MESD: MULTI-EVIDENCE SELF-DISTILLATION FOR VIDEOLLM

Anonymous authors
Paper under double-blind review

ABSTRACT

While reinforcement learning with verifiable rewards provides reliable outcome supervision for VideoLLMs, sequence-level rewards offer limited token-level guidance. On-policy self-distillation addresses this limitation by conditioning a self-teacher on privileged information to provide dense token-level supervision. However, aggregating heterogeneous evidence within a single teacher context obscures cross-evidence agreement and conflict. A further challenge lies in determining whether teacher guidance should refine reward-based updates or provide corrective supervision for failed trajectories. To address these issues, we propose MeSD, a multi-evidence self-distillation framework for VideoLLMs. MeSD constructs three evidence-conditioned teachers with shared parameters, using the ground-truth answer as a common semantic context while separately incorporating temporal and spatial evidence. Given the same student-generated prefixes, MeSD evaluates evidence-specific preferences relative to the Answer Teacher and fuses teacher-common preferences with gated teacher-specific residuals. Furthermore, MeSD introduces Verification-Guided Optimization to classify trajectories as Success, Failure, or Indeterminate. For Success and Indeterminate trajectories, MeSD refines token-level advantage magnitudes while preserving reward-derived signs. For verified failure trajectories that contain the required evidence, MeSD applies failure-conditioned distillation, using reverse-KL correction toward the fused distribution. Experiments on multiple video benchmarks demonstrate consistent gains over reinforcement learning and self-distillation baselines.

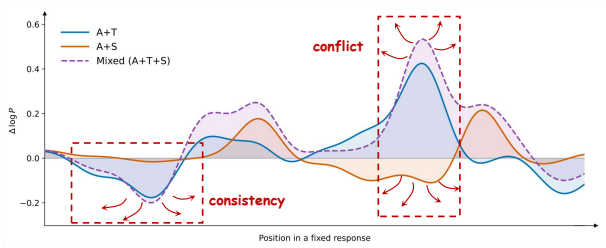
1 INTRODUCTION

Video Large Language Models (VideoLLMs) have made significant progress in video question answering and temporal understanding (Bai et al., 2025b; Zhang et al., 2025a; Ren et al., 2024; Wang et al., 2025b). However, robust reasoning over complex videos remains challenging (Meng et al., 2025; Lin et al., 2026). Such reasoning requires models to temporally identify key events, spatially localize relevant objects, and integrate cross-frame observations to derive the final answer (Meng et al., 2025). Final-answer rewards cannot reliably indicate whether intermediate steps correctly use temporal and spatial evidence, making token-level supervision critical (Lin et al., 2026).

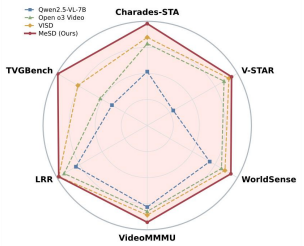
Reinforcement learning with verifiable rewards (RLVR) provides reliable outcome supervision (Guo et al., 2025a), but sequence-level rewards offer limited token-level guidance (Guo et al., 2025b). On-policy distillation (OPD) (Ma et al., 2026) provides token-level supervision on student-generated trajectories, but typically depends on a stronger teacher. On-policy self-distillation (OPSD) (Zhao et al., 2026) avoids this dependence by using the student model conditioned on privileged information as its teacher. However, matching the privileged teacher's distribution may lead to privileged information leakage and training instability. RLSD (Yang et al., 2026) therefore uses rewards to determine the signs of token-level advantages and teacher signals to refine their magnitudes.

However, applying self-distillation to video reasoning exposes two challenges. First, heterogeneous privileged evidence can induce conflicting token preferences along the same reasoning trajectory, as illustrated in Figure 1(a). Combining these sources within a single teacher context obscures their individual contributions, as well as any agreement or conflict between them. Second, existing methods like RLSD use teacher signals to adjust token-level advantage magnitudes while preserving reward-derived signs across both successful and failed trajectories. As shown in Figure 1(b), this

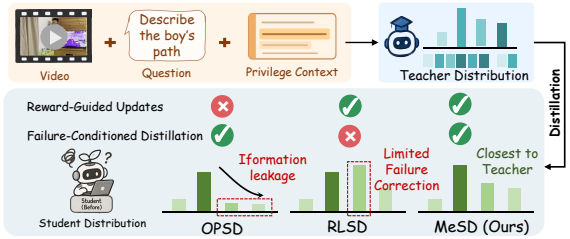
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



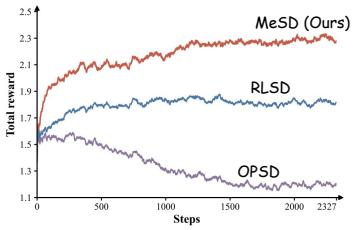
(a) Comparison of Evidence-Specific Preferences.



(c) Performance across Video Benchmarks.



(b) Comparison of Distillation Strategies.



(d) Training Reward Curves.

Figure 1: (a) Temporal and spatial evidence yield consistent or conflicting token preferences. (b) MeSD combines reward-anchored credit refinement with corrective distillation on verified failures. (c) MeSD leads on selected metrics across six benchmarks. (d) Comparison of training rewards.

merely reweights updates to sampled tokens without directly aligning the student distribution with the teacher distribution, thereby offering limited corrective guidance for failed trajectories. These challenges raise two questions: how to model heterogeneous privileged evidence and when teacher guidance should refine reward-based updates or correct failed trajectories.

To address these limitations, we propose **MeSD**, a multi-evidence self-distillation framework for VideoLLMs. MeSD integrates evidence-specific preferences and guides optimization based on trajectory verification outcomes and evidence availability. It comprises two components: Multi-Evidence Self-Teacher and Verification-Guided Optimization. The Multi-Evidence Self-Teacher employs three evidence-conditioned teachers with shared parameters: the Answer, Answer-Temporal, and Answer-Spatial Teachers. The ground-truth answer provides a common semantic context, while temporal and spatial evidence are introduced separately. Using the Answer Teacher distribution as a semantic anchor, MeSD captures the token preference shifts induced by additional temporal or spatial evidence. It then combines teacher-common preferences with gated teacher-specific residuals to yield a fused teacher distribution. The fusion selectively retains teacher-specific support while reducing residual contributions when evidence-specific preferences conflict.

Verification-Guided Optimization uses a task-specific verifier to classify trajectories as Success, Failure, or Indeterminate. For Success and Indeterminate trajectories, MeSD retains token-level advantage refinement, adjusting advantage magnitudes while preserving their reward-derived signs. MeSD applies failure-conditioned distillation only to verified failures with the required temporal or spatial evidence, using a reverse-KL objective targeting the fused teacher distribution. This objective uses teacher preferences over candidate tokens to provide corrective guidance beyond sampled-token reweighting. Overall, MeSD improves performance across video benchmarks and achieves sustained reward gains during training, as shown in Figure 1(c,d). Our main contributions are:

- We introduce MeSD, to our knowledge the first multi-teacher on-policy self-distillation framework for VideoLLMs. Its Multi-Evidence Self-Teacher uses the Answer Teacher as a reference to fuse teacher-common preferences with gated teacher-specific residuals.
- We propose Verification-Guided Optimization, which refines token-level advantage magnitudes while preserving reward-derived signs for Success and Indeterminate trajectories. It applies failure-conditioned distillation only to verified failures with the required evidence, using a reverse-KL objective targeting the fused teacher distribution.
- Extensive experiments across multiple video reasoning benchmarks demonstrate that MeSD consistently outperforms RLVR, OPSD, RLSD, and video post-training methods.

2 RELATED WORK

2.1 POST-TRAINING FOR VIDEOLLMS

RLVR improves video understanding and reasoning through verifiable outcome feedback (Guo et al., 2025a; Wu et al., 2026; Wang et al., 2025a). Video-R1 Feng et al. (2025) and Time-R1 Wang et al. (2025b) target temporal reasoning and grounding, respectively. GKD Agarwal et al. (2024) and Video-OPD Li et al. (2026) supervise student-generated trajectories using teacher distributions to mitigate training–inference mismatch. OPSD Zhao et al. (2026) and SDPO Hübotter et al. (2026) condition self-teachers on privileged solutions and environment feedback, respectively. Visual variants use regional crops (Vision-OPD Yuan et al. (2026)), annotated evidence frames (Video-OPSD Wang et al. (2026)), or denser temporal sampling (S^3T Venkatraman et al. (2026)). Open-o3 Video Meng et al. (2025) jointly supervises answers and spatiotemporal evidence through reasoning traces containing timestamps, objects, and bounding boxes. Its rewards jointly evaluate answer correctness and spatiotemporal grounding quality. Following this paradigm, MeSD provides token-level guidance for student-generated trajectories.

2.2 MULTI-TEACHER DISTILLATION AND OPTIMIZATION

MOPD Ma et al. (2026) combines independently trained specialists, while MT-SDPO He et al. (2026) conditions an EMA self-teacher on verified teacher feedback and AVSD Nguyen et al. (2026) aggregates privileged views. For optimization, VG-OPD Xu & Zhang (2026) uses verification to select eligible experts and localize token-level supervision. RLSD Yang et al. (2026) and VISD Lin et al. (2026) refine token-level credit while preserving reward-derived advantage signs, whereas H^2SD Cai et al. (2026) uses outcome-dependent asymmetric supervision. MeSD measures temporal and spatial preference shifts relative to an answer-conditioned anchor, then fuses teacher-common preferences with gated teacher-specific residuals to account for cross-evidence agreement and conflict. It uses verification outcomes and evidence availability to choose between reward-anchored refinement and reverse-KL correction, reserving the latter for eligible verified failures.

3 METHOD

3.1 OVERVIEW

We propose MeSD, a multi-evidence self-distillation framework designed for VideoLLMs. The core challenge is to integrate heterogeneous privileged evidence into token-level supervision and determine how this supervision should guide trajectories with different verification outcomes. As illustrated in Figure 2, MeSD constructs Evidence-Conditioned Teachers from answer, temporal, and spatial annotations, with shared EMA parameters. These teachers evaluate reasoning trajectories sampled by the student from video-question pairs. Teacher Distribution Fusion compares their predictions at the same student-generated prefixes and aggregates evidence-specific preferences into a fused teacher distribution. Verification-Guided Optimization uses a task-specific verifier to classify each trajectory as Success, Failure, or Indeterminate. For Success and Indeterminate trajectories, token-level advantage refinement adjusts advantage magnitudes while preserving their reward-derived signs. For verified failures with the required evidence, failure-conditioned distillation replaces the GRPO policy loss with a reverse-KL objective targeting the fused teacher distribution.

3.2 MULTI-EVIDENCE SELF-TEACHER

As illustrated in Figure 1, temporal and spatial evidence can drive consistent or conflicting token preferences along a single reasoning trajectory. Integrating these evidence sources within a single teacher context obscures their relationship and makes their individual contributions hard to distinguish. To address this, we construct distinct evidence-conditioned teachers and fuse their predictions through teacher-common preferences and gated teacher-specific residuals.

