VIDEO EVIDENCE INDEXING: LEARNING WHERE TO LOOK FROM VIDEO PREVIEWS FOR TOKEN-BUDGETED LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video question answering is limited by the high cost of visual tokens and by the fixed context width of current VLMs. A long-video question may require broad temporal coverage, but the answer is often supported by only a compact set of moments. To locate these moments efficiently, we propose token-budgeted Video Evidence Indexing (VEI): given a dense low-resolution Video Preview, the model constructs a compact high-resolution Evidence Set for final reasoning. We treat VEI as a policy that must jointly solve *evidence localization*, which finds question-relevant moments, and *budget planning*, which decides where to spend the limited high-resolution frame budget. We implement this idea with an inference pipeline: the Video Preview provides cheap global coverage, Video Evidence Indexing constructs the Evidence Set, and Answer Generation combines both inputs for final VQA. To address missing frame-level supervision, we adopt privileged self-distillation, where an answer-aware teacher guides the normal test-time policy on student-generated indexing traces. We explore previews at 1, 6, 12, and 24 visual tokens per frame, training a single policy that supports all four resolutions. Experiments show that Video Evidence Indexing improves accuracy under limited visual budgets, and self-distillation further improves both QA accuracy and temporal evidence localization.

1 INTRODUCTION

Recent advances in Video Language Models (VLMs) (Qwen Team, 2026; Bai et al., 2025a; Comanici et al., 2025) have shown strong capabilities in understanding and reasoning over complex video content. However, long-video question answering (VQA) is still largely handled by fitting as much video input as possible into a finite context width. A long video may contain thousands of frames, while the answer to a question is often supported by only a compact set of moments (Tang et al., 2025b; Wang et al., 2025a). Processing the full video at high resolution provides broad temporal coverage and fine visual detail, but the visual-token cost quickly becomes too large. Uniformly sampling fewer frames reduces this cost, but it may miss the short event, object state, or scene change that determines the answer.

Token budget matters not only because fewer visual tokens reduce inference cost, but also because context width is a hard capacity limit. On common 24–48GB GPUs, an 8B VLM can often handle only tens of thousands of tokens before memory and latency become limiting. Long-video QA therefore needs methods that operate within a fixed context width, not only methods that reduce average cost. Token-compression methods reduce visual redundancy through pruning, merging, or compression inside the model (Yao et al., 2026; Yuan et al., 2025; Li et al., 2025), but they often require architectural changes or access to internal token states. Adaptive frame selection uses relevance, coverage, attention, or inference feedback to construct compact evidence sets (Tang et al., 2025a; Chen et al., 2026b; Wang et al., 2026b; Eltahir et al., 2026). These signals offer efficient selection, while leaving open how to train explicit evidence requests for the downstream question. Agentic video-observation methods plan and refine evidence acquisition through repeated or parallel tool calls (Li et al., 2026b; Xu et al., 2026; Yang et al., 2026c). Their observation structure differs from our single preview-to-index phase.

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

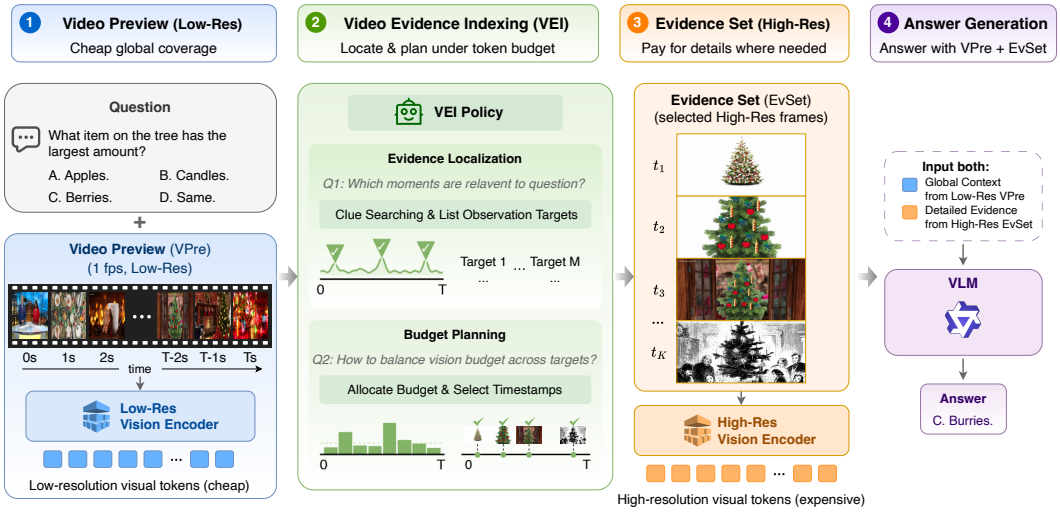


Figure 1: **Overview of Video Evidence Indexing.** A densely sampled low-resolution Video Preview (VPre) provides cheap global coverage under a strict context limit. We explore how VEI can use VPre to locate complementary evidence and reserve high-resolution detail for the moments that matter.

We study how best to use a global video preview to guide the acquisition of question-relevant detail under a limited context budget. Specifically, we explore densely sampled **Video Previews** (VPre) at four spatial resolutions $\tau_{\text{lo}} = 1, 6, 12$, and 24 visual tokens per frame (Figure 1, 2). Even when fine details are lost, low-resolution frames often preserve coarse cues such as scene layout, object presence, motion, and event boundaries. These cues may be insufficient for final answer verification, but they can help the model decide where to spend high-resolution tokens. We formulate this setting as token-budgeted **Video Evidence Indexing** (VEI). Given a densely sampled low-resolution VPre, the model performs VEI by mapping the preview and question to a compact **Evidence Set** (EvSet) under a fixed token budget. The goal is not to find a single salient keyframe, but to construct a question-relevant set of high-resolution frames that supports final reasoning while balancing coverage, diversity, and redundancy reduction.

