WHERE TO LOOK MATTERS: ON-POLICY SELF-DISTILLATION FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have made substantial progress in long-video understanding, with standard backbone models typically answering questions based on frames sampled across the full video. However, as videos become longer, the full-video context inevitably contains more question-irrelevant temporal content, which can distract the model from the evidence needed to answer a specific question. We empirically find that focusing the visual input on short annotated clue intervals containing question-relevant evidence consistently improves prediction accuracy across model scales compared with using the corresponding full videos, while requiring fewer input frames. Based on this finding, we introduce **`Clue-OPSD`**, a clue-privileged on-policy self-distillation framework for long-video understanding. During training, a full-video student learns from a self-teacher conditioned on the corresponding clue interval by aligning their next-token distributions along student-generated trajectories. `Clue-OPSD` thus uses clue intervals as privileged supervision without relying on ground-truth answer labels, while requiring no clue annotations or additional modules at inference time. Extensive experiments across multiple long-video understanding benchmarks and Qwen3.5 model scales show consistent gains over vanilla backbones while achieving competitive performance against several supervised post-training methods.

1 INTRODUCTION

Recent advances in foundation vision-language models (VLMs) have substantially improved video understanding and enabled models to process increasingly long visual contexts (Chen et al., 2024; Bai et al., 2025; Zhu et al., 2025; Team, 2026). For long-video question answering, standard backbone models typically sample frames across the full video and jointly process them with the textual query (Ranasinghe et al., 2025; Shu et al., 2025; Li et al., 2026c; Lin et al., 2026). With stronger backbones and larger visual context windows, this simple full-video paradigm has become increasingly effective.

However, as videos become longer, the full-video context inevitably contains more temporal content that is unrelated to a particular question. The model must therefore identify the useful evidence from a large amount of surrounding visual information before making a prediction, and relevant events can easily be obscured by unrelated content distributed across the video. Figure 1 illustrates this issue with a representative example. When the model processes frames sampled from the full 3,409-second video, it predicts an incorrect answer. In contrast, when the visual input is restricted to the corresponding 10-second clue interval containing the relevant event, the same model answers the question correctly.

We further examine this phenomenon systematically using CG-Bench (Chen et al., 2025), which provides annotated temporal clue intervals for individual questions. Under the same frames-per-second (FPS) and maximum-frame constraints, we compare the same VLM using frames sampled from the full video with frames sampled only from the corresponding clue interval. As shown in Table 1, across different Qwen3.5 model scales, focusing the visual input on short annotated clue intervals containing question-relevant evidence consistently improves prediction accuracy compared with using the corresponding full videos, while requiring fewer input frames. These results show that, for long-video question answering, the way relevant temporal evidence is presented to the model can substantially affect prediction quality.

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.

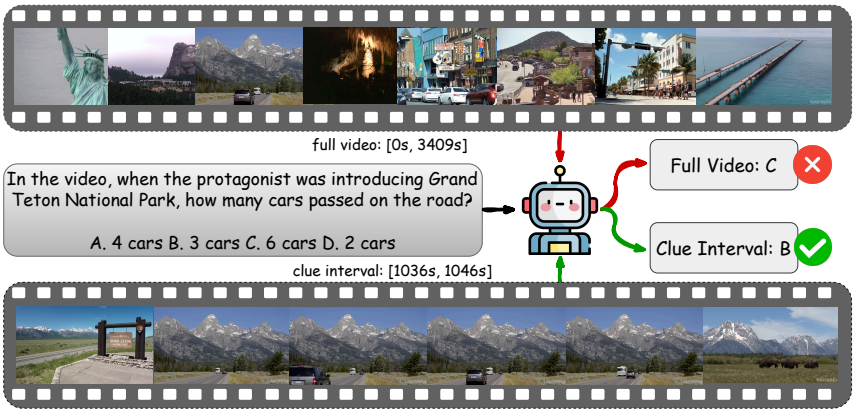

Figure 1: Illustration of the effect of clue intervals in long-video question answering. The model fails when processing frames sampled from the full 3,409-second video but answers correctly when conditioned on the corresponding 10-second clue interval containing the question-relevant evidence.

Based on this finding, we introduce **`Clue-OPSD`**, a clue-privileged on-policy self-distillation (OPSD) framework for long-video understanding. During training, `Clue-OPSD` constructs two asymmetric visual conditions from the same video: a student operates on the full-video input, while an exponential moving average (EMA) self-teacher observes only the annotated clue interval associated with the current question. The student first generates an on-policy trajectory, and both branches are evaluated along the same student-generated prefixes. We then align their next-token predictive distributions, allowing the full-video student to learn from a teacher conditioned on the corresponding clue interval. Importantly, the clue annotations are used only to construct the privileged teacher input during training. `Clue-OPSD` does not use ground-truth answer labels as supervision and does not require a separately pretrained or larger teacher model. At inference time, both the clue annotations and EMA teacher are removed, and the trained student directly processes the full video using the standard VLM inference pipeline, without additional temporal localization or auxiliary inference-time modules.

We evaluate `Clue-OPSD` across multiple long-video understanding benchmarks using Qwen3.5 models at different scales. Across model sizes and evaluation settings, `Clue-OPSD` consistently improves over the corresponding backbone models and achieves strong performance against supervised post-training baselines. The contributions of this paper are summarized as follows:

- We show that focusing the visual input on short annotated clue intervals containing question-relevant evidence can consistently outperform full-video input across different VLM scales, even with fewer input frames.
- We propose `Clue-OPSD`, a clue-privileged on-policy self-distillation framework that uses clue intervals as privileged visual supervision for a full-video student without relying on ground-truth answer labels or an external teacher model.
- We demonstrate consistent improvements across multiple Qwen3.5 model scales and long-video understanding benchmarks while retaining the standard full-video VLM inference pipeline without additional inference-time modules.

2 Related Work

2.1 Long-Video Understanding

With the rapid progress of foundation vision-language models (VLMs) (Chen et al., 2024; Bai et al., 2025; Zhu et al., 2025; Team, 2026; Wang et al., 2026; Wang & Huang, 2026), long-video understanding has improved substantially. Standard backbone models typically sample frames across the full video and process them together with the textual query (Ranasinghe et al., 2025; Shu et al., 2025; Li et al., 2026c; Lin et al., 2026). This paradigm enables direct reasoning over increasingly

long temporal contexts, while recent work has further explored more efficient representations and compression strategies for long-video inputs (Shu et al., 2025; Li et al., 2026c; Zhang et al., 2025).

