SEEING THE EVIDENCE: SPATIOTEMPORAL-PRIVILEGED ON-POLICY SELF-DISTILLATION FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

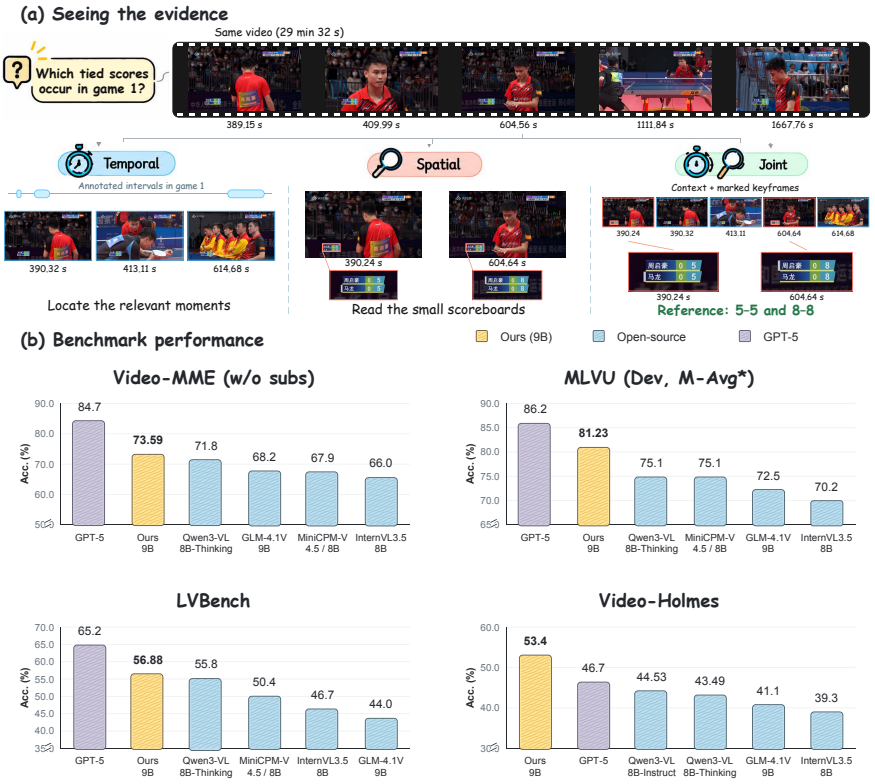

Figure 1: **Privileged visual evidence and benchmark performance.** (a) For the same video question, we construct three candidate teacher views: frames from the annotated temporal clue interval, a bounding-box-marked keyframe with its region crop, and a joint view combining both. The student retains the ordinary full-video input. (b) Accuracy (%) of ST-OPSD (Qwen3.5-9B; yellow), selected open-source models (blue), and GPT-5 (gray) on four video-understanding benchmarks. Video-MME is evaluated without subtitles, and MLVU uses the development split.

ABSTRACT

Vision-language models still struggle to identify and integrate fine-grained spatiotemporal evidence for long-video question answering. We compare different combinations of temporal and spatial evidence and find that joint spatiotemporal views improve the same model's answer accuracy over standard full-video inputs. Building on this study, we propose ST-OPSD (Spatiotemporal On-Policy Self-Distillation), a framework that transfers the model's own predictive advantage under privileged spatiotemporal views to its full-video policy. Specifically, we construct the privileged visual input by combining temporal evidence frames, bounding-box-marked keyframes, and local crops, preserving event context while highlighting spatial details. A frozen teacher initialized from the same pre-

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.

trained model as the student provides token-level supervision under this view along student-generated response trajectories. We train ST-OPSD with a forward-KL distillation objective on 3,200 carefully selected examples from CG-Bench, which we annotate with spatial bounding boxes. Experiments show that ST-OPSD achieves strong performance across four video understanding benchmarks using Qwen3.5 models at different scales.

1 INTRODUCTION

Advances in vision-language models (VLMs) have steadily improved video understanding (Bai et al., 2025; Cheng et al., 2024; Zhang et al., 2024b) and extended it to longer videos (Zhang et al., 2024a; Shen et al., 2024), yet answering questions that depend on fine-grained visual evidence remains challenging (Wu et al., 2024; Fu et al., 2025; Zhao et al., 2025; Qi et al., 2025). Uniform sampling across the full video can introduce question-irrelevant frames that distract models from critical evidence (Wang et al., 2025b). Even when relevant events are captured, models may overlook local object states and interactions or struggle to connect them across time (Liu et al., 2024; Meng et al., 2026). A key challenge is therefore to guide models toward the relevant moments and regions within the available video input and help them integrate this evidence into their answers.

Prior studies highlight the importance of grounding video question answering in question-relevant temporal intervals and spatial regions (Lei et al., 2020; Chen et al., 2025). Some works show that adding spatial hints to temporal evidence can further improve video question answering (Meng et al., 2026; Lei et al., 2020; Tang et al., 2025). Together, these studies highlight the importance of spatial details and temporal information, motivating us to emphasize keyframes and local regions to help models make accurate judgments from fine-grained evidence. This raises a natural question: can training internalize the evidence-use capabilities supported by explicit guidance, enabling models to identify and integrate question-relevant evidence more effectively from ordinary video inputs without additional spatiotemporal hints?

On-policy self-distillation (OPSD) provides a natural framework for transferring the benefits of privileged guidance into a model's behavior under ordinary inputs (Zhao et al., 2026a; Penaloza et al., 2026). It constructs a teacher from the same pretrained model as the student and gives the teacher access to additional information during training. The student generates responses from ordinary inputs, and the teacher supplies next-token distributions conditioned on the same generated prefixes. This provides dense supervision at states encountered during the student's own generation process (Agarwal et al., 2024). In text reasoning, privileged information such as reference answers or reasoning traces helps the teacher guide the student, which receives only the original question. Related work further explores demonstrations, rich feedback, and structured problem-solving guidance as teacher context (Shenfeld et al., 2026; Hübotter et al., 2026; Zhao et al., 2026b).