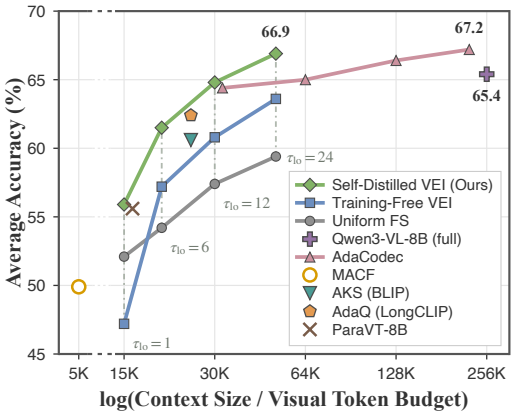


Figure 2: **Accuracy–context trade-off.** Mean accuracy across three datasets (Section 4). All methods use Qwen3-VL-8B. Comparators cover predictive compression (AdaCodec), multi-agent aggregation (MACF), encoder-based selection (AKS, AdaQ), and RL-trained video-tool acquisition (ParaVT).

Our formulation turns VQA into a policy-learning problem. A pretrained VLM has general visual reasoning ability, but it is not specifically trained to allocate a limited visual budget from a blurry preview. A strong VEI policy must jointly solve *evidence localization*, which identifies question-relevant moments, and *budget planning*, which decides where limited high-resolution frames should be spent. However, most VQA datasets provide answer labels without corresponding frame-level evidence annotations, making direct supervised fine-tuning (SFT) (Zhang et al., 2025d) difficult. Reinforcement learning with answer-level rewards, such as GRPO (Shao et al., 2024), can optimize final-answer correctness, but provides little direct guidance for individual selection decisions. A correct answer alone does not validate the preceding evidence choices, so even successful rollouts cannot reliably serve as reference labels for evidence localization and budget planning.

To guide these intermediate decisions without frame-level labels, we adopt privileged self-distillation (Zhao et al., 2026). A frozen teacher receives the video preview and correct answer, and is instructed to identify visible evidence supporting that answer. This provides pseudo-supervision for evidence localization and budget planning: the teacher can generate reference indexing traces for the student to learn from through off-policy distillation. In practice, we let the student generate

its own indexing traces while the teacher supplies token-level guidance along them, following prior work (Agarwal et al., 2024). This on-policy setting, also favored by our ablations, aligns supervision with the decisions the student actually encounters, teaching it what evidence to seek and how to distribute its limited observations. Our contributions are summarized as follows:

- We explore the use of dense Video Previews for long-video question answering across four preview token budgets, demonstrating their potential to guide evidence acquisition under limited context.
- We formulate **Video Evidence Indexing (VEI)** as a policy that couples evidence localization with budget planning, using explicit observation targets and target-wise timestamp selection to acquire complementary evidence.
- We address missing frame-level labels by deriving supervision for VEI decisions from an answer-aware teacher. Privileged self-distillation transfers this guidance across preview resolutions.

2 Related Work

Efficient video representations and memory. Token pruning and merging reduce visual redundancy (Yao et al., 2026; Yuan et al., 2025; Li et al., 2025; Du et al., 2026; Zeng et al., 2026), while learned compressors and reusable memories retain compact video context (Fei et al., 2026; Yamao et al., 2026; Yang et al., 2026a; Cai et al., 2026). AdaCodec encodes detailed reference frames and compact inter-frame changes (Hou et al., 2026); MACF distributes segment perception across agents communicating through latent messages (Chen et al., 2026a). These approaches compress representations or aggregate observations. VEI instead generates explicit evidence requests from a visual preview and reloads selected frames, preserving global coverage while reserving detailed visual tokens for question-relevant evidence.

Frame selection and budget allocation. Encoder-based selectors use question relevance, coverage, or diversity (Tang et al., 2025a; Zhang et al., 2025c; 2026b; Chen et al., 2026b; Shi et al., 2026); MEC constructs a ranking reusable across budgets. Other approaches use confidence-guided exploration (Zhu et al., 2026b), timestamped textual grounding (Wei et al., 2026), or learned contextual selectors (Buch et al., 2025; Qin et al., 2026). DAFS extracts frame scores from VLM attention and allocates candidate and resolution budgets (Wang et al., 2026b), while EcoFrame and GridProbe use inference feedback or multi-frame confidence probes (Li et al., 2026a; Eltahir et al., 2026). ReUnit allocates mixed-granularity visual layouts (Tang et al., 2026). VEI uses a dense visual preview to formulate explicit observation targets and distribute timestamps across them. This lets the policy allocate its limited frames according to what each target requires, with the aim of covering the different evidence needed to answer the question.

Reasoning-driven video acquisition. Agentic methods reason about which evidence to inspect through temporal tools and adaptive observation (Zhang et al., 2025b; Xie et al., 2025; He et al., 2025; Shen et al., 2025; Ding et al., 2025; Yang et al., 2025; Jin et al., 2025; Wang et al., 2025b). LensWalk plans temporal scope and sampling density, while ReViSe maintains a compact state across rounds (Li et al., 2026b; Xu et al., 2026). A.I.R. and VideoXAgent further adapt evidence acquisition to the question (Zou et al., 2025; Yang et al., 2026b). ParaVT learns parallel video-tool calls with reinforcement learning (Yang et al., 2026c); AutoSkill routes questions to discovered selection programs (Hu et al., 2026). VEI plans all evidence requests from one dense preview before answering, avoiding repeated search rounds while keeping the high-resolution frame budget fixed.

Privileged visual self-distillation. On-policy distillation supervises student-generated prefixes (Agarwal et al., 2024); self-distillation can derive guidance from richer teacher context (Zhao et al., 2026; Ye et al., 2026; Hübotter et al., 2026). Video-OPSD and Clue-OPSD use annotated evidence frames or intervals to strengthen video supervision (Wang et al., 2026c;a). RP-OPSD, S^2VOPD, and Vision-OPD exploit resolution, augmentation, or regional-view asymmetries (Zhu et al., 2026a; Li et al., 2026c; Yuan et al., 2026). These methods primarily distill perception and answer reasoning from visual inputs. VEI distills the evidence-acquisition policy, teaching the model what to inspect and how to distribute high-resolution budget. This helps the student use a compressed preview to acquire question-relevant detail under a limited context, without frame-level annotations.

