COCO train2014 images, and spatial-temporal grounding uses LLaVA-ST-STVG [38]. Video QA combines LLaVA-Video-178K [112], LongVILA [10], STAR [80], CLEVRER [98], and Video-Holmes [11]. Spatial intelligence uses VSI-590K [96] training scenes from ScanNet [15], ScanNet++ [97], and ARKitScenes [4], together with a SenseNova-SI-800K [5] subset whose templates resemble MMSI-Bench and MindCube [95, 75] but are not drawn from those benchmarks.

Mixture construction. All source datasets are normalized to a common schema. Examples with inaccessible media or unrecoverable annotations are discarded, and duplicates are identified by task type, media, question, and answer. The SFT and RL mixtures are sampled from the resulting pools according to the task proportions in Fig. 8, with at most two prompts retained per video. For the RL mixture, we estimate the difficulty of each candidate by performing inference with the SFT checkpoint and scoring its prediction against the annotation with the corresponding task metric. Based on these scores, each task-specific RL subset is constructed to emphasize intermediate-difficulty candidates, with candidates scoring at least 0.80 restricted to 15% of the subset.

B Training Details

The SFT and RL configurations are reported in Tab. 17.

Supervised fine-tuning. Each model is initialized from its corresponding pretrained backbone, with Qwen3.5 used by default. The vision tower and multimodal aligner remain frozen during SFT. Before RL, the language-model parameters are initialized by interpolation as $\theta_{\text{init}} = (1-\alpha)\theta_{\text{base}} + \alpha\theta_{\text{SFT}}$, while the pretrained visual parameters are retained. We set $\alpha = 0.7$ for the controlled 4B experiments and $\alpha = 0.6$ for the 9B model, using an identical initialization for all methods compared at each scale.

Reinforcement learning. Prompts are batched by task so that each rollout group shares a reward function and response format. For each prompt, the policy generates eight on-policy rollouts, and the annotation is appended as a ninth oracle rollout. All rollouts follow the task-specific answer format without chain-of-thought generation. The reward definitions are given in Appendix C, and outputs that fail format validation receive zero reward.

Optimization and selection. At the default pruning ratio $\kappa = 0.5$, sign-balanced pruning retains the oracle together with one positive and two negative on-policy rollouts. Only these four rollouts

contribute to gradient computation.

Visual processing and hardware. The default video pipeline samples at 2 fps, retains at most 128 frames, and applies a per-clip budget of 8.4 million pixels. Tracking instead uses 32 uniformly sampled frames. For video QA, segmentation, and spatial-intelligence tasks requiring higher resolution, the pixel budget ranges from 10.5 to 16.8 million. The controlled 4B comparisons use four nodes with eight NVIDIA H20 GPUs each. The full 9B RL run uses eight such nodes with full parameter sharding.

C Task Oracles and Rewards

For each task k, the adapter serializes an annotation y as an oracle rollout o_{gt} and assigns a candidate rollout o the task-aligned score $R_k(o, y)$. Let $\mathcal{A}_k$ denote the set of syntactically valid responses for task k, and define the format indicator as $F_k(o) = \mathbf{1}[o \in \mathcal{A}_k]$. A valid response contains exactly one `<answer>` block and excludes duplicate answer blocks, leaked conversational turns, and `<think>` tags. Spatial grounding also permits its task-specific fenced JSON box schema, which is verified by the spatial parser. The final score is $F_k(o)R_k(o, y)$, with no further transformation.

Temporal grounding. Following our earlier temporal-grounding formulation [45], let $[\hat{t}^s, \hat{t}^e]$ and $[t^s, t^e]$ denote the predicted and annotated intervals, respectively. The task score is temporal IoU

$$R_{\mathrm{temp}} = \frac{\max\big(0,\ \min(\hat{t}^e, t^e) - \max(\hat{t}^s, t^s)\big)}{\max(\hat{t}^e, t^e) - \min(\hat{t}^s, t^s) + \epsilon}. \tag{15}$$

Here, $\epsilon > 0$ ensures numerical stability. The oracle is the annotated interval in the model's response format.

Spatial and spatial-temporal grounding. Spatial grounding is scored by the IoU between the predicted and annotated boxes. For spatial-temporal grounding, the adapter combines temporal IoU, strict spatial IoU, temporal coverage, and framewise box quality. The corresponding oracles are serialized as an annotated box and an annotated interval with one box per frame, respectively. Highlight-style temporal queries use the same interval representation and temporal IoU score as Eq. (15).