Research in the visual domain has also explored self-distillation through different visual views and privileged visual context (Zhang et al., 2026; Zhu et al., 2026; Li et al., 2026b;a). For example, Vision-OPD uses evidence-centered image crops to transfer a model's predictive behavior under local views to full-image inputs (Yuan et al., 2026). For video, the design of teacher inputs must also account for temporal relationships among evidence. Local regions can highlight object details, while understanding evolving actions and changes in object states often depends on preceding and subsequent frames and the surrounding scene (Xiao et al., 2021; Lei et al., 2020). Focusing solely on isolated regions may obscure the temporal progression of events and the spatial context of local evidence, whereas providing temporal segments alone may leave decisive spatial details hard to notice. This raises the question of how to organize and combine temporal context with local spatial evidence to provide a more informative basis for teacher predictions.

To investigate this question, we compare three privileged views: a temporal evidence view, a spatially grounded keyframe view, and a joint spatiotemporal view, examining how different evidence presentations affect answer accuracy. Building on this comparison, we introduce **ST-OPSD**, which constructs a privileged teacher using the joint spatiotemporal view. Full frames sampled within evidence intervals preserve the progression of events, bounding-box-marked keyframes indicate the locations of relevant objects, and local crops highlight fine-grained visual details. Using 3,200 training examples selected from CG-Bench (Chen et al., 2025) and augmented with spatial bounding-box

annotations, ST-OPSD distills predictions from this joint view into a student operating on ordinary full-video inputs.

We evaluate ST-OPSD on four video understanding benchmarks using Qwen3.5 models at different scales, achieving strong performance. Our contributions are summarized as follows:

- We investigate how temporal and spatial evidence can be organized into privileged visual inputs, comparing three candidate views to assess their effects on video question-answering accuracy.
- We propose ST-OPSD, a spatiotemporal on-policy self-distillation framework that combines temporal evidence frames, bounding-box-marked keyframes, and local crops to supervise a student operating on ordinary video inputs, without requiring privileged annotations at inference time.
- We construct a training set of 3,200 examples selected from CG-Bench and augmented with spatial bounding-box annotations. Experiments across four video understanding benchmarks demonstrate strong performance with Qwen3.5 models at different scales.

2 Related Work

2.1 Video Understanding and Fine-Grained Spatio-Temporal Reasoning

Advances in multimodal foundation models have driven substantial progress in long-context modeling, efficient visual representations, and reasoning over video content (Zhang et al., 2024a; Shen et al., 2024; Li et al., 2024; Wang et al., 2024). However, holistic video comprehension does not fully reflect a model's understanding of local objects, state changes, and interactions. Recent evaluations examine temporal localization, spatial regions, and relational reasoning, revealing limitations in fine-grained spatio-temporal understanding (Xiao et al., 2024; Cheng et al., 2026). Evidence-guided evaluations further demonstrate the additional value of explicit regional evidence for video question answering (Meng et al., 2026).

To improve fine-grained understanding, existing approaches explore object-level instruction learning, spatio-temporal region encoding, and explicit region markers to associate local visual content with language and support object understanding across frames (Yuan et al., 2025; Heo et al., 2025). Another line of work incorporates visual prompting into training, combining reinforcement learning with self-distillation to retain the benefits of visual guidance without requiring additional prompts at inference (Lee et al., 2026). Our work focuses on regional information for self-teacher construction, using region-marked keyframes to supervise a student operating on ordinary video inputs.

2.2 On-Policy Distillation and Self-Distillation

On-policy distillation (OPD) obtains token-level distributional supervision from a teacher along student-generated trajectories, mitigating the mismatch between training and inference prefixes in offline distillation (Agarwal et al., 2024; Gu et al., 2024). Building on this paradigm, on-policy self-distillation (OPSD) constructs teacher and student policies from the same underlying model under asymmetric input conditions, turning training-time information advantages into learning signals (Zhao et al., 2026a; Penaloza et al., 2026). Existing approaches enrich the teacher's context with reference answers, reasoning traces, or demonstrations, enabling the student to learn privileged-conditioned behavior under ordinary inputs (Zhao et al., 2026a; Shenfeld et al., 2026).

This principle has also been extended to visual tasks through evidence-centered image crops and asymmetric visual views (Yuan et al., 2026; Zhu et al., 2026; Li et al., 2026b). Related OPD and OPSD methods further adjust token-level distillation supervision according to visual evidence (Liu et al., 2026; Bi et al., 2026). These studies establish visual evidence organization as an important design consideration for privileged self-distillation. Our work investigates region annotations on complete keyframes as a teacher condition, providing explicit regional guidance while retaining within-frame context and transferring this supervision to ordinary video inputs along student-generated trajectories.

Visual Input	Accuracy (%)	Δ (pp)
Full-video Sampling	52.20	0.00
Temporal Evidence View	58.42	+6.22
Spatially Grounded Keyframe View	63.18	+10.98
Joint Spatiotemporal View	70.18	+17.98

Table 1: Effect of visual input on the CG-Bench subset. All four conditions evaluate the same frozen Qwen3.5-9B model without training. Δ is the accuracy difference in percentage points relative to full-video sampling.