Evidence-Conditioned Teachers. Given a video-question pair x_i and a student-generated prefix $y_{i,<t}$, we construct three evidence-conditioned teachers with shared parameters $\bar{\theta}$. The ground-truth

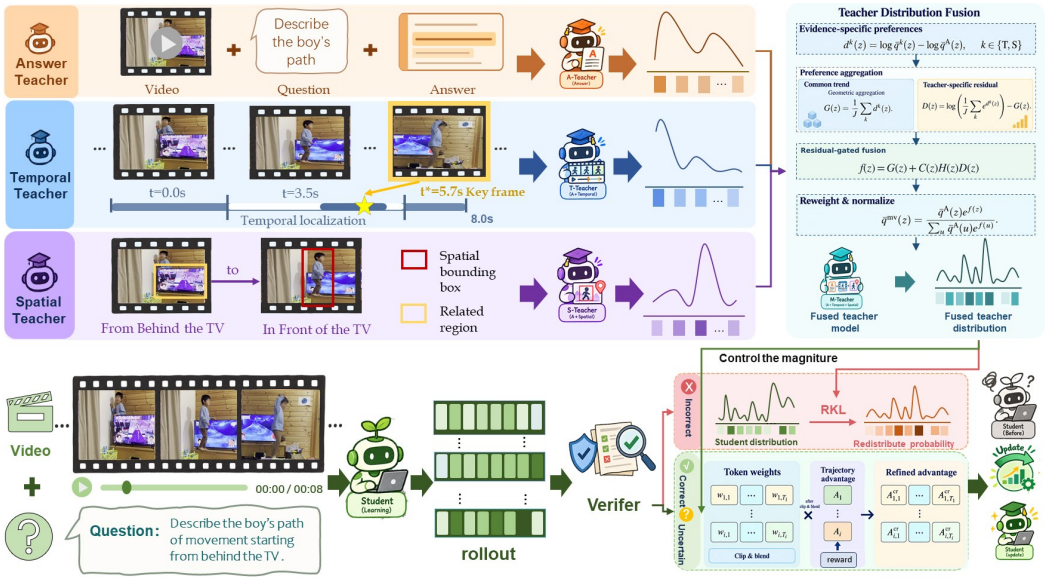


Figure 2: Overview of MeSD. Teachers sharing EMA parameters use answer, temporal, and spatial annotations as privileged information to evaluate student-generated trajectories. Using the Answer Teacher as an anchor, teacher distribution fusion aggregates evidence-specific preferences through a gated consensus–residual decomposition. Trajectory verification determines how the student learns. For failed trajectories, the student matches the fused teacher distribution through reverse-KL distillation. For other trajectories, teacher guidance adjusts the strength of token-level updates, and rewards determine whether tokens are reinforced or suppressed.

answer provides a common semantic context, with temporal and spatial evidence added separately:

$$\begin{aligned} \bar{q}^{k}_{i,t} &= \pi_{\bar{\theta}}\big(\cdot \mid x_i, E^{k}_{i}, y_{i,<t}\big), \quad k \in \{\mathrm{A},\mathrm{T},\mathrm{S}\}, \\ E^{\mathrm{A}}_{i} &= a_i, \quad E^{\mathrm{T}}_{i} = (a_i, \tau_i), \quad E^{\mathrm{S}}_{i} = (a_i, \mathcal{B}_i), \end{aligned} \tag{1}$$

where A, T, and S denote the Answer, Answer-Temporal, and Answer-Spatial Teachers, respectively. The component a_i contains only the semantic ground-truth answer, excluding temporal intervals, keyframe timestamps, and bounding box coordinates. The temporal evidence τ_i provides annotated time intervals or keyframe timestamps. For the spatial evidence, we format the ground-truth answer, object names, and transformed bounding boxes as privileged information. The teachers compute predictions for the existing student trajectories without generating additional rollouts, and these predictions are detached to prevent gradient backpropagation.

Teacher Distribution Fusion. Given these teacher distributions, we use the Answer Teacher as the anchor teacher and measure evidence-specific preferences relative to its predictions. We then combine *teacher-common preferences* with a *gated teacher-specific residual* to capture consistency and conflict across evidence sources. For each candidate token z, we define the preference change induced by temporal or spatial evidence as:

$$d^{k}(z) = \log \bar{q}^{k}(z) - \log \bar{q}^{\mathrm{A}}(z), \qquad k \in \{\mathrm{T},\mathrm{S}\}, \tag{2}$$

where $\bar{q}^{k}$ denotes $\bar{q}^{k}_{i,t}$ with the sample and token indices omitted. A positive $d^{k}(z)$ indicates that the additional evidence increases the support for token z relative to the Anchor Teacher, while a negative value indicates a decrease in support.

The geometric mean emphasizes tokens supported across all non-anchor teachers, as a low probability from any single non-anchor teacher reduces the joint support. The arithmetic mean retains stronger support from individual teachers and preserves complementary evidence not shared across teachers. We use the geometric mean to define the teacher-common preferences, and the difference between the two means in log space defines the teacher-specific residual:

$$G(z) = \tfrac{1}{J}\textstyle\sum_{k} d^{k}(z), \quad D(z) = \log\big(\tfrac{1}{J}\textstyle\sum_{k} e^{d^{k}(z)}\big) - G(z), \tag{3}$$

where J is the number of non-anchor teachers. The teacher-common preference $G(z)$ is the log ratio of the unnormalized geometric mean to the Answer Teacher distribution. The teacher-specific residual $D(z)$ is the log ratio of the arithmetic mean to the unnormalized geometric mean, capturing additional support from individual teachers. Since this support may reflect complementary evidence or conflicting preferences, we gate the residual using directional agreement and relative magnitude:

$$C(z) = \frac{|G(z)|}{\frac{1}{J}\sum_k |d^k(z)| + \epsilon}, \qquad H(z) = \frac{|G(z)|}{|G(z)| + D(z) + \epsilon}, \tag{4}$$

where ϵ is a numerical stability constant. The agreement gate $C(z)$ reduces the residual contribution when the evidence-specific preferences $d^k(z)$ exhibit opposing signs. The magnitude gate $H(z)$ limits this contribution relative to $|G(z)|$. We add the gated residual to the teacher-common preference $G(z)$, subsequently reweighting and normalizing the Answer Teacher distribution:

$$f(z) = G(z) + C(z)H(z)D(z), \qquad \bar{q}^{\mathrm{mv}}(z) = \frac{\bar{q}^{\mathrm{A}}(z)e^{f(z)}}{\sum_u \bar{q}^{\mathrm{A}}(u)e^{f(u)}}. \tag{5}$$

The fused distribution combines teacher-common preferences with gated teacher-specific residuals from available teachers. When semantic answers are unavailable, an unprivileged anchor teacher is used, and its distribution is retained if no auxiliary teacher is available.

3.3 Verification-Guided Optimization

Token-level advantage refinement preserves the signs of reward-derived advantages but provides limited corrective guidance for failed trajectories. We therefore apply token-level advantage refinement or failure-conditioned distillation according to trajectory verification outcomes.

Trajectory Verification. A task-specific verifier assigns an outcome o_i to each trajectory i based on its required output components $\mathcal{C}_i$. For each component $c \in \mathcal{C}_i$, we first determine its verification status $v(c)$ (more details in Appendix C.2). The overall outcome is defined as:

$$o_i = \begin{cases} \text{Success}, & \text{if } \forall c \in \mathcal{C}_i, v(c) = \text{correct}, \\ \text{Failure}, & \text{if } \exists c \in \mathcal{C}_i, v(c) = \text{incorrect}, \\ \text{Indeterminate}, & \text{otherwise}. \end{cases} \tag{6}$$

The status $v(c)$ is determined by task-specific rules: For multiple-choice questions, $v(c)$ is correct if the answer matches the ground truth, and incorrect if it is parseable but wrong. For temporal localization, $v(c)$ is correct when the temporal intersection over union (tIoU) is at least 0.5, and incorrect when the tIoU is zero. For temporal multiple-choice questions, both the answer and the interval must be correct. If a component cannot be parsed, or if the task lacks explicit verification rules, $v(c)$ is marked as unparseable, leading to an ***Indeterminate*** outcome. We apply failure-conditioned distillation only to verified failure trajectories that have the required temporal or spatial evidence.

Token-Level Advantage Refinement. For non-failure trajectories, we follow RLSD Yang et al. (2026) and refine token-level advantage magnitudes while preserving their reward-derived signs. We measure the fused teacher's relative support for each sampled token using the teacher-student log-probability difference:

$$\Delta_{i,t} = \mathrm{sg}\big[\log \bar{q}^{\mathrm{mv,cr}}_{i,t}(y_{i,t}) - \log \bar{p}^{\mathrm{cr}}_{i,t}(y_{i,t})\big], \tag{7}$$

where $y_{i,t}$ is the token at position t in trajectory i. The fused teacher distribution $\bar{q}^{\mathrm{mv,cr}}_{i,t}$ and student distribution $\bar{p}^{\mathrm{cr}}_{i,t}$ are conditioned on the same input and student-generated prefix $y_{i,<t}$. Both distributions are normalized over the shared candidate set, and the operator sg denotes the stop-gradient operation. We use $\Delta_{i,t}$ to rescale the trajectory advantage at each token:

$$\begin{aligned} w_{i,t} &= \exp\big(\mathrm{sign}(A_i)\Delta_{i,t}\big), \\ A^{\mathrm{cr}}_{i,t} &= A_i\Big[(1-\lambda_s) + \lambda_s \,\mathrm{clip}\Big(w_{i,t},\, 1-\epsilon_w,\, 1+\epsilon_w\Big)\Big]. \end{aligned} \tag{8}$$

where A_i is the group-relative advantage of trajectory i, $w_{i,t}$ is the token-specific weight, and $A^{\mathrm{cr}}_{i,t}$ is the refined token advantage. At training step s, $\lambda_s \in [0,1]$ controls the strength of teacher

reweighting, while $\epsilon_w \in [0, 1)$ defines the weight clipping range. We substitute $A^{\text{cr}}_{i,t}$ into the GRPO clipped objective:

$$\ell^{\text{cr}}_i = -\tfrac{1}{T_i} \sum\nolimits_{t=1}^{T_i} \min\big[\rho_{i,t} A^{\text{cr}}_{i,t}, \text{clip}(\rho_{i,t}, 1-\epsilon_{\text{pg}}, 1+\epsilon_{\text{pg}}) A^{\text{cr}}_{i,t}\big], \\ \rho_{i,t} = \pi_\theta(y_{i,t} \mid x_i, y_{i,<t}) \,/\, \pi_{\text{old}}(y_{i,t} \mid x_i, y_{i,<t}). \tag{9}$$

where $\rho_{i,t}$ is the probability ratio between the current and rollout policies. Here, T_i denotes the number of valid generated tokens, and ϵ_{pg} is the policy clipping threshold.

Failure-Conditioned Distillation. For the verified failure trajectories, we minimize the reverse KL divergence from the student to the fused teacher distribution, conditioned on the same student-generated prefixes. This reverse-KL loss replaces the GRPO policy loss for these trajectories:

$$\ell^{\text{rkl}}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\text{KL}}\big(\bar{p}^{\text{rkl}}_{i,t} \,\big\|\, \text{sg}[\widehat{q}^{\text{rkl}}_{i,t}]\big)\,, \tag{10}$$

where $\bar{p}^{\text{rkl}}_{i,t}$ is the student distribution and $\widehat{q}^{\text{rkl}}_{i,t}$ is the fused teacher distribution. Both distributions are defined over the same candidate set, and the probabilities of all other tokens are combined into a single category. During optimization, the teacher target distribution is detached, so gradients flow only through the student distribution. Importantly, this loss does not depend on the advantage A_i. It provides corrective supervision even when group-relative advantages vanish, since the reverse KL divergence inherently penalizes any discrepancy between the student and teacher distributions.

Unified Training Objective. The overall objective is:

$$\mathcal{L}_s(\theta) = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \big[(1-g_i)\ell^{\text{cr}}_i + g_i \gamma_s \ell^{\text{rkl}}_i\big]\,, \tag{11}$$

where $\mathcal{B}_s$ is the valid trajectory batch at step s and $\gamma_s \geq 0$ weights the distillation loss. We set $g_i = 1$ for trajectories that use failure-conditioned distillation with the self-teacher, and $g_i = 0$ otherwise. Teacher targets remain detached throughout training. We gradually decay λ_s during self-distillation, then set $\lambda_s = g_i = 0$ to recover GRPO. During self-distillation, the shared teacher parameters are initialized as $\bar{\theta}_0 = \theta_0$ and updated after each student update:

$$\bar{\theta}_s = \beta \bar{\theta}_{s-1} + (1-\beta)\theta_s, \qquad 0 \leq \beta < 1, \tag{12}$$

where β is the EMA decay coefficient. We gradually decay λ_s during self-distillation and recover GRPO afterward by setting $\lambda_s = g_i = 0$ for all trajectories.

4 Experiments

4.1 Experimental Setup

Implementation Details. MeSD is initialized from Open-o3-Video-SFT-7B and trained on STGR-RL (Meng et al., 2025) for one epoch. For a fair comparison, we use the same training data and annotations, with existing answer and grounding labels serving as teacher privileged information. We generate four on-policy responses per prompt using vLLM at temperature 1.0, with a maximum response length of 768 tokens. Videos are sampled at 2 FPS with at most 16 frames. The learning rate is 10^{-6}, and token weights are clipped to $[0.8, 1.2]$. The teacher reweighting coefficient λ_s decays linearly from 0.5 to zero over the first 1000 steps. Eligible failures use reverse-KL distillation in place of the GRPO policy loss, with γ_s increasing linearly from zero to 0.8 over the first 100 steps. After step 1000, teacher supervision is disabled and training continues with standard GRPO. Further experimental details are provided in Appendix D.

