VIDEO-SR1: REINFORCEMENT LEARNING WITH SELF-REWARDING VIDEO-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models increasingly support video, but reinforcement learning (RL) for video reasoning still relies largely on sparse final-answer rewards. Such supervision is poorly suited for long video-reasoning traces, where correct observations and localized perceptual or temporal errors may coexist within the same response. We introduce VIDEO-SR1, a self-rewarding framework for dense RL post-training of video-language models. Its core algorithm, *Group and Sentence Relative Policy Optimization* (GSRPO), uses the policy to apply self-rewards across each sentence in its own reasoning trace, aligns semantically corresponding sentences across responses, and performs group-relative credit assignment at the sentence level. Further, GSRPO complements the self-rewards with a counterfactual temporal-sensitivity reward that encourages reasoning to respond to changes in video frame order. Across six diverse video-reasoning benchmarks, VIDEO-SR1 significantly improves accuracy, sample efficiency, and cross-domain transfer relative to strong baselines in settings with and without the sparse ground-truth answer reward included during RL. Error analyses and ablations reveal two complementary factors underlying VIDEO-SR1's gains: the temporal-sensitivity reward promotes reasoning that responds to the video's spatiotemporal structure, while local self-rewards ensure the temporally-sensitive reasoning remains accurate and grounded in the visual evidence. We will release all code, models, and data at `https://video-sr1.github.io/`.

1 INTRODUCTION

Modern vision-language models (VLMs) have become strong single-image reasoners, but video remains substantially more challenging. Answering a video question may require tracking entities through occlusion, integrating evidence across distant frames, separating camera motion from object motion, or resolving the order and duration of events. These demands make video an especially difficult setting for post-training: successful reasoning depends not only on producing the correct answer, but also on grounding intermediate reasoning in the evolving spatiotemporal evidence.

Reinforcement learning (RL) provides a promising way to strengthen such reasoning. Recent work such as Video-R1 adapts R1-style post-training to video and introduces temporal-aware objectives for multimodal reasoning (Feng et al., 2025). However, RL for video reasoning still relies on sparse outcome supervision, where all tokens generated during reasoning receive the same reward based on the accuracy of the final answer (Shao et al., 2024; Feng et al., 2025). For a long video rationale, this uniform reward conflates distinct sources of error in perception, temporal integration, logical inference, and final answer formation. A wrong final answer caused by a localized temporal mistake results in penalizing all tokens in an otherwise sound reasoning trace, while, in other cases, unsupported intermediate claims can still be reinforced if the final answer happens to be correct.

Earlier actor–critic methods such as PPO can provide finer-grained credit through learned value functions, but in language-model RL the critic adds substantial memory and computation and is difficult to train accurately from sparse terminal rewards (Shao et al., 2024). Group Relative Policy Optimization (GRPO) avoids this cost by replacing the critic with a group-relative Monte Carlo baseline (Shao et al., 2024), but still applies a single outcome advantage across the full response. Recent single-image methods instead derive denser supervision from the VLM itself: VL-Calibration uses visual perturbation sensitivity and token entropy to estimate intrinsic certainty (Xiao et al., 2026), while Vision-SR1 derives a self-reward from whether the model's visual representation is sufficient to support subsequent reasoning (Li et al., 2026). These approaches show that self-derived signals can improve credit assignment beyond final-answer correctness, but they are designed for static images and still lack the reward granularity needed to disentangle localized perceptual, temporal, and reasoning errors within video reasoning traces.

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

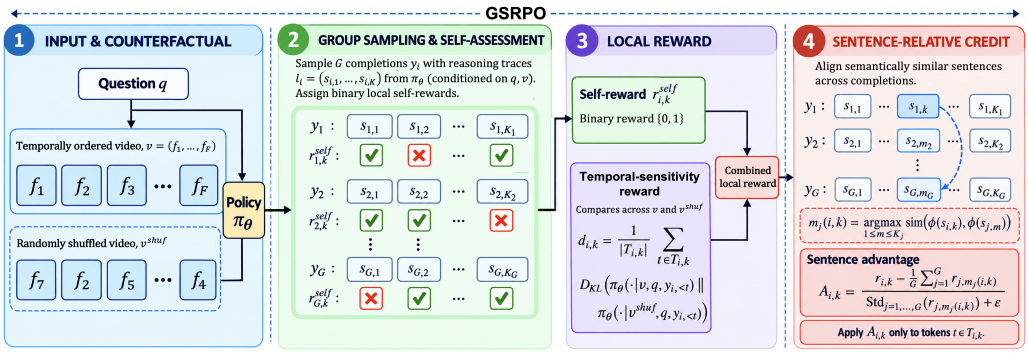

Figure 1: Overview of Group and Sentence Relative Policy Optimization (GSRPO).

We introduce **VIDEO-SR1**, a self-rewarding framework for video-language reasoning that provides dense supervision without training a separate value model. Its core algorithm, *Group and Sentence Relative Policy Optimization* (GSRPO), leverages the policy to assess the local correctness of each sentence in the reasoning trace with respect to the video, question, and preceding trace (Figure 1). Instead of assigning one reward per response, GSRPO compares semantically corresponding sentences across the group, normalizes their self-rewards locally, and applies each advantage only to that sentence's tokens. Thus, GSRPO preserves the critic-free, group-relative structure of GRPO, while transforming one sparse learning signal per response into a sequence of localized learning signals.

The improved reward density is especially useful for video, where errors are often localized in temporal, spatial, or semantic understanding. For example, a rationale may correctly identify a person pouring batter into a pan and correctly recognize a circular hand motion, but incorrectly describe that motion as clockwise rather than counter-clockwise (Figure 2). Unlike local self-rewards, a sparse answer-level reward cannot distinguish the correct observations from the isolated directional error.

Reliable self-rewarding requires the model to distinguish plausible language from genuinely grounded reasoning. We therefore precede RL with a synthetic supervised warm start. A stronger foundation VLM generates video reasoning traces containing mixtures of supported and deliberately corrupted statements, together with sentence-level assessments identifying unsupported visual claims, temporal inconsistencies, incorrect counts, entity errors, and other reasoning failures. We fine-tune the model on these examples, allowing it to learn to identify errors in mixed reasoning traces and apply self-rewards.

Finally, during RL, we complement the self-reward with a temporal-sensitivity reward. For each video, we shuffle its frames and measure how much the policy's sentence-level reasoning changes relative to the original video. We reward sensitivity up to a fixed margin, discouraging reasoning that is invariant to temporal order without encouraging arbitrary instability. Together, the temporal-sensitivity reward encourages reasoning that responds to the video's spatiotemporal structure, while the local self-rewards ensure that this sensitivity remains accurate and grounded in the visual evidence.