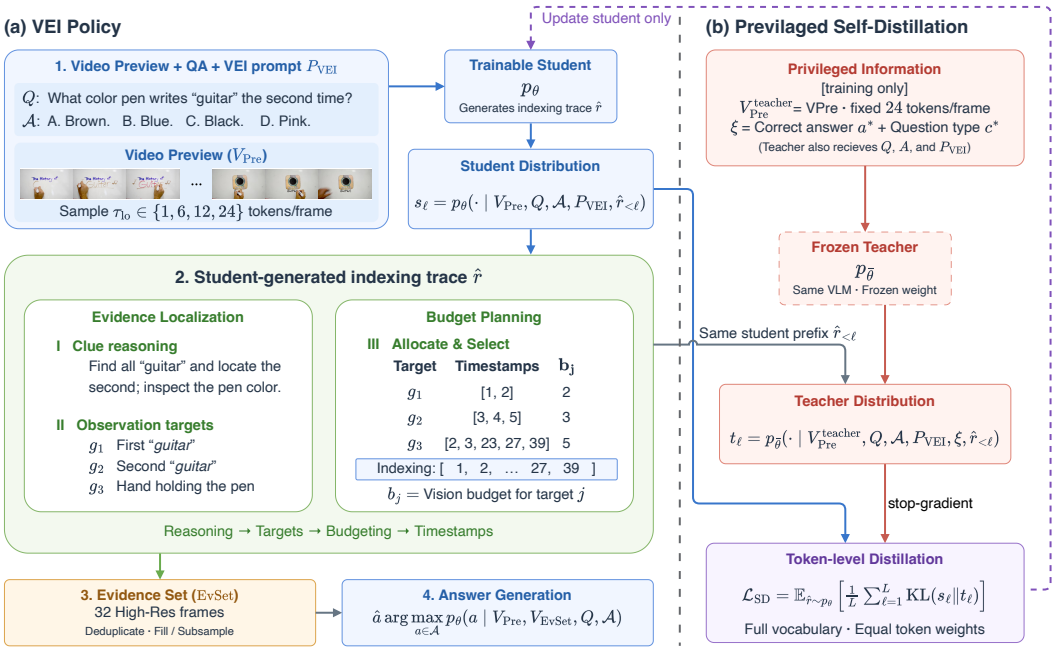


Figure 3: **Privileged self-distillation for VEI.** (a) The student generates observation targets and timestamp lists; the complete VEI prompt P_{VEI} appears in fig. 6. (b) Privileged information enables the teacher to guide the student's search toward evidence supporting the correct answer. Distilling this guidance teaches the student what to look for and how to allocate its frame budget from compressed previews, without frame-level annotations or privileged inputs at inference.

3 METHOD

We study multiple-choice long-video QA under a fixed visual-token budget B. Given a video V, a question Q, and answer choices $\mathcal{A} = \{a_1, \ldots, a_M\}$, the model must acquire the evidence needed to answer within this budget. As shown in Figure 1, our inference pipeline follows four steps: Video Preview, Video Evidence Indexing, Evidence Set construction, and Answer Generation. The VEI policy performs two coupled behaviors: *evidence localization* identifies what visual evidence to look for, and *budget planning* decides where to spend the limited high-resolution frames. We then train these behaviors through privileged self-distillation.

3.1 VIDEO PREVIEW

We construct the Video Preview by sampling the video at 1 fps and reducing each frame's spatial resolution. Let $t_1, \ldots, t_n$ be these sampling timestamps in seconds, and let $V(t_i)$ denote the original frame at t_i. The preview is

$$V_{\text{Pre}} = \left(\text{Downsize}_{\tau_{\text{lo}}}(V(t_i))\right)_{i=1}^{n}, \tag{1}$$

where $\text{Downsize}_{\tau_{\text{lo}}}$ reduces each frame's resolution so that the backbone's visual encoder produces at most τ_{lo} visual tokens. With n sampled frames, the preview uses at most $n\tau_{\text{lo}}$ visual tokens. Sampling is reduced only if required by the context limit.

The temporally dense preview provides global context at low visual-token cost. It is designed to preserve coarse event progression across nearby frames, supporting global reasoning and localization of complementary high-resolution evidence in one indexing attempt.

3.2 VIDEO EVIDENCE INDEXING

The Video Evidence Indexing Prompt P_{VEI} guides the model to identify evidence needs and distribute its high-resolution observations across them (Appendix B). Given this prompt, the Video Preview, question, and answer choices, the policy generates an indexing trace:

$$\hat{r} \sim p_\theta(\cdot \mid V_{\text{Pre}}, Q, \mathcal{A}, P_{\text{VEI}}). \tag{2}$$

The trace contains two coupled behaviors: evidence localization and budget planning.

Evidence localization. The model briefly reasons about what to inspect from the question and preview, then generates a list of observation targets:

$$G = [g_1, \ldots, g_J] \sim p_\theta(\cdot \mid V_{\mathrm{Pre}}, Q, \mathcal{A}, P_{\mathrm{VEI}}). \tag{3}$$

Each g_j describes a specific evidence need in natural language, such as an object's appearance, an action, or an event to distinguish. The prompt requests compact, diverse targets and encourages merging overlapping ones. This explicit list organizes what must be verified and guides the subsequent allocation of frames.

Budget planning. Conditioned on the target list, the model jointly plans its high-resolution observations through target-specific timestamp lists:

$$(E_1, \ldots, E_J) \sim p_\theta(\cdot \mid V_{\mathrm{Pre}}, Q, \mathcal{A}, P_{\mathrm{VEI}}, G), \qquad b_j = |E_j|. \tag{4}$$

The visual budget b_j for target g_j is implicit in its timestamp choices and list termination. Joint planning aims to balance target coverage and reduce redundancy within K frames, reserving more observations for targets requiring temporal comparison or fine detail.

Evidence Set and answer generation. We enforce the final K-frame budget by merging valid timestamps into $E = \bigcup_{j=1}^{J} E_j$, removing duplicates, and uniformly supplementing or subsampling. We reload these frames at high resolution to form V_{EvSet}. A separate call generates the answer from the preview and Evidence Set:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p_\theta(a \mid V_{\mathrm{Pre}}, V_{\mathrm{EvSet}}, Q, \mathcal{A}). \tag{5}$$

With a per-frame visual-token cap of τ_{hi} for the Evidence Set, the answering call's visual input costs at most $n\tau_{\mathrm{lo}} + K\tau_{\mathrm{hi}}$ tokens. We choose the preview and evidence budgets so that $n\tau_{\mathrm{lo}} + K\tau_{\mathrm{hi}} \leq B$.