4.2 Main Results

What, When, and Where: Grounded Reasoning on V-STAR. Table 1 demonstrates that MeSD substantially outperforms Qwen2.5-VL-7B across all evaluation dimensions. Answer accuracy achieves the largest absolute gain, increasing from 33.5% to 63.7%. Meanwhile, improvements in temporal and spatial IoU indicate that these more accurate answers are accompanied by better

Table 1: Performance comparison on V-STAR. Chain1 and Chain2 denote the what–when–where and what–where–when reasoning orders, respectively. Best and second-best results are shown in **bold** and underlined.

Model	What	When (temporal IoU)		Where (spatial IoU)		Overall	
	Acc.	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Closed-source Models							
Gemini-2-Flash (Google DeepMind, 2025)	53.0	24.5	23.8	4.6	2.2	26.9	35.6
GPT-4o (OpenAI, 2024)	60.8	16.7	12.8	6.5	3.0	26.8	38.2
Open-source General Models							
Oryx-1.5-7B (Liu et al., 2025)	20.5	13.5	14.8	10.1	3.5	15.1	13.8
InternVL-2.5-8B (Chen et al., 2024)	44.2	8.7	7.8	0.7	0.1	17.6	24.9
Qwen2.5-VL-7B (Bai et al., 2025b)	33.5	15.4	13.8	17.0	2.5	19.3	22.4
Video-specialized Models							
TRACE (Guo et al., 2025d)	17.6	19.1	17.1	0.0	0.0	12.0	13.3
VideoLLaMA3-7B (Zhang et al., 2025a)	41.9	23.0	23.1	0.9	0.2	21.7	27.0
LLaVA-Video (Zhang et al., 2025b)	49.5	10.5	12.2	1.9	1.3	20.8	27.3
VideoChat2 (Li et al., 2024)	36.2	13.7	12.5	2.5	1.0	17.0	20.3
Open-o3-Video-7B (Meng et al., 2025)	61.0	24.5	24.0	25.4	6.0	33.7	46.6
VISD-7B (Lin et al., 2026)	61.9	**26.8**	**27.2**	26.7	6.3	35.1	48.9
MeSD-7B (Ours)	**63.7**	25.8	25.0	**28.1**	**6.5**	**35.5**	**50.2**
Δ vs. Qwen2.5-VL-7B	+30.2	+10.4	+11.2	+11.1	+4.0	+16.2	+27.8

Table 2: Performance comparison on WorldSense, VideoMMMU, LRR, and TVGBench. LRR denotes LongVideo-Reason-eval. Avg denotes the average of WorldSense Overall, VideoMMMU Overall, LRR Acc., and TVGBench mIoU. Best and second-best results among open-source models are shown in **bold** and underlined, respectively.

Model	WorldSense		VideoMMMU		LRR	TVGBench	Avg
	Overall	Understanding	Overall	Perception	Acc.	mIoU	
GPT-4o	42.6	–	61.2	66.0	–	–	–
Qwen2.5-VL-7B (Bai et al., 2025b)	35.5	38.4	51.2	64.7	59.3	16.3	40.6
VideoRFT-7B (Wang et al., 2025a)	38.4	40.0	51.1	66.0	69.4	14.3	43.3
VideoR1-7B (Feng et al., 2025)	36.9	37.7	52.4	65.3	68.9	9.6	42.0
Open-o3-Video-7B (Meng et al., 2025)	38.9	39.7	52.6	67.4	69.1	19.2	45.0
VisionCoach-7B (Lee et al., 2026)	39.2	40.2	53.8	70.3	70.7	21.0	46.2
VISD-7B (Lin et al., 2026)	39.9	40.4	53.8	69.3	73.5	24.8	48.0
MeSD-7B (Ours)	**41.6**	**46.0**	**56.1**	**73.0**	**73.6**	**29.8**	**50.3**
Δ vs. Qwen2.5-VL-7B	+6.1	+7.6	+4.9	+8.3	+14.3	+13.5	+9.7

event and object localization. Compared to VISD, the advantages of MeSD lie primarily in answer accuracy and spatial grounding, achieving the best results among all evaluated models in both dimensions. Although its temporal IoU is slightly lower than that of VISD, MeSD achieves the highest aggregate scores, with an mAM of 35.5% and an mLGM of 50.2%. This highlights its ability to combine accurate answers with explicit visual grounding.

Performance across Diverse Video Tasks. As shown in Table 2, MeSD achieves the highest performance among the evaluated open-source models across all reported metrics. Compared to Qwen2.5-VL-7B, MeSD increases the average score from 40.6% to 50.3%. The most substantial benchmark-level gains occur on LRR and TVGBench, indicating stronger long-video reasoning and more precise temporal localization. Furthermore, MeSD outperforms VISD on all four benchmarks, increasing the average score from 48.0% to 50.3%. The most notable improvements are observed in temporal grounding, WorldSense understanding, and VideoMMMU perception. These advancements over a strong self-distillation baseline validate the effectiveness of MeSD.

Zero-Shot Temporal Grounding on Charades-STA. Table 3 presents the zero-shot temporal grounding results in terms of recall at IoU thresholds of 0.3, 0.5, and 0.7, alongside mIoU. MeSD outperforms all compared methods across these four metrics. Compared to VISD, it improves R@0.5 and R@0.7 by 8.6% and 7.2%, respectively, and increases mIoU from 44.6% to 49.2%. Furthermore, at the highest IoU threshold, MeSD surpasses Open-o3-Video, an achievement VISD fails to match. These results demonstrate more precise temporal alignment and validate the transferability of the evidence-based supervision in MeSD.

Table 3: *Charades-STA* results. Best and second-best results are shown in **bold** and underlined.

Model	R@0.3	R@0.5	R@0.7	mIoU
General Video LLMs				
VideoChat (Li et al., 2023)	9.0	3.3	1.3	6.5
VideoLLaMA (Zhang et al., 2023)	10.4	3.8	0.9	7.1
Video-ChatGPT (Maaz et al., 2024)	20.0	7.7	1.7	13.7
Valley (Luo et al., 2023)	28.4	1.8	0.3	21.4
Temporal Grounding Video LLMs				
TimeChat (Ren et al., 2024)	–	32.2	13.4	–
HawkEye (Wang et al., 2024)	50.6	31.4	14.5	33.7
VTG-LLM (Guo et al., 2025c)	51.2	33.8	15.7	34.4
VTimeLLM (Huang et al., 2024)	51.0	27.5	11.4	31.2
Grounded Reasoning Models				
Open-o3-video (Meng et al., 2025)	62.6	45.6	<u>24.5</u>	42.5
VISD (Lin et al., 2026)	<u>67.3</u>	<u>46.1</u>	21.6	<u>44.6</u>
MeSD (Ours)	**73.1**	**54.7**	**28.8**	**49.2**

4.3 Ablation Studies and Analysis

Table 4: Ablation study on MeSD components and training strategies. FCD denotes failure-conditioned distillation, ES denotes evidence separation, T&S denotes temporal and spatial evidence, GPA denotes gated preference aggregation, and VGO denotes verification-guided optimization. The best results in each column are highlighted in bold. More details in Appendix F.5.

Model	V-STAR		WorldSense		Video-MMMU		LRR	TVGBench
	mAM	mLGM	Overall	Understanding	Overall	Perception	Acc.	mIoU
Component Variants								
w/o FCD	30.7	41.0	38.6	38.7	53.8	68.3	71.3	23.0
w/o ES	30.9	41.3	38.3	38.4	55.4	71.0	70.2	22.4
w/o T&S	30.8	41.2	37.9	39.3	54.8	71.0	70.6	22.3
w/o GPA	30.9	41.4	39.5	39.5	53.9	69.7	70.4	22.9
Training Strategy Variants								
Base VLM	19.3	22.4	35.5	38.4	51.2	64.7	59.3	16.3
GSPO	33.7	46.6	37.5	39.7	52.3	68.0	69.4	20.8
GRPO	30.6	40.7	39.2	40.4	53.2	69.3	71.9	22.5
w/o VGO	30.6	40.9	39.5	41.0	54.1	67.7	71.2	20.7
MeSD	**35.5**	**50.2**	**41.6**	**46.0**	**56.1**	**73.0**	**73.6**	**29.8**

Effects of Evidence Separation and Complementarity. Table 4 compares MeSD with two ablated variants: a single teacher conditioned on all available evidence (w/o ES) and an answer-only teacher (w/o T&S). Both variants underperform compared to MeSD across all metrics. Notably, the TVGBench mIoU drops from 29.8% to 22.4% and 22.3%, respectively. These results support incorporating temporal and spatial evidence and modeling their preferences separately.

Effects of Gated Preference Aggregation. With the evidence-conditioned teachers fixed, we replace the gated preference aggregation with a simple arithmetic averaging of the evidence-specific preferences (w/o GPA). This modification degrades performance across all benchmarks (Table 4). Specifically, V-STAR mLGM drops from 50.2% to 41.4%, and TVGBench mIoU decreases from 29.8% to 22.9%. These results demonstrate the advantages of selectively retaining teacher-specific support while suppressing conflicting contributions.

Effects of Failure-Conditioned Distillation. Applying token-level advantage refinement to all trajectories without FCD reduces the V-STAR mLGM and TVGBench mIoU by 9.2% and 6.8%, respectively (Table 4). The training curves exhibit a similar pattern: the model without FCD reaches a lower reward plateau, whereas the MeSD reward continues to increase before stabilizing (Figure 3(a)). These results indicate that the corrective guidance provided by advantage refinement for failed trajectories is inherently limited. Failure-conditioned distillation employs the fused teacher

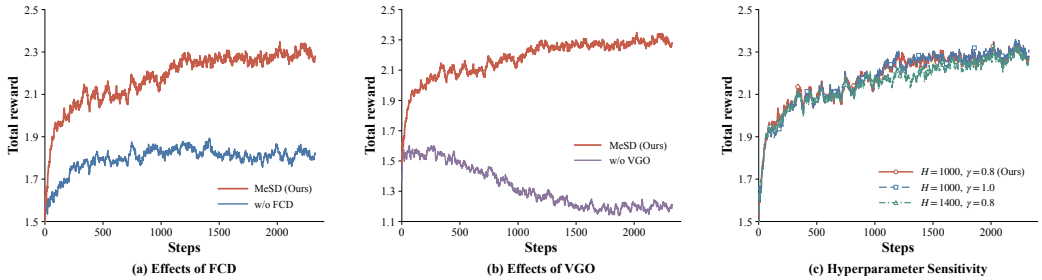


Figure 3: Training ablation curves. MeSD achieves higher training rewards than (a) credit refinement alone (w/o FCD) and (b) multi-evidence OPSD (w/o VGO). (c) Varying the teacher cutoff step H and γ yields similar reward trajectories.

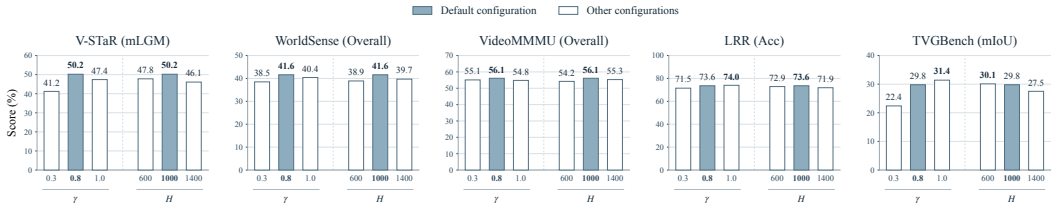


Figure 4: Sensitivity of MeSD to γ and the teacher cutoff step H across five benchmarks. Filled bars indicate the default settings ($\gamma = 0.8$, $H = 1000$).

distribution for reverse-KL correction, thereby providing supervision over candidate tokens that extends beyond sampled-token reweighting.

Effects of Optimization Strategy. MeSD outperforms GRPO and GSPO across all reported metrics (Table 4), raising V-STAR mLGM from GSPO's 46.6% to 50.2% and TVGBench mIoU from GRPO's 22.5% to 29.8%. These gains support complementing reward-based optimization with evidence-conditioned teacher supervision for spatio-temporal reasoning and general video tasks.