Visual tracking. For annotated frames $\mathcal{T}$, let $\mathrm{IoU}_t = \mathrm{IoU}(\hat{b}_t, b_t)$, with $\mathrm{IoU}_t = 0$ for missing predictions. For threshold τ, we define

$$\begin{aligned} \mathrm{AO} &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathrm{IoU}_t, \\ \mathrm{R@}\tau &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbf{1}[\mathrm{IoU}_t \geq \tau]. \end{aligned} \tag{16}$$

We use the evaluation-aligned average overlap as the task reward, $R_{\mathrm{trk}} = \mathrm{AO}$, and retain $\mathrm{R@}\tau$ only as a diagnostic metric. The oracle rollout contains the complete annotated box trajectory in the model's response format.

Mask-aware video segmentation. A segmentation rollout specifies a timestamp $\hat{t}$, a box $\hat{b}$, positive points $\hat{\mathcal{P}}^+$, and negative points $\hat{\mathcal{P}}^-$. At the selected frame, we decode the annotated mask $m_{\hat{t}}$ and derive its tight box $b(m_{\hat{t}})$. Let $\mathrm{In}(\hat{\mathcal{P}}^+, m_{\hat{t}})$ denote the fraction of positive points inside the mask and $\mathrm{Out}(\hat{\mathcal{P}}^-, m_{\hat{t}})$ the fraction of negative points outside it. The reward is

$$\begin{aligned} R_{\mathrm{seg}}^{\mathrm{vid}} = 0.35\ \mathrm{IoU}\Big(\hat{b}, b(m_{\hat{t}})\Big) + 0.10\ \mathbf{1}[\hat{t}\ \text{is valid}] \\ + 0.40\ \mathrm{In}\Big(\hat{\mathcal{P}}^+, m_{\hat{t}}\Big) + 0.15\ \mathrm{Out}\Big(\hat{\mathcal{P}}^-, m_{\hat{t}}\Big). \end{aligned} \tag{17}$$

A rollout without a valid timestamp receives zero reward. To suppress degenerate solutions that exploit individual reward components, we cap the score at 0.1 when no positive point lies inside the mask and at 0.2 when the mask-box IoU is below 0.1. This mask-aware score evaluates the complete prompt consumed by the segmentation evaluator against the annotated mask rather than against a sampled proxy box or point set.

Video QA and spatial-intelligence tasks. For multiple-choice Video QA and categorical spatial-intelligence tasks, let $\hat{a}$ and a denote the parsed predicted and annotated answer tokens, respectively. We use the exact-match score

$$R_{\mathrm{cat}} = \mathbf{1}[\hat{a} = a]. \tag{18}$$

Spatial intelligence additionally includes numerical perception tasks, such as object counting, absolute-distance estimation, object-size estimation, and room-size estimation. For a predicted scalar $\hat{v}$ and a nonzero annotated target v, their score is mean relative accuracy over $\mathcal{C} = \{0.50, 0.55, \ldots, 0.95\}$,

$$R_{\mathrm{num}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}\left[\frac{|\hat{v} - v|}{|v|} \leq 1 - c\right]. \tag{19}$$

The corresponding oracle rollout contains the annotated answer token or numerical target serialized in the model's response format.

D Full Segmentation Results

The following table extends the segmentation comparison in the main text to every RefCOCO split and reports both R@0.5 and cIoU for the image benchmarks.

E Full Spatial-Intelligence Results

The following table reports complete task-level results for MMSI-Bench and MindCube-Tiny, which the main text summarizes by their two-benchmark average.

F Evaluation on ReVSI

ReVSI [111] corrects noisy 3D-derived answers and frame-budget mismatches in VSI-Bench through expert re-annotation and budget-specific answer sets. Because Video-ORA is trained on VSI-590K [96] and leads VSI-Bench in Tab. 6, we test whether its gain persists under this corrected protocol. We use the official code and pair 64-frame inputs with the corresponding answer set and 128-frame inputs with the all-frame answer set. Baseline results are taken from the ReVSI paper. At 128 frames, Video-ORA-9B scores 58.2 in Tab. 20, the highest open-source result, with Gemini-3-Pro the only model above it. At 64 frames, it remains the strongest open-source model and improves over the matched

Table 18 Complete segmentation results. RefCOCO/+/g use cIoU on the val set, whereas MeViS and ReasonVOS report J, F, and their mean J&F.