3 METHOD

3.1 MOTIVATING STUDY: HOW SHOULD SPATIOTEMPORAL EVIDENCE BE PRESENTED?

Previous studies suggest that explicitly identifying relevant evidence can improve video question answering (Meng et al., 2026; Lei et al., 2020; Tang et al., 2025). These findings motivate a closer examination of how temporal and spatial evidence should be presented together. Frames sampled across an evidence interval retain the progression of events, while spatially localized views highlight particular objects and their details. We investigate whether combining these observations provides a more effective input for answering questions than using either view alone.

3.1.1 CANDIDATE PRIVILEGED VIEWS

Given a video, a question, annotated temporal clue intervals, and timestamped spatial bounding boxes, we construct three evidence views:

- **Temporal Evidence View.** We uniformly sample full frames within the annotated temporal intervals. This view presents the relevant event sequence together with its surrounding scene.
- **Spatially Grounded Keyframe View.** We use keyframes with spatial annotations and provide both bounding-box-marked full frames and the corresponding region crops. The marked frames indicate the locations of relevant objects, while the crops expose their local details.
- **Joint Spatiotemporal View.** We combine the interval-sampled full frames, bounding-box-marked keyframes, and region crops in chronological order. This view presents the temporal progression of relevant events alongside localized spatial evidence.

We compare these views with a full-video baseline that uniformly samples 256 frames across the video without evidence guidance. All conditions receive the same question and answer options, without the reference answer.

3.1.2 PRELIMINARY STUDY

We conduct a preliminary study on CG-Bench to examine how evidence presentation affects answer accuracy. The evaluation subset contains 5,000 questions after filtering. For each retained question, every annotated temporal clue interval contains at least one of the 256 frames uniformly sampled from the full video. This criterion ensures that the baseline input covers all annotated intervals, although it does not guarantee that every decisive visual detail is captured. We evaluate the same frozen Qwen3.5-9B model under four input conditions: full-video sampling, the Temporal Evidence View, the Spatially Grounded Keyframe View, and the Joint Spatiotemporal View. No training is performed.

All conditions use the same evaluation questions, question prompts, and decoding settings, without providing reference answers as input. We report answer accuracy and its difference from the full-video baseline in percentage points. This comparison measures the predictive effect of changing the visual input while keeping the model parameters fixed.

As shown in Table 1, the Temporal Evidence View and the Spatially Grounded Keyframe View achieve accuracies of 58.42% and 63.18%, exceeding the full-video baseline by 6.22 and 10.98 percentage points, respectively. The Joint Spatiotemporal View achieves the highest accuracy of 70.18%, improving over the baseline by 17.98 percentage points and over the two individual views by 11.76 and 7.00 percentage points.

3.2 Framework Overview

Motivated by the evidence-view comparison, we introduce ST-OPSD to transfer predictions from a joint spatiotemporal view to a model operating on ordinary video inputs. We first define the learning problem, then describe the training data and paired visual observations, and finally present the on-policy self-distillation objective.

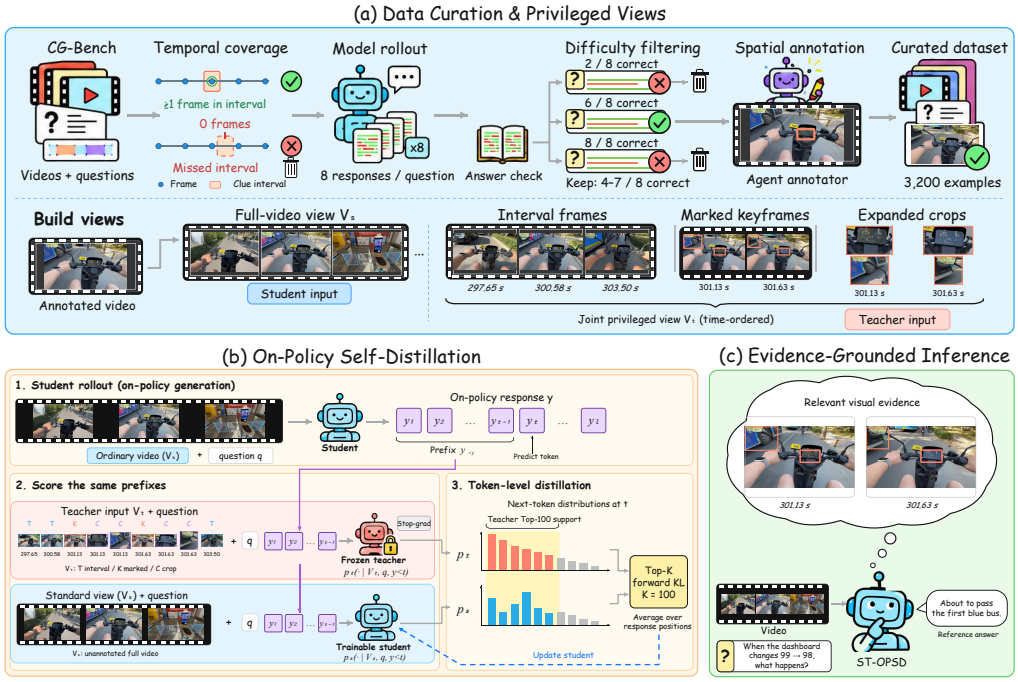

Figure 2: **Overview of ST-OPSD.** (a) We curate training examples from CG-Bench and construct a full-video student view and a privileged teacher view combining interval frames, bounding-box-marked keyframes, and enlarged region crops. (b) The student generates on-policy responses from the ordinary video. A frozen self-teacher evaluates the same response prefixes under the privileged view, and the student aligns their next-token distributions using Top-K forward KL. (c) At inference time, only the student is used with ordinary video inputs, without privileged annotations or additional modules.