Effects of Verification-Guided Optimization. Replacing VGO with vanilla OPSD (w/o VGO) while retaining the multi-evidence teachers lowers all reported metrics, with TVGBench mIoU decreasing from 29.8% to 20.7% (Table 4). Figure 3(b) further shows declining training rewards without VGO, whereas MeSD improves and stabilizes at a higher level. These results suggest that multi-evidence supervision alone is insufficient, supporting verification-guided teacher supervision.

Effects of Correction Strength and Teacher Guidance Duration. We separately vary $\gamma \in \{0.3, 0.8, 1.0\}$ (direct-correction strength after warm-up) and $H \in \{600, 1000, 1400\}$ (teacher-supervision cutoff), keeping the other parameter at its default. The three configurations in Figure 3(c) exhibit similar reward trajectories and final rewards, suggesting training robustness within the tested ranges. Figure 4 shows that $\gamma = 0.8$ and $H = 1000$ perform best on three and four benchmarks, respectively, supporting these defaults as a balanced choice across tasks.

5 Conclusion

We introduced MeSD, a multi-evidence self-distillation framework for VideoLLMs that combines evidence-aware supervision with verification-guided optimization. Using an answer-conditioned teacher as a semantic anchor, MeSD integrates temporal and spatial evidence through teacher-common preferences and gated teacher-specific residuals. It preserves reward-derived advantage signs during credit refinement and reserves reverse-KL correction for verified failures with the required evidence. Experiments demonstrate overall gains over evaluated baselines across video understanding, spatiotemporal reasoning, and temporal grounding, including improvements across Qwen3-VL model sizes. Ablations support evidence separation, gated fusion, and selective correction for effective video post-training using existing annotations.

AI USE STATEMENT

Generative AI tools were used solely to polish the language and improve the readability of the manuscript. The authors take full responsibility for the final content, including all claims, references, and reported results.

ETHICS STATEMENT

This work studies multi-evidence self-distillation for video understanding using existing training data and established benchmarks. MeSD reuses the annotations in STGR-RL without introducing additional human-written or model-generated annotations. The resulting models may inherit biases from the underlying models and datasets and may produce incorrect answers or unsupported spatiotemporal evidence. Improved benchmark performance does not establish fairness, safety, or suitability for high-stakes deployment. Applications involving videos of people should respect privacy and consent and account for potential misuse in surveillance or automated judgments about individuals.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 describe the MeSD framework, and Section 4.1 summarizes the experimental setup. Appendix B.1 documents the training data and evidence construction; Appendix B.2 and Table 6 specify the hyperparameters and training schedule; and Appendix B.3 presents the evidence-conditioned teacher prompts. Appendices C.1–C.3 detail teacher distribution computation, trajectory verification, and the analysis of credit refinement and failure correction, respectively. Appendices D.1–D.3 describe the evaluation benchmarks, baselines, and metrics. Appendix F.5 provides detailed comparisons and ablations, while Appendix F.6 presents training curves.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Qiye Cai, Yichuan Ma, Peiji Li, Yongkang Chen, Qipeng Guo, Yicheng Zou, Linyang Li, Xiaocheng Feng, and Bing Qin. H^2SD: Hybrid hindsight self-distillation. *arXiv preprint arXiv:2607.18955*, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 172842–172870, 2025. doi: 10.52202/085713-5753.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-STaR: Benchmarking video-LLMs on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9155–9164, 2026.

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 *Advances in Neural Information Processing Systems*, volume 38, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026. doi: 10.48550/arXiv.2604.05015.

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.

Google DeepMind. Gemini 2.0 Flash model card, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. *Advances in Neural Information Processing Systems*, 38:114399–114431, 2025b.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3302–3310, 2025c. doi: 10.1609/aaai.v39i3.32341.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video LLM via causal event modeling. In *International Conference on Learning Representations*, volume 2025, pp. 89594–89618, 2025d.

Xixiang He, Xingming Li, Baiqi Wu, Qiyao Sun, Xuanyu Ji, Ao Cheng, and Qingyong Hu. Learn from whoever is right: Answer-verified multi-teacher distillation for multi-domain LLMs. *arXiv preprint arXiv:2609.02548*, 2026.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 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.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. VisionCoach: Reinforcing grounded video reasoning via visual-perception prompting. *arXiv preprint arXiv:2603.14659*, 2026.

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. *arXiv preprint arXiv:2602.02994*, 2026.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

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.

Hao Lin, Kunyang Lv, Xu Jiang, Jingqi Tian, Zhongjing Du, Jiayu Ding, Qiaoman Zhang, and Hongbo Jin. VISD: Enhancing video reasoning via structured self-distillation. *arXiv preprint arXiv:2605.06094*, 2026.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations*, volume 2025, pp. 85485–85507, 2025.

Ruipu Luo, Ziwang Zhao, Min Yang, Zheming Yang, Minghui Qiu, Tao Wang, Zhongyu Wei, Yanhao Wang, and Cen Chen. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024. doi: 10.18653/v1/2024.acl-long.679.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3-Video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Zaid Khan, Anirban Das, Austin Zhang, Sambit Sahu, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. AVSD: Adaptive-view self-distillation by balancing consensus and teacher-specific privileged signals. *arXiv preprint arXiv:2605.20643*, 2026.

OpenAI. Hello GPT-4o, 2024.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. DeepVideo-R1: Video reinforcement fine-tuning via difficulty-aware regressive GRPO. In *Advances in Neural Information Processing Systems*, volume 38, pp. 153710–153737, 2025. doi: 10.52202/085713-4626.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Shravan Venkatraman, Wenshuai Zhao, Mohammad Hassan Vali, and Arno Solin. Temporal self-distillation: Learning visual state tracking in videos without supervision. *arXiv preprint arXiv:2609.04203*, 2026.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4350–4376. Curran Associates, Inc., 2025a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawk-Eye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024.

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-OPSD: Exploiting privileged visual evidence for on-policy self-distillation in video large language models. *arXiv preprint arXiv:2608.27065*, 2026.

Zhiheng Wu, Tong Wang, Shuning Wang, Naiming Liu, and Yumeng Zhang. See further, think deeper: Advancing vlm's reasoning ability with low-level visual cues and reflection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18870–18880, June 2026.

Xun Xu and Zaixi Zhang. Who teaches which token? verifier-gated multi-expert on-policy distillation for scientific reasoning. *arXiv preprint arXiv:2609.15404*, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023. doi: 10.18653/v1/2023.emnlp-demo.49.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

APPENDIX CONTENTS

A Extended Related Work **15**

A.1 Post-Training for VideoLLMs 15

A.2 On-Policy and Privileged-Context Self-Distillation 15

A.3 Evidence Aggregation and Adaptive Supervision 15

B Implementation and Training Details **16**

B.1 Training Data and Evidence Construction 16

B.2 Hyperparameters and Training Schedule 16

B.3 Evidence-Conditioned Teacher Prompts 17

C Supplementary Model Details **18**

C.1 Teacher Distribution Computation 18

C.2 Trajectory Verification 19

C.3 Analysis of Credit Refinement and Failure Correction 21

D Evaluation Details **21**

D.1 Benchmarks 22

D.2 Baselines 22

D.3 Evaluation Metrics 22

F Additional Experimental Results **23**

F.1 Video-MME-v2 Results 23

F.2 Qwen3-VL Detailed Results 23

F.3 Effects of Correction Selection. 24

F.4 Training Efficiency. 25

F.5 Detailed Comparisons and Ablations 25

F.6 Training Curves 27

G Case Study **28**

H Limitations and Future Work **29**

A Extended Related Work

A.1 Post-Training for VideoLLMs

RLVR supports VideoLLM post-training through outcome-level feedback on model-generated trajectories. Video-R1 (Feng et al., 2025) encourages temporal reasoning, VideoRFT (Wang et al., 2025a) rewards consistency between reasoning and visual evidence, and DeepVideo-R1 (Park et al., 2025) improves optimization through a regressive objective and difficulty-aware augmentation. For localization, Time-R1 (Wang et al., 2025b) targets temporal grounding, while VideoChat-R1 (Li et al., 2025) strengthens spatiotemporal perception through multi-task reinforcement fine-tuning. Video-OPD (Li et al., 2026) provides dense teacher guidance for efficient temporal grounding, and VisionCoach (Lee et al., 2026) combines visual prompting, grounding rewards, and self-distillation to improve evidence localization.

Open-o3 Video (Meng et al., 2025) unifies answer and spatiotemporal supervision through reasoning traces containing timestamps, objects, and bounding boxes, with rewards jointly evaluating answer correctness and grounding quality. MeSD follows this paradigm using the same training data and annotations. It organizes answer and grounding information into evidence-conditioned self-teacher contexts for token-level guidance along student-generated trajectories.

A.2 On-Policy and Privileged-Context Self-Distillation

On-policy distillation addresses the mismatch between distillation sequences and student-generated sequences at inference. **Generalized Knowledge Distillation (GKD)** (Agarwal et al., 2024) provides teacher-distribution supervision at student-generated prefixes. **Video-OPD** (Li et al., 2026) applies this principle using a stronger external teacher. **On-Policy Self-Distillation (OPSD)** (Zhao et al., 2026) uses the same model under different contexts, conditioning the teacher on privileged information, such as reference solutions, to supervise student rollouts. **Self-Distillation Policy Optimization (SDPO)** (Hübotter et al., 2026) converts textual environment feedback into dense supervision through a feedback-conditioned self-teacher. Both derive teacher guidance from contextual information without a separately trained, stronger model.

Multimodal self-distillation extends this principle to visual evidence. **Vision-OPD** (Yuan et al., 2026) transfers crop-conditioned predictions to a student processing full images. **Video-OPSD** (Wang et al., 2026) conditions its self-teacher on annotated evidence frames and weights distillation by token reliance on visual evidence. **S³T** (Venkatraman et al., 2026) uses denser temporal sampling to supervise a sparse-view student without labels or rewards. **VISD** (Lin et al., 2026) incorporates structured feedback from a video-aware judge to diagnose answer consistency and spatiotemporal grounding. These methods construct visual self-teachers through distinct contextual evidence. MeSD organizes existing answer and grounding annotations into separate contexts to explicitly compare and fuse evidence-conditioned teacher preferences.

A.3 Evidence Aggregation and Adaptive Supervision

Multi-teacher on-policy distillation combines complementary capabilities on student-generated trajectories. **MOPD** (Ma et al., 2026) merges independently trained specialists into one student, while **MT-SDPO** (He et al., 2026) conditions an EMA self-teacher on verified teacher feedback. **AVSD** (Nguyen et al., 2026) aggregates privileged views for self-distillation. A single video-evidence context leaves source contributions and cross-evidence agreement, complementarity, and conflict implicit. MeSD separates evidence-conditioned views and fuses temporal and spatial preference shifts relative to an answer-conditioned semantic anchor.

For adaptive supervision, **VG-OPD** (Xu & Zhang, 2026) uses verification to select eligible experts and localize token-level supervision. **RLSD** (Yang et al., 2026) and **VISD** (Lin et al., 2026) refine token credit while preserving reward-derived update directions; **H²SD** (Cai et al., 2026) explores outcome-dependent asymmetric self-distillation. Failure feedback alone does not identify preferred alternatives. Based on verification and evidence availability, MeSD uses its fused teacher distribution for reward-anchored credit refinement on successful and indeterminate trajectories, and reverse-KL correction only on eligible, verified failures.

B Implementation and Training Details

B.1 Training Data and Evidence Construction

Table 5: Task composition and annotation types of STGR-RL. Answer denotes the semantic answer used in the answer context.

Task type	Samples	Answer	Temporal	Spatial
General video QA (MCQ)	13,000	Option	–	–
General video QA (free-form)	2,000	Text	–	–
Temporal QA	2,280	–	Time intervals	–
Temporal QA (MCQ)	2,904	Option	Time intervals	–
Visual QA	5,000	Object name	–	Object boxes
Joint spatio-temporal QA	12,047	Text	Keyframe timestamps	Object boxes
Total	**37,231**			

We follow Open-o3-Video (Meng et al., 2025) and use its STGR-RL training data, which covers video question answering and temporal and spatial grounding. For a fair comparison, we retain the same training data and existing annotations without introducing any additional human-written or model-generated annotations.

