ANCHORPLAY: RELIABLE SELF-PLAY FOR VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-play can expand training supervision for video grounding without additional human annotations, but it also risks reinforcing its own localization errors: agreement among predictions does not establish that an event's boundaries are correct. We introduce ANCHORPLAY, an end-to-end self-play framework that jointly optimizes a shared model as a questioner and a grounding solver. The central idea is to decouple question evolution from the construction of localization targets. A frozen reference model proposes evidence anchors, whose reliability is assessed by comparing predictions across transformed views after mapping them into a common coordinate system. Conditioned on accepted anchors, the questioner learns to generate queries matched to the solver's capabilities, while the solver learns to localize the corresponding objects or events. Both roles are updated in the same optimization step, allowing the curriculum to evolve around targets supplied by a reference model independent of the evolving policy. Reliability-weighted reinforcement learning emphasizes better-supported targets, while grounding-preservation regularization encourages retention of reliable localization behavior from the initial model. Experiments on temporal and spatio-temporal grounding benchmarks show improvements over initial checkpoints across both general-purpose and task-adapted backbones. Although self-play training uses only grounding tasks, ANCHORPLAY also improves performance on additional video and image understanding tasks, demonstrating transfer beyond its training objectives.

1 INTRODUCTION

A video model may recognize that a person opens a cupboard while predicting an interval that begins before the action or extends into an unrelated event. Temporal grounding requires both understanding a language query and locating the event's temporal extent (Gao et al., 2017). Recent vision–language models incorporate temporal localization alongside general visual reasoning (Feng et al., 2026), but improving localization still calls for supervision with accurate boundaries. Self-play offers a way to construct training tasks from unlabeled inputs. As illustrated in Figure 1, however, its usefulness depends on the targets that determine whether a prediction is rewarded. This raises a central question: *how can a model learn to ground through self-play when the supervision it generates may itself be imprecise?*

Self-play constructs adaptive supervision by coupling task generation with solver feedback (Liu et al., 2025; He et al., 2025), while reliability mechanisms such as RISE's quality supervisor further screen generated supervision (Xu et al., 2026). Recent video approaches illustrate two sources of localization targets. Video-Zero generates questions around evidence segments and obtains temporal targets from the evolving solver's rollout consensus (Zhang et al., 2026). Other approaches use targets defined by construction: EvoVid provides the boundaries of sampled video windows as temporal supervision (Huang et al., 2026b), and VideoSSR constructs pretext tasks with known answers (He et al., 2026). These approaches reduce reliance on human labels, but their targets offer different assurances about where the queried event actually begins and ends.

For grounding, *task difficulty* and *reference uncertainty* describe different properties of a training example. A query can be challenging because it distinguishes similar events, even when its target interval is precise; an easy query can still carry inaccurate start or end times. When the evolving solver supplies its own targets, agreement among its predictions can reinforce the same boundary error. Targets defined by construction avoid this dependence, but a sampled window's boundaries need not coincide with the extent of the event referred to by a question. Consequently, neither

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

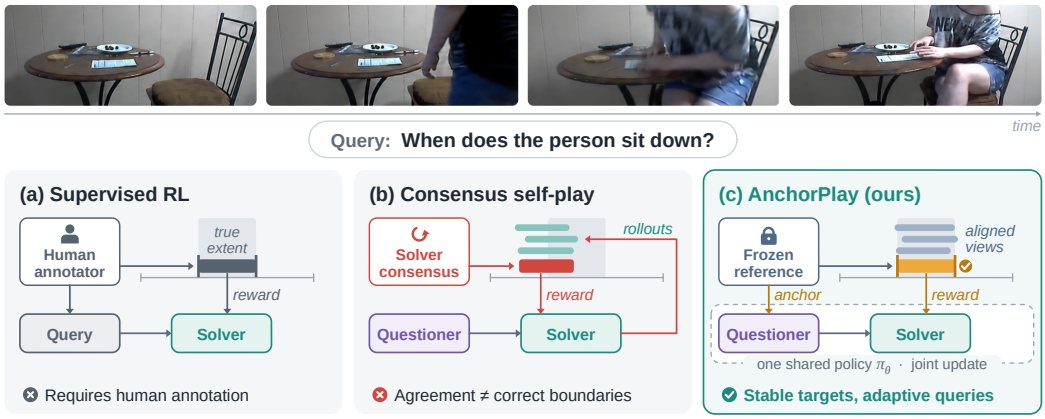


Figure 1: **Three sources of supervision for video grounding.** ANCHORPLAY uses screened reference anchors to guide evolving questions and solver learning within a jointly optimized policy.

answer agreement nor a known construction interval alone establishes accurate event boundaries. We therefore separate the evolving question curriculum from the source and reliability assessment of the localization targets used to train the solver.

We introduce ANCHORPLAY, an end-to-end self-play framework for video grounding (Figure 1(c)). A frozen copy of the initial model proposes candidate evidence anchors. For temporal grounding, each anchor specifies an event interval, whose reliability is assessed by localizing the same event across transformed views and comparing predictions in the original video's coordinates. The questioner conditions on accepted anchors to generate queries matched to the solver's capabilities, while the solver learns to localize their referents against these anchors. The two roles share one policy and are optimized in the same step. Estimated anchor reliability weights the solver's learning signal, and a grounding-preservation objective encourages retention of reliable localization behavior from the initial model. This separation allows the question curriculum to evolve while keeping its supervision tied to a frozen reference and screened for cross-view consistency.

We evaluate ANCHORPLAY with Qwen3-VL-8B-Instruct and OneThinker-SFT-8B, covering both a general-purpose initialization and a checkpoint already adapted to visual reasoning tasks. Both backbones improve over their initial checkpoints on Charades-STA, ActivityNet, and ANet-RTL, with further gains in spatio-temporal grounding. These improvements are obtained using only grounding tasks during self-play, without additional human-provided questions or localization annotations. Additional gains on other video understanding and image tasks indicate that the benefits can extend beyond the tasks used for self-play training.

2 RELATED WORK

Learning for video grounding. Video grounding requires models to associate language with precise temporal intervals or spatio-temporal regions. Recent reinforcement learning methods optimize these structured predictions using localization rewards: Time-R1 improves temporal grounding with rewards computed against annotated intervals (Wang et al., 2025), while VideoChat-R1 extends reinforcement fine-tuning to spatio-temporal perception, including temporal grounding and tracking (Li et al., 2025b). OneThinker incorporates grounding alongside other image and video tasks within a unified reasoning model (Feng et al., 2026). Annotation-free grounding has also been studied outside self-play. For example, language-free training uses visual representations in CLIP's shared embedding space as substitutes for language features, learning from sampled intervals without task-specific query–interval annotations (Kim et al., 2023). These approaches establish complementary ways to acquire localization ability. We study how to further improve this ability through an adaptive question curriculum with reliability-screened localization targets.

Self-play in language and vision–language models. Self-play couples task generation with learning, while different frameworks obtain feedback from different sources. R-Zero co-evolves a challenger and a solver using self-generated questions (Huang et al., 2026a), whereas Absolute Zero constructs tasks that can be checked through code execution (Zhao et al., 2025a). SPICE grounds task generation in an external document corpus and jointly optimizes the challenger and reasoner as two roles of

one shared model (Liu et al., 2025). For visual inputs, V-Zero and VisPlay generate questions from unlabeled images and use solver consensus to construct pseudo answers (Wang et al., 2026; He et al., 2025). RISE further improves supervision through frequent role alternation, skill balancing, and a supervisor that checks question validity and pseudo-answer correctness (Xu et al., 2026). These studies establish the importance of both adaptive curricula and supervision quality. ANCHORPLAY adopts shared-policy joint optimization and focuses on a distinct reliability requirement: whether the geometric target used to reward a localization prediction is sufficiently supported by visual evidence.

Self-evolution and self-supervision for video. Recent approaches incorporate temporal information directly into self-generated supervision. Video-Zero generates questions around evidence spans and derives temporal targets primarily by taking the median start and end times of solver rollouts that agree with the majority answer (Zhang et al., 2026). EvoVid encourages temporally dependent questions through frame-order perturbations and uses the boundaries of sampled video windows as localization targets (Huang et al., 2026b). VideoSSR constructs anomaly-grounding, object-counting, and temporal-jigsaw tasks with verifiable targets determined by their construction (He et al., 2026). For temporal localization, rollout-derived targets change with the evolving solver, while a sampled window specifies a known segment without necessarily identifying the full extent of the event described by a query. Construction-based pretext tasks provide verifiable objectives, but do not by themselves define an adaptive curriculum of natural grounding queries. ANCHORPLAY obtains candidate localization targets from a frozen reference model and screens them through coordinate-aligned predictions across multiple views. Accepted anchors both condition question generation and score solver predictions, allowing the curriculum to evolve without deriving its targets from the current solver's rollouts.