More recently, *thinking-with-videos* methods have explored a more active inference paradigm by searching for, retrieving, or temporally grounding question-relevant video segments and iteratively inspecting them during reasoning (Wang et al., 2024; Yang et al., 2025; Yuan et al., 2025; Liu et al., 2026; Yang et al., 2026; Zhang et al., 2026). While effective, these approaches typically require multiple rounds of temporal search, segment inspection, or tool interaction, leading to more complex and time-consuming inference. In contrast, our work uses annotated temporal clue intervals only during training as privileged visual supervision, while retaining the standard full-video VLM pipeline at inference time.

2.2 On-Policy Distillation and Self-Distillation

On-policy distillation (OPD) trains a student on trajectories sampled from its current policy while using a teacher to provide token-level distribution supervision (Agarwal et al., 2024). By matching teacher and student predictions on student-generated trajectories, OPD reduces the mismatch between training-time supervision and the states encountered during autoregressive inference. Recent studies have extended this paradigm through different teacher constructions, conditioning contexts, and supervision strategies (Yu et al., 2026; Tan & Hong, 2026; Sang et al., 2026; Song & Zheng, 2026; Ye et al., 2026; Hou et al., 2026). In particular, on-policy self-distillation (OPSD) (Zhao et al., 2026) allows the same underlying model to serve as both teacher and student under asymmetric conditioning, where privileged information available only to the teacher provides the supervision signal.

This idea has recently been extended to multimodal learning. Vision-OPD (Yuan et al., 2026) uses evidence-centered image crops as privileged visual context for a full-image student, while Visual-OPSD (Li et al., 2026b) and Visual Contrastive Self-Distillation (Liang et al., 2026) explore cross-modal and contrastive forms of asymmetric visual supervision. Video-OPD (Li et al., 2026a) applies on-policy distillation to temporal video grounding. Our work focuses instead on long-video understanding, where question-relevant temporal clue intervals are used as privileged visual context to supervise a student operating directly on the full video.

3 Method

3.1 Preliminaries

Video Question Answering. Given a raw video $\mathcal{V}$ and a question q, we first convert the video into a sequence of temporally sampled frames:

$$V = \{v_1, \ldots, v_N\}, \tag{1}$$

where N denotes the number of sampled frames used as visual input to the model. A vision-language model (VLM) parameterized by θ then autoregressively generates an answer $y = (y_1, \ldots, y_T)$ according to:

$$p_\theta(y \mid V, q) = \prod_{t=1}^{T} p_\theta\left(y_t \mid V, q, y_{<t}\right), \tag{2}$$

where $y_{<t}$ denotes the previously generated tokens.

On-Policy Distillation. On-policy distillation (OPD) performs token-level knowledge distillation along trajectories generated by the current student policy. Given an input x, the student first samples an autoregressive response:

$$\hat{y} \sim p_{\theta_S}(\cdot \mid x), \tag{3}$$

where θ_S denotes the parameters of the student model. The sampled response is then used as the shared autoregressive trajectory for both the student and teacher. At each decoding step t, conditioned on the same prefix $\hat{y}_{<t}$, the two models produce next-token distributions:

$$\begin{cases} p_S^t = p_{\theta_S}\left(\cdot \mid x, \hat{y}_{<t}\right), \\ p_T^t = p_{\theta_T}\left(\cdot \mid x, \hat{y}_{<t}\right), \end{cases} \tag{4}$$

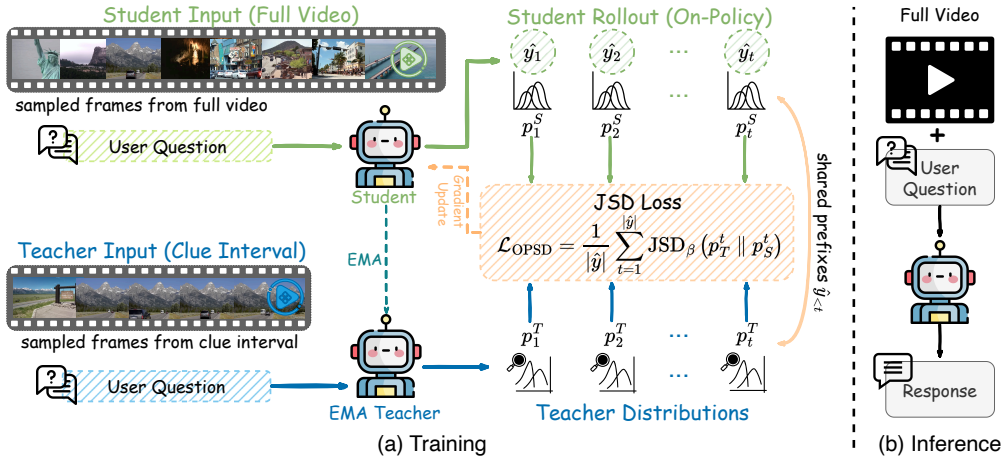

Figure 2: Overview of `Clue-OPSD`. (a) During training, the student processes frames sampled from the full video and generates an on-policy trajectory, while the EMA self-teacher is conditioned on frames sampled from the corresponding clue interval. Both branches are evaluated along the same student-generated prefixes, and their next-token distributions are aligned using the JSD objective. The teacher parameters are updated through an exponential moving average of the student. (b) At inference time, the clue interval and teacher are removed, and only the student is used with the standard full-video VLM pipeline.

where θ_T denotes the parameters of the teacher model, and p_S^t and p_T^t represent the student and teacher next-token probability distributions at decoding step t, respectively. The student is optimized by minimizing the token-level discrepancy between these two distributions:

$$\mathcal{L}_{\text{OPD}} = \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} D\left(p_S^t, p_T^t\right), \tag{5}$$

where $D(\cdot, \cdot)$ denotes a distribution-level distillation objective. By evaluating the teacher on student-generated prefixes, OPD provides dense supervision directly along the trajectories induced by the current student policy.