Models	Image Segmentation			Video Segmentation					
	RefCOCO [101]	RefCOCO+ [101]	RefCOCOg [54]	MeViS [17]			ReasonVOS [3]		
	cIoU	cIoU	cIoU	J	F	J&F	J	F	J&F
PixelLM-7B [61]	73.0	66.3	69.3	–	–	–	–	–	–
LISA-7B [35]	74.1	62.4	66.4	35.1	39.4	37.2	29.1	33.1	31.1
VISA-13B [89]	72.4	59.8	65.5	–	–	44.5	–	–	–
Seg-R1-7B [99]	74.3	62.6	71.0	–	–	–	–	–	–
Sa2VA-4B [103]	78.9	71.7	74.1	–	–	46.2	–	–	–
MomentSeg-7B [16]	77.8	68.2	70.8	–	–	–	–	–	–
VideoSeg-R1-7B [88]	78.2	71.8	73.1	–	–	–	–	–	–
ReferFormer [84]	–	–	–	29.8	32.2	31.0	30.2	35.6	32.9
VideoLISA-3.8B [3]	–	–	–	41.3	47.6	44.4	45.1	49.9	47.5
Veason-R1-7B [25]	–	–	–	48.4	56.0	52.2	56.0	63.8	59.9
Qwen3-VL-8B [2]	73.8	65.1	70.1	19.4	26.4	22.9	16.6	22.7	19.6
OneThinker-8B [21]	75.8	67.1	70.8	48.8	56.7	52.7	51.1	58.7	54.9
Qwen3.5-4B [71]	74.7	64.0	69.0	23.5	31.8	27.7	18.0	23.3	20.7
Qwen3.5-9B [71]	75.2	65.6	70.0	28.9	35.3	32.1	19.0	24.0	21.5
Video-ORA-4B	76.7	68.7	73.0	53.5	61.5	57.5	**60.1**	67.4	**63.8**
Video-ORA-9B	**79.4**	**72.6**	**75.1**	**57.7**	**64.8**	**61.3**	59.8	**67.6**	63.7

Table 19 Complete results on MMSI-Bench [95] and MindCube-Tiny [75]. Scores are accuracies; C/O/R denote the MMSI-Bench camera/object/region categories.

Models	MMSI-Bench												MindCube-Tiny			
	Avg.	Positional Relationship						Attribute		Motion		MSR	Avg.	Rot.	Amg.	Ard.
		C-C	O-O	R-R	C-O	O-R	C-R	Meas.	Appr.	Cam.	Obj.					
Proprietary Models																
Gemini-2.5 [13]	38.0	38.7	34.0	**40.7**	44.2	38.8	41.0	**62.5**	30.3	39.2	25.0	33.3	57.6	88.0	44.9	63.2
Grok-4 [85]	37.8	36.6	35.1	39.5	34.9	**45.9**	50.6	21.9	22.7	**40.5**	**43.4**	**38.4**	**63.6**	93.0	54.4	61.6
GPT-5 [67]	**41.8**	**41.9**	33.0	35.8	49.8	42.4	**68.7**	54.7	37.4	28.3	40.8	36.4	56.3	**94.5**	38.2	68.4
Open-source General Models																
Qwen3-VL-8B [2]	31.1	28.0	**37.2**	32.1	31.4	35.3	38.5	37.5	15.2	27.0	28.9	29.8	29.4	29.5	28.6	31.2
InternVL3-8B [117]	28.0	22.6	22.3	34.6	31.4	42.4	33.7	25.0	19.7	20.3	34.2	24.8	41.5	36.5	38.1	53.6
Qwen3.5-4B [71]	31.6	29.0	31.9	24.7	37.2	34.1	45.8	42.2	22.7	28.4	26.3	28.8	46.1	42.0	41.5	60.4
Qwen3.5-9B [71]	31.7	29.0	35.1	29.6	32.6	31.8	41.0	45.3	27.3	24.3	29.0	28.8	41.2	33.5	38.3	54.4
Open-source Spatial Intelligence Models																
SpaceR-7B [56]	27.4	25.8	31.9	29.6	25.6	31.8	22.9	26.6	28.8	16.2	34.2	27.3	38.0	35.0	34.2	49.2
ViLaSR-7B [83]	30.2	29.0	35.1	28.4	39.5	40.0	44.6	31.2	16.7	17.6	31.6	23.2	35.1	35.5	31.0	44.4
VST-7B-SFT [94]	32.5	39.8	36.2	35.8	37.2	29.4	33.7	29.7	**47.0**	36.5	35.5	18.2	39.7	37.0	35.9	50.8
Cambrian-S-7B [96]	27.1	24.7	26.6	24.7	47.7	22.4	31.3	32.8	24.2	12.2	30.3	24.2	37.9	33.0	39.0	39.2
Video-ORA-4B	31.9	31.2	35.1	34.6	43.0	27.1	44.6	39.1	21.2	27.0	29.0	25.8	48.6	37.0	49.8	54.8
Video-ORA-9B	37.9	**41.9**	30.9	30.9	**57.0**	41.2	50.6	39.1	21.2	36.5	23.7	**38.4**	57.4	44.0	**57.2**	**68.8**

Qwen3.5-9B backbone from 51.6 to 55.8, with gains of 2.5 to 7.6 points on six of seven tasks. The matched training gain therefore survives the ReVSI corrections. However, Video-ORA drops by 14.9 points from VSI-Bench to ReVSI, compared with 6.3 points for the backbone, so its leading score does not imply greater robustness to the correction. Relative direction is the only task that does not improve over the backbone.