3 METHOD

ANCHORPLAY establishes reliable geometric targets before learning which queries to ask about them. As shown in Figure 2, a frozen reference model proposes evidence anchors, which are screened using predictions from aligned views. One shared policy then plays two roles: a *Query Designer* generates queries conditioned on the accepted anchors, and a *Localizer* predicts their geometry from the video. Both roles are trained jointly in the same optimizer step, allowing the query curriculum to evolve while its scoring targets remain independent of the Localizer's current rollouts.

3.1 GROUNDING WITH A SHARED POLICY

We use a corpus $\mathcal{D}$ of videos without their accompanying query–answer or localization annotations. Training covers temporal grounding and spatio-temporal grounding, denoted by $\tau \in \mathcal{T} = \{\mathrm{TG}, \mathrm{STG}\}$. Temporal grounding returns an interval specifying when an event occurs; spatio-temporal grounding additionally localizes the referred subject within the relevant frames. We write g for the corresponding geometric target and $\widehat{g}(y)$ for the structured prediction extracted from a response y.

The trainable policy π_θ is initialized from π_{θ_0}. Role prompts distinguish the Query Designer (D) from the Localizer (L):

$$\begin{aligned} q &\sim \pi_\theta(\cdot \mid x, g, \tau, \mathrm{D}), \\ y &\sim \pi_\theta(\cdot \mid x, q, \tau, \mathrm{L}). \end{aligned} \tag{1}$$

The anchor is available to the Designer, whereas the Localizer must infer the answer from the video and query alone. Both roles share all policy parameters, following the shared-model self-play formulation of SPICE (Liu et al., 2025). We also retain a frozen reference $\pi_{\mathrm{ref}} = \pi_{\theta_0}$ to propose anchors, assess query validity, and provide the reference behavior for regularization.

3.2 CONSTRUCTING RELIABLE EVIDENCE ANCHORS

A target source independent of current rollouts. Several localization rollouts can agree on the same incorrect interval; scoring them against their own consensus would reward that shared error. We instead obtain candidate anchors from π_{ref}. Each anchor specifies the geometry of an event or subject in the video, which subsequent queries must continue to refer to. The frozen reference prevents the target generator from co-adapting with the Localizer; a retained anchor supplies the same target whenever its associated queries are scored.

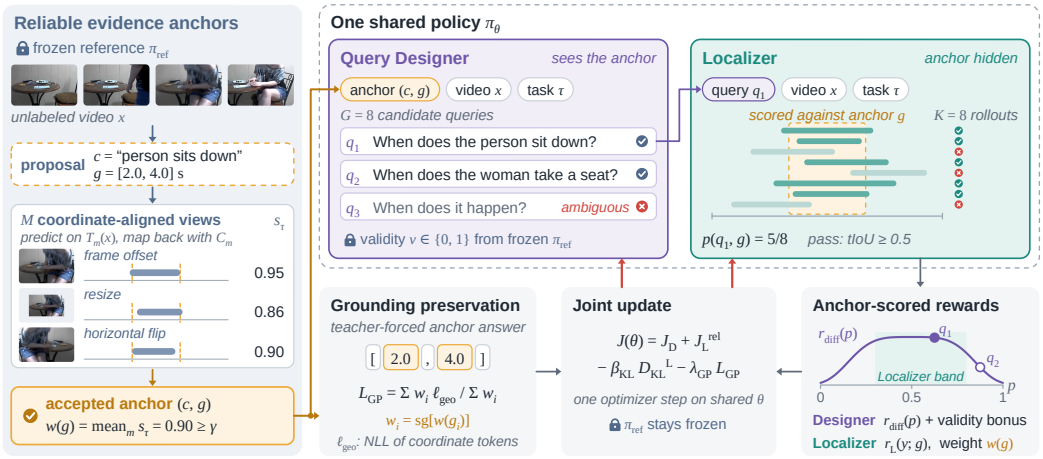


Figure 2: **Overview of AnchorPlay.** A frozen reference proposes localization anchors from unlabeled videos and screens them through coordinate-aligned multi-view consistency. The Query Designer generates anchor-conditioned queries, while the Localizer predicts their geometry without seeing the anchors. Anchor-based feedback rewards query difficulty and localization quality, with reliability weighting applied to Localizer learning. Both roles share one policy and are jointly optimized with a grounding-preservation objective.

Consistency after coordinate alignment. A single reference prediction may still have inaccurate boundaries. For each candidate g, we query the reference on multiple views $T_1(x), \ldots, T_M(x)$ that retain the same referent. Let $\widehat{g}_m$ be its prediction on view m, and let C_m map valid predictions back to the original video's coordinates. The aligned predictions define an anchor reliability score:

$$\widetilde{g}_m = C_m(\widehat{g}_m), \qquad w(g) = \frac{1}{M} \sum_{m=1}^{M} s_\tau(g, \widetilde{g}_m), \qquad w(g) \geq \gamma. \tag{2}$$

Here $s_\tau \in [0, 1]$ measures geometric overlap, and the last condition, with $\gamma \in (0, 1]$, determines acceptance. The reference is queried about the same event or subject in every view; predictions with invalid geometry receive zero agreement. Alignment is essential: predictions made under different temporal or spatial coordinates cannot be compared directly. The score measures consistency of the visual evidence across views; it is an estimate of target reliability, rather than a guarantee of correctness. The view construction and geometric scorers are described in Appendix B.

3.3 Evolving Queries around Anchored Evidence

Query generation and validity. The Designer samples candidate queries conditioned on an accepted anchor and the video. A usable query must preserve the anchored referent and its event extent: changing the wording must not silently change the correct localization target. The frozen reference evaluates parsed queries for visual relevance, unambiguous reference, and consistency with the anchored event extent, returning $v \in \{0, 1\}$. Invalid queries retain feedback for Designer training but do not trigger Localizer rollouts. This separates the opportunity to improve question generation from the cost of answering unsuitable questions.

Difficulty measured against the anchor. For each valid query, the Localizer samples $K = 8$ responses. Let $s_\tau(\widehat{g}, g)$ be the task-specific geometric score and $\chi_\tau(\widehat{g}, g)$ indicate whether the localization passes the task's overlap thresholds. The empirical pass rate is

$$p(q, g) = \frac{1}{K} \sum_{k=1}^{K} \chi_\tau\big(\widehat{g}(y_k), g\big). \tag{3}$$

For temporal grounding, s_τ is temporal intersection over union (tIoU); spatio-temporal grounding uses the corresponding scorer for its structured localization output. Unlike rollout agreement, p measures success against an independently supplied target.

Following the variance-based curriculum in SPICE (Liu et al., 2025), we reward valid queries using $r_{\mathrm{diff}}(p) = \exp\{-[p(1-p) - 0.25]^2/(2\delta)\}$, with $\delta = 0.01$. This reward peaks at intermediate

success, where the current Localizer sometimes succeeds and sometimes fails. An additional validity bonus completes the Designer reward (Appendix C). For Localizer training, each video contributes one selected candidate with $p \in (0.3, 0.8)$. The difficulty criterion changes with the shared policy, allowing the curriculum to evolve around accepted evidence without redefining the scoring target from the policy's answers. If no candidate lies in this band, the Localizer policy loss is skipped for that video; valid query–anchor pairs can still contribute to grounding preservation.

3.4 Joint Learning with Reliable Targets

Localization feedback. The Localizer receives a geometric reward together with output-format and structural-validity terms:

$$r^{\mathrm{L}}(y; g) = (1 - \lambda_f) s_\tau\big(\widehat{g}(y), g\big) + \lambda_f s_{\mathrm{fmt}}(y) + s_{\mathrm{str}}(y), \qquad \lambda_f = 0.1. \tag{4}$$

The format term checks the reasoning and answer delimiters, while the structure term checks the task's geometric answer schema.