3.2.1 Problem Formulation

We consider video question answering with spatiotemporal annotations available during training. Let

$$\mathcal{D} = \{(V_i, q_i, \mathcal{T}_i, \mathcal{B}_i)\}_{i=1}^{N} \tag{1}$$

denote the training set, where V_i is a video and q_i is its associated question. The temporal annotations $\mathcal{T}_i = \{[s_{ij}, e_{ij}]\}_{j=1}^{M_i}$ identify evidence intervals, with s_{ij} and e_{ij} denoting their start and end times. The spatial annotations $\mathcal{B}_i = \{(\tau_{ik}, b_{ik})\}_{k=1}^{K_i}$ specify bounding boxes at selected timestamps, where $b_{ik} = (x_{ik}^{\min}, y_{ik}^{\min}, x_{ik}^{\max}, y_{ik}^{\max})$ contains normalized box coordinates.

For each example, we construct two visual observations:

$$V_i^S = S(V_i), \qquad V_i^T = \Phi(V_i; \mathcal{T}_i, \mathcal{B}_i), \tag{2}$$

where S samples frames across the full video and Φ constructs a joint spatiotemporal view using the annotations. The student receives (V_i^S, q_i), and the teacher receives (V_i^T, q_i). Our objective is to use the teacher's privileged observations during training to improve the student's predictions under ordinary video inputs.

3.2.2 Training Data and Paired View Construction

We construct our training set from CG-Bench using its questions and annotated temporal clue intervals. For each video, we uniformly sample 256 frames and retain a question only if every annotated clue interval contains at least one sampled frame. This criterion ensures coverage of all annotated intervals, although it does not guarantee that every decisive visual detail is captured. We then use Qwen3.5-9B to generate eight responses per question from the sampled full-video input without privileged information. Each response is evaluated against the ground-truth answer, and we retain questions answered correctly in four to seven of the eight responses. This selects examples for which the model demonstrates some ability to answer but remains inconsistent. We obtain timestamped spatial bounding boxes through an annotation agent powered by GPT-6-astra, with the annotation prompts provided in the appendix. Our final training set contains 3,200 examples. Representative training examples and their corresponding privileged visual views are shown in Appendix A.

The student view V^S consists of the 256 uniformly sampled full-video frames. To construct the teacher view V^T, we first uniformly sample full frames within each annotated interval to retain scene context and the progression of relevant events. At each spatially annotated keyframe, we overlay bounding boxes on the full frame to indicate the locations of relevant objects. We also extract a separate crop for each box, expanding it on each side by 25% of its width or height along the corresponding axis and clipping the expanded region to the image boundaries. These crops highlight local details while retaining nearby visual context.

We combine the interval-sampled frames, bounding-box-marked keyframes, and region crops in chronological order. Each observation is paired with its source timestamp, and marked frames and crops receive explicit view-type indicators. Views originating from the same frame share the same timestamp. Both branches receive the same question. Reference answers are used for data filtering and are excluded from both branches' inputs.

3.2.3 On-Policy Self-Distillation

We initialize the trainable student π_θ and the frozen teacher $\pi_{\bar{\theta}}$ from the same pretrained model. For clarity, we omit the example index below. Given the ordinary visual observation V^S and question q, the student samples an on-policy response:

$$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_L) \sim \pi_\theta(\cdot \mid V^S, q), \tag{3}$$

where L is the response length. This response supplies the trajectory on which distillation is performed.

The teacher evaluates the student-generated response under V^T using teacher forcing. At each output position t, both branches compute next-token distributions conditioned on the same response prefix:

$$\begin{aligned} p_t^S &= \pi_\theta(\cdot \mid V^S, q, \hat{y}_{<t}), \\ p_t^T &= \pi_{\bar{\theta}}(\cdot \mid V^T, q, \hat{y}_{<t}), \end{aligned} \tag{4}$$

where $\hat{y}_{<t}$ denotes the tokens generated before position t. This shared prefix aligns the teacher's token-level supervision with states encountered during the student's own generation process.

We optimize the student using a teacher Top-K truncated forward-KL objective. Let

$$\mathcal{K}_t = \mathrm{TopK}_K(p_t^T) \tag{5}$$

denote the set of tokens with the highest teacher probabilities at position t. The token-level loss is

$$\ell_t = \sum_{v \in \mathcal{K}_t} p_t^T(v) \log \frac{p_t^T(v)}{p_t^S(v)}. \tag{6}$$

Both distributions are normalized over the full vocabulary, and their probabilities are used without renormalization within $\mathcal{K}_t$. We use $K = 100$. The overall objective is

$$\mathcal{L}_{\text{ST-OPSD}}(\theta) = \mathbb{E}_{(V,q,\mathcal{T},\mathcal{B})\sim\mathcal{D}}\mathbb{E}_{\hat{y}\sim\pi_\theta(\cdot|V^S,q)}\left[\frac{1}{L}\sum_{t=1}^{L}\ell_t\right]. \tag{7}$$

During optimization, the sampled response and teacher distributions are treated as fixed, and gradients propagate only through the student distributions. The teacher parameters remain unchanged throughout training. At inference time, the student answers questions using only the sampled full-video input and the question, without privileged annotations or access to the teacher.

4 Experiments

4.1 Experimental Setup

Training data. We train on 3,200 multiple-choice examples curated from CG-Bench (Chen et al., 2025), following the procedure in Section 3. Each example contains a video, a question with answer options, annotated temporal clue intervals, and keyframe bounding boxes. All post-training methods use the same examples, and the training videos do not overlap with those in any evaluation benchmark.