3.2 Empirical Observation: Temporal Clues Improve Prediction

Model	Full Video	Clue Interval	Δ
Qwen3.5-2B	45.33	58.63	+13.30
Qwen3.5-4B	50.50	64.10	+13.60
Qwen3.5-9B	53.47	68.43	+14.96
Mean Frames	1024.00	68.15	-955.85

Table 1: Accuracy (%) on 1,000 sampled CG-Bench questions using the full video or the annotated clue interval. Δ denotes the change from full-video input to clue-interval input. The last row reports the mean number of input frames under the same sampling configuration, with a maximum of 1,024 frames at 2 FPS.

Although current VLMs typically process the entire video for long-video question answering, whether the full temporal context is always beneficial for answering a given question remains unclear. To investigate this, we conduct a preliminary comparison on 1,000 randomly sampled multiple-choice questions from CG-Bench (Chen et al., 2025). In addition to question-answer annotations, CG-Bench provides clue intervals that contain the visual evidence required to answer each question. This enables us to directly compare model predictions under two visual conditions: the full video and the corresponding annotated clue interval. For each example, we keep the question, answer options, and decoding configuration identical across the two settings.

As shown in Table 1, using the annotated clue interval consistently yields higher accuracy than using the full video across different model scales. Both settings use the same frames-per-second (FPS) and maximum number of frames, while the clue interval is typically much shorter and therefore contains

fewer input frames than the full video. The improvement shows that focusing the visual input on the question-relevant temporal interval leads to more effective prediction than sampling frames from the entire video.

This observation naturally motivates our approach: instead of requiring clue annotations at inference time, we use them only during training to construct a privileged teacher and distill its predictions into a student that operates on the full video.

3.3 CLUE-OPSD: CLUE-PRIVILEGED ON-POLICY SELF-DISTILLATION

Building on the above observation, we introduce `Clue-OPSD`, an on-policy self-distillation framework for long-video understanding, which transfers the predictive behavior induced by clue-focused visual inputs to a student that operates on the full video. The detailed architecture is shown in Figure 2. Specifically, the student receives the full-video input, while the teacher is conditioned on the annotated clue interval under the same question and answer options. Rather than relying on a separately pretrained or larger teacher, we maintain the teacher as an exponential moving average (EMA) of the student parameters. As a result, the supervision gap is introduced by the visual context rather than model capacity: the teacher predicts from question-relevant temporal evidence, while the student learns to reproduce such predictions from the full video. The algorithm of `Clue-OPSD` is presented in Algorithm 1.

Asymmetric Visual Contexts. For each training instance, let $V = \{v_1, \ldots, v_N\}$ denote the sequence of N frames sampled from the full video, and let $V^\star = \{v_1^\star, \ldots, v_M^\star\}$ denote the sequence of M frames sampled from the corresponding annotated clue interval. Both inputs follow the same FPS and maximum-frame constraint. Therefore, we have $M \leq N$, with $M = N$ when both the clue interval and the full video reach the maximum number of sampled frames. The student is conditioned on V, while the teacher is conditioned on $V^\star$. Both branches receive exactly the same textual input, including the question q and its candidate answer options, such that the only difference between them lies in the visual context.

The student, parameterized by θ_S, first generates an on-policy response $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ from the full-video input:

$$\hat{y} \sim p_{\theta_S}(\cdot \mid V, q), \tag{6}$$

where T denotes the number of generated tokens. The generated response is then fixed and used as the shared autoregressive trajectory for both branches. At decoding step t, let $\hat{y}_{<t}$ denote the prefix consisting of all tokens generated before step t. The student and teacher next-token distributions are respectively defined as:

$$\begin{cases} p_S^t = p_{\theta_S}\left(\cdot \mid V, q, \hat{y}_{<t}\right), \\ p_T^t = p_{\theta_T}\left(\cdot \mid V^\star, q, \hat{y}_{<t}\right), \end{cases} \tag{7}$$

where θ_T denotes the teacher parameters, and p_S^t and p_T^t denote the student and teacher next-token probability distributions at step t, respectively.

On-Policy Distribution Alignment. Given the student-generated trajectory $\hat{y}$, we align the student and teacher next-token distributions at each decoding step using generalized Jensen–Shannon divergence (JSD). Specifically, the distillation objective is:

$$\mathcal{L}_{\mathrm{OPSD}} = \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \mathrm{JSD}_\beta\left(p_T^t \parallel p_S^t\right), \tag{8}$$

where $|\hat{y}|$ denotes the length of the student-generated response. For each decoding step, the generalized JSD is defined as:

$$\mathrm{JSD}_\beta\left(p_T^t \parallel p_S^t\right) = \beta \cdot \mathrm{KL}\left(p_T^t \parallel m^t\right) + (1-\beta) \cdot \mathrm{KL}\left(p_S^t \parallel m^t\right), \tag{9}$$

where

$$m^t = \beta \cdot p_T^t + (1-\beta) \cdot p_S^t \tag{10}$$

is the mixture distribution at step t, and β controls the relative contribution of the two KL terms. During optimization, the teacher distribution is treated as a fixed target and gradients are propagated only through the student branch.

Following Vision-OPD (Yuan et al., 2026), we adopt top-K logit distillation to reduce the memory and computation required for full-vocabulary distribution matching. Specifically, we retain the top-K tokens selected from the student distribution, together with the corresponding teacher logits and the remaining tail probability, and compute the distillation objective on this compressed distribution.

By minimizing $\mathcal{L}_{\mathrm{OPSD}}$, the student is encouraged to reproduce the clue-conditioned teacher distribution while operating on the full-video input. Since the shared trajectory is generated by the current student policy, the resulting token-level supervision remains on-policy throughout training.

EMA Self-Teacher. We construct the teacher as an EMA of the student rather than introducing a separately pretrained model, following (Zhao et al., 2026). After each optimization step, the teacher parameters are updated as

$$\theta_T \leftarrow (1-\alpha)\cdot\theta_T + \alpha\cdot\theta_S, \tag{11}$$

where α denotes the EMA update coefficient. The teacher is not optimized by back-propagation; instead, it evolves only through the EMA update and serves as a slowly changing reference for distillation. This design preserves the self-distillation setting while providing a more stable teacher distribution than directly reusing the current student parameters. Combined with the clue-privileged visual input $V^\star$, the EMA teacher provides the student with a stable prediction target conditioned on question-relevant temporal evidence.