We evaluate VIDEO-SR1 across two regimes, spanning six diverse video-reasoning benchmarks: VLM4D, VideoEspresso, NExT-QA, STAR, Video-MME, and MVBench. First, we remove ground-truth answer rewards entirely and train using only signals that can be derived from the model itself. In this setting, VIDEO-SR1 significantly outperforms strong baselines for self-derived rewards, including VL-Calibration and UMPIRE, and remains effective under cross-domain transfer for benchmarks where training and evaluation data come from different video domains. Second, when sparse ground-truth answer rewards are restored, VIDEO-SR1 continues to improve accuracy, sample efficiency, and cross-domain transfer over answer-supervised approaches including GRPO, Video-R1, and Vision-SR1. Our error analyses and ablations reveal that these gains arise from complementary components: the temporal-sensitivity reward encourages reasoning to respond to the video's spatiotemporal structure, while sentence-level self-rewards keep that sensitivity accurate and grounded in visual evidence by suppressing unsupported intermediate claims. Together, these results show that localized self-rewards provide effective supervision for video reasoning both with and without ground-truth answer rewards.

Our contributions are threefold. First, we introduce VIDEO-SR1, a self-rewarding framework for video-language reasoning that derives local supervision without a learned critic or external reward model. Second, we propose GSRPO, which performs group-relative credit assignment at the level of semantically aligned reasoning sentences, together with a counterfactual temporal-sensitivity reward.

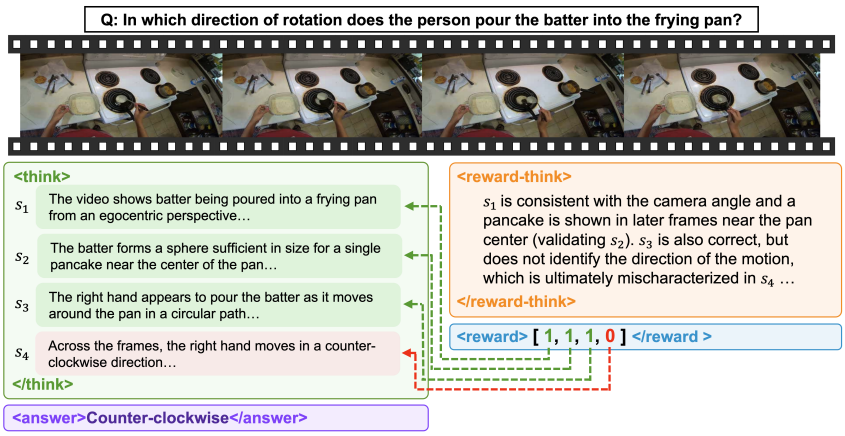

Figure 2: Example of VIDEO-SR1 self-rewarding reasoning.

Third, we show that VIDEO-SR1 significantly improves accuracy, sample efficiency, and cross-domain transfer relative to strong baselines across six benchmarks, in settings with and without the sparse ground-truth answer reward included during RL. Further, extensive error analyses and ablations elucidate two complementary factors underlying VIDEO-SR1's gains: the temporal-sensitivity reward encourages reasoning that responds to the video's spatiotemporal dynamics, while local self-rewards ensure the temporally-sensitive reasoning remains accurate and grounded in the visual evidence.

2 VIDEO-SR1

We first describe the RL objective used to train VIDEO-SR1, followed by the synthetic supervised warm start used to initialize reliable self-rewards. Let a training instance be $x = (v, q)$, where $v = (f_1, \ldots, f_F)$ is a video represented by sampled frames and q is a query. For each prompt, the policy π_θ samples a group of G responses. Response i contains a reasoning trace $l_i = (s_{i,1}, \ldots, s_{i,K_i})$, a final answer a_i, a self-reward rationale z_i, and sentence-level self-rewards $r_i^{\text{self}} = (r_{i,1}^{\text{self}}, \ldots, r_{i,K_i}^{\text{self}})$.

2.1 GROUP AND SENTENCE RELATIVE POLICY OPTIMIZATION

Standard GRPO samples G responses $y_i = (l_i, a_i)$ and assigns a single scalar reward r_i to each response. The corresponding group-relative advantage is

$$A_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\text{Std}_{j=1,\ldots,G}(r_j) + \epsilon}. \tag{1}$$

Ignoring PPO clipping terms for readability, the policy objective can be written as

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} A_i \sum_{t=1}^{T_i} \log \pi_\theta \left(y_{i,t} \mid x, y_{i,<t}\right) - \beta D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right). \tag{2}$$

Thus, every token in a response is updated using the same normalized outcome advantage, regardless of which intermediate statements were correct or incorrect.

In GSRPO, we instead assign a local reward $r_{i,k}$ to each reasoning sentence $s_{i,k}$. To construct a meaningful group-relative baseline for a sentence when sampled traces may differ in length, organization, and level of detail, we align sentences across responses according to semantic similarity. Let $\phi(s)$ denote the sentence representation used for matching, and define

$$m_j(i,k) = \arg\max_{1 \le m \le K_j} \text{sim}\left(\phi(s_{i,k}), \phi(s_{j,m})\right), \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes semantic similarity. Thus, $m_j(i,k)$ identifies the sentence in response j that is most semantically similar to sentence $s_{i,k}$. The comparison group for $s_{i,k}$ is therefore formed from sentences that express similar observations or reasoning steps, even when they occur at different positions in their respective traces. We implement ϕ using a frozen sentence encoder and measure

similarity with cosine similarity. Specifically, for each reasoning sentence $s_{i,k}$, we compute an embedding $h_{i,k} = \phi(s_{i,k})$, and define

$$\mathrm{sim}\left(\phi(s_{i,k}), \phi(s_{j,m})\right) = \frac{h_{i,k}^{\top} h_{j,m}}{\|h_{i,k}\|_2 \|h_{j,m}\|_2}. \tag{4}$$

We compute the sentence-level group-relative advantage as

$$A_{i,k} = \frac{r_{i,k} - \frac{1}{G}\sum_{j=1}^{G} r_{j,m_j(i,k)}}{\mathrm{Std}_{j=1,\ldots,G}\left(r_{j,m_j(i,k)}\right) + \epsilon}. \tag{5}$$

Let $\mathcal{T}_{i,k}$ denote the token indices belonging to sentence $s_{i,k}$. Our sentence-level objective is

$$J_{\mathrm{sent}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\sum_{k=1}^{K_i} A_{i,k} \sum_{t \in \mathcal{T}_{i,k}} \log \pi_\theta\left(y_{i,t} \mid x, y_{i,<t}\right) - \beta D_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\right). \tag{6}$$

Consequently, a token in sentence $s_{i,k}$ is reinforced or suppressed according to the quality of that sentence relative to *semantically comparable* reasoning across the sampled group. The method therefore accommodates traces with different numbers of sentences and different reasoning orders without imposing an artificial positional correspondence. Like GRPO, GSRPO requires no learned critic and retains within-group normalization.