Reliability-weighted policy learning. We first center rewards within each role's rollout group and normalize them using task-wise EMA reward scales (Feng et al., 2026), obtaining advantages A_i^{D} and A_i^{L}. The Designer has its own scale statistics, while Localizer statistics are maintained separately for the two grounding tasks. Let $\ell_{it}^{\mathrm{L}}(\theta)$ be the clipped GRPO policy surrogate for token t of Localizer rollout i. We apply reliability *after* advantage normalization:

$$\mathcal{J}_{\mathrm{L}}^{\mathrm{rel}}(\theta) = \frac{1}{N_{\mathrm{L}}} \sum_{i \in \mathcal{B}_{\mathrm{L}}} \mathrm{sg}[w(g_i)] \sum_{t=1}^{|y_i|} \ell_{it}^{\mathrm{L}}(\theta), \qquad N_{\mathrm{L}} = \sum_i |y_i|. \tag{5}$$

Here sg denotes stop-gradient. Hard screening removes unreliable anchors, while continuous weighting attenuates the influence of less reliable accepted examples. Applying the weight after normalization preserves this distinction; multiplying every reward in a group by the same number can otherwise be canceled by within-group standardization.

Preserving reliable grounding behavior. We also preserve reliable coordinate predictions through a lightweight answer-level objective. Let $a(g_i)$ serialize an accepted anchor in the task's answer schema, and let $\mathcal{C}_i$ index its coordinate tokens. Using a direct localization prompt, we teacher-force this answer. Let $\ell_{\mathrm{geo}}(\theta; i)$ be the mean negative log-likelihood of its coordinate tokens. We minimize

$$\mathcal{L}_{\mathrm{GP}} = \frac{\sum_{i \in \mathcal{B}_A} w_i \, \ell_{\mathrm{geo}}(\theta; i)}{\sum_{i \in \mathcal{B}_A} w_i}, \tag{6}$$

where $w_i = \mathrm{sg}[w(g_i)]$ and $\mathcal{B}_A$ contains valid queries paired with accepted anchors. The target supervises the geometric answer without prescribing a reference reasoning trace. It supplies a retention signal even when sampled localization rewards within a group are identical and the relative policy advantage vanishes. A separate KL penalty constrains the Localizer's full response distribution toward π_{ref}.

One update for both roles. Writing $\mathcal{J}_{\mathrm{D}}$ for the Designer's GRPO surrogate (Shao et al., 2024), the joint objective is

$$\mathcal{J}(\theta) = \mathcal{J}_{\mathrm{D}}(\theta) + \mathcal{J}_{\mathrm{L}}^{\mathrm{rel}}(\theta) - \beta_{\mathrm{KL}} \mathcal{D}_{\mathrm{KL}}^{\mathrm{L}} - \lambda_{\mathrm{GP}} \mathcal{L}_{\mathrm{GP}}, \qquad \beta_{\mathrm{KL}} = 10^{-2}. \tag{7}$$

Both policy surrogates contribute gradients to the same θ in one optimizer step. We use token-level aggregation (Yu et al., 2026) with a separate token-count denominator for each role, so their relative weight is not determined solely by the difference in response lengths. The explicit KL term applies to Localizer trajectories; both roles are nevertheless affected through their shared parameters. Sampled targets and screening decisions are held fixed during the policy update. The complete surrogate definitions appear in Appendix C.

3.5 Why Evidence Anchoring Helps

The design separates two questions: *is the supervision reliable?* and *is the query useful at the current difficulty level?* Multi-view screening addresses the first, while anchor-based pass rates guide the second. Their connection can be made precise for temporal grounding. For a valid query, let $g^\star$ be its unknown true interval. The triangle inequality for Jaccard distance (Kosub, 2019) implies, for any predicted interval $\widehat{g}$,

$$|\mathrm{tIoU}(\widehat{g}, g) - \mathrm{tIoU}(\widehat{g}, g^\star)| \le 1 - \mathrm{tIoU}(g, g^\star). \tag{8}$$

Thus, anchor error bounds the distortion of the geometric reward. More accurate anchors provide a tighter connection between optimizing the available reward and improving true localization. Reliability weighting can further reduce an expected distortion bound when its scores predict lower target error (Appendix D). Meanwhile, the frozen target source leaves the query curriculum free to adapt through joint self-play. This explains the complementary roles of anchoring and curriculum evolution; it does not assume that multi-view agreement establishes ground truth or guarantees improvement after optimization.

4 Experiments

4.1 Experimental Setup

Training data and supervision. We construct the adaptation corpus from the temporal-grounding and spatio-temporal-grounding portions of OneThinker's released RL training data (Feng et al., 2026), comprising x unique videos. We retain visual inputs and task identities but discard the original queries, ground-truth intervals, bounding boxes, and reasoning traces. The Query Designer supplies training queries; the frozen reference supplies localization anchors screened through multi-view consistency. Only these two video-grounding tasks are used for adaptation. Image understanding, referring-expression grounding, segmentation, tracking, and video QA are evaluated without additional training.

Initializations and checkpoint protocol. We use Qwen3-VL-8B-Instruct (Bai et al., 2025), OneThinker-SFT-8B (Feng et al., 2026), and Qwen3.5-9B (Team, 2026). Each run is compared with its own initialization and uses its own frozen reference. The same checkpoint is used across all benchmarks within a run; these comparisons are not matched in training duration across backbones. The full OneThinker-8B is an annotated-training reference for temporal grounding, rather than our initialization. Annotation-free training refers to the adaptation stage, not the initial models' pretraining or supervised training histories.

Implementation. The Query Designer and Localizer share one policy and are jointly optimized with EMA-GRPO (Feng et al., 2026). We use eight NVIDIA H100 GPUs, 128 videos per step, a learning rate of 2×10^{-6}, and a Localizer KL coefficient of 0.01. Each video produces $G = 8$ query candidates, with $K = 8$ Localizer rollouts per valid query. One eligible query per video contributes to the Localizer policy loss. Optimizer steps belong to a continuous joint-training run, rather than separate Designer and Localizer training stages. Additional configuration is given in Appendix E.

Benchmarks and metrics. Temporal grounding uses Charades-STA (Gao et al., 2017), ActivityNet Captions (Krishna et al., 2017), and ANet-RTL (Huang et al., 2024), with temporal mIoU and R@$\{0.3, 0.5, 0.7\}$. Spatio-temporal grounding is evaluated on the STVG task of ST-Align (Li et al., 2025a), which builds on VidSTG (Zhang et al., 2020) with textual descriptions refined using GPT-4-turbo. We use the released validation set and follow OneThinker's evaluation implementation (Feng et al., 2026), reporting mean temporal IoU (tIoU) and mean spatial IoU (sIoU) separately in the main comparison. Image understanding uses VLMEvalKit (Duan et al., 2024); other transfer tasks follow the metric definitions in the appendix. All scores are percentages. We report task-specific results and within-task averages, without averaging heterogeneous metrics into a single overall score.

4.2 Main Grounding Results

Grounding improves across initializations. Tables 1 and 2 show gains in both temporal and spatio-temporal grounding. OneThinker-SFT improves average temporal mIoU by 11.06 points, showing that self-play can build on an already supervised model. Qwen3-VL and Qwen3.5 improve across the temporal benchmarks as well (+8.25 and +7.63 average mIoU), and all three initializations gain on STVG in both components, the largest being Qwen3.5 (tIoU 27.45 $\rightarrow$ 45.35). These paired comparisons support the effectiveness of adapting grounding behavior through self-generated tasks. They do not require the original training queries or target geometry.

Why anchoring helps. The method connects changing training questions to a reference that does not co-evolve with the policy. The Designer formulates a query around an accepted anchor, whereas the Localizer must recover its geometry without seeing the target. This gives the interaction a concrete localization objective beyond agreement between the two roles. Screening rejects inconsistent anchors, and reliability weighting moderates uncertain accepted targets. Difficulty feedback then changes which supported queries the learner encounters. The observed improvements are consistent

Table 1: **Temporal grounding across initializations.** ANCHORPLAY improves end-to-end grounding scores without using the original training annotations.