3.3 Privileged Self-Distillation

The lack of frame-level evidence annotations in most VQA datasets makes direct supervision of intermediate VEI decisions difficult. To improve evidence localization and budget planning without explicit annotations, we adopt privileged self-distillation (Zhao et al., 2026), using answer-aware token-level guidance to train the VEI policy (Figure 3).

Student indexing traces. For each example, we uniformly sample the student's preview budget τ_{lo} from $\{1, 6, 12, 24\}$ visual tokens per frame. The student receives $V_{\mathrm{Pre}}, Q, \mathcal{A}, P_{\mathrm{VEI}}$ and generates an L-token indexing trace with clue reasoning, observation targets, and timestamp lists:

$$\hat{r} = [z_1, \ldots, z_L] \sim p_\theta(\cdot \mid V_{\mathrm{Pre}}, Q, \mathcal{A}, P_{\mathrm{VEI}}). \tag{6}$$

Here, z_ℓ is the token at position ℓ. This trains one VEI policy for all four preview resolutions. Our rollout-source ablation favors student-generated (on-policy) traces over teacher-generated (off-policy) traces among the tested settings (Section 4.5).

Privileged teacher. The teacher starts from the same backbone as the student and remains frozen. It receives a preview of the same video fixed at 24 tokens per frame, denoted $V_{\mathrm{Pre}}^{\mathrm{teacher}}$, along with the question, choices, and VEI prompt. Its privileged information $\xi = (a^\star, c^\star)$ contains the correct answer $a^\star$ and optional complementary information $c^\star$, such as the question type. The teacher uses this context to score the student's existing trace rather than generate a separate reference. At each token position ℓ, their next-token distributions are

$$\begin{aligned} s_\ell &= p_\theta(\cdot \mid V_{\mathrm{Pre}}, Q, \mathcal{A}, P_{\mathrm{VEI}}, \hat{r}_{<\ell}), \\ t_\ell &= p_{\bar{\theta}}(\cdot \mid V_{\mathrm{Pre}}^{\mathrm{teacher}}, Q, \mathcal{A}, P_{\mathrm{VEI}}, \xi, \hat{r}_{<\ell}), \end{aligned} \tag{7}$$

where $\bar{\theta}$ denotes the frozen teacher parameters and $\hat{r}_{<\ell} = [z_1, \ldots, z_{\ell-1}]$ is the shared prefix. We train the student distribution s_ℓ toward the teacher distribution t_ℓ at each prefix. This provides answer-aware self-correction: the student first seeks evidence, then learns from the teacher's guidance on its own decisions, including those following a weak target or an unhelpful timestamp.

Distillation objective. We minimize reverse KL between student and teacher distributions over the indexing trace, excluding final-answer generation:

$$\mathcal{L}_{\mathrm{SD}} = \mathbb{E}_{\hat{r}\sim p_\theta(\cdot|V_{\mathrm{Pre}},Q,\mathcal{A},P_{\mathrm{VEI}})}\left[\frac{1}{L}\sum_{\ell=1}^{L}\mathrm{KL}(s_\ell\|t_\ell)\right]. \tag{8}$$

The divergence is computed over the full vocabulary and averaged equally across token positions, without special weights for timestamps. Gradients update only the student; sampled tokens and teacher distributions are held fixed during each update. Among the tested objectives, our ablation favors reverse KL over forward KL and generalized Jensen–Shannon divergence (Section 4.5).

The same objective provides guidance for both VEI behaviors. At reasoning and target positions, it guides how the student describes the evidence needed by the question. At timestamp positions, it guides where to look. Guidance over list endings is central to budget planning: whether to request another frame or close the list determines how many frames are proposed for that target. Together, these signals are designed to transfer evidence localization and budget allocation without per-target quota labels or a separate allocation loss. At inference, the student follows the same VEI pipeline with no privileged information.

4 Experiments

4.1 Implementation Details

Backbone and variants. We use Qwen3-VL-8B (Bai et al., 2025a) as the backbone model. **Training-Free VEI** denotes Training-Free Video Evidence Indexing, which uses the VEI inference pipeline without parameter updates. **Self-Distilled VEI** denotes Self-Distilled Video Evidence Indexing, which trains the preview-based indexing policy with privileged self-distillation and uses the trained checkpoint for both indexing and answering at test time.

Datasets. We evaluate on VideoMME, LVBench, and MLVU (Fu et al., 2025; Wang et al., 2025a; Zhou et al., 2025). VideoMME contains videos from roughly one minute to one hour. LVBench focuses on long videos and provides temporal evidence windows for localization analysis. For MLVU, we evaluate on the 502-question multiple-choice test set. We report percentage accuracy.

Inference protocol and settings. For low-budget settings, the model first samples the video at 1 fps and compresses each preview frame to a fixed visual-token cap $\tau_{\mathrm{lo}} \in \{1, 6, 12, 24\}$. This dense low-resolution input serves as the Video Preview. We reduce the preview sampling rate only when required by the context limit. For VEI methods, the model indexes question-relevant evidence timestamps from the Video Preview (VPre), and the selected Evidence Set (EvSet) is then re-injected at higher resolution for final answer prediction. The uniform baseline uses the same VPre and high-resolution frame budget, replacing EvSet with uniformly sampled frames. Unless otherwise stated, the additional EvSet budget is 32 frames with a maximum of 128 visual tokens per frame.

The full-context Qwen3-VL baseline uses at most 2048 frames and 640 tokens per frame, the default setting in the original report. The preview budgets $\tau_{\mathrm{lo}} \in \{1, 6, 12, 24\}$ are chosen to cover a wide range of compression levels while avoiding distorted aspect ratios from very small intermediate token grids. For example, 2 tokens can force a 1:2 or 2:1 grid, which distorts many videos. For evaluation, we use deterministic decoding with temperature 0.0, repetition penalty 2.0, thinking disabled, and a maximum generation length of 1024 tokens. Our evaluations use a single run per setting.