Self-reward generation. After producing its reasoning trace, the same policy reflects on each sentence in that trace. For every $s_{i,k}$, it evaluates whether the statement is supported by the video, temporally consistent with the observed events, logically consistent with the preceding reasoning, and relevant to the query. The model then predicts a binary sentence reward $r_{i,k}^{\mathrm{self}} \in \{0, 1\}$ (Figure 2).

Temporal-sensitivity reward. To complement the self-rewards with a signal that encourages temporal grounding, we compare the policy's predictions under the original video and a frame-shuffled counterfactual. For each $v = (f_1, \ldots, f_F)$, we sample a permutation σ and construct $v^{\mathrm{shuf}} = (f_{\sigma(1)}, \ldots, f_{\sigma(F)})$, which largely preserves frame-level visual content while disrupting temporal order. For each reasoning sentence $s_{i,k}$, we measure the average token-level divergence between predictions conditioned on the original and shuffled videos:

$$d_{i,k} = \frac{1}{|\mathcal{T}_{i,k}|} \sum_{t \in \mathcal{T}_{i,k}} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid v, q, y_{i,<t}) \,\|\, \pi_\theta(\cdot \mid v^{\mathrm{shuf}}, q, y_{i,<t})\right). \tag{7}$$

We convert this sensitivity into a bounded reward,

$$r_{i,k}^{\mathrm{temp}} = \min\left(\frac{d_{i,k}}{\tau}, 1\right), \tag{8}$$

where τ is a target sensitivity margin. This rewards reasoning whose predictions respond to corrupted temporal order without encouraging unbounded instability. In particular, a temporal statement that receives high self-rewards but changes little under frame shuffling is less likely to be grounded in the video's temporal structure.

In practice, we split each rollout group of size G into two equal subsets. For $G/2$ responses, we use the original video ordering and compute the sentence-level self-reward described above. For the remaining $G/2$ responses, we replace the video with a frame-shuffled counterfactual and use the resulting change in model predictions to compute the temporal-sensitivity reward. Thus, the total rollout budget remains fixed at G responses per prompt.

Optional sparse answer reward. Our primary setting excludes ground-truth answer rewards. When answer supervision is available, we combine it with the local self-reward and temporal reward. For multiple-choice benchmarks, r_i^{ans} is exact-match correctness. For open-ended tasks, we use the benchmark's official correctness function.

2.2 Synthetic Self-Reward Warm Start

Before RL, we warm-start the policy on synthetic video reasoning examples generated by a stronger foundation VLM. This stage teaches the model to distinguish locally correct from locally incorrect reasoning before its own judgments are used as rewards.

Data generation. We construct synthetic examples from video reasoning training data. For each (v, q) pair, a strong foundation VLM first generates a correct rationale and answer. We then construct corrupted traces by modifying one or more reasoning sentences while leaving the remainder of the trace plausible. Corruptions include temporal reversals, entity swaps, incorrect counts, hallucinated objects or actions, incorrect causal relations, and unsupported conclusions. The resulting traces therefore contain mixtures of correct and incorrect reasoning rather than uniformly bad responses.

The foundation VLM subsequently evaluates each sentence in the context of the complete trace, producing a short self-evaluation rationale and a binary sentence-level reward. Uncorrupted, supported sentences receive reward 1, while deliberately corrupted sentences receive reward 0. We apply automatic filters for malformed outputs and answer inconsistencies and manually review 1,000 randomly sampled examples to assess data quality. We provide detailed examples of the synthesized video data at `https://video-sr1.github.io/generated_video_data`.

Supervised objective. Each supervised example consists of a video v, query q, reasoning trace l, answer a, self-evaluation rationale z, and sentence rewards $r^{\text{self}} = (r_1^{\text{self}}, \ldots, r_K^{\text{self}})$. We serialize these components into a single target $y = [l; a; z; r^{\text{self}}]$. Importantly, although the synthetic traces contain deliberately corrupted reasoning sentences, we do *not* train the policy to reproduce these corruptions.

The corrupted sentences are included in the autoregressive context so that the model can learn to detect and assess them, but their tokens are masked from the supervised language-modeling loss. Thus, the model receives supervision for generating supported reasoning, the final answer, the self-evaluation rationale, and the sentence-level reward predictions, while deliberately incorrect reasoning serves only as context for learning self-assessment.

Formally, let $\mathcal{C}$ denote the token positions belonging to deliberately corrupted reasoning sentences, and define the token-level loss mask

$$m_t = \begin{cases} 0, & t \in \mathcal{C}, \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

We optimize the masked autoregressive negative log-likelihood

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(v,q,y)\sim\mathcal{D}} \left[\frac{1}{\sum_{t=1}^{|y|} m_t} \sum_{t=1}^{|y|} m_t \log \pi_\theta \left(y_t \mid v, q, y_{<t} \right) \right]. \tag{10}$$

This masking allows the synthetic corruptions to provide negative examples for self-assessment without turning them into imitation targets for the reasoning policy.

3 Experiments

We evaluate VIDEO-SR1 in two settings. We first exclude the sparse ground-truth answer rewards during RL to test whether useful video-reasoning supervision can be derived entirely from the model itself. We then restore the answer reward and compare against methods that depend on answer-level supervision. Across both settings, we evaluate final task accuracy, sample efficiency, calibration between perceived and actual reasoning quality, qualitative failure modes, and temporal sensitivity.

Benchmarks. We evaluate on six video-reasoning benchmarks: **VLM4D**, which emphasizes spatiotemporal reasoning over dynamic scenes (Zhou et al., 2025); **VideoEspresso**, which targets fine-grained video question answering and reasoning across diverse video tasks (Han et al., 2024); **NExT-QA**, which focuses on causal and temporal reasoning in video question answering (Xiao et al., 2021); **STAR**, which evaluates situated reasoning over video with an emphasis on interactions and temporal dynamics (Wu et al., 2024); **MVBench**, which contains 20 temporally focused video-understanding tasks (Li et al., 2024); and the short video subset of **Video-MME**, which includes manually annotated questions spanning diverse domains (Fu et al., 2025).