Training and Inference. During training, the updated student parameters are synchronized with the rollout model after each optimization step so that subsequent responses are generated by the current student policy. The clue interval annotations and EMA teacher are used only for training. At inference time, both are discarded, and the student directly processes the full video following the standard VLM inference pipeline, requiring no additional clue information at inference time.

4 Experiments

4.1 Experimental Setting

Data Construction. We construct our training data from CG-Bench (Chen et al., 2025), a long-video understanding benchmark that provides question-answer pairs together with annotated temporal clue intervals indicating the evidence required to answer each question. We focus on the multiple-choice questions (MCQs) and retain only samples with valid temporal clue annotations. From the resulting set, we randomly sample 5,000 question-answer instances for training, covering 1,206 distinct long videos. To avoid data leakage, we further verify that the videos used for training do not overlap with those in any of our evaluation benchmarks. This ensures that the reported results reflect generalization to unseen videos rather than memorization of training content.

Models and Baselines. We conduct experiments on the Qwen3.5 model family (Team, 2026), including Qwen3.5-2B, Qwen3.5-4B, and Qwen3.5-9B, to evaluate the effectiveness of our proposed `Clue-OPSD` across different model scales. For each model size, we compare `Clue-OPSD` with the corresponding vanilla Qwen3.5 model, as well as models trained on the same data using SFT, GRPO, and standard OPSD. For SFT, the ground-truth answer is directly used as the target response for teacher-forced likelihood optimization. GRPO uses the ground-truth answer to determine the reward for sampled responses, while standard OPSD conditions the teacher on the ground-truth answer as privileged information and keeps the student on the original full-video input. In contrast, `Clue-OPSD` uses only temporal clue annotations as privileged supervision and does not use ground-truth answers during training.

Training Details. We initialize both the student and EMA teacher from the same pretrained Qwen3.5 checkpoint and perform full-parameter fine-tuning. All models are trained on 8 NVIDIA B200 GPUs with a learning rate of 2×10^{-6}, an effective batch size of 32, and up to 300 optimization steps. We adopt a cosine learning-rate schedule with a warmup ratio of 0.03. We use generalized JSD with $\beta = 0.5$, top-K logit distillation with $K = 100$, and a distillation temperature of 1.0. The EMA update coefficient is set to $\alpha = 0.05$. The student weights are synchronized with the vLLM rollout engine after every optimization step to ensure on-policy sampling. During rollout, we use a sampling temperature of 1.0, top-$p = 0.95$, and top-$k = 20$, with a maximum generation length of

Model	Setting	Video-MME (w/o Subtitles)			Video-MME (w/ Subtitles)			Average
		Short	Medium	Long	Short	Medium	Long	
Qwen3.5-2B	Vanilla	72.40	57.93	46.73	75.30	65.22	57.24	62.47
	SFT	66.15	57.00	47.33	70.48	62.56	53.92	59.57
	GRPO	71.33	61.22	50.77	75.11	66.11	58.00	63.76
	OPSD	69.70	59.44	46.67	72.93	64.67	52.78	61.03
	Clue-OPSD	**72.78**	**62.67**	**51.33**	**75.85**	**67.78**	**59.11**	**64.92**
Qwen3.5-4B	Vanilla	78.02	67.09	60.16	80.33	74.49	70.56	71.78
	SFT	76.78	68.48	58.74	79.18	75.41	67.85	71.07
	GRPO	77.96	69.15	60.70	80.37	74.00	69.89	72.01
	OPSD	**80.37**	69.85	61.56	**82.45**	74.63	71.19	73.34
	Clue-OPSD	80.31	**72.09**	**62.29**	82.29	**75.65**	**72.48**	**74.19**
Qwen3.5-9B	Vanilla	81.04	73.42	64.47	83.11	77.91	73.42	75.56
	SFT	78.37	72.18	61.48	81.22	77.11	70.82	73.53
	GRPO	81.67	74.11	65.22	84.66	78.07	74.00	76.29
	OPSD	81.82	74.52	65.41	84.59	77.92	74.65	76.49
	Clue-OPSD	**82.30**	**74.89**	**66.33**	**85.33**	**78.33**	**74.67**	**76.98**

Table 2: Results on Video-MME across different video durations under settings with and without subtitles. Each experiment is repeated three times, and the mean performance is reported.

2,048 tokens. Both the student and teacher videos are sampled at 2 FPS with at most 256 frames during training.

Benchmarks. We evaluate our method on five video understanding benchmarks: Video-MME (Fu et al., 2025), LVBench (Wang et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and MMVU (Zhao et al., 2025). LVBench, LongVideoBench, and MLVU primarily evaluate long-video understanding over extended temporal contexts, while Video-MME provides a broader evaluation across different video durations. For Video-MME, we report performance on the short, medium, and long subsets under both with- and without-subtitle settings. For MLVU, we report the multiple-choice macro-average (M-Avg) following the standard evaluation protocol. For MMVU, we report the overall benchmark accuracy. Together, these benchmarks cover a diverse range of video durations, temporal reasoning requirements, and question-answering settings.

Inference Details. At inference time, the EMA teacher and clue annotations are removed, and the student directly processes the full video. We sample videos at 2 FPS and use benchmark-specific maximum frame budgets: 768, 1,024, and 2,048 frames for the short, medium, and long subsets of Video-MME, respectively; 2,048 frames for LVBench, LongVideoBench, and MLVU; and 1,024 frames for MMVU. Following the official Qwen3.5 recommended configuration for non-thinking inference, we use a sampling temperature of 0.7, top-$p = 0.8$, top-$k = 20$, a presence penalty of 1.5, and a repetition penalty of 1.0, with a maximum generation length of 2,048 tokens. The same decoding configuration is used across all evaluated models and benchmarks. We repeat each evaluation three times and report the mean performance over the three runs.

4.2 Experimental Results

Overall Performance. Table 3 compares `Clue-OPSD` with both publicly reported VLMs and controlled Qwen3.5-based baselines. Across all three model scales, `Clue-OPSD` consistently improves over the corresponding vanilla backbones on the four evaluated benchmarks. The gains are particularly pronounced for smaller models. For Qwen3.5-2B, `Clue-OPSD` improves MLVU, LVBench, LongVideoBench, and MMVU by 6.94, 3.45, 9.81, and 3.41 points, respectively. Similar improvements are observed for Qwen3.5-4B, with gains of 8.23 points on MLVU and 5.42 points on LVBench over the vanilla model. These results show that clue-privileged self-distillation provides consistent benefits across different long-video reasoning settings and model capacities.