Self-distillation training. Self-Distilled VEI is trained on the MLVU training set. To fit training into available GPU memory, we keep videos shorter than 20 minutes, yielding 2044 multiple-choice questions and 495 short-answer questions. The frozen teacher uses 24 visual tokens per preview frame and receives the correct answer, supplemented by the question type when available. For each example, the student uses a preview budget sampled uniformly from $\{1, 6, 12, 24\}$ visual tokens per frame, without privileged information. We train for 1,500 steps with a batch size of 8 using full-vocabulary reverse-KL distillation. Training uses LoRA on four H100 GPUs and takes 40 hours. Hyperparameters appear in Table 10; prompts are in Appendix B.

Method	Context Limit	Preview Size (#Frame/Tok)	VideoMME w/o sub. Short	Medium	Long	Overall	LVBench Train	MLVU$_{MCQ}$ Test	Average Score	%
Use 100% Context Width Across All Benchmarks										
Qwen3-VL-8B	256K	$2048/640_{Tok}$	77.8	70.7	64.7	71.0	57.6	67.7	65.4	100.0
Use 5.9% Context Width Across All Benchmarks										
Uniform FS	15K	$1fps/1_{Tok}$	73.1	60.0	58.4	63.8	40.2	52.2	52.1	79.6
Training-Free VEI	15K	$1fps/1_{Tok}$	70.4	53.9	51.4	58.6	36.7	46.2	47.2	72.1 (−7.5)
Self-Distilled VEI	15K	$1fps/1_{Tok}$	75.8	61.8	54.2	63.9	44.9	58.8	55.9	85.4 (+5.8)
Use 7.8% Context Width Across All Benchmarks										
Uniform FS	20K	$1fps/6_{Tok}$	76.2	62.7	57.2	65.4	42.1	55.2	54.2	82.9
Training-Free VEI	20K	$1fps/6_{Tok}$	73.2	64.1	55.9	64.4	46.1	61.2	57.2	87.5 (+4.6)
Self-Distilled VEI	20K	$1fps/6_{Tok}$	77.0	65.8	56.2	66.3	49.6	68.5	61.5	94.0 (+11.1)
Use 11.7% Context Width Across All Benchmarks										
Uniform FS	30K	$1fps/12_{Tok}$	75.6	66.4	58.9	67.0	46.3	59.0	57.4	87.7
Training-Free VEI	30K	$1fps/12_{Tok}$	75.6	67.2	59.0	67.3	51.6	63.5	60.8	92.9 (+5.2)
Self-Distilled VEI	30K	$1fps/12_{Tok}$	77.8	67.2	59.9	68.3	53.7	72.5	64.8	99.1 (+11.4)
Use 19.5% Context Width Across All Benchmarks										
Uniform FS	50K	$1fps/24_{Tok}$	75.6	65.9	60.6	67.3	50.7	60.2	59.4	90.8
Training-Free VEI	50K	$1fps/24_{Tok}$	77.7	70.3	63.6	70.5	53.5	66.7	63.6	97.1 (+6.3)
Self-Distilled VEI	50K	$1fps/24_{Tok}$	78.9	70.5	63.8	71.1	57.3	72.3	66.9	102.2 (+11.4)

Table 1: **Main results across preview budgets.** One self-distilled policy supports all four preview resolutions, outperforming uniform sampling even at one token per frame under the same 32-frame EvSet budget. The 1 fps preview rate is reduced when the input exceeds the context limit.

4.2 Main Results

Table 1 shows that Training-Free VEI improves over uniform sampling on all three datasets at 12 and 24 preview tokens per frame, demonstrating that our VEI policy works out of the box on a pretrained VLM. However, the pretrained model struggles to select useful evidence from one-token previews.

Privileged self-distillation strengthens VEI under extreme preview compression, with substantial one-token gains over uniform sampling on LVBench and MLVU and comparable VideoMME overall accuracy. At 24 tokens per frame, Self-Distilled VEI reaches 102.2% of the full-context baseline's average accuracy with less than a quarter of its context budget. Our privileged self-distillation training enables VEI to make effective use of highly compressed Video Previews.

Cross-backbone evaluation. Table 2(d) shows that VEI's gains extend to LLaVA-OneVision-2 and Penguin-VL. Training-Free VEI improves all three benchmarks on both backbones, and Self-Distilled VEI further improves each result, supporting VEI beyond Qwen3-VL backbone.

High-resolution evidence budget. We vary the high-resolution evidence budget over $K = 8, 16, 32$ frames on VideoMME (Table 3). Self-Distilled VEI remains robust under tighter budgets, surpassing 32-frame uniform sampling with only eight selected frames.

Method	$K = 8$	$K = 16$	$K = 32$
Uniform FS	66.26	67.19	67.33
Self-Distilled VEI	69.52	70.24	71.07

Table 3: **Evidence-budget sensitivity on VideoMME.**

Comparison with related works. Our CLIP† baseline matches Training-Free VEI's preview and 32-frame high-resolution evidence budget (Table 2(d)). VEI's higher average accuracy

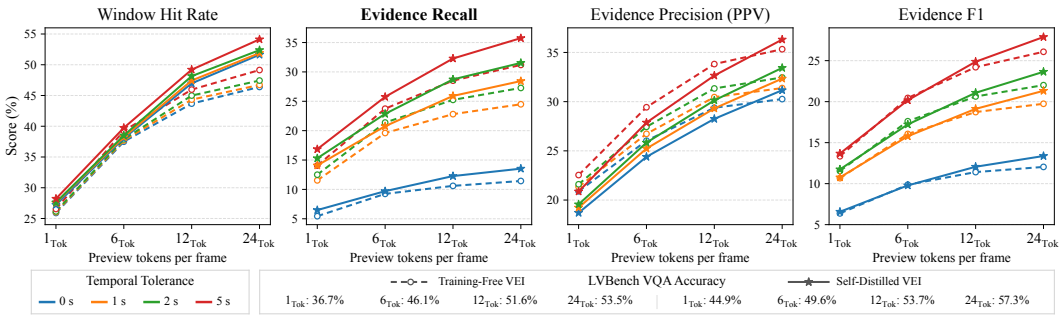


Figure 4: **Evidence localization on LVBench.** Training-Free and Self-Distilled VEI are compared across four preview budgets. Evidence Recall is the key metric, which measures how much of the annotated evidence the VEI policy recovers in its indexing.