Baselines. We compare against **VL-Calibration**, which estimates intrinsic visual certainty from sensitivity to visual perturbations and token entropy (Xiao et al., 2026), and **UMPIRE**, a multimodal uncertainty estimator based on sampled-response semantic diversity and internal modality confidence (Lau et al., 2026). We adapt each confidence or uncertainty score into an intrinsic RL reward. When answer labels are available, we additionally compare against **GRPO** (Shao et al., 2024), **Video-R1**, which introduces temporal-aware T-GRPO (Feng et al., 2025), and **Vision-SR1**, which self-evaluates whether its visual representation remains sufficient for reasoning when the original visual input is removed (Li et al., 2026). All methods use the same backbone model, `Qwen3.5-9B`, with matched rollout budgets and evaluation settings. RL runs for 5,000 steps, with the checkpoint that performs best on the held-out validation set selected for final evaluation on the test set. The final evaluation results show the mean, with bars for SEM, across three independent evaluation runs. We also plot the test accuracy evaluated every 500 steps.

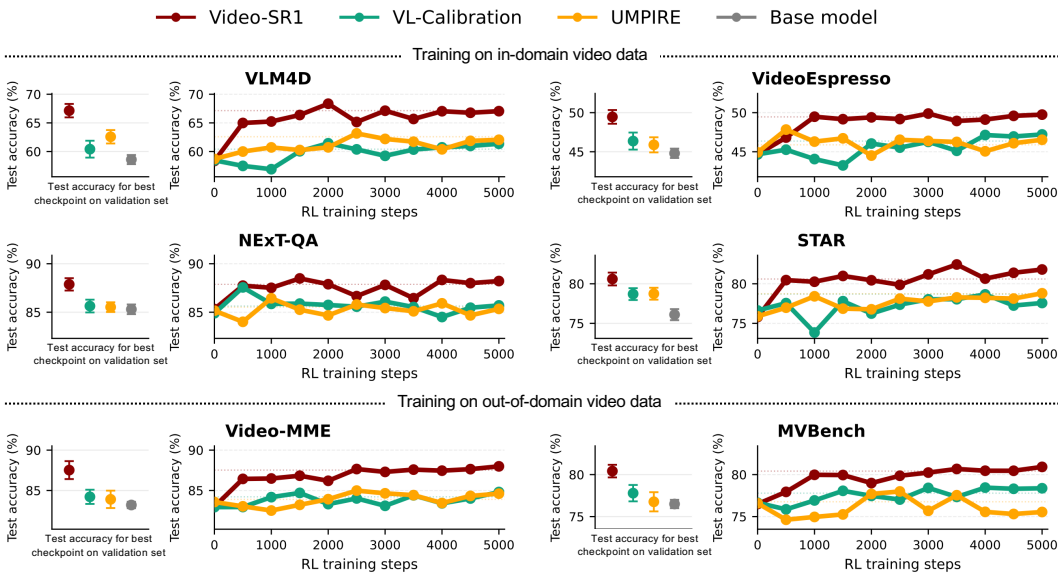

Figure 3: **RL without ground-truth answer rewards.** Learning curves and final test accuracy (mean ± SEM) when using in-domain versus out-of-domain video data during training.

3.1 RL WITHOUT GROUND-TRUTH ANSWER REWARDS

We first exclude all ground-truth signal about final-answer correctness during RL. GRPO, Video-R1, and Vision-SR1 are therefore omitted because they require answer rewards. For benchmarks with training splits (VLM4D, VideoEspresso, NExT-QA, and STAR), all models use half of the in-domain training examples for warm-starting and the remaining half for RL. For VIDEO-SR1, the warm-start data is converted into the synthetic self-assessment examples described in Section 2.2. VL-Calibration and UMPIRE instead receive standard question–answer supervised fine-tuning. RL then operates on the held-out training videos without access to their ground-truth answers. MVBench and Video-MME do not provide training splits and are therefore used to test cross-domain transfer, where all training examples come from combining all the other training datasets. During RL, all models are limited to $G = 16$ responses with the same token budgets.

VIDEO-SR1 improves accuracy, sample efficiency, and cross-domain transfer using only self-derived rewards. Figure 3 shows consistent gains over all baselines across all six benchmarks. On the in-domain evaluations, VIDEO-SR1 reaches 67.1% on VLM4D, 49.5% on VideoEspresso, 87.9% on NExT-QA, and 80.6% on STAR, improving over the strongest baseline by 4.5, 3.1, 2.3, and 1.9 percentage points, respectively. The gains persist under cross-domain transfer: VIDEO-SR1 reaches 86.8% on Video-MME-Short and 80.4% on MVBench, outperforming the strongest baseline by 3.5 and 2.6 points. This cross-domain transfer provides evidence that VIDEO-SR1's sentence-level self-rewarding and temporal grounding improve generalization beyond the post-training distribution. Finally, the learning curves further show that VIDEO-SR1 surpasses the final performance of all baselines within 1,000 RL steps and continues improving thereafter. Together, these results show that sentence-local self-rewards and temporal sensitivity provide an effective learning signal even when excluding all ground-truth rewards. Tables showing full results are provided in Appendix D.1.

3.2 RL WITH GROUND-TRUTH ANSWER REWARDS

We next restore the sparse ground-truth answer reward r_i^{ans}. This enables comparison with methods that depend on this reward (GRPO, Video-R1, and Vision-SR1), while testing whether VIDEO-SR1 remains strong when final-answer supervision is already available.

VIDEO-SR1 is boosted by the sparse ground-truth rewards, further outperforming all baselines on accuracy, sample efficiency, and cross-domain transfer. As shown in Figure 4, VIDEO-SR1 again achieves the highest accuracy across all six benchmarks. It reaches 70.8% on VLM4D, 52.1% on VideoEspresso, 89.2% on NExT-QA, and 82.2% on STAR, exceeding the strongest baseline on each benchmark by 4.5, 3.3, 2.0, and 2.9 percentage points, respectively. The same advantage transfers out of domain: VIDEO-SR1 obtains 87.2% on Video-MME-Short and 81.5% on MVBench, gains of 3.1 and 2.5 points over the strongest competing method. Across all six benchmarks, VIDEO-SR1 significantly improves on the strongest baseline. Its learning curves also show faster improvement, surpassing the best final baseline performance after substantially fewer RL steps. These results indicate that sparse and dense rewards resolve different parts of the credit-assignment problem: the answer reward identifies whether the trajectory ultimately succeeds, while GSRPO determines which intermediate claims within that trajectory should be reinforced or suppressed.