Method	Frames	Charades-STA				ActivityNet				ANet-RTL			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Published supervised reference													
OneThinker-8B†	128	81.18	66.13	39.33	56.74	57.32	38.23	22.62	40.80	52.84	37.99	20.96	37.59
Initialized from Qwen3-VL-8B-Instruct													
Base Model	128	57.93	33.49	13.33	36.75	39.84	26.17	15.31	29.13‡	37.12	27.51	17.03	27.17
ANCHORPLAY	128	**76.37**	**54.89**	**25.91**	**49.43**	**48.33**	**30.87**	**17.81**	**34.95**‡	**47.60**	**30.13**	**18.78**	**33.43**
Initialized from OneThinker-SFT-8B													
Base Model	128	79.73	64.49	36.77	55.15	28.72	15.67	7.70	21.23	26.20	13.10	6.99	18.55
ANCHORPLAY	128	**81.75**	**65.54**	**37.77**	**56.65**	**54.04**	**35.01**	**20.44**	**38.34**	**46.72**	**34.50**	**19.65**	**33.12**
Initialized from Qwen3.5-9B													
Base Model	128	82.08	69.54	46.32	59.26	61.55	43.74	28.96	45.16	50.02	38.27	25.11	38.42
ANCHORPLAY	128	**84.54**	**70.64**	**48.70**	**61.82**	**74.70**	**52.86**	**32.29**	**53.17**	**70.65**	**51.95**	**38.79**	**50.74**

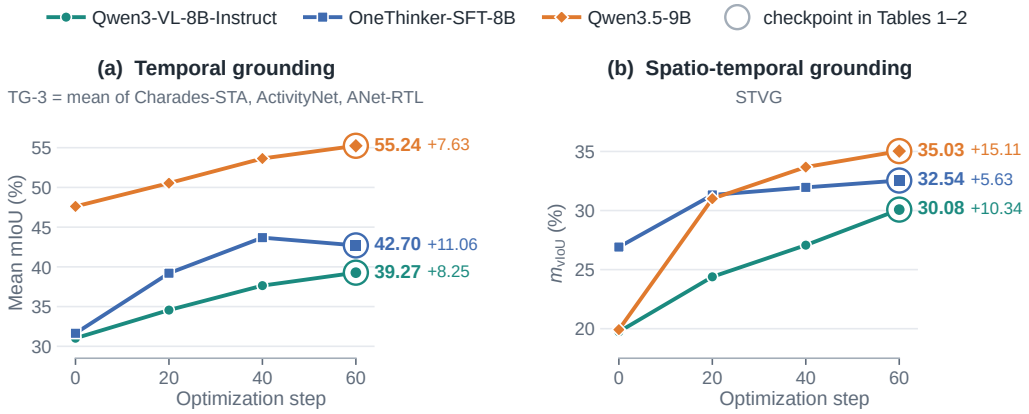


Figure 3: **Grounding dynamics during joint self-play.** TG-3 averages temporal mIoU across Charades-STA, ActivityNet, and ANet-RTL, while STVG Mean averages tIoU and sIoU. Grounding improves overall, with non-monotonic temporal performance for OneThinker-SFT.

with this combination of stable evidence and adaptive practice; the component analysis separates these mechanisms rather than attributing all gains to increased question difficulty.

Comparison with annotated training. The full OneThinker-8B provides context for the temporal results. From its SFT initialization, ANCHORPLAY narrows the gaps on ActivityNet and ANet-RTL using model-generated supervision. This suggests that task construction and geometric feedback can develop localization behavior without further human-provided targets. The published model has a different training history and is therefore a reference, while the paired initialization remains the direct baseline for each adaptation run.

Table 2: **STVG.** ANCHORPLAY improves tube localization across all three initializations.

Backbone	tIoU		sIoU	
	Base	ANCHORPLAY	Base	ANCHORPLAY
Qwen3-VL-8B	25.83	**37.68**	13.65	**22.47**
OneThinker-SFT-8B	31.22	**39.13**	22.60	**25.96**
Qwen3.5-9B	27.45	**45.35**	12.39	**24.71**

Progress is not uniformly monotonic. Figure 3 traces the joint-training process. Qwen3 improves through its stable checkpoints, but OneThinker-SFT shows a trade-off: mean temporal performance peaks at step 40, whereas STVG continues to improve at step 60. A frozen reference constrains the source of the targets; it does not guarantee monotonic improvement as the policy and generated query distribution change. We therefore retain the complete checkpoint record in Appendix Table 7, rather than combining the best score from each benchmark into one model.

Table 3: **Transfer to referring-expression grounding (mIoU, %).** Both Qwen initializations improve across all eight evaluated splits.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	testA	testB	val	testA	testB	val	test	val
Qwen3-VL-8B-Instruct								
Base Model	84.82	78.56	83.31	82.70	72.72	78.78	80.58	80.85
ANCHORPLAY	84.44	79.66	82.78	82.33	73.20	78.54	80.75	80.74
Qwen3.5-9B								
Base Model	78.68	71.37	76.62	77.85	67.33	73.86	76.17	76.69
ANCHORPLAY	85.21	82.58	84.42	82.29	76.32	80.00	82.20	83.14

Table 4: **Segmentation and video reasoning.** Qwen3 improves on the segmentation tasks, while video-QA transfer is mixed.

Image segmentation			*Video segmentation*			*Video QA*		
cIoU	**Base**	ANCHORPLAY	$\mathcal{J}\&\mathcal{F}$	**Base**	ANCHORPLAY	*Acc.*	**Base**	ANCHORPLAY
Seg-RefCOCO	73.23	74.90	MeViS	23.69	32.01	MMVU	68.48	68.64
Seg-RefCOCO+	66.40	67.11	ReasonVOS	19.31	27.36	Video-MMMU	64.22	63.78
Seg-RefCOCOg	68.78	70.30				VideoMathQA	25.24	25.48

4.3 GENERALIZATION BEYOND VIDEO GROUNDING

We emphasize Qwen3 for transfer analysis and also report other initializations where comparable task results are available. Every task uses the checkpoint selected in Sec. 4.1.

Transfer to image understanding. Figure 4(a) highlights five image-based multimodal understanding and reasoning benchmarks: MMMU (Yue et al., 2024), MMBench (Liu et al., 2024), AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), and MMT-Bench (Ying et al., 2024). Qwen3 improves on each displayed task, with gains of 4.67 and 1.63 points on MMMU and MMT-Bench, respectively. Video-grounding adaptation thus transfers to some image-understanding capabilities without image training examples. A plausible connection is the shared requirement to associate language with relevant visual evidence.

Transfer to localization and segmentation. On Qwen3, every evaluated RefCOCO, RefCOCO+, and RefCOCOg split (Yu et al., 2016; Mao et al., 2016) improves, as do all three image-segmentation tasks and both video-segmentation benchmarks, MeViS (Ding et al., 2023) and ReasonVOS (Bai et al., 2024). These tasks share a need to identify the region or sequence of regions specified by a query, making them closely related to the evidence selection practiced during grounding. Table 3 focuses on the two Qwen initializations, both of which improve across the eight referring splits. These results use box mIoU, without image-grounding examples in the adaptation corpus. Qwen3.5's gains also reflect output-schema compliance, as discussed below. Table 4 reports Qwen3's segmentation and video-QA results. Transfer remains initialization-dependent: the full OneThinker-SFT results in Appendix Table 11 show lower referring-grounding scores after adaptation.

Video reasoning and tracking reveal limits. Video-QA changes are mixed across MMVU (Zhao et al., 2025b), Video-MMMU (Hu et al., 2025), and VideoMathQA (Rasheed et al., 2025); the appendix additionally reports Video-MME (Fu et al., 2025) and LongVideoReason (Chen et al., 2025) where available. On GOT-10k (Huang et al., 2021), Qwen3's higher aggregate AO coincides with static predictions on 99.4% of clips, and every reported adapted model remains below the first-frame-copy control. These results do not establish improved motion tracking. Grounding a queried event or region does not necessarily teach continuous object motion estimation, and the tracking decomposition in Appendix F.1 makes this distinction explicit.

4.4 COMPONENT ANALYSIS AND DIAGNOSTICS

Separating output validity from localization. The grounding scorer requires structured coordinate outputs, so its end-to-end score reflects both format compliance and geometry. This matters especially for Qwen3.5: its initial schema compliance is only 5.6% on Charades-STA. The large temporal gains therefore cannot be attributed wholly to better event localization. Qwen3 has higher initial compliance but is not entirely free of this effect. Appendix Table 8 records available schema-

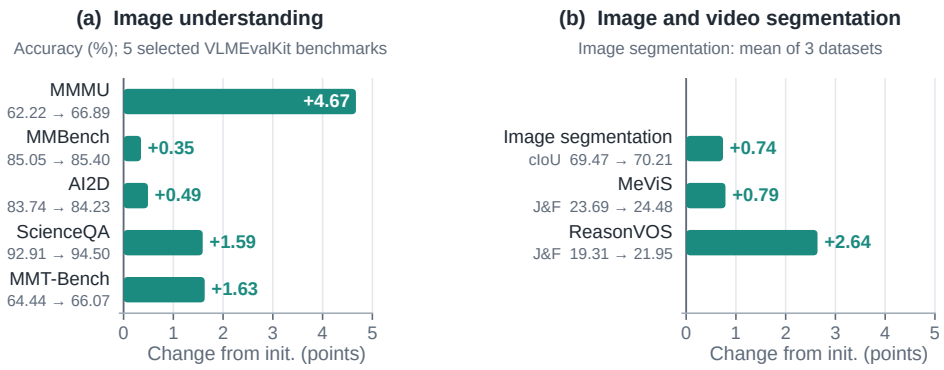


Figure 4: **Grounding-only adaptation transfers selectively.** Qwen3 improves on the five selected image-understanding benchmarks and the displayed localization and segmentation evaluations.

validity rates, providing context for interpreting the grounding scores. The main comparisons report improvements in end-to-end grounding performance.