Model	Size	Token Budget	VideoMME		LVBench	MLVU	AVG
			Long	Overall			
(a) Proprietary Models							
GPT-4o	–	128K	65.3	71.9	64.4	64.6	67.0
Gemini-2.5-Pro	–	1M	–	87.0	69.2	81.2	79.1
(b) Open-Source / Public Pretrained Models							
LLaVA-OneVision (Li et al., 2024)	7B	32K	46.7	58.2	26.9	50.5	45.2
LongVA (Zhang et al., 2024)	7B	128K	47.6	54.3	36.2	43.9	44.8
InternVL2.5 (Chen et al., 2025)	8B	64K	–	66.9	38.4	53.7	53.0
Qwen2.5-VL (Bai et al., 2025b)	7B	128K	51.6	65.1	45.3	54.8	55.1
NVILA (Liu et al., 2026)	8B	64K	54.8	64.2	44.0	54.7	54.3
VideoLLaMA3 (Zhang et al., 2025a)	7B	64K	54.1	66.2	45.3	56.9	56.1
Qwen3-VL (Bai et al., 2025a)	8B	256K	**64.7**	71.0	**57.6**	67.7	65.4
(c) Evidence-Acquisition Methods							
MACF (Chen et al., 2026a)	8B	5K	–	60.4	40.2	49.2	49.9
ParaVT-8B (Yang et al., 2026c)	8B	16K	–	62.1	39.8	65.0	55.6
AKS + BLIP (Tang et al., 2025a)	8B	25K	59.0	68.6	49.0	64.3	60.6
AdaQ + LongCLIP (Zhang et al., 2026b)	8B	25K	59.4	69.6	51.4	66.2	62.4
AdaCodec (Hou et al., 2026)	8B	64K	–	70.6	57.2	67.2	65.0
(d) VEI Cross-Backbone Evaluation							
LLaVA-OneVision-2-8B (An et al., 2026)							
Uniform FS	8B	50K	56.0	65.9	47.5	57.4	56.9
Training-Free VEI	8B	50K	58.3	68.4	53.5	64.9	62.3
Self-Distilled VEI	8B	50K	58.8	68.7	55.8	65.1	63.2
Penguin-VL-8B (Zhang et al., 2026a)							
Uniform FS	8B	50K	51.7	63.9	39.3	56.0	53.1
Training-Free VEI	8B	50K	57.2	66.6	44.5	60.6	57.2
Self-Distilled VEI	8B	50K	57.7	67.2	47.4	64.0	59.5
Qwen3-VL-8B (Bai et al., 2025a)							
Uniform FS	8B	50K	60.6	67.3	50.7	60.2	59.4
CLIP† FS	8B	50K	62.1	68.9	53.9	65.5	62.8
Training-Free VEI	8B	50K	63.6	70.5	53.5	66.7	63.6
Self-Distilled VEI	8B	50K	63.8	**71.1**	57.3	**72.3**	**66.9**

Table 2: **Comparison with existing VLMs and evidence-acquisition methods.** Prior results in (a–c) use published evaluations and budgets; methods in (c) use Qwen3-VL-8B. CLIP† is our retrieval baseline (Appendix H). Bold marks the best across (b–d).

(63.6% vs. 62.8%) supports structured indexing beyond relevance retrieval under matched visual budgets. Self-Distilled VEI reaches 66.9% at 50K context, exceeding AdaCodec's reported 65.0% at 64K visual tokens and approaching 67.2% at 224K. Broader comparisons in (a–c) use different observation protocols; Uniform FS and CLIP† provide matched visual-budget controls. AKS scores follow AdaQ's Qwen3-VL-8B evaluation (Zhang et al., 2026b). Cost assumptions appear in Appendix C.

4.3 Evidence Localization and Budget Planning

We separate the two VEI behaviors with a nested ablation. *Base FS* directly selects timestamps from the question and preview. Adding evidence localization (EL) introduces explicit observation targets but retains one global timestamp list. Adding budget planning (BP) instead produces a timestamp list for each target. All variants use the same final evidence budget and uniform supplementation rule.

Table 4 shows that listing targets alone brings only a modest gain without training and none after distillation. Target-wise selection adds 2.4 and 2.6 percentage points, respectively, while reducing uniform supplementation under the same final budget and filling rule. VEI remains strongest on questions with comparable selected-frame counts (Appendix A), supporting the value of connecting each target to concrete frame choices. Appendix F separately tests the contribution of high-resolution evidence by removing the Evidence Set while retaining the preview.

4.4 Evidence Quality and Answer Correctness

Figure 4 evaluates generated timestamps before filling or truncation against LVBench's evidence windows, expanded by one second per side and discretized into seconds, with tolerance δ. **Window hit rate** is the fraction of questions with a selection within δ seconds of an evidence window.

Configuration	Training-Free		Self-Distilled	
	Accuracy	Filled frames	Accuracy	Filled frames
Base FS	63.2	17.6	69.7	17.5
+ EL	64.3	16.9	69.7	16.5
+ EL + BP (VEI)	66.7	10.7	72.3	9.8

Table 4: **VEI component ablation on MLVU.** Filled frames denotes the mean number of uniformly sampled frames added to reach the 32-frame evidence budget.

Method	Hit rate	Evidence recall	Correct given hit	Correct given miss
Training-Free VEI	48.8	31.2	65.6	41.6
Self-Distilled VEI	53.3	35.8	69.5	43.2
Privileged teacher	54.6	36.3	69.5	43.1

Table 5: **Evidence localization and QA on LVBench.** Scores (%) use 24-token previews and a 5 s tolerance on 1,549 questions; recall measures generated timestamps as in Figure 4. The privileged teacher provides an empirical upper reference using answer information unavailable at inference.

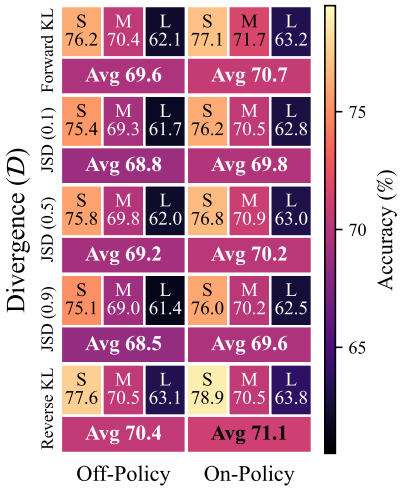


Figure 5: **Rollout policy and distillation objective ablation.** On-policy distillation with reverse KL achieves the highest overall VideoMME accuracy.