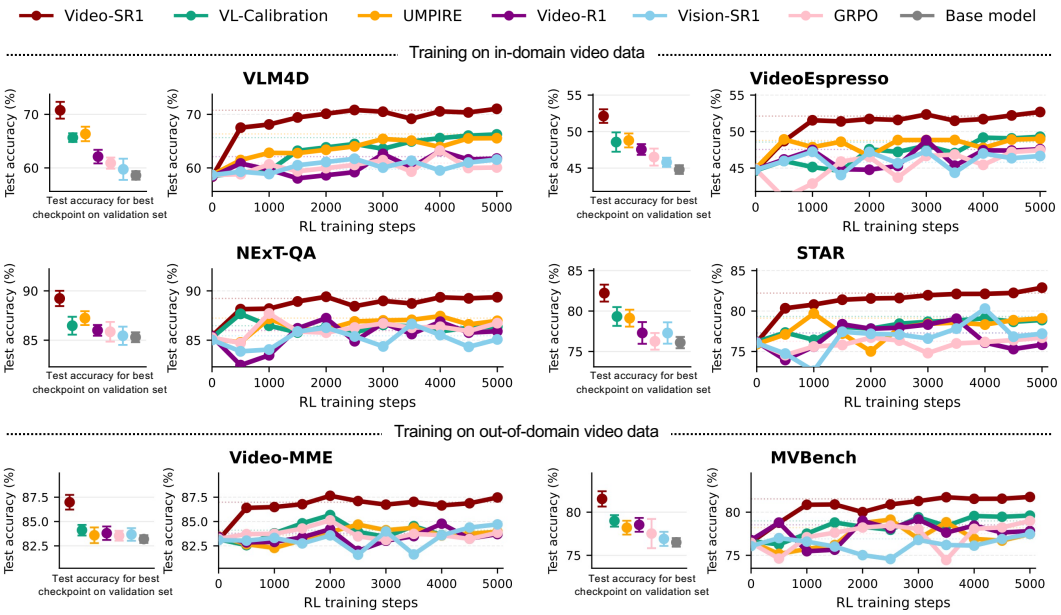

Figure 4: **RL with ground-truth answer rewards.** Learning curves and final test accuracy (mean ± SEM) when using in-domain versus out-of-domain video data during training.

3.3 Error Analysis

The results above establish that VIDEO-SR1 produces better final performance and sample efficiency, but do not reveal *why* it is more effective. We therefore analyze whether the learned self-rewards remain aligned with actual task performance, whether improved answer accuracy is accompanied by more reliable reasoning, and whether the reasoning genuinely depends on the video's temporal structure. Across these diagnostics, VIDEO-SR1 exhibits a consistent pattern: its internal notion of reasoning quality better tracks true correctness, it produces fewer locally unsupported claims, and its predictions are more sensitive to temporal perturbations of the video.

3.3.1 True Task Accuracy versus Perceived Reasoning Accuracy

We first test whether the model's own estimate of reasoning quality remains calibrated to actual task success during RL. We define *perceived reasoning accuracy* as the mean sentence-level self-reward assigned to a generated trace and compare it with true final-answer accuracy on each benchmark.

VIDEO-SR1's perceived reasoning quality tracks true task performance more closely than competing self-derived rewards. As shown in Figure 5a, the gap between perceived and true accuracy is consistently smaller for VIDEO-SR1 across benchmarks. In contrast, VL-Calibration and UMPIRE show more vulnerability to reward-hacking behavior, where high confidence is assigned to reasoning traces even when final-answer accuracy is substantially lower. This distinction is important for self-rewarded RL: a reward that becomes increasingly detached from true correctness can reinforce precisely the behaviors that make the model appear confident without making it more accurate. However, we see improvement with VL-Calibration when ground-truth rewards are included, as the method explicitly uses these rewards to improve the calibration of the self-derived confidence estimates. Nevertheless, VIDEO-SR1 retains the closest correspondence between perceived reasoning quality and actual task success.

Figure 5b examines how this relationship evolves during optimization. **VIDEO-SR1's ratio of true accuracy to perceived reasoning accuracy remains nearly constant throughout RL**, whereas the ratio declines for the competing self-derived rewards. The latter pattern indicates growing overconfidence: the reward signal continues to increase even when true task performance does not improve proportionally. In contrast, the relative stability of VIDEO-SR1 suggests that optimizing its self-reward does not rapidly decouple the reward from the behavior it is intended to measure. This is especially important in the setting without ground-truth answer rewards, where there is otherwise no external signal during RL to correct reward miscalibration. Including the ground-truth reward again substantially improves VL-Calibration, but VIDEO-SR1 remains the most stable.

Finally, Figure 5c compares perceived reasoning accuracy for responses with correct versus incorrect final answers. **VIDEO-SR1 assigns substantially different self-rewards to successful and unsuccessful traces**, while the distributions for baselines overlap strongly. This suggests VIDEO-SR1's local assessments are more discriminative of whether the underlying reasoning leads to a correct answer. Together, these results indicate that VIDEO-SR1's self-rewards provide a substantially more stable and informative training signal than baselines.

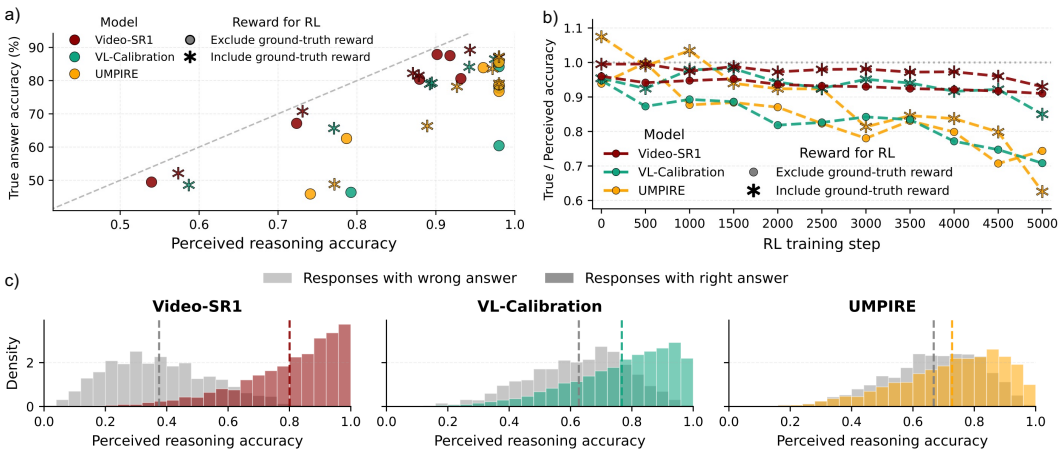

Figure 5: **Analysis of true versus perceived accuracy. a)** True answer accuracy versus average perceived reasoning accuracy (scaled from 0 to 1) on test set of each benchmark. **b)** Ratio of true test accuracy to perceived reasoning accuracy during RL, averaged across benchmarks. **c)** Perceived reasoning accuracy for correct and incorrect responses, averaged across benchmarks.