Relative-direction analysis. ReVSI augments relative-direction evaluation with backward questions, in which the observer faces away from the orienting object. At 128 frames, Video-ORA achieves 91.5% accuracy on forward questions but only 8.3% on backward questions. Cambrian-S and VLM3R show similar imbalances at 77.0% versus 20.0% and 84.3% versus 14.6%, whereas Gemini-3-Pro is nearly symmetric at 55.7% versus 56.4%. Under matched inputs, Video-ORA raises forward accuracy over Qwen3.5-9B from 77.4% to 85.8% but lowers backward accuracy from 18.4% to 9.1%. These opposing changes leave the aggregate score nearly unchanged, matching VSI-590K's forward-only templates and motivating backward targets for future spatial training.

Table 20 Results on ReVSI [111]. Numerical and multiple-choice tasks use MRA and accuracy, respectively; Avg. is the reported aggregate over seven tasks, and the last column gives the original VSI-Bench average. All baseline rows, including their VSI-Bench averages, are quoted from [111], which evaluates each model at its native frame setting and restricts proprietary models to a 1,093-question subset, so these values differ from our own measurements in Tab. 6. The backbone and Video-ORA rows are evaluated by us.

Method	Frames	Numerical Question				Multiple-Choice Question			Avg.	VSI-Bench Avg.
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan		
Chance (frequency)	all	52.2	40.1	17.4	20.9	25.8	31.9	30.2	31.4	34.0
Proprietary models										
GPT-5.2 [67]	64	56.2	41.5	73.9	**63.0**	48.4	34.9	38.2	50.9	49.2
Gemini-3-Flash [26]	1 fps	**65.7**	53.1	77.6	52.8	64.6	47.9	41.8	57.6	55.9
Gemini-3-Pro [26]	1 fps	60.1	54.7	**79.3**	51.9	**68.1**	**56.0**	**56.4**	**60.9**	60.5
Open-source general models										
LLaVA-Video-7B [112]	64	31.3	1.4	52.5	16.7	38.3	33.3	38.4	30.3	36.3
LLaVA-Video-72B [112]	64	40.1	29.6	59.3	27.9	39.6	24.8	43.0	37.8	39.7
InternVL3.5-8B [76]	64	43.3	54.6	64.2	47.6	45.0	36.3	44.4	47.9	55.4
InternVL3.5-38B [76]	64	43.8	60.6	70.2	58.4	57.4	45.9	42.7	54.1	60.8
Qwen3-VL-8B [2]	64	40.4	52.3	69.0	45.1	57.1	39.5	40.5	49.1	57.4
Qwen3-VL-32B [2]	64	46.9	65.0	70.4	55.8	53.8	34.0	47.3	53.3	61.8
Open-source spatial intelligence models										
SpaceR-7B [56]	32	30.7	34.5	52.0	18.6	22.8	34.5	20.2	30.5	43.5
Spatial-MLLM-4B [81]	16	41.5	40.0	53.1	–	30.7	39.2	–	40.9	53.8
VST-7B [94]	4 fps	35.4	52.6	67.9	47.2	49.2	36.9	35.4	46.4	65.2
Cambrian-S-7B [96]	128	48.4	60.5	65.5	46.7	37.1	48.5	37.0	49.1	67.5
VLM3R-7B [19]	32	41.6	61.6	64.8	52.5	46.5	49.5	34.1	50.1	60.9
Qwen3.5-9B [71] (backbone)	64	38.8	58.1	67.9	45.2	58.9	47.9	44.4	51.6	57.9
Video-ORA-9B	64	46.4	65.3	70.4	50.2	62.3	47.4	48.8	55.8	73.1
Video-ORA-9B	128	48.8	**69.1**	72.7	51.1	66.3	49.9	49.8	58.2	73.1

G Qualitative Examples

Fig. 9 through Fig. 15 present qualitative results across all seven task families. All examples are held-out evaluation samples, and the predictions are generated by the released Video-ORA checkpoint. Each card shows the original prompt, sampled frames, and serialized outputs from the annotation (`GT`) and model (`PRED`). Amber dashed marks denote annotations, whereas green marks denote model predictions, following the visual convention used throughout the paper. Their overlap indicates agreement. For tracking and spatial-temporal grounding, long structured outputs are abbreviated to the entries at the displayed timestamps. The complete outputs contain one box for every second in the target interval, as required by the prompt.