Evidence recall measures the fraction of evidence seconds covered within δ of a selection; **evidence precision** measures the fraction of selections within δ seconds of the evidence window. **Evidence F1** is their harmonic mean; these three metrics are averaged per question. Recall is central here because it measures how much supporting evidence is recovered. At one token per frame and $\delta = 5$ s, self-distillation raises recall from 14.1% to 16.9%, supporting improved evidence coverage as an explanation for the QA gain from highly compressed previews.

Table 5 uses the privileged teacher as an empirical upper reference for answer-aware localization. Self-Distilled VEI approaches the teacher's hit rate and evidence recall, and matches its answer accuracy on hit cases without privileged inputs. Across all three policies, hits are associated with substantially higher answer accuracy, but do not guarantee correctness.

4.5 Privileged Self-Distillation Analysis

Privileged guidance. The teacher diagnostic in Table 5 reveals the evidence selections available under answer-aware conditioning; final answering remains non-privileged. During training, the teacher instead scores the student's prefixes as described in Section 3.3, transferring this guidance without supplying a separate gold trace or frame-level labels.

Rollout policy and objective. Figure 5 compares student-generated traces with teacher-generated traces. Student rollouts perform better in most settings, consistent with the benefit of supervising prefixes the student actually produces. This matters for VEI because later timestamps depend on earlier target choices. Reverse KL gives the strongest result in this comparison, followed by forward KL, while the JSD variants are slightly lower. We therefore use student rollouts and reverse KL for privileged self-distillation. The multi-resolution training setting additionally exposes one student to all four preview resolutions, while the teacher retains its 24-token preview.

5 Conclusion

We explore how dense, compressed Video Previews can guide evidence acquisition for long-video QA under limited context. VEI couples evidence localization with budget planning to construct a compact, question-relevant Evidence Set. We adopt privileged self-distillation to learn these decisions without frame-level annotations, producing one policy across four preview resolutions, including one token per frame. Our results demonstrate the value of using cheap global coverage to direct high-resolution observations toward the evidence needed for answering.

AI USE STATEMENT

The original manuscript was written by the authors. During revision, we used AI tools to edit figures, refine language and grammar. The authors reviewed and selectively incorporated AI-assisted suggestions, checking their consistency with the intended method and reported results. We take responsibility for the final text, claims, figures, and other content of this work.

REPRODUCIBILITY STATEMENT

Section 3 describes the VEI pipeline and privileged self-distillation objective. Section 4.1 specifies the datasets, preprocessing, inference settings, and training setup; Appendix B provides the indexing prompt, and Table 10 lists training hyperparameters. We plan to publicly release the source code and the weights of our best-performing multi-resolution self-distilled model. These materials are being prepared and are not yet publicly available.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

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, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082, June 2025.

Weitong Cai, Hang Zhang, Yukai Huang, Yiqiao Xie, Shan Gao, Jiankang Deng, Songcen Xu, Jifei Song, and Zhensong Zhang. Caption-once, Frames-on-Demand: Visual-Need Routing for Budget-Aware Agentic Long Video Understanding, 2026. URL https://arxiv.org/abs/2609.11899.

Kerui Chen, Jinglu Wang, Jianrong Zhang, Ming Li, Yan Lu, and Hehe Fan. Scaling Video Understanding via Compact Latent Multi-Agent Collaboration, 2026a. URL https://arxiv.org/abs/2605.00444.

Wang Chen, Yu Chen, Xiang Wang, Shuai Li, Jinfa Huang, and Xiawu Zheng. One Ranking, Any Budget: Matryoshka Evidence-to-Context Frame Selection for Long-Video Understanding, 2026b. URL https://arxiv.org/abs/2608.05707.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning, 2025. URL https://arxiv.org/abs/2512.22315.

Junhao Du, Jialong Xue, Anqi Li, Jincheng Dai, and Guo Lu. Unified Spatiotemporal Token Compression for Video-LLMs at Ultra-Low Retention, 2026. URL https://arxiv.org/abs/2603.21957.

Mohamed Eltahir, Lama Ayash, Ali Habibullah, Tanveer Hussain, and Naeemullah Khan. GridProbe: Posterior-Probing for Adaptive Test-Time Compute in Long-Video VLMs, 2026. URL https://arxiv.org/abs/2605.10762.

Junjie Fei, Jun Chen, Zechun Liu, Yunyang Xiong, Chong Zhou, Wei Wen, Junlin Han, Mingchen Zhuge, Saksham Suri, Qi Qian, Shuming Liu, Lemeng Wu, Raghuraman Krishnamoorthi, Vikas Chandra, Mohamed Elhoseiny, and Chenchen Zhu. Small Vision-Language Models are Smart Compressors for Long Video Understanding, 2026. URL https://arxiv.org/abs/2604.08120.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting, 2025. URL https://arxiv.org/abs/2509.24304.

Haowen Hou, Zhen Huang, Zheming Liang, Qingyi Si, Chenglin Li, Shuai Dong, Kele Shao, Ruilin Li, Dianyi Wang, Nan Duan, and Jiaqi Wang. AdaCodec: A Predictive Visual Code for Video MLLMs, 2026. URL https://arxiv.org/abs/2606.02569.

Jian Hu, Zixu Cheng, Da Li, Wei Li, Ziquan Liu, and Shaogang Gong. One Skill Does Not Fit All: Automatic Discovery and Taxonomy-Guided Routing of Frame-Selection Skills for Long-Video Question Answering, 2026. URL https://arxiv.org/abs/2609.12517.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management, 2025. URL https://arxiv.org/abs/2512.04540.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Ke Li, Jiayu Chen, Maoliang Li, Zihao Zheng, Hailong Zou, Hengyi Zhang, Xuanzhe Liu, and Xiang Chen. When and Where to Look: Adaptive Visual Evidence Scheduling for Efficient Long Video Understanding, 2026a. URL https://arxiv.org/abs/2608.03918.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. LensWalk: Agentic Video Understanding by Planning How You See in Videos, 2026b. URL https://arxiv.org/abs/2603.24558.