3.3.2 Reasoning Error Types

We next test whether the improvements in accuracy correspond to improvements in the reasoning traces themselves. We manually inspect 100 randomly sampled responses per model and benchmark and categorize failures as *total hallucination*, *perceptual error*, *reasoning error*, or *other*, following the protocol in Appendix C. Figure 6a shows that **Video-SR1 reduces all error types relative to the baselines.** Video-SR1 lowers the frequency of all major error categories. This is consistent with the intended division of labor between its two reward components: the sentence-level self-reward penalizes unsupported local claims, while the temporal-sensitivity objective encourages the policy to condition its reasoning on the ordered visual evidence rather than on language priors or static appearance alone.

A particularly revealing failure mode appears when the final answer is correct despite flawed intermediate reasoning. As shown in the middle panel of Figure 6a, such traces remain common for the baselines: a response can arrive at the correct answer while still containing one or more incorrect intermediate statements. Under a sparse answer-level reward, these errors are invisible to the optimizer because the entire trajectory receives positive credit. **Video-SR1 substantially reduces these "right answer, wrong reasoning" cases.** This directly reflects the advantage of sentence-local credit assignment: an accidentally correct final prediction cannot cause every preceding statement to be reinforced. The result also agrees with the calibration analysis above, where Video-SR1's perceived reasoning quality more closely tracks true task success.

Finally, we evaluate whether the reasoning actually depends on temporal order by shuffling the video frames for 50 randomly selected examples and checking whether the model remains inappropriately unchanged. As shown in the right panel of Figure 6a, Video-SR1 is temporally insensitive on only about $7/50$ examples, substantially fewer than all baselines. This provides complementary evidence that the temporal-sensitivity reward is changing what information the model relies on: the model is less likely to produce the same reasoning when the event order in the visual input has been disrupted.

Together, these analyses support two complementary mechanisms behind Video-SR1's gains. **The temporal-sensitivity reward encourages the policy to rely on the video's ordered spatiotemporal evidence, while the local self-rewards ensure that this temporally responsive reasoning remains accurate and grounded.**

4 Ablations

We ablate the two central components of Video-SR1: the sentence-level self-reward and the temporal-sensitivity reward. Full benchmark accuracies are reported in Appendix E. For Figure 6b, we repeat the diagnostic error analysis. Removing sentence-level self-rewards increases all major error types, and sharply increases the number of correct-answer traces containing incorrect intermediate statements. This confirms that local rewards are primarily responsible for tying reasoning quality to final correctness. Removing the temporal-sensitivity reward produces a different pattern. Generic reasoning errors change relatively little, but total hallucination and other perceptual errors involving motion direction, state changes, and temporal order increase. Temporally insensitive responses also rise sharply. Thus, we again see the two components are complementary: the temporal reward specifically encourages reasoning to depend on the ordered visual evidence, while the sentence-level self-reward improves accuracy and grounding.

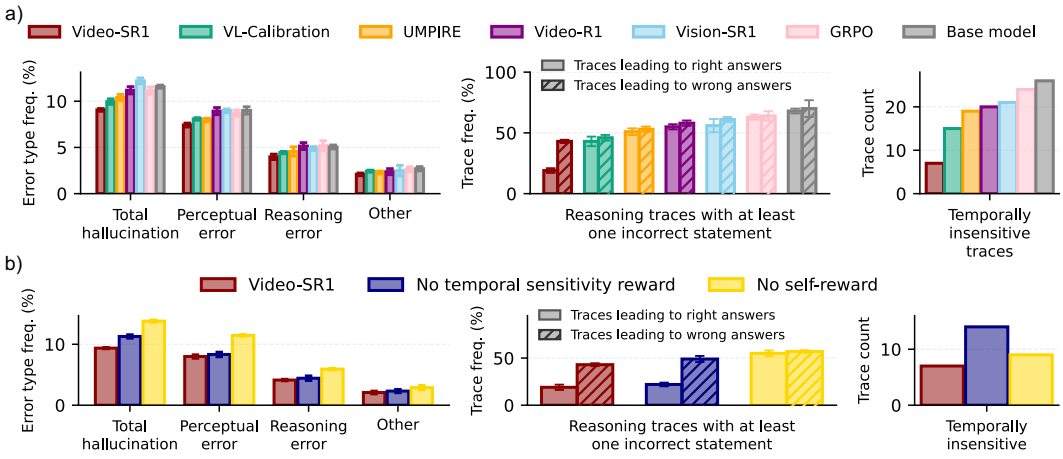

Figure 6: **Analysis of error types. a)** VIDEO-SR1 versus baselines. **b)** VIDEO-SR1 versus ablations. **Left:** Primary error types among 100 randomly sampled traces. **Middle:** Percentage of reasoning traces with incorrect statements among traces that lead to correct versus incorrect final answers. **Right:** Temporally insensitive responses after applying shuffling to 50 randomly sampled videos.

5 RELATED WORK

RL for multimodal and video reasoning. GRPO replaces PPO's learned critic with within-group normalization of outcome rewards (Shao et al., 2024) and is widely used for multimodal RL post-training. Visual-RFT and VLM-R1 extend verifiable-reward RL to VLMs in single-image settings (Liu et al., 2025; Shen et al., 2025). For video, Video-R1 introduces temporal-aware GRPO, called T-GRPO (Feng et al., 2025); TW-GRPO adds token weighting and softer rewards (Dang et al., 2025), VideoP2R separates perception and reasoning rewards (Jiang et al., 2025), and recent multi-task RL targets data-efficient video reasoning (Wang et al., 2026). GSRPO instead derives group-relative credit from the policy's own sentence judgments, aligns semantically corresponding reasoning sentences across rollouts, and applies each advantage only to that sentence's tokens.

Process supervision, self-reward, and uncertainty. Step-level supervision improves credit assignment in language reasoning, from human-generated labels in PRM800K (Lightman et al., 2023) to automatically constructed step rewards in Math-Shepherd (Wang et al., 2024). Self-Rewarding Language Models use the model itself as a judge (Yuan et al., 2024). In multimodal settings, Vision-SR1 self-evaluates visual-representation sufficiency (Li et al., 2026), UMPIRE estimates uncertainty from sampled-response diversity and internal confidence (Lau et al., 2026), VL-Calibration combines visual perturbations with token entropy (Xiao et al., 2026), and ViSS-R1 derives self-supervision from transformed-video pretext tasks (Fang et al., 2025). In contrast, VIDEO-SR1 uses local self-rewards for each reasoning sentence together with a frame-order counterfactual, providing critic-free process credit that is both sentence-local and explicitly sensitive to the temporal structure of the given video.