Temporal Grounding

STRUCTURED PERCEPTION · 3 EXAMPLES

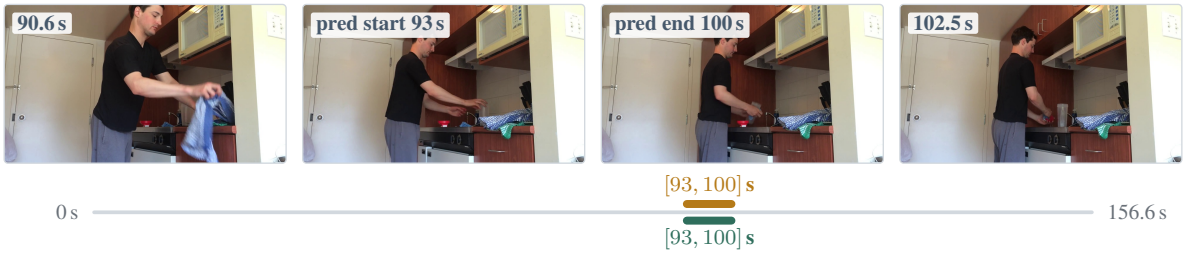
Localize a queried event in an untrimmed video.

<video> + query → <answer> t_s to t_e </answer>

EXAMPLE 1 · ACTIVITYNET

Frames are decoded at the predicted boundaries and 2.5 s outside them.

<video> To accurately pinpoint the event "He washed a cup" in the video, determine the precise time period of the event. Provide the start and end times in seconds within <answer> </answer> tags.



GT <answer> 93 to 100 </answer>

PRED <answer> 93 to 100 </answer>

EXAMPLE 2 · CHARADES-STA

A two-second event inside a 36-second indoor clip.

<video> To accurately pinpoint the event "A person sneezes" in the video, determine the precise time period of the event. Provide the start and end times in seconds within <answer> </answer> tags.

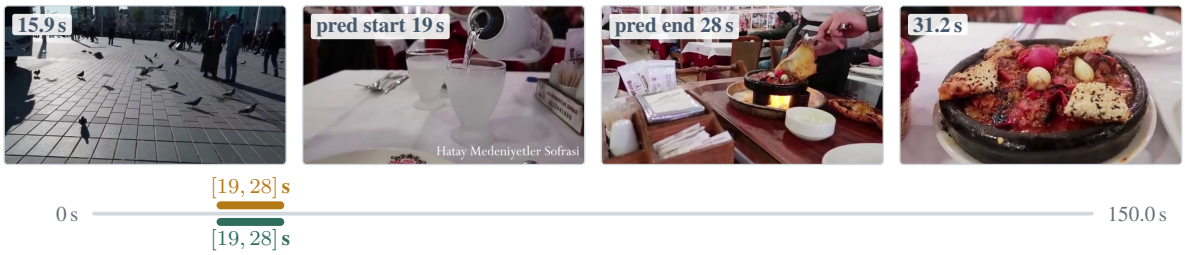
GT <answer> 29 to 31 </answer>

PRED <answer> 29 to 31 </answer>

EXAMPLE 3 · QVHIGHLIGHTS

A vlog whose target segment covers six percent of the timeline.

<video> To accurately pinpoint the event "A chef is cooking" in the video, determine the precise time period of the event. Provide the start and end times in seconds within <answer> </answer> tags.



GT <answer> 19 to 28 </answer>

PRED <answer> 19 to 28 </answer>

Figure 9 Temporal-grounding qualitative examples. Frames are decoded at the predicted boundaries and just outside them, and the timeline carries the annotated span above the axis and the predicted span below it. Video-ORA recovers all three intervals exactly, including a two-second event and a segment covering six percent of the timeline.

Spatial Grounding

STRUCTURED PERCEPTION · 3 EXAMPLES

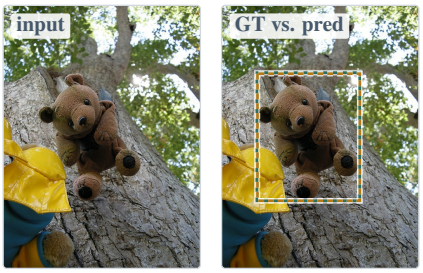
Locate a referred object and return its normalized image coordinates.