Models and baselines. We evaluate ST-OPSD on Qwen3.5-2B, Qwen3.5-4B, and Qwen3.5-9B. For each backbone, we compare the pretrained model with SFT, GRPO (Shao et al., 2024), standard OPSD (Zhao et al., 2026a), and ST-OPSD. All post-training methods use the same 3,200 training examples. SFT directly uses the ground-truth answer option as the target for teacher-forced likelihood optimization. GRPO uses only an outcome reward determined by whether the sampled option matches the ground-truth answer. Standard OPSD provides the ground-truth answer to its teacher as textual privileged information; both its teacher and student receive the ordinary video input. In ST-OPSD, the teacher instead receives the joint spatiotemporal visual view, and neither branch is conditioned on the ground-truth answer during distillation.

Evaluation benchmarks. We evaluate on four video question-answering benchmarks: Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), LVBench (Wang et al., 2025a), and Video-Holmes (Cheng et al., 2025). The model generates a multiple-choice answer for each question, and we compute accuracy from the selected option. At evaluation time, ST-OPSD receives only the video and question; temporal clue intervals, bounding boxes, region crops, and the teacher are unavailable. We apply the same frame-sampling and answer-extraction protocol to all models compared within each backbone.

Implementation details. The student and frozen teacher are initialized from the same pretrained checkpoint. The student generates answer-option responses from the ordinary video input, and the teacher evaluates the student-generated prefixes under the privileged visual input. We optimize the student with the teacher Top-K truncated forward-KL objective described in Section 3, using $K = 100$. No privileged visual annotations or additional modules are used at inference time.

4.2 Main Results

Performance across model scales. Table 2 reports the results on four video understanding benchmarks. ST-OPSD improves the Qwen3.5-9B backbone on every benchmark, with gains of 1.78, 2.21, 3.10, and 2.45 percentage points on Video-MME, MLVU, LVBench, and Video-Holmes, respectively. The 4B model also improves on all four benchmarks; the largest gains occur on MLVU (+6.85 points) and Video-Holmes (+3.65 points). These results show that supervision from the privileged spatiotemporal view can improve a student evaluated using ordinary video inputs.

Comparison with post-training baselines. For the 4B and 9B backbones, ST-OPSD outperforms both GRPO and standard OPSD on all four benchmarks under the same training data and evaluation protocol. The results at 2B are less consistent: ST-OPSD improves the pretrained model on Video-MME, MLVU, and LVBench, but its Video-Holmes accuracy decreases from 39.30% to 37.51%. It

Method	Video-MME	MLVU	LVBench	Video-Holmes	Avg.
Qwen3.5-2B	58.96	67.66	43.00	39.30	52.23
+ SFT	59.03	67.00	41.45	39.85	51.83
+ GRPO	**60.19**	**70.33**	44.67	**46.54**	**55.43**
+ OPSD	58.26	68.45	44.80	43.98	53.87
+ **ST-OPSD**	59.15	67.85	**45.64**	37.51	52.54
Δ vs. base	**+0.19**	**+0.19**	**+2.64**	**-1.79**	**+0.31**
Qwen3.5-4B	67.59	71.21	51.65	46.76	59.30
+ SFT†	67.59	69.83	51.65	48.29	59.34
+ GRPO	66.41	73.23	48.22	47.85	58.93
+ OPSD	68.04	74.56	50.81	45.18	59.65
+ **ST-OPSD**	**69.15**	**78.06**	**52.61**	**50.41**	**62.56**
Δ vs. base	**+1.56**	**+6.85**	**+0.96**	**+3.65**	**+3.26**
Qwen3.5-9B	71.81	79.02	53.78	50.95	63.89
+ SFT†	71.81	80.40	53.78	49.70	63.92
+ GRPO	71.30	77.60	51.58	50.90	62.85
+ OPSD	69.70	76.91	52.29	47.69	61.65
+ **ST-OPSD**	**73.59**	**81.23**	**56.88**	**53.40**	**66.28**
Δ vs. base	**+1.78**	**+2.21**	**+3.10**	**+2.45**	**+2.39**

Table 2: Accuracy (%) on four video understanding benchmarks. Avg. is the unweighted mean across benchmarks, rounded to two decimal places. Within each backbone, bold and underlined method scores indicate the best and second-best distinct values, respectively; ties share the same formatting. Δ denotes the change from the pretrained model in percentage points. †The 4B and 9B SFT rows contain illustrative placeholder values pending measured evaluation results.

also does not consistently outperform the other post-training methods at this scale. Thus, the benefits of the privileged teacher view are clear for the 4B and 9B models in these experiments, but are less reliable for the 2B model.

Teacher view	Video-MME	MLVU	LVBench	Video-Holmes	Avg.
Temporal Evidence	69.74	75.99	51.59	50.08	61.85
Spatially Grounded Keyframe View	71.54	78.03	55.26	51.20	64.01
ST-OPSD	**73.59**	**81.23**	**56.88**	**53.40**	**66.28**

Table 3: Effect of privileged teacher views on student accuracy (%) across four video benchmarks. Only the teacher's visual input changes; Avg. denotes the mean score.

4.3 Ablation Studies

Effect of privileged visual views. We investigate how temporal and regional evidence should be presented to the privileged teacher. Table 3 compares three visual views defined in Section 3: (1) the Temporal Evidence View, consisting of full frames sampled from the annotated clue intervals; (2) the Spatially Grounded Keyframe View, consisting of one bounding-box-marked keyframe and its corresponding region crop; and (3) the Joint Spatiotemporal View, which combines the interval-sampled frames, marked keyframe, and crop. The temporal view preserves the progression of relevant events, whereas the keyframe view makes local object details more explicit. The joint view provides both forms of evidence to the teacher. For all three variants, we use the same training examples, backbone, distillation objective, and optimization settings. The student always receives the same 256-frame ordinary video input; only the teacher's visual view changes.