Compared with supervised post-training baselines, `Clue-OPSD` also achieves competitive performance despite not using ground-truth answer labels during training. On Qwen3.5-2B and Qwen3.5-4B, `Clue-OPSD` outperforms SFT, GRPO, and standard OPSD across all four benchmarks. For Qwen3.5-9B, `Clue-OPSD` achieves the best results on LVBench, LongVideoBench, and MMVU, while remaining competitive with OPSD on MLVU. Notably, these improvements are obtained using only temporal clue annotations as privileged supervision, suggesting that question-relevant tempo-

Model	Setting	MLVU	LVBench	LongVideoBench	MMVU
Frontier Proprietary Models					
GPT-5	Reported	77.7	65.2	72.6	–
Gemini 3 Pro	Reported	75.7	77.0	75.9	77.5
Gemini 2.5 Pro	Reported	–	75.7	76.8	75.8
Gemini 2.5 Flash	Reported	75.1	64.9	73.1	70.3
Strong Open-Source VLMs					
GLM-4.1V-9B	Reported	56.6	44.0	65.7	–
MiniCPM-V-4.5-8B	Reported	60.6	50.4	63.9	–
Molmo2-8B	Reported	60.2	52.8	67.5	–
InternVL3.5-8B	Reported	53.2	43.4	62.1	–
InternVL3.5-38B	Reported	77.0	–	65.7	–
InternVL3-78B	Reported	79.5	–	65.7	–
Qwen3.5-Based Models					
Qwen3.5-2B	Vanilla	59.34	48.43	51.28	55.84
	SFT	65.74	50.14	51.78	55.04
	GRPO	65.89	50.43	60.23	57.59
	OPSD	65.67	50.48	57.44	57.92
	Clue-OPSD	**66.28**	**51.88**	**61.09**	**59.25**
Qwen3.5-4B	Vanilla	68.45	54.45	61.58	63.84
	SFT	71.54	55.59	58.14	63.36
	GRPO	75.98	59.22	64.72	65.92
	OPSD	75.75	59.00	64.94	65.56
	Clue-OPSD	**76.68**	**59.87**	**66.62**	**66.99**
Qwen3.5-9B	Vanilla	74.91	60.31	66.72	72.96
	SFT	73.28	57.78	59.51	69.76
	GRPO	80.93	60.96	68.42	73.87
	OPSD	**81.05**	61.30	67.93	73.76
	Clue-OPSD	80.27	**61.53**	**68.49**	**74.88**

Table 3: Comparison with frontier proprietary models, strong open-source VLMs, and controlled Qwen3.5-based baselines on long-video understanding benchmarks. External model results are taken from publicly reported evaluations. All Qwen3.5-based results are averaged over three runs.

ral evidence provides an effective learning signal beyond direct answer supervision. In addition, Clue-OPSD remains competitive with substantially larger open-source and proprietary VLMs on several benchmarks, while retaining the original Qwen3.5 architecture and standard full-video inference pipeline.

Divergence Objective	MLVU	LVBench	LongVideoBench	MMVU
Forward KL ($\mathrm{KL}(p_T \| p_S)$)	76.21	59.14	65.84	66.41
Reverse KL ($\mathrm{KL}(p_S \| p_T)$)	75.86	58.93	66.08	66.62
JSD ($\beta = 0.5$)	**76.68**	**59.87**	**66.62**	**66.99**

Table 4: Effect of different divergence objectives for token-level distribution alignment on Qwen3.5-4B. All other training and inference configurations are kept unchanged.

Performance across Video Durations. Table 2 reports detailed results on Video-MME across short, medium, and long videos under both subtitle settings. Clue-OPSD consistently improves the average performance of Qwen3.5-2B, 4B, and 9B from 62.47, 71.78, and 75.56 to 64.92, 74.19, and 76.98, respectively. The improvements are especially clear on the medium- and long-video subsets, where identifying relevant temporal evidence becomes increasingly important. For example, on Qwen3.5-2B without subtitles, Clue-OPSD improves the medium- and long-video accuracy from 57.93 and 46.73 to 62.67 and 51.33, respectively. Similar trends hold for larger models and when subtitles are available.

Compared with the supervised baselines, Clue-OPSD achieves the highest average Video-MME accuracy at all three model scales. In particular, it improves over the strongest competing post-training baseline by 1.16, 0.85, and 0.49 points for Qwen3.5-2B, 4B, and 9B, respectively. The

Teacher Visual Input	Additional Privilege	MLVU	LVBench	LongVideoBench	MMVU
Full Video	GT Answer	75.75	59.00	64.94	65.56
Full Video	Textual Clue	73.66	58.70	64.17	63.57
Random Interval	-	52.32	37.85	51.78	57.44
Clue Interval	-	**76.68**	**59.87**	**66.62**	**66.99**

Table 5: Comparison of different privileged teacher constructions for on-policy distillation on Qwen3.5-4B. "Textual Clue" provides the annotated temporal interval in the prompt while retaining the full-video visual input. "Random Interval" uses a randomly sampled interval with the same duration as the ground-truth clue interval. The clue-interval setting corresponds to the default configuration of `Clue-OPSD`.

consistent gains across video durations and subtitle settings further indicate that the benefit of clue-privileged training is not tied to a particular evaluation condition.

4.3 Ablation Studies

Effect of Divergence Objective. We study the effect of different divergence objectives used for token-level distribution matching between the teacher and student. Specifically, we compare forward KL divergence, reverse KL divergence, and Jensen–Shannon divergence (JSD) with $\beta = 0.5$. We keep all other training configurations unchanged and conduct this ablation on Qwen3.5-4B. As shown in Table 4, JSD achieves the best performance across the four benchmarks. These results suggest that JSD provides a more balanced distribution-matching objective for clue-privileged on-policy self-distillation. We therefore adopt JSD with $\beta = 0.5$ as the default divergence objective in all main experiments.