<image> + expression → box JSON

EXAMPLE 1 · REFCOCO testB

IoU 0.98 against the annotated box.

<image> Locate “the teddy bear” in the image. Return its bounding box inside <answer>...</answer>.



GT <answer>[{"bbox_2d": [182, 260, 701, 743]}]</answer>

PRED <answer>[{"bbox_2d": [179, 256, 700, 743]}]</answer>

EXAMPLE 2 · REFCOCO+ testA

IoU 0.99; the expression must separate two people in the scene.

<image> Locate “woman on phone” in the image. Return its bounding box inside <answer>...</answer>.

GT <answer>[{"bbox_2d": [410, 53, 737, 988]}]</answer>

PRED <answer>[{"bbox_2d": [409, 56, 739, 987]}]</answer>

EXAMPLE 3 · REFCOCOg test

IoU 1.00; the target is truncated at the right image border.

<image> Locate “a bus” in the image. Return its bounding box inside <answer>...</answer>.

GT <answer>[{"bbox_2d": [631, 0, 1000, 520]}]</answer>

PRED <answer>[{"bbox_2d": [630, 0, 1000, 520]}]</answer>

Figure 10 Spatial-grounding qualitative examples. Each card pairs the raw input with a copy carrying both boxes in coordinates normalized to $[0, 1000]$. The two outlines are nearly indistinguishable, with an IoU of 0.98 or above on all three referring expressions.

Mask-Aware Segmentation

STRUCTURED PERCEPTION · 3 EXAMPLES

Predict the box and signed point prompts that drive the segmenter.

<video>/<image> + expression → box + signed points

EXAMPLE 1 · REFCOCO IMAGE

Predicted box within 1% of the tight box of the annotated mask.

<image> Given the referring expression "ketchup", provide one bounding box, along with positive points located inside the object and negative points located outside, as segmentation hints.

GT run-length mask, 426x640; tight box [439, 0, 732, 247]

PRED
```
<answer>{"boxes": [435, 0, 740, 245],
"positive_points": [[526, 100], [608, 150], [573, 50]],
"negative_points": [[154, 125], [353, 50], [859, 200]]}</answer>
```

EXAMPLE 2 · MEVIS VIDEO

Panels are zoomed on the target, which the cue "parking" disambiguates.

<video> Given the referring expression "parking white car", provide one bounding box, along with positive points located inside the object and negative points located outside, as segmentation hints.

GT run-length mask, 960x540; tight box [812, 135, 917, 252]

PRED
```
<answer>{"time": 0.0, "boxes": [814, 142, 917, 247],
"positive_points": [[845, 175], [878, 195], [895, 165]],
"negative_points": [[750, 200], [950, 150], [800, 100]]}</answer>
```

EXAMPLE 3 · REASONVOS VIDEO

The expression requires reasoning rather than a literal category name.

<video> Given the referring expression "Considering our plans for a self-driving trip with the family this weekend, which vehicle would be the most appropriate?", provide one bounding box together with positive and negative point hints.

GT run-length mask, 960x540; tight box [367, 331, 759, 696]

PRED
```
<answer>{"time": 0.0, "boxes": [357, 332, 757, 715],
"positive_points": [[450, 450], [550, 500], [650, 400]],
"negative_points": [[150, 250], [800, 400], [200, 600]]}</answer>
```

Figure 11 Mask-aware segmentation qualitative examples. The left panel carries the predicted box with its positive (dots) and negative (crosses) point prompts, and the right panel tints the annotation mask decoded from its run-length target. The model never emits a mask itself, so the comparison is between the predicted hints and the tight box of the annotated mask, which agree to within one percent.

Visual Tracking

STRUCTURED PERCEPTION · 3 EXAMPLES

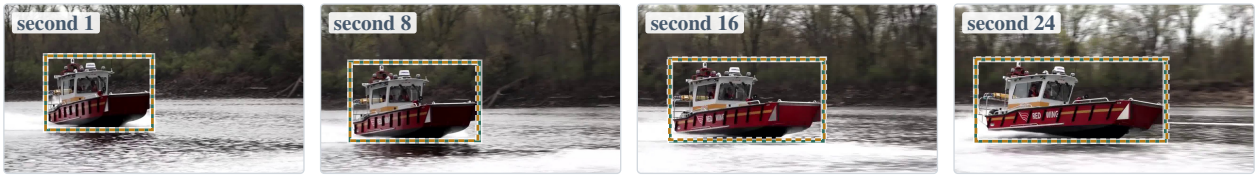
Propagate an initial target box through a video at one-second intervals.