We construct the privileged teacher contexts directly from the available annotations. The Answer context contains the semantic answer, while the Answer+Temporal and Answer+Spatial contexts additionally include the available temporal annotations and object-level bounding boxes, respectively. Temporal evidence specifies relevant intervals or keyframe timestamps, and spatial evidence preserves the associations between annotated objects, bounding boxes, and their corresponding frames. Thus, MeSD reorganizes the existing supervision into separate evidence-conditioned contexts without expanding the annotation set.

Table 5 summarizes the task composition and annotation types of the training set. It contains 37,231 samples, including 32,231 video-based samples and 5,000 image-based spatial grounding samples. Annotation availability varies across tasks: general video QA provides semantic answers, temporal QA supplies time intervals, and joint spatio-temporal QA includes keyframe timestamps and object-level bounding boxes. For pure temporal QA, the answer is itself a time interval and therefore contributes temporal evidence rather than a separate semantic answer.

Table 6: Training configuration of MeSD.

Training and rollout		**Self-distillation**	
Parameter	**Value**	**Parameter**	**Value**
Training epochs	1	EMA update rate	0.01
Optimizer	AdamW	Initial strength λ_0	0.5
Learning rate	10^{-6}	Annealing steps	1,000
LR schedule	Cosine	Token-weight clipping	$[0.8, 1.2]$
Prompt batch size	16	Correction strength γ	0.8
Rollouts per prompt	4	Correction warm-up	100 steps
Sampling temperature	1.0	Teacher cutoff H	1,000 steps
Video sampling	2 FPS, $\leq$ 16 frames		
Max. response length	768 tokens		

B.2 Hyperparameters and Training Schedule

Table 6 summarizes the main hyperparameters for our MeSD-7B experiments. We use a staged training schedule that combines early self-teacher guidance with subsequent reward-based optimization.

Prompt Template for Answer Teacher

```
<|im_start|>system
A conversation between user and assistant. The user provides a video and asks a
question, and the Assistant solves it. The assistant MUST first think about the
reasoning process in the mind and then provide the user with the answer. The reasoning
process and answer are enclosed within <think> </think> and <answer> </answer> tags,
respectively. All reasoning must be grounded in visual evidence from the video. When
you mention any related object, person, or specific visual element in the reasoning
process, you must strictly follow the following format:
`<obj>object_name</obj><box>bounding_box</box>at<t>time_in_seconds</t>s`. The answer
part only requires a text response; tags like <obj>, <box>, <t> are not needed.
<|im_end|>
<|im_start|>user
{original_visual_input_and_question}
[Hidden hindsight evidence: answer view]
Use only the verified fields listed below as hidden supervision.
Verified semantic answer: {semantic_answer}
<|im_end|>
<|im_start|>assistant
{student_generated_prefix}
```

Figure 5: Answer teacher prompt with privileged semantic answers.

During the first 1,000 steps, trajectories eligible for failure-conditioned correction are optimized with reverse-KL distillation, while the remaining trajectories retain reward-anchored credit refinement.

For credit refinement, the mixing coefficient λ_s is initialized to 0.5 and linearly annealed over the first 1,000 steps, gradually reducing the influence of teacher-based token reweighting. The token weights are clipped to $[0.8, 1.2]$ to bound the modulation while preserving the sign of the reward-derived advantage. For eligible verified failures, the correction coefficient γ_s is linearly increased to 0.8 over the first 100 steps and then held constant while teacher supervision remains active. This warm-up gradually introduces distribution-level correction during early training.

The shared self-teacher is updated by an exponential moving average with an update rate of 0.01, corresponding to an EMA decay coefficient of 0.99. After step 1,000, both teacher-based credit refinement and failure-conditioned distillation are disabled. All subsequent updates use standard GRPO, completing training under the original reward objective.

B.3 Evidence-Conditioned Teacher Prompts

We construct three teacher contexts: Answer, Answer+Temporal, and Answer+Spatial. All contexts retain the original task-specific system prompt, visual input, and question, while appending different privileged evidence to the user message. The Answer context provides the verified semantic answer. The Answer+Temporal context additionally includes annotated temporal intervals or keyframe timestamps, whereas the Answer+Spatial context adds object-level bounding boxes associated with the corresponding image or annotated keyframe. These boxes are expressed in the pixel coordinates of the processed input image.

The accompanying prompt templates illustrate the three contexts for joint spatio-temporal question answering. Braced placeholders are populated from existing annotations; unavailable fields are omitted, and keyframe or spatial-target entries are repeated when multiple annotations are available. All three contexts are evaluated by the same EMA self-teacher on the same student-generated prefix, without generating additional teacher rollouts. The privileged evidence is used only during training and is not provided to the student at inference time.

Prompt Template for Temporal Teacher

```
<|im_start|>system
A conversation between user and assistant. The user provides a video and asks a
question, and the Assistant solves it. The assistant MUST first think about the
reasoning process in the mind and then provide the user with the answer. The reasoning
process and answer are enclosed within <think> </think> and <answer> </answer> tags,
respectively. All reasoning must be grounded in visual evidence from the video. When
you mention any related object, person, or specific visual element in the reasoning
process, you must strictly follow the following format:
`<obj>object_name</obj><box>bounding_box</box>at<t>time_in_seconds</t>s`. The answer
part only requires a text response; tags like <obj>, <box>, <t> are not needed.
<|im_end|>
<|im_start|>user
{original_visual_input_and_question}
[Hidden hindsight evidence: temporal view]
Use only the verified fields listed below as hidden supervision.
Verified semantic answer: {semantic_answer}
Verified keyframe timestamp: {timestamp}s
<|im_end|>
<|im_start|>assistant
{student_generated_prefix}
```

Figure 6: Answer+Temporal teacher prompt with semantic answers and keyframe timestamps.

Prompt Template for Spatial Teacher

```
<|im_start|>system
A conversation between user and assistant. The user provides a video and asks a
question, and the Assistant solves it. The assistant MUST first think about the
reasoning process in the mind and then provide the user with the answer. The reasoning
process and answer are enclosed within <think> </think> and <answer> </answer> tags,
respectively. All reasoning must be grounded in visual evidence from the video. When
you mention any related object, person, or specific visual element in the reasoning
process, you must strictly follow the following format:
`<obj>object_name</obj><box>bounding_box</box>at<t>time_in_seconds</t>s`. The answer
part only requires a text response; tags like <obj>, <box>, <t> are not needed.
<|im_end|>
<|im_start|>user
{original_visual_input_and_question}
[Hidden hindsight evidence: spatial view]
Use only the verified fields listed below as hidden supervision.
Verified semantic answer: {semantic_answer}
Verified spatial target: anchor=annotated_keyframe_image,
coordinate_space=processed_pixel_xyxy, object={object_name}, bbox=[{x1},{y1},{x2},{y2}]
<|im_end|>
<|im_start|>assistant
{student_generated_prefix}
```

Figure 7: Answer+Spatial teacher prompt with semantic answers and object-level bounding boxes.

C Supplementary Model Details

This section provides additional details on teacher distribution computation, trajectory verification, and the two optimization branches, using the notation established in the main text.

C.1 Teacher Distribution Computation

At position t in trajectory i, the student predicts the next token according to $\pi_\theta(\cdot \mid x_i, y_{i,<t})$. The evidence-conditioned teacher distributions $\bar{q}^k_{i,t}$, $k \in A, T, S$, are defined in the main text using their respective contexts E^k_i. These teachers share parameters $\bar{\theta}$ and predict the next token conditioned

on the same student-generated prefix $y_{i,<t}$. Below, the superscripts cr and rkl distinguish the probability representations used for credit refinement and failure-conditioned distillation, respectively.

Shared candidate support. Let $\mathcal{K}_i \subseteq \{A, T, S\}$ denote the available teacher views for trajectory i. For each optimization branch $b \in \{\mathrm{cr}, \mathrm{rkl}\}$, we construct

$$\mathcal{U}_{i,t}^{b} = \{y_{i,t}\} \cup \mathrm{TopK}_{K_b}\big(\pi_\theta(\cdot \mid x_i, y_{i,<t})\big) \quad \cup \bigcup_{k \in \mathcal{K}_i} \mathrm{TopK}_{K_b}\big(\bar{q}_{i,t}^{k}\big), \tag{13}$$

where TopK_{K_b} returns the token IDs with the K_b highest probabilities in a distribution. The union removes duplicate IDs, excludes unavailable views, and always includes the sampled token $y_{i,t}$. We set $K_{\mathrm{cr}} = 16$ and $K_{\mathrm{rkl}} = 32$; these values specify the number of top-ranked candidates selected from each distribution, not the size of the final candidate set. Candidate selection is performed without gradient tracking.

Probabilities for credit refinement. We renormalize the student and available teacher distributions over the shared candidate set $\mathcal{U}_{i,t}^{\mathrm{cr}}$:

$$\begin{aligned} \bar{p}_{i,t}^{\mathrm{cr}}(z) &= \frac{\pi_\theta(z \mid x_i, y_{i,<t})}{\sum_{u \in \mathcal{U}_{i,t}^{\mathrm{cr}}} \pi_\theta(u \mid x_i, y_{i,<t})}, \\ \bar{q}_{i,t}^{k,\mathrm{cr}}(z) &= \frac{\bar{q}_{i,t}^{k}(z)}{\sum_{u \in \mathcal{U}_{i,t}^{\mathrm{cr}}} \bar{q}_{i,t}^{k}(u)}, \qquad k \in \mathcal{K}_i, \end{aligned} \tag{14}$$

for $z \in \mathcal{U}_{i,t}^{\mathrm{cr}}$. We apply the teacher distribution fusion described in the main text to $\bar{q}_{i,t}^{k,\mathrm{cr}}$, obtaining $\bar{q}_{i,t}^{\mathrm{mv,cr}}$. Accordingly, all fusion quantities, $d^k(z)$, $G(z)$, $D(z)$, $C(z)$, $H(z)$, and $f(z)$, are computed from these candidate-normalized teacher distributions.

The fused teacher distribution $\bar{q}_{i,t}^{\mathrm{mv,cr}}$ and the student distribution $\bar{p}_{i,t}^{\mathrm{cr}}$ define the log-probability difference $\Delta_{i,t}$ used in the main text. This difference is detached before computing the token weight $w_{i,t}$ and the refined advantage $A_{i,t}^{\mathrm{cr}}$. The policy ratio $\rho_{i,t}$, however, uses the full-vocabulary probabilities of the sampled token under the current and rollout policies, rather than their candidate-normalized probabilities.

Probabilities for failure-conditioned distillation. We define the coarse probability space $\Omega_{i,t} = \mathcal{U}_{i,t}^{\mathrm{rkl}} \cup \{\perp\}$, where the tail category $\perp$ groups all tokens outside the candidate set. The student distribution on this space is

$$\bar{p}_{i,t}^{\mathrm{rkl}}(z) = \begin{cases} \pi_\theta(z \mid x_i, y_{i,<t}), & z \in \mathcal{U}_{i,t}^{\mathrm{rkl}}, \\ 1 - \displaystyle\sum_{u \in \mathcal{U}_{i,t}^{\mathrm{rkl}}} \pi_\theta(u \mid x_i, y_{i,<t}), & z = \perp. \end{cases} \tag{15}$$

Each teacher distribution $\bar{q}_{i,t}^{k,\mathrm{rkl}}$ is constructed in the same way from $\bar{q}_{i,t}^{k}$. This representation preserves the full-vocabulary probability of each explicit candidate and assigns the remaining probability mass to $\perp$, rather than renormalizing over the candidate set alone.

We apply teacher distribution fusion to $\bar{q}_{i,t}^{k,\mathrm{rkl}}$ over $\Omega_{i,t}$, including the tail category, to obtain $\bar{q}_{i,t}^{\mathrm{mv,rkl}}$. We then apply a probability floor and renormalize the fused distribution to obtain the target used in the main-text reverse-KL objective:

$$\widehat{q}_{i,t}^{\mathrm{rkl}}(z) = \frac{\max\{\bar{q}_{i,t}^{\mathrm{mv,rkl}}(z), \epsilon_{\mathrm{tar}}\}}{\sum_{u \in \Omega_{i,t}} \max\{\bar{q}_{i,t}^{\mathrm{mv,rkl}}(u), \epsilon_{\mathrm{tar}}\}}, \tag{16}$$

where $\epsilon_{\mathrm{tar}} = 10^{-8}$. The target $\widehat{q}_{i,t}^{\mathrm{rkl}}$ is detached, while gradients flow through the student distribution $\bar{p}_{i,t}^{\mathrm{rkl}}$, including its tail mass. The probability floor is applied only to the teacher target. Thus, the two optimization branches use the same fusion rule but construct their fused targets separately, using their respective candidate sets and probability representations.