Divergence Objective	Video-MME	MLVU	LVBench	Video-Holmes
JSD ($\beta = 0.5$)	69.95	74.61	53.12	51.60
Reverse KL	68.88	73.56	55.51	49.30
Forward KL	**73.59**	**81.23**	**56.88**	**53.40**

Table 4: Comparison of divergence objectives on four video understanding benchmarks. Forward and reverse KL correspond to $\mathrm{KL}(p_T \| p_S)$ and $\mathrm{KL}(p_S \| p_T)$, respectively.

Effect of the distillation objective. We study the effect of the distribution-matching objective between the privileged teacher and the student. Using Qwen3.5-9B, we compare teacher-to-student forward KL, student-to-teacher reverse KL, and Jensen–Shannon divergence (JSD) with $\beta = 0.5$. As shown in Table 4, forward KL achieves the highest accuracy on all four benchmarks. On MLVU, it reaches 81.23%, compared with 74.61% for JSD and 73.56% for reverse KL. On LVBench, where reverse KL is more competitive, forward KL still obtains 56.88%, compared with 55.51%. The consistent results support our choice of the teacher Top-K truncated forward-KL objective for the main experiments.

4.4 Analysis

Role of Privileged Visual Supervision. The post-training methods differ in the information used to form their supervision. SFT optimizes the ground-truth option, while GRPO rewards the correctness of a sampled answer; neither explicitly identifies the video evidence supporting that answer. Standard OPSD provides a teacher distribution, but its teacher receives the ground-truth answer as text while viewing the same ordinary video as the student. The answer specifies what to predict without making the supporting frames or regions easier to inspect. In ST-OPSD, the frozen self-teacher instead observes frames from the clue intervals, which preserve the progression of relevant events, together with a bounding-box-marked keyframe and region crop that expose local details. The ground-truth answer is not supplied to the teacher during distillation, and the student learns from the teacher's distribution using only the ordinary video. This difference in teacher conditioning is consistent with ST-OPSD outperforming GRPO and answer-conditioned OPSD on all four benchmarks at 4B and 9B. The mixed 2B results show that the benefit is not uniform across model scales.

Attention to Question-Relevant Regions. To examine whether ST-OPSD changes the student's attention to question-relevant visual evidence, we compare the pretrained Qwen3.5-9B backbone and the trained ST-OPSD-9B student on the same held-out questions. Both models receive identical sampled video frames, questions, and answer options, without privileged annotations. At the answer-option prediction position, we extract attention to visual tokens from selected full-attention layers and project it onto the corresponding video frames. We visualize only annotated frames that can be aligned with frames actually seen by the student. The bounding boxes are overlaid for analysis and are not part of either model's input. In the representative cases, the trained student's attention appears better aligned with question-relevant visual evidence, alongside correct predictions where the backbone answers incorrectly. Attention outside the annotated regions remains visible, indicating that this alignment is not exclusive. These qualitative observations suggest that ST-OPSD encourages attention to relevant evidence, while the heatmaps alone do not establish that those regions causally determine the answers. Detailed case visualizations and paired predictions are provided in Appendix C.

5 Conclusion

We presented ST-OPSD, a spatiotemporal on-policy self-distillation framework for long-video understanding. Our preliminary study shows that combining temporal evidence frames, bounding-box-marked keyframes, and local crops improves the same model's answer accuracy over ordinary full-video inputs. ST-OPSD uses this joint view to construct a privileged teacher and transfers its predictions to a full-video student through token-level forward-KL distillation. Trained on 3,200 curated and spatially annotated CG-Bench examples, ST-OPSD achieves strong performance across four video understanding benchmarks.

AI USE STATEMENT

We used GPT-6-astra to generate timestamped bounding-box annotations for the training data. Generative AI tools also assisted with experimental design, result interpretation, translation, literature review, and drafting and editing parts of the manuscript and its figure and table presentation. The annotation procedure and prompts are described in the appendix. The authors take responsibility for the final content, results, and artifacts.

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, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

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

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 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*, pp. 9155–9164, 2026.

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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3919–3930. IEEE, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. 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. In *European Conference on Computer Vision*, pp. 620–637. Springer, 2026.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8211–8225, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026a.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026b.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*, 2026.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Zitian Tang, Shijie Wang, Junho Cho, Jaewook Yoo, and Chen Sun. How can objects help video-language understanding? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21994–22003. IEEE, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 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.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 2024.

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.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980. IEEE, 2025.

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

Yan Zhang, Daiqing Wu, Huawen Shen, Can Ma, and Yu Zhou. Learn where to click from yourself: On-policy self-distillation for gui grounding. *arXiv preprint arXiv:2605.00642*, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

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*, 2026a.

Xuyang Zhao, Liting Zhang, Zichen Xu, Zhihu Wang, Xu Caiyue, Shiwan Zhao, and Qicheng Li. Is more privileged information better? from solution traces to problem-solving structure in self-distilled reasoning. *arXiv preprint arXiv:2608.01589*, 2026b.

Yiming Zhao, Yu Zeng, Yukun Qi, YaoYang Liu, Xikun Bao, Lin Chen, Zehui Chen, Qing Miao, Chenxi Liu, Jie Zhao, et al. V2p-bench: Evaluating video-language understanding with visual prompts for better human-model interaction. *arXiv preprint arXiv:2503.17736*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. Rp-opsd: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026.