<video> + initial box → box trajectory

EXAMPLE 1 · **GOT-10k val 000001**

Mean IoU 0.97 over all 32 predicted seconds; 4 shown.

<video> Given the bounding box [139, 306, 495, 743] of the target object in the first frame, track this object and output one bounding box per second, up to second 32, inside <answer>...</answer>.



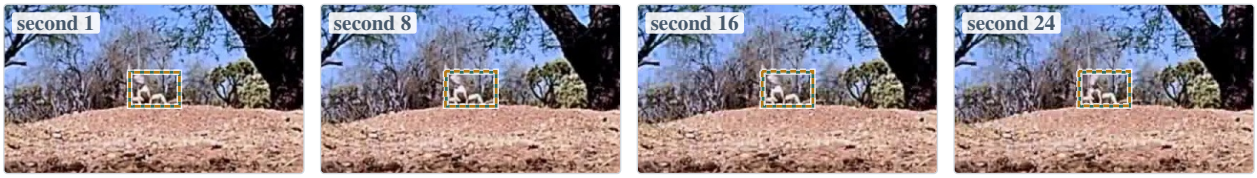
GT <answer>{"boxes": {"1": [139,306,495,743], "8": [98,344,524,807], "16": [107,326,623,801], "24": [73,325,709,810], ...}}</answer>

PRED <answer>{"boxes": {"1": [139,306,495,743], "8": [98,344,530,808], "16": [110,328,620,811], "24": [72,328,704,811], ...}}</answer>

EXAMPLE 2 · **GOT-10k val 000042**

Mean IoU 0.96; frames are zoomed on a small, nearly static target.

<video> Given the bounding box [331, 603, 386, 669] of the target object in the first frame, track this object and output one bounding box per second, up to second 32, inside <answer>...</answer>.



GT <answer>{"boxes": {"1": [331,603,386,669], "8": [330,601,386,669], "16": [330,601,386,669], "24": [330,601,386,668], ...}}</answer>

PRED <answer>{"boxes": {"1": [331,603,386,669], "8": [331,603,386,669], "16": [331,603,386,669], "24": [331,603,386,669], ...}}</answer>

EXAMPLE 3 · **GOT-10k val 000112**

Mean IoU 0.96 while the target grows and drifts upward.

<video> Given the bounding box [407, 526, 665, 863] of the target object in the first frame, track this object and output one bounding box per second, up to second 32, inside <answer>...</answer>.

GT <answer>{"boxes": {"1": [407,526,665,863], "8": [446,494,746,901], "16": [374,390,762,954], "24": [357,285,789,956], ...}}</answer>

PRED <answer>{"boxes": {"1": [407,526,665,863], "8": [445,494,744,891], "16": [378,391,764,951], "24": [361,291,790,951], ...}}</answer>

Figure 12 Visual-tracking qualitative examples. Four of the thirty-two predicted seconds are shown, covering a target that doubles in width, one that barely moves, and one that drifts upward while growing. Mean IoU over the full trajectory stays at 0.96 or above in all three sequences.

Spatial-Temporal Grounding

Localize a referred event jointly in time and space.

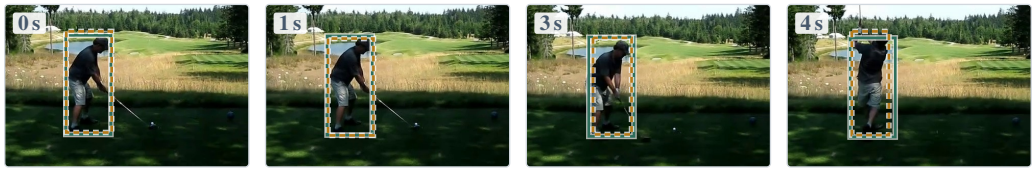
STRUCTURED PERCEPTION · 3 EXAMPLES

<video> + query → time span + per-second boxes

EXAMPLE 1 · HUMAN MOTION

tIoU 0.94 and mean box IoU 0.87; frames are cropped vertically.

<video> Who holds the bat on the grass? Find the corresponding time period, and give the spatial bounding box for each integer second inside it.



GT <answer>{"time": [0.0, 5.105], "boxes": {"0": [252,433,438,744], ...}}</answer>

PRED <answer>{"time": [0.0, 4.8], "boxes": {"0": [250,445,435,750], ...}}</answer>

EXAMPLE 2 · OUTDOOR SCENE

tIoU 0.94; the prediction keeps one box for the whole span.

<video> Who is behind the adult on the wild? Find the corresponding time period, and give the spatial bounding box for each integer second inside it.