C.2 Trajectory Verification

We distinguish three decisions: outcome verification assigns a trajectory label, data validation determines whether the training example is usable, and correction selection determines whether to apply failure-conditioned distillation.

Outcome verification. Unlike sequence-level rewards, which provide graded feedback for policy optimization, verification assigns Success, Failure, or Indeterminate. For supported tasks, a trajectory is labeled Success if all required components pass, Failure if any component explicitly fails, and Indeterminate otherwise. Table 7 summarizes the task-specific criteria and outcome-based eligibility for reverse-KL (RKL) distillation.

Table 7: Task-specific outcome verification and RKL eligibility under the strict setting.

Task	**Success**	**Failure**	**RKL eligibility**
Video MCQ	Correct option	Incorrect option	Incorrect option
Temporal localization	$\mathrm{tIoU} \geq 0.5$	$\mathrm{tIoU} = 0$	$\mathrm{tIoU} = 0$
Temporal MCQ	Correct option **and** $\mathrm{tIoU} \geq 0.5$	Incorrect option **or** $\mathrm{tIoU} = 0$	Incorrect option **and** $\mathrm{tIoU} = 0$
Spatial localization	Not assigned	Not assigned	Not eligible
Other composite or free-form tasks	Not assigned	Not assigned	Not eligible

The verifier requires exactly one nonempty `<answer>` block. Multiple-choice outputs must contain a parseable option from A–H. Temporal outputs must contain exactly one interval in the expected timestamp format, with finite endpoints and an end time strictly greater than the start time. For temporal multiple-choice QA, the option and interval are extracted separately from the `Correct Option:` field and the answer block. An unparseable component satisfies neither the success nor the failure criterion; another explicitly incorrect component can still establish Failure. Missing or malformed intervals are never treated as zero-overlap predictions. Tasks without explicit verification rules retain their task-specific rewards for policy optimization.

Data validity. Malformed model outputs are distinct from invalid training data. The former may remain valid trajectories with uncertain outcomes, whereas unparseable ground-truth options, invalid ground-truth intervals, and visual loading failures are flagged as invalid. Visual loading failures additionally receive a skip flag. Any Indeterminate label attached to an invalid row is a placeholder, not an uncertain prediction on valid data.

Required reward metadata is also validated. Spatial localization requires annotation and processed image sizes. Temporal–spatial free-form QA requires frame sizes, nonempty keyframe annotations, an object annotation mapping, and training progress for reward scheduling. Under strict validation, missing or malformed required metadata raises an error rather than silently yielding a zero reward.

Correction selection. A Failure label alone does not guarantee RKL eligibility. In particular, temporal multiple-choice QA requires both an incorrect option and a valid interval with zero tIoU under the strict setting. Responses with only one incorrect component are therefore labeled Failure but excluded from correction.

Let v_i indicate data validity, e_i denote outcome-based eligibility from Table 7, and h_i indicate whether the teacher evidence required to construct a correction target is available. During teacher supervision, the correction indicator is

$$g_i = v_i e_i h_i. \tag{17}$$

Evidence availability is checked independently of the verification outcome.

For valid trajectories, $g_i = 1$ selects failure-conditioned distillation in place of the policy loss, whereas $g_i = 0$ retains token-level advantage refinement. This fallback includes Success and Indeterminate trajectories, failures excluded by the strict eligibility rule, and otherwise eligible failures lacking the required teacher evidence. If no usable teacher signal is available for refinement, the original reward-derived advantage is retained.

C.3 Analysis of Credit Refinement and Failure Correction

We examine the complementary roles of the refined advantage $A^{\mathrm{cr}}_{i,t}$ and the failure-conditioned loss ℓ^{rkl}_i. Credit refinement adjusts reward-derived token credit, whereas failure correction provides a distribution-level learning signal for eligible verified failures.

Sign preservation and bounded refinement. The clipped token weight lies in $[1-\epsilon_w, 1+\epsilon_w]$. Thus, for $A_i \neq 0$, the refined advantage satisfies

$$1 - \lambda_s \epsilon_w \leq \frac{A^{\mathrm{cr}}_{i,t}}{A_i} \leq 1 + \lambda_s \epsilon_w. \tag{18}$$

For $\lambda_s \in [0,1]$ and $\epsilon_w \in [0,1)$, the lower bound is strictly positive, yielding

$$\begin{aligned} \operatorname{sign}(A^{\mathrm{cr}}_{i,t}) &= \operatorname{sign}(A_i), \\ \left|A^{\mathrm{cr}}_{i,t} - A_i\right| &\leq \lambda_s \epsilon_w |A_i|. \end{aligned} \tag{19}$$

When $A_i = 0$, refinement gives $A^{\mathrm{cr}}_{i,t} = 0$, so both properties in Eq. equation 19 still hold. These results establish sign preservation and a bounded change in the advantage coefficients; they do not imply that the aggregate parameter-gradient direction is unchanged. With the initial settings $\lambda_s = 0.5$ and $\epsilon_w = 0.2$, the ratio $A^{\mathrm{cr}}_{i,t}/A_i$ lies in $[0.9, 1.1]$ for nonzero A_i. This interval contracts to $\{1\}$ as λ_s approaches zero. The teacher-derived refinement factors are detached throughout optimization.

Correction beyond multiplicative refinement. Identical rewards within a rollout group yield zero group-relative advantages. Multiplicative refinement then gives $A^{\mathrm{cr}}_{i,t} = 0$ regardless of $\Delta_{i,t}$, so the corresponding credit-refined policy-loss terms provide no gradient signal. In contrast, ℓ^{rkl}_i is independent of A_i.

To characterize this distinction, write the per-token reverse-KL loss as

$$\begin{aligned} \ell^{\mathrm{rkl}}_{i,t} &= D_{\mathrm{KL}}\left(\bar{p}^{\mathrm{rkl}}_{i,t} \,\middle\|\, \operatorname{sg}[\widehat{q}^{\mathrm{rkl}}_{i,t}]\right), \\ \ell^{\mathrm{rkl}}_i &= \frac{1}{T_i} \sum_{t=1}^{T_i} \ell^{\mathrm{rkl}}_{i,t}. \end{aligned} \tag{20}$$

For this analysis, hold the coarse alphabet $\Omega_{i,t}$ and the detached teacher target fixed. Parameterize a strictly positive coarse student distribution using auxiliary logits $\eta_{i,t}(z)$:

$$\bar{p}^{\mathrm{rkl}}_{i,t}(z) = \frac{\exp(\eta_{i,t}(z))}{\sum_{u \in \Omega_{i,t}} \exp(\eta_{i,t}(u))}. \tag{21}$$

Differentiating with respect to these logits gives

$$\frac{\partial \ell^{\mathrm{rkl}}_{i,t}}{\partial \eta_{i,t}(z)} = \bar{p}^{\mathrm{rkl}}_{i,t}(z) \left[\log \frac{\bar{p}^{\mathrm{rkl}}_{i,t}(z)}{\widehat{q}^{\mathrm{rkl}}_{i,t}(z)} - \ell^{\mathrm{rkl}}_{i,t}\right]. \tag{22}$$

This gradient does not depend on A_i. Since all student probabilities are positive, the gradient vanishes in every coordinate if and only if the log-probability ratio is constant across $\Omega_{i,t}$. Normalization then implies $\bar{p}^{\mathrm{rkl}}_{i,t} = \widehat{q}^{\mathrm{rkl}}_{i,t}$. Thus, for eligible verified failures, distribution mismatch can supply a correction signal even when credit refinement provides none. This conclusion concerns the distribution-level objective; it does not guarantee a correct answer or a nonzero aggregate gradient with respect to the model parameters.

Selective correction and return to GRPO. For valid trajectories, g_i selects the optimization branch: $g_i = 1$ replaces ℓ^{cr}_i with $\gamma_s \ell^{\mathrm{rkl}}_i$, whereas $g_i = 0$ retains credit refinement. The latter includes failures excluded by the strict rules in Section C.2. When no usable teacher signal is available, $\Delta_{i,t} = 0$, so $w_{i,t} = 1$ and $A^{\mathrm{cr}}_{i,t} = A_i$. After teacher supervision is disabled, setting $\lambda_s = 0$ and $g_i = 0$ for all trajectories reduces $\mathcal{L}_s(\theta)$ to standard GRPO.

D Evaluation Details

Our evaluation examines whether multi-evidence self-distillation improves answer correctness and spatio-temporal localization, and whether these gains extend to broader video understanding tasks. We describe the benchmarks, comparison methods, and evaluation metrics below.

D.1 BENCHMARKS

Joint spatiotemporal reasoning. V-STaR (Cheng et al., 2026) jointly evaluates question answering and the localization of supporting temporal and spatial evidence. It contains two reasoning chains: what–when–where and what–where–when. We report results for both chains to assess MeSD under different evidence prediction orders and distinguish gains in answer accuracy from those in temporal and spatial grounding.

Temporal grounding. We use TVGBench (Wang et al., 2025b) and Charades-STA (Gao et al., 2017) to evaluate temporal localization beyond joint reasoning. TVGBench contains temporal grounding examples from multiple source datasets, while Charades-STA requires identifying the video interval described by a natural-language query. These benchmarks complement V-STaR by directly assessing how well predicted intervals align with annotated events.

General and long-video understanding. We use WorldSense (Hong et al., 2026), VideoMMMU (Hu et al., 2025), and LongVideo-Reason-eval (LRR) (Chen et al., 2025) to assess whether evidence-based training also improves video question answering. WorldSense covers recognition, understanding, and reasoning. VideoMMMU evaluates knowledge acquisition from professional videos through perception, comprehension, and adaptation, while LRR tests reasoning over long videos. We also evaluate on Video-MME-v2 (Fu et al., 2026), which uses groups of related questions to assess capability consistency and reasoning coherence. Together, these benchmarks assess whether MeSD's benefits extend beyond explicit grounding.

D.2 BASELINES

General and video-specialized models. We include Qwen2.5-VL-7B (Bai et al., 2025b), InternVL-2.5-8B (Chen et al., 2024), and Oryx-1.5-7B (Liu et al., 2025), alongside video-specialized models such as VideoChat2 (Li et al., 2024), Video-LLaMA3 (Zhang et al., 2025a), LLaVA-Video (Zhang et al., 2025b), and TRACE (Guo et al., 2025d). We also report results for GPT-4o (OpenAI, 2024) and Gemini-2-Flash (Google DeepMind, 2025) as proprietary references where available. For Charades-STA, we include temporal grounding methods such as TimeChat (Ren et al., 2024), HawkEye (Wang et al., 2024), VTG-LLM (Guo et al., 2025c), and VTimeLLM (Huang et al., 2024). These comparisons position MeSD relative to a broad range of video understanding and localization methods.

Video post-training and self-distillation. Our main post-training baselines include VideoRFT (Wang et al., 2025a), Video-R1 (Feng et al., 2025), Open-o3-Video (Meng et al., 2025), Vision-Coach (Lee et al., 2026), and VISD (Lin et al., 2026). Our training setup builds on Open-o3-Video's grounded video reasoning framework. VISD serves as a closely related self-distillation baseline, using teacher-derived signals to refine token-level credit while preserving reward-derived update directions. Comparing MeSD with VISD assesses the added value of separating evidence-conditioned teacher preferences and introducing distribution-level correction for eligible, verified failures.

Optimization and component comparisons. We also include GRPO and GSPO as reward-based optimization baselines. Component ablations individually assess evidence separation, temporal and spatial supervision, gated preference aggregation, and failure-conditioned distillation. These comparisons complement the model-level results by identifying the contributions of individual MeSD components.