Effect of Privileged Clue Utilization. We investigate how temporal clue annotations should be incorporated into the privileged teacher. Besides the default `Clue-OPSD` setting, where the teacher directly observes the annotated clue interval, we compare a full-video teacher with ground-truth answers, a full-video teacher with the clue interval provided only as textual side information, and a teacher conditioned on a randomly sampled interval with the same duration. As shown in Table 5, directly conditioning the teacher on the annotated clue interval achieves the best performance across all four benchmarks. For example, on MLVU, the clue-interval teacher reaches 76.68, compared with 73.66 for the textual-clue variant and 52.32 for the random-interval variant. The large drop with random intervals shows that the gain does not come from using a shorter visual input alone, but from exposing the teacher to question-relevant temporal evidence. Moreover, the clue-interval teacher consistently outperforms the answer-privileged OPSD baseline. These results support using clue interval as the privileged condition for on-policy self-distillation.

Robustness to Clue Boundary Precision. Although `Clue-OPSD` uses temporal clue annotations as privileged supervision, accurately localizing the exact boundaries of question-relevant evidence can be difficult in practice. We therefore evaluate the sensitivity of our method to coarser clue annotations and find that expanding each interval by up to ± 60s results in only a 1.78-point average performance drop, with full results reported in Appendix C.

5 Conclusion

In this work, we studied how temporal clue annotations can be used as privileged supervision for long-video understanding. Our empirical analysis shows that short question-relevant clue intervals can provide more effective visual context than the corresponding full videos, even with fewer input frames. Building on this finding, we introduced `Clue-OPSD`, a clue-privileged on-policy self-distillation framework that uses a clue-conditioned self-teacher to supervise a full-video student without ground-truth answer labels. Extensive experiments across multiple benchmarks and Qwen3.5 model scales demonstrate consistent improvements over the original backbones and strong performance against supervised post-training baselines. These results validate the effectiveness of temporal clues as privileged supervision for improving full-video VLMs.

AI USE STATEMENT

In this work, we used generative AI tools only to improve the readability and linguistic clarity of the manuscript. Generative AI tools were not used to develop research ideas, design the methodology or experiments, generate or analyze data, implement the proposed methods, or interpret experimental results. All research contributions, experimental design, implementation, analysis, and conclusions were produced by the authors. We reviewed and verified all AI-assisted edits and take full responsibility for the final content of this work.

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.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

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.

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, et al. Dash: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models. *arXiv preprint arXiv:2608.06243*, 2026.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026a.

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

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026c.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, et al. Unleashing hour-scale video training for long video-language understanding. *Advances in Neural Information Processing Systems*, 38:17523–17552, 2026.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael Ryoo. Understanding long videos with multimodal language models. In *International Conference on Learning Representations*, volume 2025, pp. 11810–11835, 2025.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. On-policy self-distillation for reasoning compression. *arXiv e-prints*, pp. arXiv–2603, 2026.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169. IEEE, 2025.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models. *arXiv preprint arXiv:2605.17497*, 2026.

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

Kaishen Wang and Heng Huang. Unsafe by reciprocity: How generation-understanding coupling undermines safety in unified multimodal models. *arXiv preprint arXiv:2603.27332*, 2026.

Kaishen Wang, Tong Zheng, Xuehao Cui, Ruibo Chen, Tianyi Xiong, and Heng Huang. Mitigating factual hallucination in large reasoning models via mixed-mode advantage regularization. *arXiv preprint arXiv:2607.05861*, 2026.

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, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

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.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20168–20179. IEEE, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Fangxu Yu, Zinan Lin, Xiaodong Liu, Weijia Xu, Michael Xu, Tianyi Zhou, and Jianfeng Gao. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026.

Huaying Yuan, Zheng Liu, Junjie Zhou, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv e-prints*, pp. arXiv–2506, 2025.

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.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069. IEEE, 2025.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026.

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

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489. IEEE, 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.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

This appendix provides additional implementation details, analyses, and discussions that complement the main paper. Appendix A presents the evaluation prompts used across all benchmarks. Appendix B provides the complete training procedure of `Clue-OPSD`. Appendix C reports additional ablations on the sensitivity of `Clue-OPSD` to the temporal precision of clue annotations. Appendix D further discusses the difference between clue-conditioned supervision and alternative forms of answer-based or privileged supervision. Finally, Appendix E provides a local output-space interpretation of `Clue-OPSD`, showing that its JSD objective can be viewed as a Fisher-weighted predictive consistency penalty between the clue-conditioned teacher and the full-video student.

A PROMPTS

We use the official evaluation prompts for most benchmarks whenever available. For completeness and reproducibility, we present the prompt templates used for benchmark evaluation below.

Prompt for Video-MME Evaluation without Subtitles

```
Select the best answer to the following multiple-choice question
based on the video.  Respond with only the letter (A, B, C, or D) of
the correct option.

Question:  {question}
{choices}
The best answer is:
```

Figure 3: Prompt template used for Video-MME evaluation without subtitles.

Prompt for Video-MME Evaluation with Subtitles

```
This video's subtitles are listed below:
{subtitles}

Select the best answer to the following multiple-choice question
based on the video.  Respond with only the letter (A, B, C, or D) of
the correct option.

Question:  {question}
{choices}

The best answer is:
```

Figure 4: Prompt template used for Video-MME evaluation with subtitles.

Prompt for LVBench Evaluation

```
Carefully watch the video and pay attention to the cause and
sequence of events, the detail and movement of objects and the
action and pose of persons.

Based on your observations, select the best option that accurately
addresses the question.

Question: {question}
{choices}

Answer with the option's letter from the given choices directly and
only give the best option.
```

Figure 5: Prompt template used for LVBench evaluation.

Prompt for LongVideoBench Evaluation

```
{question}
{options}

Answer with the option's letter from the given choices directly.
```

Figure 6: Prompt template used for LongVideoBench evaluation.

Prompt for MLVU Evaluation

```
Carefully watch this video and pay attention to every detail.
Based on your observations, select the best option that accurately
addresses the question.

Question: {question}
Options: {options}

Only give the best option.
```

Figure 7: Prompt template used for MLVU evaluation.

Prompt for MMVU Evaluation

```
{question}
{options}

Answer with the option's letter from the given choices directly.
```

Figure 8: Prompt template used for MMVU evaluation.

B ALGORITHM

Algorithm 1 summarizes the training procedure of `Clue-OPSD`. For each training instance, the student first generates an on-policy response from the full-video input. The same student-generated prefixes are then used to evaluate both the full-video student and the clue-conditioned EMA teacher.