APPENDIX

A TRAINING DATA EXAMPLES

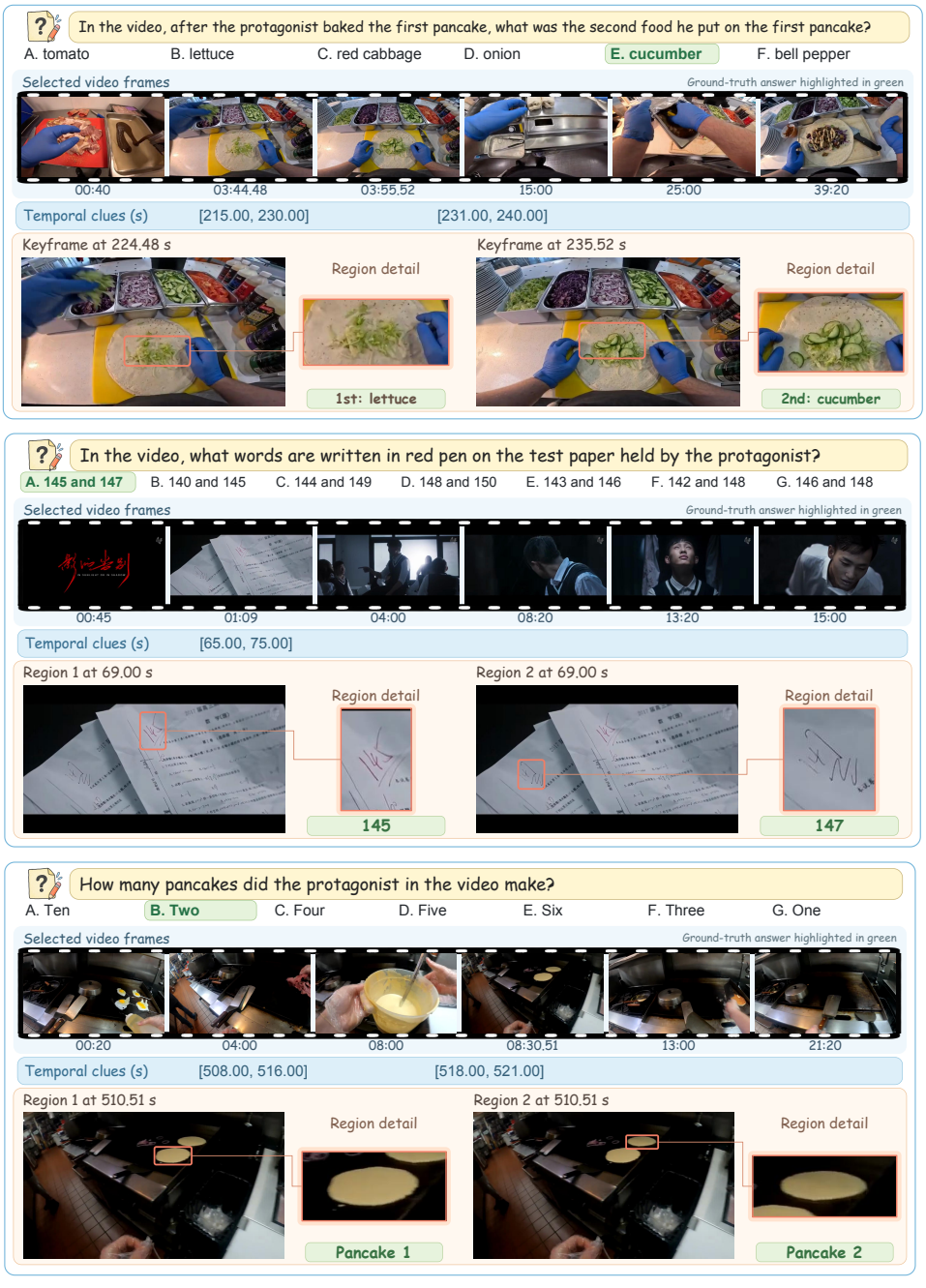

Figure 3: Representative training examples showing temporal clue intervals, marked keyframes, and enlarged question-relevant regions.

B TRAINING IMPLEMENTATION

The training configurations for SFT, GRPO, and OPSD-based methods are summarized in Tables 5–7. All experiments were conducted on 16 NVIDIA H200 GPUs, each equipped with 141 GB of memory.

Table 5: Key hyperparameters for SFT.

Name	Value
Fine-tuning mode	Full
Training target	Gold option
Trainable modules	Language model
Frozen modules	Vision + projector
Cutoff length	40,960
Maximum video frames	256
Epochs	2
Global batch size	32
Micro batch / GPU	1
Learning rate	1.0×10^{-5}
LR schedule	Cosine
Warmup ratio	0.05

Table 6: Key hyperparameters for GRPO.

Name	Value
Trainable modules	Full model
Reward	Outcome (0/1)
Rollouts / question	4
Questions / update	32
Mini-batch size	32
PPO epochs	1
Learning rate	2.0×10^{-6}
PPO clip ratio	0.2
Maximum prompt length	40,960
Maximum response length	2,048

Table 7: Key settings for OPSD-based methods.

Name	Value
Student fine-tuning	Full model
Teacher	Frozen self-teacher
Teacher Top-K	100
Rollouts / question	1
Training batch size	32
Learning rate	2.0×10^{-6}
Maximum prompt length	40,960
Maximum response length	2,048
Optimization steps	100

C CASE STUDY

Figures 4–6 compare Qwen3.5-9B and ST-OPSD-9B on three representative questions. Dashed boxes mark question-relevant regions; green and red indicate correct and incorrect predictions, respectively.