D.3 EVALUATION METRICS

Answer accuracy. We report answer accuracy for WorldSense, VideoMMMU, and LRR, together with the category-level results presented in the main tables. V-STaR answer correctness is assessed using the Qwen2.5-72B-Instruct evaluator in our evaluation pipeline. Answer accuracy measures whether the final response is correct, while the grounding metrics below assess the predicted evidence separately.

Temporal and spatial overlap. For a predicted interval I_p and a reference interval I_g, temporal intersection over union is defined as

$$\mathrm{tIoU}(I_p, I_g) = \frac{|I_p \cap I_g|}{|I_p \cup I_g|}. \tag{23}$$

We report mean IoU on TVGBench and Charades-STA. For Charades-STA, we additionally report $R@0.3$, $R@0.5$, and $R@0.7$, measuring the fraction of queries whose predicted interval reaches the corresponding IoU threshold. Spatial grounding on V-STaR is measured using bounding-box IoU with the matching and aggregation rules of its evaluation pipeline.

Joint reasoning scores. For V-STaR, let $\mathcal{C}$ denote the set of evaluated reasoning chains and $\mathcal{D}$ the set of evaluation dimensions: answer accuracy, temporal IoU, and spatial IoU. Let $m_{c,d}$ denote the component score for dimension $d \in \mathcal{D}$ in chain $c \in \mathcal{C}$, normalized to $[0, 1]$. Answer accuracy is shared across chains, whereas temporal and spatial IoU are chain-specific. We compute

$$\mathrm{mAM} = \frac{1}{|\mathcal{C}|\,|\mathcal{D}|} \sum_{c\in\mathcal{C}} \sum_{d\in\mathcal{D}} m_{c,d}, \tag{24}$$

$$\mathrm{mLGM} = -\frac{1}{|\mathcal{C}|\,|\mathcal{D}|} \sum_{c\in\mathcal{C}} \sum_{d\in\mathcal{D}} \log(1 - m_{c,d}). \tag{25}$$

We report the individual dimensions alongside these aggregates to distinguish answer correctness from temporal and spatial grounding. Both aggregates are multiplied by 100 for display. mLGM is a logarithmic score rather than a percentage accuracy and is not bounded by 100.

Grouped video understanding scores. For Video-MME-v2, we report the grouped non-linear score, Level 1–3 scores, capability consistency, reasoning coherence, and average question-level accuracy. The grouped scores account for correctness across related questions and their dependency structure, providing information beyond isolated answer accuracy.

Cross-benchmark average. The average reported for WorldSense, VideoMMMU, LRR, and TVG-Bench is the unweighted arithmetic mean of WorldSense overall accuracy, VideoMMMU overall accuracy, LRR accuracy, and TVGBench mIoU. All four scores use the 0–100 scale. Category-level scores are excluded from this average to avoid counting the same benchmark multiple times.

F Additional Experimental Results

F.1 Video-MME-v2 Results

Table 8: Video-MME-v2 results with 64 sampled frames and no subtitles. Bold and underlining mark the best and second-best scores; Δ gives the gain over Qwen2.5-VL-7B.

Model	Non-Lin Score	Level 1	Level 2	Level 3	Consistency	Coherence	Avg Acc
Qwen2.5-VL-7B (*Baiet al.*, 2025*b*)	7.9	11.4	6.7	6.7	8.2	7.3	21.3
VideoRFT-7B (*Wanget al.*, 2025*a*)	8.2	11.8	7.4	6.8	8.8	7.2	21.7
Video-R1-7B (*Fenget al.*, 2025)	7.7	11.1	6.7	6.4	8.0	7.2	20.7
Open-o3-Video-7B* (Meng et al., 2025)	7.8	11.0	6.9	6.6	8.3	6.9	21.2
VISD-7B (Lin et al., 2026)	<u>9.4</u>	<u>13.4</u>	<u>8.2</u>	<u>7.9</u>	<u>10.2</u>	<u>8.0</u>	<u>23.8</u>
MeSD-7B (Ours)	**12.7**	**18.7**	**10.0**	**11.2**	**12.2**	**13.7**	**28.4**
Δ vs. Qwen-VL	+4.8	+7.3	+3.3	+4.5	+4.0	+6.4	+7.1

On Video-MME-v2, MeSD reaches 28.4 average accuracy, exceeding VISD by 4.6 points and Qwen2.5-VL-7B by 7.1 points. Its gains also extend to the non-linear score and each reported component, indicating stronger performance under the grouped-question evaluation.

F.2 Qwen3-VL Detailed Results

Tables 9 and 10 compare each MeSD model with its base Qwen3-VL counterpart at 4B, 8B, and 32B parameters. They report both spatiotemporal grounding and the four-benchmark average.

Table 9: V-STAR results across Qwen3-VL model sizes. Chain1 follows what–when–where and Chain2 follows what–where–when. Bold marks the better score within each base/MeSD pair; Δ is the MeSD–base point difference.

Model	What	When (temporal IoU)		Where (spatial IoU)		Overall	
	Acc.	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Qwen3-VL-4B (Bai et al., 2025a)	22.7	16.8	16.2	15.9	2.8	16.2	18.0
Qwen3-VL-MeSD-4B	**62.1**	**23.1**	**23.8**	**22.5**	**6.4**	**33.3**	**46.6**
Δ vs. Qwen3-VL-4B	+39.4	+6.3	+7.6	+6.6	+3.6	+17.2	+28.6
Qwen3-VL-8B (Bai et al., 2025a)	31.9	21.7	21.1	5.5	0.7	18.8	21.9
Qwen3-VL-MeSD-8B	**60.0**	**25.5**	**25.9**	**25.4**	**6.3**	**33.8**	**46.4**
Δ vs. Qwen3-VL-8B	+28.1	+3.8	+4.8	+19.9	+5.6	+15.0	+24.5
Qwen3-VL-32B (Bai et al., 2025a)	47.5	24.5	25.7	30.9	7.5	30.6	38.6
Qwen3-VL-MeSD-32B	**66.0**	**29.2**	**29.4**	**31.1**	**11.4**	**38.9**	**55.8**
Δ vs. Qwen3-VL-32B	+18.5	+4.7	+3.7	+0.2	+3.9	+8.3	+17.2

Table 10: Results across Qwen3-VL model sizes on WorldSense, VideoMMMU, LRR, and TVG-Bench. Bold marks the better score within each pair, including ties; Δ is the MeSD–base difference. Avg and Δ are computed before rounding.

Model	WorldSense				VideoMMMU				LRR	TVGBench	Avg
	Overall	Recog.	Underst.	Reason.	Overall	Percep.	Adapt.	Compreh.	Acc.	mIoU	
Qwen3-VL-4B (Bai et al., 2025a)	39.7	39.4	40.9	38.9	55.2	76.7	38.7	50.3	69.3	13.7	44.5
Qwen3-VL-MeSD-4B	**41.8**	**42.7**	**41.0**	**41.1**	**59.7**	**78.7**	**45.3**	**55.0**	**75.2**	**29.9**	**51.7**
Δ vs. Qwen3-VL-4B	+2.1	+3.3	+0.1	+2.2	+4.5	+2.0	+6.6	+4.7	+5.9	+16.2	+7.2
Qwen3-VL-8B (Bai et al., 2025a)	41.4	40.1	45.5	38.9	61.3	76.7	**48.7**	58.7	70.2	17.1	47.5
Qwen3-VL-MeSD-8B	**44.1**	**42.8**	**48.1**	**41.9**	**62.3**	**79.0**	47.7	**60.3**	**76.1**	**30.2**	**53.2**
Δ vs. Qwen3-VL-8B	+2.7	+2.7	+2.6	+3.0	+1.0	+2.3	−1.0	+1.6	+5.9	+13.1	+5.7
Qwen3-VL-32B (Bai et al., 2025a)	44.6	44.4	**48.0**	**41.4**	65.2	84.0	51.3	60.3	74.4	31.2	53.9
Qwen3-VL-MeSD-32B	**46.1**	**47.7**	47.9	**41.4**	**72.8**	**85.7**	**58.7**	**74.0**	**79.4**	**39.7**	**59.5**
Δ vs. Qwen3-VL-32B	+1.5	+3.3	−0.1	0.0	+7.6	+1.7	+7.4	+13.7	+5.0	+8.5	+5.7

Across the 4B, 8B, and 32B Qwen3-VL models, MeSD raises V-STAR mLGM by 28.6, 24.5, and 17.2 points, respectively. The corresponding four-benchmark averages rise by 7.2, 5.7, and 5.7 points. TVGBench mIoU improves at every size, although the 8B model loses 1.0 point on VideoMMMU Adaptation and the 32B model loses 0.1 point on WorldSense Understanding. The overall gains therefore coexist with small category-level trade-offs.

F.3 Effects of Correction Selection.

Table 11 compares correction selection strategies while keeping the evidence-conditioned teachers and gated preference aggregation fixed. Applying token-level advantage refinement to all trajectories without FCD reduces V-STAR mLGM from 50.2 to 41.0 and TVGBench mIoU from 29.8 to 23.0. These results support the contribution of failure-conditioned distillation beyond advantage refinement. To examine the role of verification-guided selection, the random-selection variant uniformly selects valid trajectories with the required teacher evidence, regardless of verification outcome. Within each supported task and batch, the number selected matches the number eligible for correction under its current rollouts, while the remaining trajectories retain token-level advantage refinement. This comparison tests whether selecting verified failures provides more effective corrective guidance than random selection.

Table 11: Effect of correction selection. CR denotes credit refinement.

Variant	Optimization Routing		V-STAR		WorldSense		Video-MMMU		LRR	TVGBench
	RKL Selection	Other Trajectories	mAM	mLGM	Overall	Understanding	Overall	Perception	Acc.	mIoU
w/o FCD	None	CR (all)	30.7	41.0	38.6	38.7	53.8	68.3	71.3	23.0
Random	Random subset	CR (unselected)	33.1	45.6	40.1	42.4	55.0	70.7	72.5	26.4
MeSD	Verified failures (strict)	CR (unselected)	**35.5**	**50.2**	**41.6**	**46.0**	**56.1**	**73.0**	**73.6**	**29.8**

F.4 Training Efficiency.

To improve training efficiency, we adopt vLLM as the inference backend during training. As shown in Table 12, at the same training budget of 2,327 steps, MeSD reduces training time from 61.36 to 20.96 hours compared with VISD no-feedback, achieving a $2.93\times$ speedup. The corresponding compute cost decreases from 981.70 to 335.38 A800 GPU-hours, a reduction of 65.8%.

Table 12: Training efficiency comparison at the same number of training steps.

Method	Steps	Time (h)	A800 GPU-hours
VISD no-feedback	2,327	61.36	981.70
MeSD	2,327	**20.96**	**335.38**

F.5 Detailed Comparisons and Ablations

Tables 13 and 14 expand the main-text comparison into seven V-STAR metrics and the category-level WorldSense and VideoMMMU scores, alongside LRR and TVGBench. Chain1 follows what–when–where, and Chain2 follows what–where–when. Scores are reported to one decimal place; "–" denotes an unreported result. Bold and underlining mark the best and second-best scores among the displayed models on V-STAR and among open-source models elsewhere.

MeSD attains the highest V-STAR mAM (35.5) and mLGM (50.2), compared with 35.1 and 48.9 for VISD. VISD remains stronger on both temporal IoU chains (26.8 and 27.2 versus 25.8 and 25.0). On the other four benchmarks, MeSD improves over VISD in WorldSense overall score (41.6 versus 39.9), VideoMMMU overall score (56.1 versus 53.8), LRR accuracy (73.6 versus 73.5), and TVGBench mIoU (29.8 versus 24.8).

Table 13: Detailed V-STAR comparison.