Algorithm 1 Training Procedure of `Clue-OPSD`

Require: Training set $\mathcal{D} = \{(\mathcal{V}, q, I^\star)\}$, pretrained VLM parameters θ, EMA coefficient α, JSD coefficient β, top-K value K
Ensure: Trained student parameters θ_S

Initialize student and teacher: $\theta_S \leftarrow \theta$, $\theta_T \leftarrow \theta$
Initialize rollout model with θ_S
for each optimization step **do**
  Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$
  Initialize batch loss $\mathcal{L} \leftarrow 0$
  for each training instance $(\mathcal{V}, q, I^\star) \in \mathcal{B}$ **do**
    Sample full-video frames V from $\mathcal{V}$
    Sample clue frames $V^\star$ from the annotated clue interval $I^\star$
    Generate an on-policy response from the full-video student: $\hat{y} \sim p_{\theta_S}(\cdot \mid V, q)$
    for $t = 1, \ldots, |\hat{y}|$ **do**
      Compute the student distribution: $p_S^t = p_{\theta_S}(\cdot \mid V, q, \hat{y}_{<t})$
      Compute the teacher distribution on the same prefix: $p_T^t = p_{\theta_T}(\cdot \mid V^\star, q, \hat{y}_{<t})$
      Construct top-K compressed student and teacher distributions
    end for
    Accumulate the instance-level distillation loss:

$$\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \mathrm{JSD}_\beta \left(p_T^t \parallel p_S^t \right)$$

  end for
  Average the loss over the mini-batch: $\mathcal{L} \leftarrow \mathcal{L}/|\mathcal{B}|$
  Update the student parameters θ_S by back-propagating $\mathcal{L}$
  Update the EMA teacher: $\theta_T \leftarrow (1-\alpha)\theta_T + \alpha\theta_S$
  Synchronize the rollout model with the updated student θ_S
end for
return θ_S

Their next-token distributions are aligned through the JSD distillation objective, with gradients applied only to the student. After each optimization step, the teacher is updated as an exponential moving average of the latest student parameters, and the rollout model is synchronized accordingly.

C Ablation Study

Sensitivity to Clue Interval Precision. We further investigate how sensitive `Clue-OPSD` is to the temporal precision of clue annotations. In practice, precisely identifying the temporal boundaries of question-relevant evidence may require additional annotation effort, while coarser temporal annotations are easier to obtain. To simulate increasingly imprecise clue annotations, we expand each annotated clue interval by 10, 30, and 60 seconds on both temporal boundaries, respectively, while clipping the expanded interval to the valid video range. These expanded intervals preserve the original question-relevant evidence while introducing increasing amounts of surrounding temporal context. We train `Clue-OPSD` on Qwen3.5-4B under each setting, with all other training configurations kept unchanged.

As shown in Table 6, the original clue intervals achieve the best average performance of 67.54, while the expanded settings remain competitive, reaching 66.02, 66.66, and 65.76 under $\pm$10s, $\pm$30s, and $\pm$60s expansion, respectively. Notably, the degradation is not monotonic with respect to the expansion range: the $\pm$30s setting outperforms $\pm$10s on average and remains close to the original setting on LVBench and MMVU. Even when the clue interval is expanded by 60 seconds on both sides, the average performance decreases by only 1.78 points. These results suggest that

Table 6: Sensitivity of Clue-OPSD to the temporal precision of clue annotations on Qwen3.5-4B. Each annotated clue interval is expanded by the specified duration on both temporal boundaries, while all other training configurations are kept unchanged.

Teacher Clue Interval	MLVU	LVBench	LongVideoBench	MMVU	Average
Original	76.68	59.87	66.62	66.99	67.54
Expanded $\pm$10s	74.34	59.46	65.32	64.96	66.02
Expanded $\pm$30s	74.09	59.78	65.92	66.83	66.66
Expanded $\pm$60s	73.56	59.01	64.32	66.13	65.76

Clue-OPSD does not critically depend on highly precise temporal boundaries. Instead, retaining the question-relevant evidence within the privileged teacher input appears to be more important than tightly localizing its exact temporal extent.

D Discussion

The advantage of Clue-OPSD over SFT, GRPO, and standard OPSD can be attributed to the form of supervision provided during training. SFT directly optimizes the ground-truth answer, but does not explicitly expose which visual content in a long video supports that answer. GRPO performs on-policy optimization, yet its supervision is mainly determined by final-answer correctness and is therefore relatively sparse over the generated trajectory. Standard OPSD provides dense on-policy distribution supervision, but the privileged information is typically centered on the ground-truth answer rather than the visual evidence supporting it. In contrast, Clue-OPSD conditions the teacher on the question-relevant clue interval, so the supervision is generated from a visual context in which the evidence needed for the current question is directly emphasized.

This distinction also helps explain why answer-privileged OPSD can underperform Clue-OPSD despite having access to the ground-truth answer. The answer specifies *what* the correct prediction is, but does not indicate *which visual evidence* should support that prediction. By contrast, the clue interval changes the teacher's visual condition itself, allowing its next-token distribution to reflect predictions made from concentrated question-relevant evidence rather than from the full-video context. Importantly, this supervision can still be useful even when the clue-conditioned model in Table 1 is not perfectly accurate, since Clue-OPSD does not distill its final answer as a hard pseudo-label. Instead, it matches dense token-level predictive distributions induced by an evidence-focused visual condition along student-generated trajectories. Moreover, the teacher used during training is an EMA of the evolving student rather than a fixed copy of the pretrained model. Our ablations support this interpretation: replacing the clue with a random interval causes a large drop, while providing the clue only as textual side information is also consistently weaker than directly conditioning the teacher on the clue interval. These results suggest that the benefit of Clue-OPSD comes not from shorter inputs or privileged information alone, but from using question-relevant visual evidence as the privileged condition for on-policy self-distillation.

E Theoretical Interpretation of Clue-OPSD

We provide a local analysis of Clue-OPSD directly in the predictive distribution space. Specifically, it characterizes the local geometry of the distillation objective when the teacher and student predictions are locally aligned, which is precisely the regime encouraged by distribution matching during training.