GT <answer>{"time": [0.0, 15.983], "boxes": {"0": [281,300,436,862], ...}}</answer>

PRED <answer>{"time": [0.0, 15.0], "boxes": {"0": [275,308,434,846], ...}}</answer>

EXAMPLE 3 · CONVERSATION

tIoU 0.88 and mean box IoU 0.89; the span is rounded outward.

<video> Who speaks to the other adult? Find the corresponding time period, and give the spatial bounding box for each integer second inside it.

GT <answer>{"time": [0.417, 5.667], "boxes": {"1": [253,352,484,998], ...}}</answer>

PRED <answer>{"time": [0.0, 6.0], "boxes": {"0": [253,364,484,997], ...}}</answer>

Figure 13 Spatial-temporal grounding qualitative examples. The timeline compares the annotated and predicted intervals while the boxes compare the referred entity at four integer seconds. Temporal IoU stays above 0.87; the wildlife clip shows a typical failure mode, where the model keeps a single box for the entire span instead of tracking the subject.

Video Question Answering

ANSWER-ONLY REASONING · 3 EXAMPLES

Reason over long-form narrative, fine-grained detail, and attributes.

<video> + question → option letter

EXAMPLE 1 · VIDEO-MME long

Correct; the count is aggregated over a 10-minute broadcast.

<video> What is the largest run differential in the first inning of the game in the video?
A. 4. B. 3. C. 2. D. 1.
Answer with the option letter inside <answer>...</answer>.

GT <answer>A</answer> 4.

PRED <answer>A</answer> 4.

EXAMPLE 2 · VIDEO-MME medium

Correct; the deciding detail is visible only briefly.

<video> What sentence best describes the performance?
A. The magic is not wonderful enough to make the audiences cheer up. B. One of the judges stops the performance. C. The magician does the magic with his eyes closed. D. The magician needs other tools besides leathers.

GT <answer>C</answer> The magician does the magic with his eyes closed.

PRED <answer>C</answer> The magician does the magic with his eyes closed.

EXAMPLE 3 · VIDEO-MME short

Correct; a short clip with a single salient attribute question.

<video> What color is the cat?
A. Yellow. B. White. C. Black. D. Blue.
Answer with the option letter inside <answer>...</answer>.

GT <answer>A</answer> Yellow.

PRED <answer>A</answer> Yellow.

Figure 14 Video question-answering qualitative examples. One short, one medium, and one long Video-MME item, spanning attribute recognition, a briefly visible detail, and a count that has to be aggregated over a ten-minute broadcast.

Spatial Intelligence

ANSWER-ONLY REASONING · 3 EXAMPLES

Integrate egocentric views into metric and relational scene judgments.

<video> + spatial question → option letter or value

EXAMPLE 1 · VSI-BENCH appearance order

Correct over a 110-second scan of an unseen room.

<video> What will be the first-time appearance order of the following categories in the video: ceiling light, cup, heater, door?
A. cup, door, heater, ceiling light B. ceiling light, door, cup, heater C. heater, cup, door, ceiling light D. ceiling light, cup, heater, door

GT <answer>A</answer> cup, door, heater, ceiling light

PRED <answer>A</answer> cup, door, heater, ceiling light

EXAMPLE 2 · VSI-BENCH absolute distance

Correct to the annotated value; the answer is a metric quantity.

<video> Measuring from the closest point of each object, what is the direct distance between the toilet and the bathtub? Answer with a number in meters within <answer>...</answer> tags.

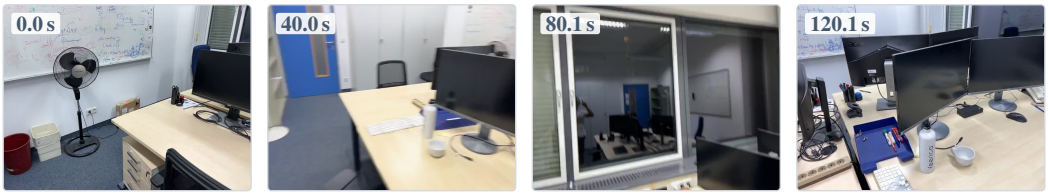
GT <answer>0.4</answer> metres

PRED <answer>0.4</answer> metres

EXAMPLE 3 · VSI-BENCH relative direction

Correct; anchor, facing target, and queried object are never co-visible.

<video> If I am standing by the telephone and facing the trash can, is the cup to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it.
A. right B. back C. left



GT <answer>C</answer> left

PRED <answer>C</answer> left

Figure 15 Spatial-intelligence qualitative examples. Sampled egocentric scans of unseen rooms support first-appearance order, a metric distance regressed in metres, and a relative direction whose anchor and query object are never visible in the same frame.