Evidence anchoring. Table 5 separates the source, acceptance, and weighting of training targets. All variants use Qwen3, the same 40-step budget, and the same nominal rollout limits. TG averages the three temporal benchmarks; STVG uses m_{vIoU}. *Current-policy targets* replaces the frozen anchor proposer with the evolving policy. This tests whether a fixed evidence source matters when both training roles change. *No consistency screen* retains multi-view scoring but removes reliability-based rejection. *Uniform target weights* preserves screening and sets accepted target weights to one in both losses. Acceptance does not make all remaining targets equally reliable: the last two variants distinguish removing inconsistent supervision from controlling the influence of residual uncertainty.

Table 5: **Component contributions.** The ablations isolate evidence anchoring and joint learning.

Variant	TG	STVG
ANCHORPLAY	39.27	30.08
Evidence anchors		
Current-policy targets	24.91	7.88
No consistency screen	31.01	6.30
Uniform target weights	34.53	16.00
Joint self-play		
No Designer RL	34.56	16.42
No difficulty reward	35.47	14.27
No grounding preservation	37.55	24.42

Adaptive queries and grounding retention. *No Designer RL* removes the direct query-generation objective; Localizer updates still change the shared parameters. *No difficulty reward* retains validity feedback but removes the pass-rate-based reward. A supported query may still be too easy to provide useful practice or too difficult to yield successful rollouts, so this comparison tests learner-aware task selection beyond validity. *No grounding preservation* removes coordinate-level retention while retaining response-level KL. The former reinforces accepted geometry, whereas the latter constrains policy drift. Their separation tests whether geometric supervision helps maintain grounding behavior as the query distribution changes.

5 CONCLUSION

We presented ANCHORPLAY, an end-to-end self-play framework for improving video grounding without using the original training annotations. A shared policy jointly learns to design queries and localize their referents, supported by reliable evidence anchors from a frozen reference. Multi-view screening and reliability-weighted feedback connect self-generated tasks to visual evidence, while difficulty-aware query generation and grounding preservation support learning as the task distribution evolves. Across three model initializations, ANCHORPLAY improves both temporal and spatio-temporal grounding, with additional transfer to selected image-understanding, localization, and segmentation tasks. These results suggest that effective self-play for grounding benefits from both adaptive task generation and an evidence source that remains stable as the learner changes.

AI USE STATEMENT

Generative AI tools (Claude and Codex) were used as discussion and editing aids to brainstorm alternative explanations for observed experimental results, provide feedback on experimental controls and presentation, and improve the paper grammar. They were not used to implement the method, execute experiments, or generate the reported results.

ETHICS STATEMENT

This work uses existing video datasets and publicly released model checkpoints, without recruiting new participants or collecting new human-subject annotations. The adaptation stage uses videos without their original grounding annotations and updates a shared policy through joint self-play. Annotation-free adaptation does not imply that the initial models were pretrained without annotated data. Generated queries and localization targets may inherit errors and biases from the underlying models, and cross-view consistency does not eliminate systematic localization errors. Our benchmark evaluation does not establish the reliability of these predictions for consequential decisions involving people.

REPRODUCIBILITY STATEMENT

Section 3 describes evidence-anchor construction, query generation, and joint policy optimization. Appendices B and C provide the geometric scoring rules and additional training-objective details. Section 4.1 and Appendix E describe the adaptation data, model initializations, training configuration, and evaluation procedures. Appendix D presents the assumptions and derivation of the geometric reward-distortion bounds. Appendix G reports checkpoint-level results for each initialization, while Appendix F examines output validity and transfer behavior.

REFERENCES

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.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6833–6859, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling RL to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2694–2703, 2023.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Videossr: Video self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26530–26540, 2026.

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

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In *International Conference on Learning Representations*, volume 2026, pp. 130770–130790, 2026a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pp. 202–218, 2024.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. doi: 10.1109/TPAMI.2019.2957464.

Shiqi Huang, Ziyue Wang, Zhongrong Zuo, Han Qiu, Qi She, and Bihan Wen. Evovid: Temporal-centric self-evolution for video large language models. *arXiv preprint arXiv:2605.21931*, 2026b.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, 2016.

Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Language-free training for zero-shot video grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2539–2548, 2023.

Sven Kosub. A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters*, 120:36–38, 2019. doi: 10.1016/j.patrec.2018.12.007.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. LLaVA-ST: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8592–8603, 2025a.

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*, 2025b.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Khan. VideoMathQA: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.

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

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Han Wang, Yi Yang, Jingyuan Hu, Minfeng Zhu, and Wei Chen. V-zero: Self-improving multimodal reasoning with zero annotation. *arXiv preprint arXiv:2601.10094*, 2026.

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

Chaoran Xu, Yingmao Miao, Pengfei Zhang, Hao Dou, Lei Sun, and Xiangxiang Chu. Rise: Reliable improvement in self-evolving vision-language models. *arXiv preprint arXiv:2605.20914*, 2026.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. *arXiv preprint arXiv:2404.16006*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, 2016.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ruixu Zhang, Deyi Ji, Lanyun Zhu, Xuanyi Liu, Yuxin Meng, Ruihang Chu, and Yujiu Yang. Video-zero: Self-evolution video understanding. *arXiv preprint arXiv:2605.14733*, 2026.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10665–10674. IEEE, 2020.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, et al. Absolute zero: Reinforced self-play reasoning with zero data, 2025. *URL https://arxiv. org/abs/2505.03335*, 27, 2025a.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8489, 2025b.

A Appendix

B Anchor Construction and Geometric Scoring

Anchor representation. The frozen reference proposes a short description c identifying an event or subject, together with its geometry g and task type τ. The description fixes the referent for subsequent view-based checks. Temporal anchors have the form $g = [s, e]$, with $0 \le s < e \le T_x$, where T_x is the video duration. Spatio-temporal anchors additionally contain boxes B_t at the evaluation timestamps within that interval. Coordinates are mapped to the original video's time axis and frame size before agreement is measured.

View construction. Views preserve the full video duration and event order while varying frame-sampling offsets, spatial resizing, or horizontal reflection. Every sampled frame retains its original timestamp. For spatio-temporal anchors, all views retain a common set of output frames at fixed timestamps; sampling offsets vary only the additional context frames. This makes box agreement comparable without temporal interpolation. For temporal outputs, C_m converts any view-relative time or frame index to the original timestamp. For spatial outputs, it additionally undoes resizing and reflection. Crops that truncate the candidate event are excluded. The reference receives the same referent description c on every view and predicts geometry without being shown g. Missing or malformed predictions contribute zero agreement in Eq. (2); they are not silently omitted from the denominator.

Geometric scores. For two temporal intervals I and $\widehat{I}$,

$$\mathrm{tIoU}(\widehat{I}, I) = \frac{|\widehat{I} \cap I|}{|\widehat{I} \cup I|}. \tag{9}$$

For spatio-temporal outputs, boxes are compared at common timestamps after coordinate alignment. Let $\mathcal{F}$ be the evaluation timestamps in the union of the predicted and target intervals, and set a box to the empty set outside its interval. We define

$$\begin{aligned} \overline{\mathrm{sIoU}}(\widehat{g}, g) &= \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \mathrm{IoU}(\widehat{B}_t, B_t), \\ s_{\mathrm{TG}}(\widehat{g}, g) &= \mathrm{tIoU}(\widehat{I}, I), \\ s_{\mathrm{STG}}(\widehat{g}, g) &= \tfrac{1}{2}\left[\mathrm{tIoU}(\widehat{I}, I) + \overline{\mathrm{sIoU}}(\widehat{g}, g)\right]. \end{aligned} \tag{10}$$

Missing required boxes or invalid coordinates receive zero spatial overlap; an empty timestamp set receives zero score. Both task scores lie in $[0, 1]$. Temporal success requires tIoU at least h_{T}. Spatio-temporal success requires both tIoU at least h_{T} and mean spatial IoU at least h_{S}, preventing a good score in one dimension from concealing failure in the other. The same geometric scoring convention is used for anchor agreement and Localizer feedback.