For clarity, we consider a fixed autoregressive prefix $h_t = (q, \hat{y}_{<t})$ and omit the decoding-step superscript t. Let

$$g_T = g_{\theta_T}(V^\star, h_t), \qquad g_S = g_{\theta_S}(V, h_t) \tag{12}$$

denote the teacher and student next-token logits, respectively, where $V^\star$ is the clue-interval input and V is the full-video input. Their corresponding predictive distributions are

$$p_T = \mathrm{softmax}(g_T), \qquad p_S = \mathrm{softmax}(g_S). \tag{13}$$

Because the softmax function is invariant to adding the same constant to all logits, only the component of the logit difference orthogonal to the all-ones direction affects the predictive distribution. Let $d = |\mathcal{V}|$ denote the vocabulary size and define the centering operator

$$P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top. \tag{14}$$

We then define the effective logit discrepancy as

$$\Delta g = P(g_S - g_T). \tag{15}$$

Since adding a constant offset does not affect the softmax distribution, p_S can equivalently be written as the softmax of $g_T + \Delta g$. The following result therefore characterizes the generalized Jensen–Shannon divergence used by `Clue-OPSD` in the local neighborhood of the teacher prediction. For analytical clarity, we consider the full-distribution JSD objective; the top-K logit distillation used in practice serves as its compressed computational approximation.

Proposition 1 (Local Geometry of `Clue-OPSD`). Suppose that the effective logit discrepancy $\|\Delta g\|$ is sufficiently small. For the generalized Jensen–Shannon divergence

$$\mathrm{JSD}_\beta(p_T \| p_S) = \beta\,\mathrm{KL}(p_T \| m) + (1-\beta)\,\mathrm{KL}(p_S \| m), \tag{16}$$

where

$$m = \beta p_T + (1-\beta) p_S, \tag{17}$$

we have

$$\mathrm{JSD}_\beta(p_T \| p_S) = \frac{\beta(1-\beta)}{2} \Delta g^\top F(p_T) \Delta g + O(\|\Delta g\|^3), \tag{18}$$

where

$$F(p_T) = \mathrm{diag}(p_T) - p_T p_T^\top \tag{19}$$

is the Fisher information matrix of the categorical predictive distribution. For the symmetric case used in our experiments, $\beta = 0.5$, this reduces to

$$\mathrm{JSD}(p_T \| p_S) = \frac{1}{8} \Delta g^\top F(p_T) \Delta g + O(\|\Delta g\|^3). \tag{20}$$

Proof. Let

$$\epsilon = p_S - p_T. \tag{21}$$

Because the constant component of the original logit difference has been removed by the projection P, a first-order expansion of the softmax around g_T gives

$$\epsilon = F(p_T)\Delta g + O(\|\Delta g\|^2). \tag{22}$$

Since

$$m = p_T + (1-\beta)\epsilon, \tag{23}$$

a second-order expansion of the KL divergence around p_T yields

$$\mathrm{KL}(p_T \| m) = \frac{(1-\beta)^2}{2} \sum_i \frac{\epsilon_i^2}{p_{T,i}} + O(\|\epsilon\|^3), \tag{24}$$

while

$$\mathrm{KL}(p_S \| m) = \frac{\beta^2}{2} \sum_i \frac{\epsilon_i^2}{p_{T,i}} + O(\|\epsilon\|^3). \tag{25}$$

Combining the two terms gives

$$\mathrm{JSD}_\beta(p_T \| p_S) = \frac{\beta(1-\beta)}{2} \epsilon^\top \mathrm{diag}(p_T)^{-1} \epsilon + O(\|\epsilon\|^3). \tag{26}$$

Using

$$F(p_T)\,\mathrm{diag}(p_T)^{-1} F(p_T) = F(p_T), \tag{27}$$

together with $\epsilon = F(p_T)\Delta g + O(\|\Delta g\|^2)$, we obtain

$$\mathrm{JSD}_\beta(p_T \| p_S) = \frac{\beta(1-\beta)}{2} \Delta g^\top F(p_T) \Delta g + O(\|\Delta g\|^3), \tag{28}$$

which establishes the result.

Separating Visual-Context Shift from EMA Teacher Lag. The teacher and student in `Clue-OPSD` differ in both their visual conditions and, due to the EMA update, their parameters. To make these two effects explicit, we decompose the effective logit discrepancy as

$$\Delta g = \Delta g_{\mathrm{ctx}} + \Delta g_{\mathrm{ema}}, \tag{29}$$

where

$$\Delta g_{\mathrm{ctx}} = P\left(g_{\theta_S}(V, h_t) - g_{\theta_S}(V^\star, h_t)\right) \tag{30}$$

measures the predictive shift caused by replacing the clue-focused visual condition with the full-video condition under the same student parameters, and

$$\Delta g_{\mathrm{ema}} = P\left(g_{\theta_S}(V^\star, h_t) - g_{\theta_T}(V^\star, h_t)\right) \tag{31}$$

captures the discrepancy introduced by the EMA parameter lag.

Substituting this decomposition into Proposition 1 shows that, locally, the `Clue-OPSD` objective penalizes the Fisher-weighted predictive discrepancy

$$\left(\Delta g_{\mathrm{ctx}} + \Delta g_{\mathrm{ema}}\right)^\top F(p_T)\left(\Delta g_{\mathrm{ctx}} + \Delta g_{\mathrm{ema}}\right). \tag{32}$$

This decomposition separates the predictive variation associated with the change in visual context from that introduced by the EMA parameter lag. Because the EMA teacher evolves smoothly from the student, the latter term is expected to remain controlled during training. When the EMA-induced discrepancy is relatively small, the local JSD penalty primarily reflects the predictive shift between the clue-focused and full-video visual conditions.

Interpretation. This analysis gives an output-space interpretation of `Clue-OPSD`. The clue-conditioned teacher provides an evidence-focused predictive anchor, whereas the student must make predictions from the full video. Locally, minimizing their JSD is equivalent to minimizing a Fisher-weighted squared discrepancy between their effective logits. Therefore, `Clue-OPSD` encourages predictive consistency between the clue-focused teacher and the full-video student and suppresses excessive prediction shifts associated with the change in visual context. Importantly, this interpretation does not require the clue-only and full-video representations themselves to be close; the local assumption instead concerns the effective predictive discrepancy, making the analysis particularly relevant to the regime approached as distillation progressively aligns the teacher and student predictions.