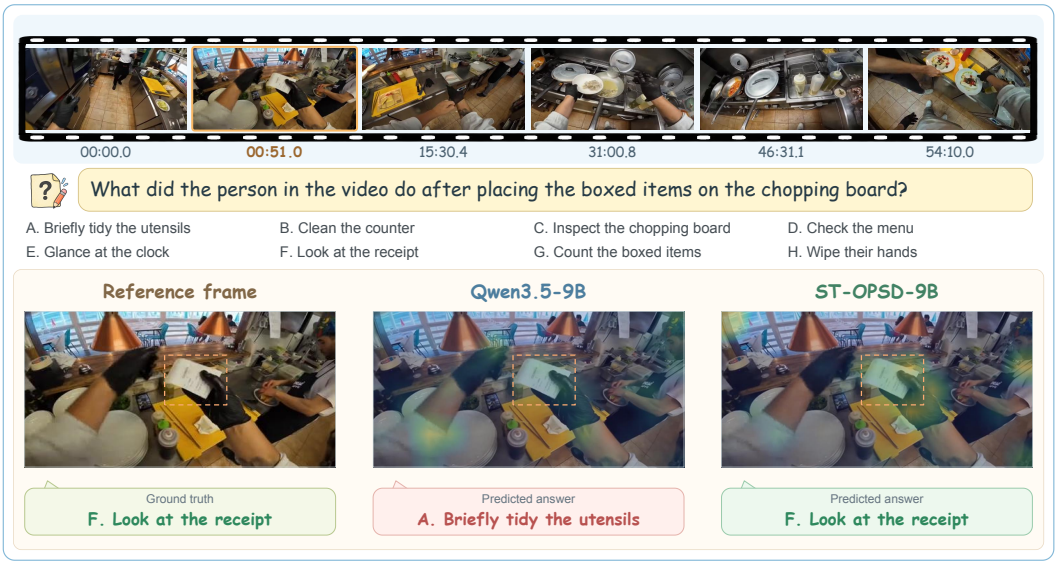

Figure 4: Inference-time attention comparison for case 1.

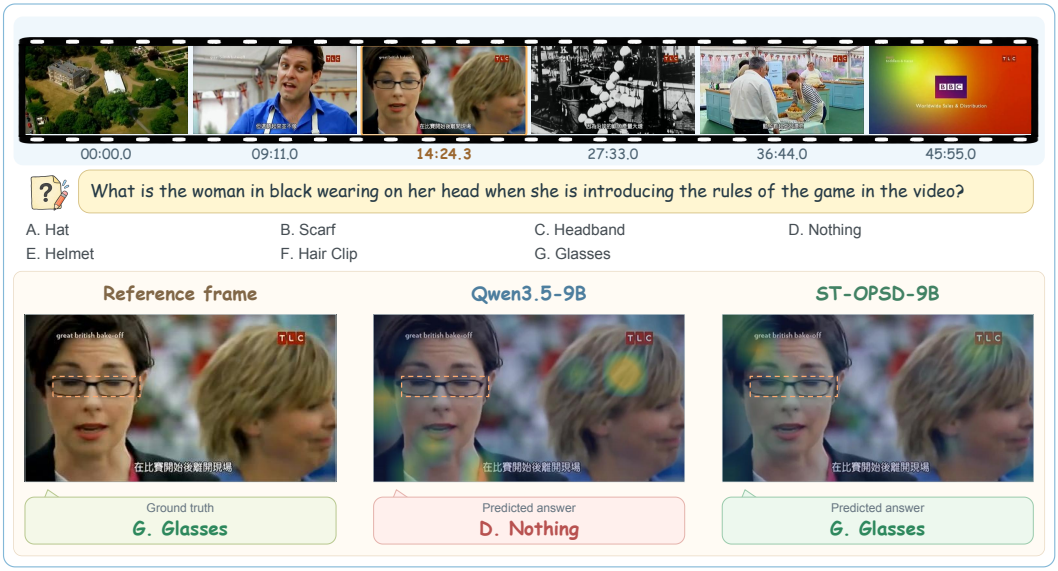

Figure 5: Inference-time attention comparison for case 2.

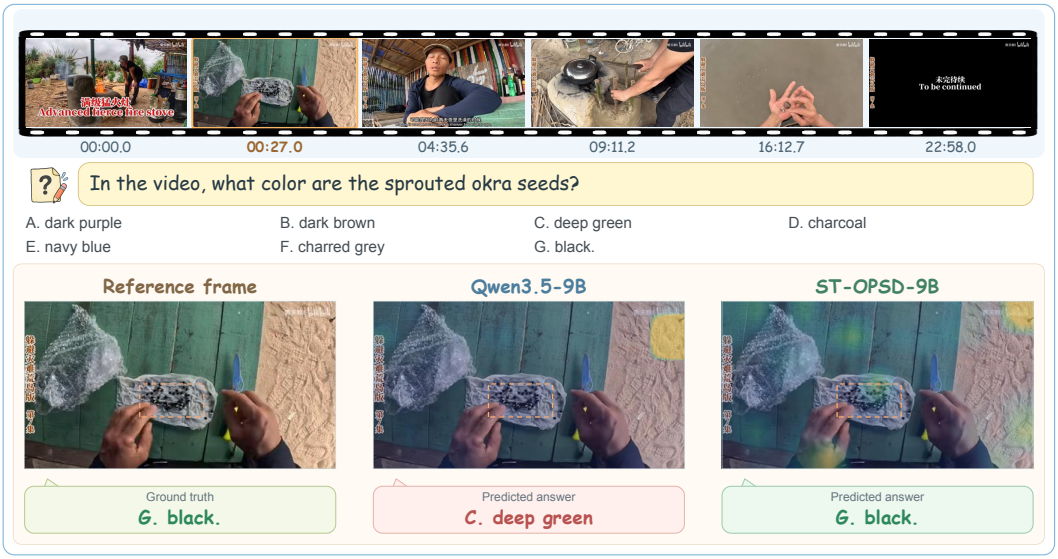

Figure 6: Inference-time attention comparison for case 3.