Query validity and selection. The Designer receives (x, c, g, τ) and generates queries that retain the referent and event extent specified by the anchor. The frozen validity judge rejects queries with unsupported premises, ambiguous subjects, changed temporal scope, or coordinates that reveal the answer directly in the question. Among a video's valid candidates with $p \in (0.3, 0.8)$, one is sampled uniformly for the Localizer's policy loss. If none qualifies, Designer feedback is retained and the Localizer policy loss is skipped for that video. Valid query–anchor pairs can still contribute to grounding preservation, independently of the pass-rate gate.

C Additional Self-Play Details

Designer reward. We use the variance-based difficulty reward of SPICE (Liu et al., 2025) with a validity bonus. The complete Designer reward is

$$r^{\mathrm{D}} = \begin{cases} -0.1, & \text{unparseable output,} \\ 0, & \text{parsed output with } v = 0, \\ r_{\mathrm{diff}}(p) + 0.2, & \text{parsed output with } v = 1, \end{cases} \qquad r_{\mathrm{diff}}(p) = \exp\left(-\frac{[p(1-p) - 0.25]^2}{2(0.01)}\right). \tag{11}$$

Invalid candidates do not require Localizer rollouts because their Designer rewards can be assigned before estimating p. Coverage of the two task types is controlled by task-balanced anchor sampling.

Within a Designer prompt group, the task is fixed, so a task-dependent additive reward would vanish under group centering.

Role-specific advantage normalization. For a rollout group with role $r \in \{\mathrm{D}, \mathrm{L}\}$, let R_i^r denote the unweighted reward for that role and let $\overline{R}^r$ be its group mean. The EMA-normalized advantage has the form

$$A_i^r = \frac{R_i^r - \overline{R}^r}{\sigma^{\mathrm{EMA}}_{\kappa(r,\tau)} + \epsilon}, \tag{12}$$

where $\kappa(\mathrm{D}, \tau) = \mathrm{D}$ and $\kappa(\mathrm{L}, \tau) = \tau$. Thus, the Designer and the two Localizer task types use separate scale statistics. These advantages enter the clipped GRPO surrogates in Eq. (7). The joint objective denotes role-assigned policy surrogates evaluated on sampled trajectories; it does not require differentiation through anchor generation, query screening, or the other role's sampled output.

Clipped policy surrogates. For a sampled response z_i and its role-specific input u_i^r, let

$$\begin{aligned} \rho_{it}^r(\theta) &= \frac{\pi_\theta(z_{it} \mid u_i^r, z_{i,<t})}{\pi_{\mathrm{old}}(z_{it} \mid u_i^r, z_{i,<t})}, \\ \ell_{it}^r(\theta) &= \min\left\{\rho_{it}^r(\theta) A_i^r,\ \mathrm{clip}(\rho_{it}^r(\theta), 1-\varepsilon_c, 1+\varepsilon_c) A_i^r\right\}. \end{aligned} \tag{13}$$

The Designer surrogate is $\mathcal{J}_{\mathrm{D}} = N_{\mathrm{D}}^{-1} \sum_{i \in \mathcal{B}_{\mathrm{D}}} \sum_t \ell_{it}^{\mathrm{D}}$, where N_{D} counts its non-padding response tokens. The Localizer surrogate is given in Eq. (5), with w applied outside reward normalization. Its denominator is the unweighted token count. Consequently, reliability attenuates the total policy contribution of a low-confidence batch instead of rescaling that batch back to unit average weight. If a role has no eligible trajectories, its surrogate is set to zero.

Grounding-preservation targets. For each pair in $\mathcal{B}_A$, $a(g)$ is a canonical serialization of the frozen-reference anchor. The direct localization prompt requests only this structured answer. The set $\mathcal{C}_i$ includes tokens encoding timestamps and box coordinates; schema punctuation and fixed field names are conditioning context, rather than supervised geometric tokens. The coordinate loss is

$$\ell_{\mathrm{geo}}(\theta; i) = -\frac{1}{|\mathcal{C}_i|} \sum_{t \in \mathcal{C}_i} \log \pi_\theta\big(a_{it} \mid x_i, q_i, \tau_i, a_{i,<t}, \mathrm{L}\big). \tag{14}$$

The same Localizer parameters are used for this auxiliary forward pass. Weights are detached and normalized across valid examples, as in Eq. (6). If $\mathcal{B}_A$ is empty, the preservation loss is zero. This auxiliary loss and both policy losses are accumulated before a single shared-parameter optimizer update.

Response-level regularization. We define $\mathcal{D}_{\mathrm{KL}}^{\mathrm{L}}$ as the token-averaged conditional divergence $D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h) \| \pi_{\mathrm{ref}}(\cdot \mid h))$ over Localizer rollout prefixes h. It covers the full response, including reasoning and answer tokens. The preservation loss instead concentrates on reference coordinate targets for accepted anchors and can use queries outside the Localizer policy's pass-rate band.

D Anchor Error and Geometric Reward Distortion

Setting. Fix a video, a valid query, an anchor interval g, and the true event interval $g^\star$. The anchor and query must refer to the same event with the same intended temporal extent. All predicted and target intervals in the following argument are measurable with positive, finite length. For intervals a, b, write $d_J(a, b) = 1 - \mathrm{tIoU}(a, b)$. Jaccard distance obeys the triangle inequality (Kosub, 2019).

Pointwise reward distortion. Applying the triangle inequality in both directions gives

$$d_J(\widehat{g}, g) \le d_J(\widehat{g}, g^\star) + d_J(g^\star, g), \tag{15}$$

$$d_J(\widehat{g}, g^\star) \le d_J(\widehat{g}, g) + d_J(g, g^\star). \tag{16}$$

Hence $|d_J(\widehat{g}, g) - d_J(\widehat{g}, g^\star)| \le d_J(g, g^\star)$, which proves Eq. (8). Because the format and structure terms of Eq. (4) do not depend on the target, the corresponding full-reward difference is bounded by $(1 - \lambda_f) d_J(g, g^\star)$.

Conditional benefit of reliability weighting. Consider a fixed policy and a distribution over accepted, selected query–anchor pairs. Define the unknown anchor error $e = d_J(g, g^\star)$, and let $w \in [0, 1]$ with $\mathbb{E}[w] > 0$. For the normalized weighted expectation of geometric scores, the pointwise inequality

Table 6: **Joint self-play configuration.** Both roles share a policy and are optimized in the same training step.

Setting	Value
Hardware	$8\times$ NVIDIA H100
Videos per optimization step	128
Learning rate	2×10^{-6}
Designer candidates per video (G)	8
Candidate attempts per retry batch	16
Maximum retry batches	4
Localizer rollouts per valid query (K)	8
Pass-rate band	$(0.3, 0.8)$
Localizer / Designer KL coefficient	0.01 / 0
Grounding-preservation coefficient	0.25
Reference views / acceptance threshold	9 / 5
Temporal / spatial success thresholds	0.5 / 0.5
Sampling temperature / maximum frames	1.0 / 128
Maximum response tokens (Designer / Localizer)	4096 / 4096
Main checkpoint (Qwen3 / OneThinker-SFT / Qwen3.5)	60 / 60 / 60
Independent runs per backbone	1
VLMEvalKit version / extraction judge	v0.2rc1 / deepseek-v4-flash

yields

$$\left|\frac{\mathbb{E}[w\ \mathrm{tIoU}(\widehat{g}, g)]}{\mathbb{E}[w]} - \frac{\mathbb{E}[w\ \mathrm{tIoU}(\widehat{g}, g^\star)]}{\mathbb{E}[w]}\right| \leq \frac{\mathbb{E}[we]}{\mathbb{E}[w]} = \mathbb{E}[e] + \frac{\mathrm{Cov}(w, e)}{\mathbb{E}[w]}. \tag{17}$$

When reliability is negatively correlated with anchor error, the right-hand side is smaller than the unweighted bound $\mathbb{E}[e]$ on the same accepted-example distribution. This states a condition under which reliability weighting improves the reward-distortion bound; it does not establish that multi-view scores satisfy that condition automatically.

Scope. This analysis concerns the geometric score before group centering, EMA normalization, and policy clipping. It is not an identity for the implemented GRPO gradient or a guarantee of final performance. It also does not establish improvement on examples rejected by the screen, nor does it cover every possible spatio-temporal scorer. Systematically wrong but consistent anchors remain possible. Finally, freezing the reference prevents the target generator from co-adapting with current policy rollouts; it does not require the distribution of selected videos, anchors, or queries to stay fixed.