Model	What	When (tIoU)		Where (sIoU)		Overall	
	Acc.	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Gemini-2-Flash (Google DeepMind, 2025)	53.0	24.5	23.8	4.6	2.2	26.9	35.6
GPT-4o (OpenAI, 2024)	60.8	16.7	12.8	6.5	3.0	26.8	38.2
Oryx-1.5-7B (Liu et al., 2025)	20.5	13.5	14.8	10.1	3.5	15.1	13.8
InternVL-2.5-8B (Chen et al., 2024)	44.2	8.7	7.8	0.7	0.1	17.6	24.9
Qwen2.5-VL-7B (Bai et al., 2025b)	33.5	15.4	13.8	17.0	2.5	19.3	22.4
TRACE (Guo et al., 2025d)	17.6	19.1	17.1	0.0	0.0	12.0	13.3
VideoLLaMA3-7B (Zhang et al., 2025a)	41.9	23.0	23.1	0.9	0.2	21.7	27.0
LLaVA-Video (Zhang et al., 2025b)	49.5	10.5	12.2	1.9	1.3	20.8	27.3
VideoChat2 (Li et al., 2024)	36.2	13.7	12.5	2.5	1.0	17.0	20.3
Open-o3-Video-7B (Meng et al., 2025)	61.0	24.5	24.0	25.4	6.0	33.7	46.6
VISD-7B (Lin et al., 2026)	<u>61.9</u>	**26.8**	**27.2**	<u>26.7</u>	<u>6.3</u>	<u>35.1</u>	<u>48.9</u>
MeSD-7B (Ours)	**63.7**	<u>25.8</u>	<u>25.0</u>	**28.1**	**6.5**	**35.5**	**50.2**

Table 14: Detailed comparison on WorldSense, VideoMMMU, LRR, and TVGBench.

Model	WorldSense			
	Overall	Recognition	Understanding	Reasoning
GPT-4o	42.6	–	–	–
Qwen2.5-VL-7B	35.5	31.8	38.4	38.7
Open-o3-Video-7B	38.9	36.8	39.7	40.5
VISD-7B	39.9	**39.4**	40.4	40.5
MeSD-7B (Ours)	**41.6**	38.9	**46.0**	**42.3**

Model	VideoMMMU				LRR	TVGBench
	Overall	Perception	Adaptation	Comprehension	Accuracy	mIoU
GPT-4o	61.2	66.0	–	–	–	–
Qwen2.5-VL-7B	51.2	64.7	38.7	48.7	59.3	16.3
Open-o3-Video-7B	52.6	67.4	41.3	49.1	69.1	19.2
VISD-7B	53.8	69.3	42.3	49.7	73.5	24.8
MeSD-7B (Ours)	**56.1**	**73.0**	**44.3**	**51.3**	**73.6**	**29.8**

Component and training-strategy ablations.

Tables 15 and 16 report the full component and training-strategy ablations. The variant names follow Table 4: w/o FCD removes failure-conditioned distillation; w/o ES uses one teacher with all evidence; w/o T&S keeps answer evidence only; w/o GPA averages teacher preferences; and w/o VGO replaces verification-guided optimization with multi-evidence OPSD. Bold marks the best score in each column.

Every ablated variant scores below MeSD on V-STAR mLGM and TVGBench mIoU. Removing failure-conditioned distillation reduces these scores from 50.2 to 41.0 and from 29.8 to 23.0, respectively. Replacing verification-guided optimization with OPSD yields a TVGBench mIoU of 20.7. These declines support the contribution of targeted correction and outcome-aware training, although the w/o ES variant slightly exceeds MeSD on VideoMMMU Adaptation (44.7 versus 44.3).

Table 15: Detailed V-STAR ablations.

Model	What	When (tIoU)		Where (sIoU)		Overall	
	Acc.	Chain1	Chain2	Chain1	Chain2	mAM	mLGM
Component Variants							
w/o FCD	56.4	22.0	20.9	23.7	5.0	30.7	41.0
w/o ES	56.5	22.1	21.7	23.4	5.2	30.9	41.3
w/o T&S	56.4	22.3	21.4	23.5	5.1	30.8	41.2
w/o GPA	56.5	22.3	21.7	23.6	5.1	30.9	41.4
Training Strategy Variants							
Base VLM	33.5	15.4	13.8	17.0	2.5	19.3	22.4
GSPO	61.0	24.5	24.0	25.4	6.0	33.7	46.6
GRPO	55.7	21.9	21.5	23.6	5.3	30.6	40.7
w/o VGO	56.6	22.4	19.4	23.2	5.2	30.6	40.9
MeSD	**63.7**	**25.8**	**25.0**	**28.1**	**6.5**	**35.5**	**50.2**

Table 16: Detailed ablations on WorldSense, VideoMMMU, LRR, and TVGBench.

Model	WorldSense			
	Overall	Recognition	Understanding	Reasoning
Component Variants				
w/o FCD	38.6	38.2	38.7	39.0
w/o ES	38.3	38.4	38.4	38.1
w/o T&S	37.9	37.3	39.3	37.3
w/o GPA	39.5	37.9	39.5	38.6
Training Strategy Variants				
Base VLM	35.5	31.8	38.4	38.7
GSPO	37.5	36.8	39.7	40.5
GRPO	39.2	38.7	40.4	38.8
w/o VGO	39.5	38.8	41.0	38.9
MeSD	**41.6**	**38.9**	**46.0**	**42.3**

Model	VideoMMMU				LRR	TVGBench
	Overall	Perception	Adaptation	Comprehension	Accuracy	mIoU
Component Variants						
w/o FCD	53.8	68.3	43.0	50.0	71.3	23.0
w/o ES	55.4	71.0	**44.7**	50.7	70.2	22.4
w/o T&S	54.8	71.0	42.0	**51.3**	70.6	22.3
w/o GPA	53.9	69.7	43.0	49.0	70.4	22.9
Training Strategy Variants						
Base VLM	51.2	64.7	38.7	48.7	59.3	16.3
GSPO	52.3	68.0	41.3	49.1	69.4	20.8
GRPO	53.2	69.3	40.0	50.3	71.9	22.5
w/o VGO	54.1	67.7	44.3	50.3	71.2	20.7
MeSD	**56.1**	**73.0**	44.3	**51.3**	**73.6**	**29.8**

F.6 Training Curves

Reward rises quickly and then levels off at all three Qwen3-VL sizes (Figure 8). MeSD also maintains higher reward than the w/o ES, w/o T&S, and w/o GPA variants (Figure 9), consistent with the ablation results in Tables 15 and 16.

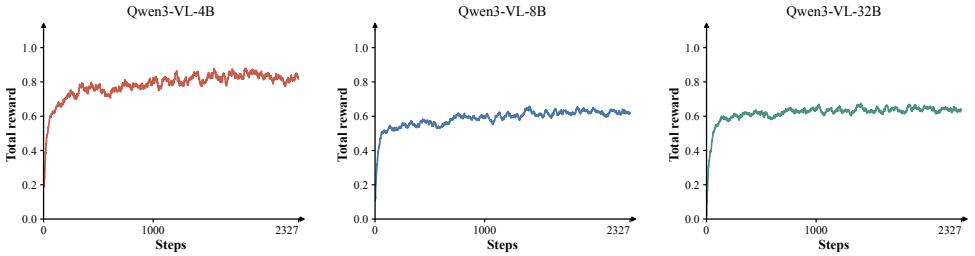


Figure 8: Training reward for Qwen3-VL models at 4B, 8B, and 32B.

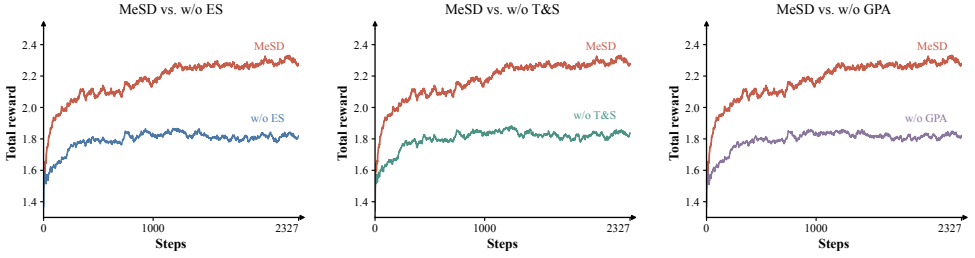


Figure 9: Training reward for MeSD and the w/o ES, w/o T&S, and w/o GPA component variants.

G CASE STUDY

Object–action grounding. Figure 10 shows MeSD identifying the baby as the actor and the toy as the object being bitten. The generated reasoning describes the action at multiple timestamps, and the final answer is consistent with the ground truth. Separate references to the baby and the toy link the textual description to model-predicted regions, allowing readers to inspect their correspondence in the sampled frames.



MeSD *Question:* who bites the toy in the room?
Ground Truth: a baby bites a toy in the room

<think>
The video shows a <obj>baby</obj><box>[173,73,637,448]</box>at<t>1.5</t>s lying on its back in a bouncer,
holding an orange <obj>toy ball</obj><box>[280,96,408,218]</box>at<t>1.5</t>s.
The baby brings the <obj>toy ball</obj><box>[280,96,408,218]</box>at<t>3.0</t>s to its mouth and starts biting it.
This action continues as the <obj>baby</obj><box>[173,73,637,448]</box>at<t>5.9</t>s is still seen chewing on the
<obj>toy ball</obj><box>[280,96,408,218]</box>at<t>5.9</t>s.
Towards the end of the sequence, the <obj>baby</obj><box>[173,73,637,448]</box>at<t>10.4</t>s is still holding the
<obj>toy ball</obj><box>[130,96,378,231]</box>at<t>10.4</t>s while lying down.
</think>

<answer>A baby bites the toy in the room.</answer>

Figure 10: Object–action grounding. MeSD identifies the baby biting the toy and includes model-predicted boxes and timestamps in its reasoning.

Person–interaction grounding. Figure 11 shows MeSD distinguishing the man from the child and describing how he holds the child at 2.4 s and 9.8 s. The final answer identifies the adult in white and is semantically consistent with the ground truth. Explicit references to both participants clarify their respective roles and provide spatial and temporal cues for assessing the described interaction.



MeSD Question: who hugs the child?
Ground truth answer: an adult in white clothes hugs a child.

<think>
The video shows a <obj>man</obj><box>[46,13,336,280]</box>at<t>2.4</t>s
holding a <obj>child</obj><box>[222,43,364,280]</box>at<t>2.4</t>s while they look at an underwater mural.
Throughout the segment, the <obj>man</obj><box>[46,13,336,280]</box>at<t>9.8</t>s continues to hold the
<obj>child</obj><box>[222,43,364,280]</box>at<t>9.8</t>s,
providing support as they walk past the mural.
</think>

<answer>A man in a white shirt hugs the child.</answer>

Figure 11: Person–interaction grounding. MeSD identifies the man holding the child, with predicted regions corresponding to the two participants referenced in its reasoning.

Answer correctness and temporal grounding. Figure 12 shows that MeSD produces an answer accepted by the evaluator for the identity question, while temporal localization remains less precise. Both reasoning chains predict 19–21 s, compared with the annotated interval of 0–5.6 s, yielding zero temporal IoU. This contrast highlights the complementary value of evaluating answer correctness and temporal grounding separately. MeSD's explicit temporal predictions make this discrepancy directly inspectable, providing a concrete target for further refinement.



Q1: who holds the green sports ball in the big room?
Ground truth answer: a child holds a green sports ball in a big room.
<think>
The video shows a <obj>green sports ball</obj><box>[596,188,644,248]</box>at<t>21.2</t>s on the floor of a large room filled with children and adults. A <obj>girl</obj><box>[463,160,578,336]</box>at<t>21.2</t>s is seen holding the <obj>green sports ball</obj><box>[596,188,644,248]</box>at<t>21.2</t>s while sitting on the floor among other children.</think>
<answer>A girl sits on the floor holding the green sports ball.</answer>
Result: Correct √

Q2: When does the child hold the green sports ball in the big room?
Predicted interval (both chains): [19.0, 21.0] s
Result: Incorrect ✗ (tIoU = 0)

Figure 12: Temporal grounding failure. MeSD's answer to the identity question is judged correct, but both reasoning chains predict a temporal interval that does not overlap with the ground truth.

H Limitations and Future Work

While MeSD effectively integrates answer, temporal, and spatial annotations, the current study explores only a subset of the evidence that could support video reasoning. Its multi-evidence formulation offers a natural basis for incorporating richer supervision, but the potential benefits of additional annotation types remain unexplored. The present evaluation therefore leaves open how greater annotation diversity and finer granularity could further improve self-distillation.

Future work could introduce complementary annotations, such as object trajectories and event relations, to construct additional evidence-conditioned teacher views. Extending the fusion mechanism to account for differences in evidence granularity and reliability could help integrate these sources effectively. Systematically examining the contributions of different annotation types and their combinations would further clarify how richer supervision supports video reasoning. Such extensions may provide more detailed guidance for complex temporal and spatial reasoning and broaden the applicability of MeSD.