6 CONCLUSIONS

We introduced VIDEO-SR1, a self-rewarding framework for video-language reasoning that replaces sparse answer-based supervision with localized credit assignment over intermediate reasoning. Its core algorithm, GSRPO, combines sentence-level self-rewards, which assess the local correctness and grounding of each reasoning step, with a temporal-sensitivity reward that favors reasoning that is responsive to changes in video frame order, and uses these signals for semantically aligned, group-relative sentence-level credit assignment. Across six diverse benchmarks, VIDEO-SR1 improves accuracy, sample efficiency, and cross-domain transfer both with and without ground-truth answer rewards. Our error analyses and ablations further reveal complementary effects of the two rewards: the temporal-sensitivity reward encourages reasoning to depend on the video's ordered spatiotemporal evidence, while sentence-level self-rewards ensure this sensitivity remains accurate and grounded in the visual evidence. Together, these mechanisms yield fewer reasoning errors, better alignment between perceived and true reasoning quality, and stronger temporal sensitivity. More broadly, our results suggest that a central limitation of RL for video reasoning is not only reward quality, but reward *resolution*: assigning credit at the level of intermediate claims can make each rollout more informative and provide a stronger training signal for grounded multimodal reasoning.

AI USE STATEMENT

As described in the methods, generative AI tools were used to generate the synthetic reasoning traces, controlled corruptions, and sentence-level assessments used for the self-reward warm start described in Section 2.2. Generative AI was also used to assist with polishing portions of the manuscript. All synthetic data were automatically filtered and a random subset was manually audited. All AI-assisted manuscript content was revised and reviewed by the authors. The authors take responsibility for the final content of the paper and for any artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies post-training methods for video-language models and inherits risks associated with generative multimodal systems, including hallucination, bias, privacy leakage from training videos, and overconfident use in high-stakes settings. Synthetic training data is produced by a foundation VLM and may reproduce its systematic biases or annotation errors. Our warm-start filtering and manual auditing reduce, but do not eliminate this risk. Further, self-rewarding can also amplify model-specific misconceptions if not handled purposefully. We will document data provenance, teacher-model usage, compute requirements, and human annotation procedures in the final release, and we will not release any restricted raw video content beyond what source licenses permit.

REPRODUCIBILITY STATEMENT

The main paper specifies the GSRPO objective, sentence-level advantage computation, temporal counterfactual reward, evaluation regimes, and baseline definitions. The Appendix includes full implementation details, hyperparameters, data-generation filters, frame sampling, reward weights, seed counts, compute, and benchmark-specific evaluation commands. We will release all trained models, data, and code at `https://video-sr1.github.io/`.

REFERENCES

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

Bo Fang, Yuxin Song, Qiangqiang Wu, Haoyuan Sun, Wenhao Wu, and Antoni B. Chan. Viss-r1: Self-supervised reinforcement video reasoning. *arXiv preprint arXiv:2511.13054*, 2025.

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. *Advances in Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2503.21776`.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. VideoEspresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. *arXiv preprint arXiv:2411.14794*, 2024. URL `https://arxiv.org/abs/2411.14794`.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. Videop2r: Video understanding from perception to reasoning. *arXiv preprint arXiv:2511.11113*, 2025.

Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for multimodal large language models with incoherence-adjusted semantic volume. *arXiv preprint arXiv:2602.24195*, 2026. URL `https://arxiv.org/abs/2602.24195`.

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. *arXiv preprint arXiv:2311.17005*, 2024. URL `https://arxiv.org/abs/2311.17005`.

Zongxia Li, Wenhao Yu, Zhenwen Liang, Chengsong Huang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2508.19652v2`.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Xiaodong Wang, Zhirong Wu, Langling Huang, Yuxi Zheng, and Peixi Peng. Incentivizing versatile video reasoning in mllms via data-efficient reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5444–5454, 2026.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. VL-Calibration: Decoupled confidence calibration for large vision-language models reasoning. *arXiv preprint arXiv:2604.09529*, 2026. URL https://arxiv.org/abs/2604.09529.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 2024.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards spatiotemporal awareness in vision language models. *arXiv preprint arXiv:2508.02095*, 2025. URL https://arxiv.org/abs/2508.02095.

A Implementation Details

Base model. All methods use the same video-capable VLM initialization and identical video preprocessing, decoding, and optimization settings unless a baseline requires otherwise. We use `Qwen3.5-9B` as the backbone model and fine-tune all language-model parameters. Baselines are initialized from the same supervised checkpoint used for their corresponding comparison.

Video preprocessing. Videos are decoded at their native frame rate and uniformly sampled to a maximum of 15 frames per example. Frames are resized according to the base model's default visual preprocessing. The same sampled frames are used across methods. For the temporal counterfactual, we draw a random non-identity permutation of these frames independently for each training example.

Supervised warm start. For benchmarks with training splits, including VLM4D, VideoEspresso, NExT-QA, and STAR, all models use about half of the in-domain training examples for warm-starting and the remaining half for RL. For VIDEO-SR1, the warm-start data is converted into the synthetic self-assessment examples described in Section 2.2, while VL-Calibration and UMPIRE instead receive standard question–answer supervised fine-tuning. For the synthetic data generation, a stronger foundation VLM first generates a rationale and answer and then produces variants containing one or more localized corruptions. We specifically use `Gemini-3.1-Pro` as the foundation VLM. Corruption types include temporal reversal, entity substitution, incorrect counting, hallucinated objects or actions, incorrect causal relations, and unsupported conclusions. The teacher subsequently assigns a binary correctness label to every reasoning sentence and provides a short assessment.

Malformed examples, examples with unparsable sentence labels, and examples whose generated final answer is inconsistent with the source supervision are removed. Deliberately corrupted sentence tokens are included in the autoregressive context but masked from the language-modeling loss as described in Section 2.2. We generate 500 examples for VLM4D, 100,000 examples for VideoEspresso, 15,000 examples for NExT-QA, and 22,500 examples for STAR. We will release all training data, along with the code used to generate it at `https://video-sr1.github.io`. We also provide detailed examples of the synthesized video data at `https://video-sr1.github.io/generated_video_data`. To enable fair comparison, for baseline models, we use the same number of source examples for standard supervised fine-tuning. Fine-tuning is run for a single epoch over the full training mixture. We use the AdamW optimizer with a base learning rate of 1×10^{-5} with cosine decay. Global batch size is 64, distributed across 2 NVIDIA H100 GPUs.

RL configuration. For each prompt, we sample $G = 16$ responses using temperature 1.0 and a maximum generation length of 2048 tokens. Training runs for 5,000 optimization steps, with evaluation every 500 steps. We select the checkpoint with the best performance on the validation set for the final evaluation. Unless otherwise stated, hyperparameters are shared across methods.