E Additional Experimental Protocol

Grounding-only adaptation data. We filter `onethinker_rl_train.json` to temporal and spatio-temporal grounding and retain the visual inputs without their supplied query, answer, or reasoning fields. The pool contains x unique videos and x/x temporal/spatio-temporal records. A video can support both task types. OneThinker's source collection includes ActivityNet-train, Vtg_it, DiDeMo, Queryd, Charades-train, and HiREST for temporal grounding, and LLaVA-ST-STVG for spatio-temporal grounding. These source names describe the released collection; the filtered RL manifest determines the actual training records. STVG is the evaluation benchmark used in our experiments and is not interchangeable with the published OneThinker ST-Align evaluation.

Metric extraction and aggregation. For the general OneThinker-suite tasks, we use the benchmark-level `overall/acc` result, rather than averaging a per-sample `accuracy` field across mixed answer formats. Image segmentation uses `image/cIoU` and video segmentation uses `video/J&F` from the segmentation evaluation artifacts; a generic accuracy field is not used for either. Referring-expression results are box mIoU, not R@0.5. For image understanding, we use the VLMEvalKit aggregate accuracy for each benchmark; MathVista uses its score-file accuracy. Answer extraction and equivalence checking follow the benchmark-specific protocol. In particular, MathVista's absolute score can depend on the answer-equivalence judge. No claim of judge-invariant ordering is made.

Table 7: **Grounding throughout joint training.** Continued optimization does not improve every benchmark monotonically.

Checkpoint	Charades	ActivityNet	ANet-RTL	STVG
Qwen3-VL-8B-Instruct				
Base Model	36.75	29.13‡	27.17	19.40
Step 20	44.31	31.03‡	28.34	24.39
Step 40	45.26	34.32‡	33.34	27.07
Step 60§	49.43	34.95	33.43	30.08
OneThinker-SFT-8B				
Base Model	55.15	21.23	18.55	26.91
Step 20	55.49	34.96	27.16	31.32
Step 40	55.79	40.04	35.21	31.96
Step 60	56.65	38.34	33.12	32.54
Qwen3.5-9B				
Base Model	59.26	45.16	38.42	19.30
Step 20	60.28	47.83	43.52	31.00
Step 40	61.34	51.41	48.16	33.68
Step 60	61.82	53.17	50.74	35.03

Shading identifies the checkpoint used in the main comparisons. §Qwen3 step 60 is provisional: resumed training used a different machine and the reported reward curve is discontinuous at resumption. ‡Earlier report.

Reported task means retain the evaluation report's precision, and rounding can produce a 0.01 difference from means recomputed using the displayed individual scores.

STVG metric. Let T_p and T_g denote predicted and reference temporal supports, with predicted and reference boxes b_t^p and b_t^g. Tube overlap combines temporal extent and spatial alignment:

$$\mathrm{vIoU} = \frac{1}{|T_p \cup T_g|} \sum_{t \in T_p \cap T_g} \mathrm{IoU}(b_t^p, b_t^g). \tag{18}$$

We report its mean across evaluated instances as m_{vIoU}. It is not the arithmetic mean of temporal IoU and spatial IoU.

Checkpoint and source accounting. The main results use one checkpoint per run, with no per-benchmark selection. Table 7 reports the full progression. The Qwen3 step-60 checkpoint was resumed on a different machine and its reported reward trajectory is discontinuous at the resume point; its scores are retained as provisional results, separate from the primary comparisons. OneThinker-SFT step 60 is reported even where step 40 scores higher. Qwen3 ActivityNet, Video-MME, and LongVideoReason values at steps 0/20/40 are retained from the earlier report and explicitly marked. They are absent from the newer complete-column table because not all checkpoint evaluations are included there. We do not combine task means defined over different benchmark sets. TG-2 in the training figure and checkpoint table uses Charades-STA and ANet-RTL for every backbone; other three-task temporal averages are identified explicitly.

Ablation controls. The component ablations start from Qwen3-VL-8B-Instruct and use the same 40-step budget, video pool, and nominal rollout limits. Replacing the anchor source retains the frozen query-validity judge. Removing consistency screening preserves multi-view scoring and continuous weighting. Uniform weighting retains the screening rule, while setting weights to one in both the Localizer policy loss and the grounding-preservation loss. Subsequent generated query distributions may differ as the policies learn. Removing Designer RL does not freeze the shared model's query-generation behavior. Removing difficulty reward retains validity feedback and the Localizer's selection band. Removing grounding preservation leaves the response-level KL active. Effective sample counts and rollout-token costs are recorded alongside performance to distinguish data-yield effects from component effects.

Reliability and curriculum diagnostics. A separate labeled audit set of x videos is excluded from the adaptation pool. Its annotations only measure anchor quality. Reliability quintiles are formed before the gate; coverage is the accepted fraction within a quintile. Target tIoU is measured against the audit annotation, not against the reference model's own output. Valid-query rate is the accepted

Table 8: **Output validity affects grounding scores.** Low schema compliance confounds the interpretation of localization gains.

Benchmark	Base validity (%)	ANCHORPLAY validity (%)
Qwen3-VL-8B-Instruct		
Charades-STA	93.2	100.0
ANet-RTL	91.7	99.1
Qwen3.5-9B		
RefCOCO val	90.3	99.9
RefCOCO testA	90.9	99.9
RefCOCO testB	87.1	99.9
RefCOCO+ val	90.7	99.9
RefCOCO+ testA	92.2	99.9
RefCOCO+ testB	88.4	99.7
RefCOCOg val	92.4	99.8
RefCOCOg test	92.1	99.9
Charades-STA	5.6	100.0
ANet-RTL	51.5	99.6

fraction among parsed queries; in-band rate is the fraction of valid queries with pass rate in $(0.3, 0.8)$; flat-group rate is the fraction of eligible groups whose unweighted Localizer rewards have zero within-group variance.

F Output Validity and Transfer Diagnostics

Schema sensitivity. Table 8 separates output validity from benchmark score. When invalid outputs receive zero overlap, the reported mean can be written as

$$\mathbb{E}[\text{score}] = \Pr(\text{valid})\,\mathbb{E}[\text{IoU} \mid \text{valid}]. \tag{19}$$

Thus, improved output compliance can raise end-to-end scores without an equivalent increase in conditional localization quality. A geometry-only comparison should use the same examples on which both models produce valid outputs, supplemented by an identical coordinate-recovery parser applied to both models. Conditional scores on different valid-output subsets would introduce a further selection effect. The validity audit is particularly relevant to Qwen3.5's temporal and image-grounding results; it does not turn a rescaled overall score into a measured valid-only result.

F.1 Tracking Decomposition

Table 9 and Figure 5 compare AO with the prevalence of static predictions. The first-frame-copy control obtains 46.26 AO. All three adapted models fall below it, and their static-output rates increase relative to initialization. For Qwen3, only one evaluated clip has non-static predictions at the primary checkpoint. This makes a higher aggregate AO insufficient evidence for improved tracking. The non-static subset is defined by the model's own predictions and changes across checkpoints; its conditional AO is descriptive rather than a comparison on a fixed subset. Tracking remains an evaluation-only transfer task.

G Complete Per-Run Results

Tables 10–12 retain all supplied checkpoint scores, including the provisional Qwen3 continuation. The full image-understanding evaluation includes MMStar (Chen et al., 2024) and MathVista (Lu et al., 2024), in addition to the five tasks displayed in the main-text figure. Benchmarks are grouped by task and metric. Their differing coverage across models precludes comparing an overall average computed over all available rows.

Table 9: **Tracking exposes static-prediction collapse.** Aggregate AO can increase even when predictions almost always remain fixed.

Checkpoint	AO	Static (%)	Non-static n	Non-static AO
Qwen3-VL-8B-Instruct				
Base Model	33.01	51.7	73	29.95
Step 20	45.79	98.3	3	13.20
Step 40	45.99	99.4	1	14.84
Step 60§	46.17	99.4	1	41.34
OneThinker-SFT-8B				
Base Model	47.74	55.6	80	45.04
Step 20	46.22	80.6	35	32.12
Step 40	46.13	92.2	14	23.90
Step 60	46.15	88.3	21	29.52
Qwen3.5-9B				
Base Model	42.57	77.6	38	47.88
Step 20	46.17	93.3	12	34.69
Step 40	46.39	94.4	10	26.74
Step 60	46.48	93.9	11	31.23
First-frame-copy control	46.26	100.0	0	—

Non-static subsets depend on model predictions and differ across checkpoints; their AO values do not compare a fixed subset. §Provisional Qwen3 checkpoint.

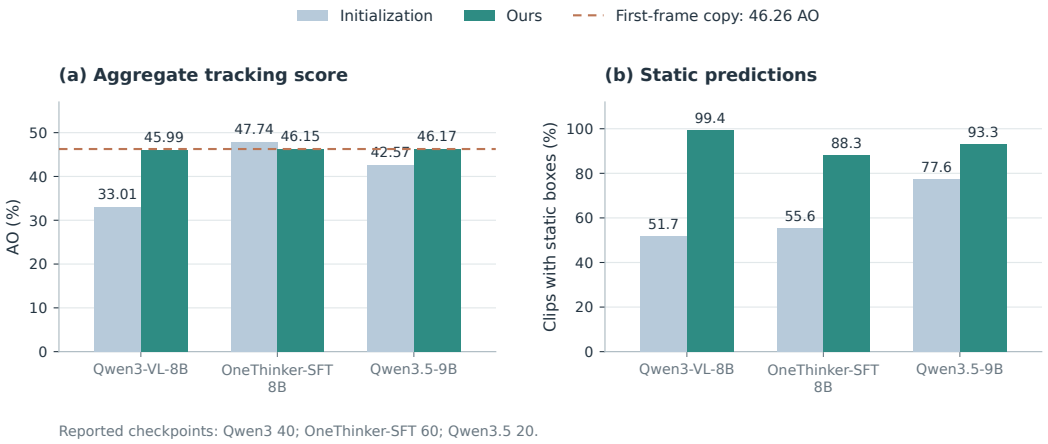


Figure 5: **Higher aggregate overlap can accompany static outputs.** The adapted models have higher static-prediction rates and remain below the first-frame-copy AO control.

Table 10: **Full evaluation record: Qwen3-VL-8B-Instruct.** All checkpoints use the same within-run initialization.

Benchmark	**Base**	**20**	**40**	**60**[§]
Referring grounding (mIoU)				
RefCOCO val	83.31	83.40	83.59	82.78
RefCOCO testA	84.82	85.08	85.18	84.44
RefCOCO testB	78.56	79.37	80.02	79.66
RefCOCO+ val	78.78	78.96	79.35	78.54
RefCOCO+ testA	82.70	82.87	83.17	82.33
RefCOCO+ testB	72.72	73.73	74.41	73.20
RefCOCOg val	80.85	80.74	81.32	80.74
RefCOCOg test	80.58	81.35	81.54	80.75
Image segmentation (cIoU)				
Seg-RefCOCO	73.23	75.06	74.90	74.93
Seg-RefCOCO+	66.40	67.30	67.11	66.24
Seg-RefCOCOg	68.78	70.20	70.30	69.46
Temporal grounding (mIoU)				
Charades-STA	36.75	44.31	45.26	49.43
ActivityNet	29.13[‡]	31.03[‡]	34.32[‡]	34.95
ANet-RTL	27.17	28.34	33.34	33.43
Spatio-temporal grounding				
STVG	19.74	24.39	27.07	30.08
Tracking (AO)				
GOT-10k	33.01	45.79	45.99	46.17
Video QA (accuracy)				
MMVU	68.48	69.60	67.20	68.64
Video-MMMU	64.22	64.33	63.78	64.44
VideoMathQA	25.24	25.48	25.48	21.67
Video-MME	65.22[‡]	66.00[‡]	66.22[‡]	65.89
LongVideoReason	71.60[‡]	72.30[‡]	73.40[‡]	74.20
Video segmentation ($\mathcal{J}\&\mathcal{F}$)				
MeViS	23.69	25.70	29.58	32.01
ReasonVOS	19.31	21.21	20.47	27.36
Image understanding (accuracy)				
MMMU (val)	62.22	65.56	65.67	66.89
MMBench (dev-en)	85.05	85.48	85.40	85.40
MMStar	70.07	70.20	68.80	69.93
AI2D	83.74	84.29	84.23	84.23
ScienceQA	92.91	93.21	94.15	94.50
MMT-Bench (val)	64.44	66.87	67.48	66.07
MathVista (mini)	73.80	76.60	73.80	74.40

§Provisional resumed checkpoint. ‡Earlier report, retained separately from the latest complete-column evaluation.

Table 11: **Full evaluation record: OneThinker-SFT-8B.** All checkpoints use the same within-run initialization.

Benchmark	**Base**	**20**	**40**	**60**
Referring grounding (mIoU)				
RefCOCO val	83.98	82.66	82.87	82.15
RefCOCO testA	85.31	83.18	83.66	83.12
RefCOCO testB	80.67	79.66	79.50	78.91
RefCOCO+ val	79.13	77.43	77.69	76.77
RefCOCO+ testA	82.34	79.63	80.19	79.40
RefCOCO+ testB	74.53	73.85	73.58	73.21
RefCOCOg val	81.88	80.67	80.78	80.03
RefCOCOg test	82.08	80.58	80.62	80.18
Image segmentation (cIoU)				
Seg-RefCOCOg	71.77	71.02	70.03	71.00
Temporal grounding (mIoU)				
Charades-STA	55.15	55.49	55.79	56.65
ActivityNet	21.23	34.96	40.04	38.34
ANet-RTL	18.55	27.16	35.21	33.12
Spatio-temporal grounding				
STVG	26.91	31.32	31.96	32.54
Tracking (AO)				
GOT-10k	47.74	46.22	46.13	46.15
Video QA (accuracy)				
MMVU	69.60	69.44	68.48	68.80
Video-MMMU	64.56	63.00	63.56	64.33
VideoMathQA	29.52	31.19	35.24	31.19
Video-MME	64.59	64.04	65.96	65.19
LongVideoReason	75.40	78.40	77.90	78.70
Image understanding (accuracy)				
MMMU (val)	64.89	66.11	64.00	64.89
MMBench (dev-en)	85.14	84.97	85.74	85.74
MMStar	68.13	69.73	69.40	68.67
AI2D	83.42	83.78	84.10	83.78
ScienceQA	94.89	95.04	95.29	95.39
MMT-Bench (val)	67.54	68.66	68.05	67.99
MathVista (mini)	76.90	76.20	75.80	76.70

Table 12: **Full evaluation record: Qwen3.5-9B.** All checkpoints use the same within-run initialization.

Benchmark	Base	20	40	60
Referring expression comprehension (mIoU)				
RefCOCO val	76.62	84.10	84.31	84.42
RefCOCO testA	78.68	85.08	85.17	85.21
RefCOCO testB	71.37	81.49	82.20	82.58
RefCOCO+ val	73.86	79.47	79.82	80.00
RefCOCO+ testA	77.85	81.79	82.12	82.29
RefCOCO+ testB	67.33	75.14	75.90	76.32
RefCOCOg val	76.69	82.31	82.83	83.14
RefCOCOg test	76.17	81.90	82.12	82.20
Referring image segmentation (cIoU)				
RefCOCO	75.06	68.31	68.18	68.12
RefCOCO+	66.58	59.45	59.28	59.19
RefCOCOg	71.92	62.71	62.80	62.84
Temporal grounding (mIoU)				
Charades-STA	59.26	60.28	61.34	61.82
ActivityNet	45.16	47.83	51.41	53.17
ANet-RTL	38.42	43.52	48.16	50.74
Spatio-temporal grounding				
STVG	19.92	31.00	33.68	35.03
Object tracking (AO)				
GOT-10k	42.57	46.17	46.39	46.48
Video QA / reasoning (accuracy)				
MMVU	75.36	75.84	73.61	72.55
Video-MMMU	72.89	73.11	72.60	72.34
VideoMathQA	32.38	35.95	36.02	36.05
Video-MME	69.67	71.15	71.40	71.51
LongVideoReason	71.20	77.40	78.62	79.30
Referring video segmentation ($\mathcal{J}\&\mathcal{F}$)				
MeViS	13.84	25.16	29.32	31.35
ReasonVOS	12.90	20.36	19.67	19.33
Image understanding — VLMEvalKit (accuracy)				
MMMU (val)	75.33	71.78	71.90	71.96
MMBench (dev-en)	89.26	89.18	89.09	89.05
MMStar	76.80	76.13	74.61	73.86
AI2D	88.67	89.64	89.58	89.55
ScienceQA	97.52	97.32	98.19	98.56
MMT-Bench (val)	70.55	70.26	70.92	71.25
MathVista (mini)	82.20	80.10	77.54	76.16