Sentence matching. Reasoning traces are segmented using sentence boundaries in the generated output. We embed each sentence using a frozen `BAAI/bge-base-en-v1.5` (Xiao et al., 2023) and use cosine similarity for the matching operation in Equation 3. Matching is performed independently for every sentence against each other response in the rollout group. We add $\epsilon = 10^{-6}$ to the group standard deviation for numerical stability.

Self-reward inference. After generating the rationale and final answer, the policy receives the original video, query, and generated trace and predicts one binary label per reasoning sentence. The assessment evaluates whether each sentence is visually supported, temporally consistent, logically consistent with the preceding trace, and relevant to the question. The resulting binary prediction is used directly as $r_{i,k}^{\text{self}}$; no external reward model is used.

A.1 Prompts

The following templates summarize the prompts used for training. The released code will contain the exact formatting required by the base model.

Reasoning generation.

> Given the video and question, reason about the relevant visual and temporal evidence step by step. Write the reasoning as a sequence of concise sentences, then provide the final answer.

Sentence-level assessment.

> Given the video, question, and reasoning trace, evaluate each reasoning sentence independently. Determine whether it is supported by the video, temporally and logically consistent

with the preceding reasoning, and relevant to the question. Return a short assessment and a binary label (1 = supported/correct, 0 = unsupported/incorrect) for every sentence.

Synthetic corruption.

Modify one or more sentences in the reasoning trace to create plausible but incorrect intermediate reasoning while preserving the overall format. Use errors such as incorrect temporal order, entity identity, count, action, causal relation, or unsupported visual detail.

B Evaluation Protocol

Benchmarks and splits. We evaluate on VLM4D, VideoEspresso, NExT-QA, STAR, Video-MME, and MVBench using their official test data and metrics. VLM4D, VideoEspresso, NExT-QA, and STAR provide the training data used for the in-domain experiments. For Video-MME-Short and MVBench, no target-benchmark examples are used for RL, yielding a cross-domain evaluation of transfer from the combined training splits from the other datasets.

Answer scoring. For multiple-choice tasks, the generated response is deterministically mapped to the corresponding answer option and scored by exact correctness. For other tasks, we use the benchmark's official evaluation procedure. All methods use identical decoding and answer extraction.

Baselines. We reproduce GRPO, Video-R1, Vision-SR1, VL-Calibration, and UMPIRE using the same base model and rollout budget wherever applicable. For VL-Calibration and UMPIRE, their intrinsic confidence or uncertainty statistic is converted into a reward such that larger rewards correspond to greater estimated correctness. In experiments with ground-truth supervision, every method additionally receives the same answer-correctness reward.

C Error Analysis Protocol

Analysis of error types. For each model and benchmark, we randomly sample 100 responses from the evaluation set for manual analysis. Reviewers inspect the video, question, generated reasoning, and final answer and annotate whether the trace contains: (1) completely hallucinated visual content, (2) another perceptual error, including incorrect entities, attributes, counts, or temporal relations, (3) a reasoning or inference error despite correctly perceived evidence, or (4) another failure. If a trace contains multiple errors, we count the error type that occurs first.

Temporal sensitivity. For 50 randomly sampled examples, we evaluate each model on the original video and a randomly frame-shuffled version. A response is considered temporally insensitive when the final answer and relevant temporal reasoning remain unchanged. The same shuffled videos and annotation criterion are used for all models.

Human review. The synthetic warm-start audit uses 1,000 randomly sampled generated examples. Reviewers verify that sentence-level labels agree with the video and that injected corruptions are genuinely incorrect while the remaining trace remains plausible. Overall, we observed 96% audit accuracy in the final version.

D Extended Results Tables

D.1 Results for RL without ground-truth answer rewards

Table 1: Test accuracy on benchmarks when training on video data from in-domain training split.

Model	VLM4D			VideoEspresso	NExT-QA	STAR
	Real	Synthetic	Avg			
Base model	56.2	60.9	58.6	44.8	85.3	76.1
UMPIRE	62.3	62.9	62.6	45.9	85.5	78.7
VL-Calibration	58.1	62.7	60.4	46.4	85.6	78.7
Video-SR1	**64.6**	**69.7**	**67.1**	**49.5**	**87.9**	**80.6**

Table 2: Test accuracy on benchmarks when training on video data from out-of-domain training split.

Model	Video-MME-Short	MVBench
Base model	83.2	76.5
UMPIRE	82.7	76.8
VL-Calibration	83.3	77.8
Video-SR1	**86.8**	**80.4**

D.2 Results for RL with ground-truth answer rewards

Table 3: Test accuracy on benchmarks when training on video data from in-domain training split.

Model	VLM4D			VideoEspresso	NExT-QA	STAR
	Real	Synthetic	Avg			
Base model	56.2	60.9	58.6	44.8	85.3	76.1
GRPO	60.6	61.1	60.9	46.5	85.9	76.3
Vision-SR1	55.2	64.3	59.7	45.8	85.4	77.3
Video-R1	59.1	65.2	62.1	47.6	86.0	77.4
UMPIRE	65.3	67.4	66.3	48.8	87.2	79.1
VL-Calibration	61.2	70.1	65.7	48.6	86.5	79.3
Video-SR1	**69.8**	**71.7**	**70.8**	**52.1**	**89.2**	**82.2**

Table 4: Test accuracy on benchmarks when training on video data from out-of-domain training split.

Model	Video-MME-Short	MVBench
Base model	83.2	76.5
GRPO	83.5	77.5
Vision-SR1	83.7	76.9
Video-R1	83.8	78.5
UMPIRE	83.6	78.2
VL-Calibration	84.1	79.0
Video-SR1	**87.2**	**81.5**

E Ablations

We perform ablations for the two central components of Video-SR1: 1) the sentence-level self-reward and 2) the temporal-sensitivity reward.

Table 5: Test accuracy when excluding ground-truth answer rewards during RL

Model	VLM4D	VideoEspresso	NExT-QA	STAR	Video-MME-Short	MVBench
Base model	58.6	44.8	85.3	76.1	83.2	76.5
No self-reward	56.7	43.1	84.6	74.6	81.1	74.8
No temp.-sens. reward	62.8	45.7	86.2	79.4	82.4	77.5
Video-SR1	**67.1**	**49.5**	**87.9**	**80.6**	**86.8**	**80.4**

Table 6: Test accuracy when including ground-truth answer rewards during RL

Model	VLM4D	VideoEspresso	NExT-QA	STAR	Video-MME-Short	MVBench
Base model	58.6	44.8	85.3	76.1	83.2	76.5
No self-reward	60.7	45.3	85.7	76.4	83.5	76.9
No temp.-sens. reward	66.8	49.4	87.9	80.3	84.2	78.9
Video-SR1	**70.8**	**52.1**	**89.2**	**82.2**	**87.2**	**81.5**