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, 2026c. URL https://arxiv.org/abs/2608.14144.

Yulin Li, Haokun Gui, Ziyang Fan, Junjie Wang, Bin Kang, Bin Chen, and Zhuotao Tian. Less is more, but where? dynamic token compression via LLM-guided keyframe prior. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uhFx1RGD1g.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2026. URL https://arxiv.org/abs/2412.04468.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient Frame Selection for Long Video Understanding via Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16944–16953, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Qin_Efficient_Frame_Selection_for_Long_Video_Understanding_via_Reinforcement_Learning_CVPR_2026_paper.html.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

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

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in, 2025. URL https://arxiv.org/abs/2512.14273.

Hongchang Shi, Jinpeng Hu, Ao Wang, Wenzheng Zhou, Hui Ma, Feng Li, and Zenglin Shi. MarKey: Marginal Utility Guided Greedy Keyframe Selection for Long Video Understanding, 2026. URL https://arxiv.org/abs/2609.15408.

Biao Tang, Xu Chen, Shuxiang Gou, Jingyi Yuan, Yuhan Zhang, Deyu Meng, and Chenqiang Gao. ReUnit: Multi-Granularity Visual Unitization for Long Video Understanding, 2026. URL https://arxiv.org/abs/2606.12125.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive Keyframe Sampling for Long Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, Los Alamitos, CA, USA, June 2025a. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02711. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02711.

Yolo Y. Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2025b. URL https://arxiv.org/abs/2312.17432.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to Look Matters: On-Policy Self-Distillation for Long-Video Understanding, 2026a. URL https://arxiv.org/abs/2608.25356v2.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025a. URL https://arxiv.org/abs/2406.08035.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient Frame Selection for Long Videos at Test Time with Attention-Based MLLM Selectors, 2026b. URL https://arxiv.org/abs/2607.15689.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding, 2025b. URL https://arxiv.org/abs/2512.05774.

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-OPSD: Exploiting Privileged Visual Evidence for On-Policy Self-Distillation in Video Large Language Models, 2026c. URL https://arxiv.org/abs/2608.27065.

Fan Wei, Siru Zhong, Runmin Dong, Miao Yang, Zhaoyang Luo, and Haohuan Fu. Ground, Cover, and Refine: Evidence-Centric Frame Selection for Long-Video Question Answering, 2026. URL https://arxiv.org/abs/2608.01660.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding, 2025. URL https://arxiv.org/abs/2508.20478.

Chenwei Xu, Zhen Ye, Shang Wu, Weijian Li, Zihan Wang, Zhuofan Xia, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Towards Sparse Video Understanding and Reasoning, 2026. URL https://arxiv.org/abs/2602.13602.

Sosuke Yamao, Natsuki Miyahara, Yuankai Qi, and Shun Takeuchi. Question-guided Visual Compression with Memory Feedback for Long-Term Video Understanding, 2026. URL https://arxiv.org/abs/2603.15167.

Qing Yang, Pengcheng Huang, Xinze Li, Zhenghao Liu, Yukun Yan, Yu Gu, Ge Yu, Gang Li, and Maosong Sun. MemoryCard: Topic-Aware Multi-Modal Clue Compression for Long-Video Question Answering, 2026a. URL https://arxiv.org/abs/2606.05917v2.

Sen Yang, Boqiang Duan, Jing Yang, Weihao Bo, Jie Liu, Boyuan Tong, Ze Feng, Wenkang Zhang, Jingdong Wang, and Hua Wu. Online Video Agent Harness for Long Video Understanding, 2026b. URL https://arxiv.org/abs/2609.12818.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing "thinking with long videos" via native tool calling, 2025. URL https://arxiv.org/abs/2511.20785.

Zuhao Yang, Kaichen Zhang, Sudong Wang, Keming Wu, Zhongyu Yang, Bo Li, Xiaojuan Qi, Shijian Lu, Xingxuan Li, and Lidong Bing. ParaVT: Taming the Tool Prior Paradox for Parallel Tool Use in Agentic Video Reinforcement Learning, 2026c. URL https://arxiv.org/abs/2605.20342.

Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, Qidong Huang, Haotian Wang, Feng Wu, Yuanxing Zhang, Pengfei Wan, Zhouchen Lin, and Xu Sun. Towards efficient multimodal large language models: A survey on token compression. *TechRxiv*, 2026(0112), 2026. doi: 10.36227/techrxiv.176823010.07236701/v1. URL https://www.techrxiv.org/doi/abs/10.36227/techrxiv.176823010.07236701/v1.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

Bowen Yuan, Sisi You, and Bing-Kun Bao. Dtoma: Training-free dynamic token manipulation for long video understanding. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 2314–2322. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/258. URL https://doi.org/10.24963/ijcai.2025/258. Main Track.

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, 2026. URL https://arxiv.org/abs/2605.18740v4.

Yuhui Zeng, Wang Chen, Jinfa Huang, Tianyu Xie, Yongdong Luo, Jiayi Ji, Xiawu Zheng, and Jiebo Luo. WaveZip: Wavelet-Driven Space-Time Decoupling for Video Token Condensation, 2026. URL https://arxiv.org/abs/2607.23265.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, and Leoweiliang. Penguin-VL: Exploring the Efficiency Limits of VLM with LLM-based Vision Encoders. *arXiv preprint arXiv:2603.06569*, 2026a. URL https://arxiv.org/abs/2603.06569.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025b. URL https://arxiv.org/abs/2508.04416.

Kun Zhang, Chenxin Fang, Tao Chen, Baiyang Song, Yunhang Shen, Yiyi Zhou, and Rongrong Ji. Towards fast and effective long video understanding of multimodal large language models via adaptive quasi-gaussian sampling, 2026b. URL https://arxiv.org/abs/2606.24187.

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, 2024. URL https://arxiv.org/abs/2406.16852.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025c. URL https://arxiv.org/abs/2506.22139.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2025d. URL https://arxiv.org/abs/2308.10792.

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, 2026. URL https://arxiv.org/abs/2601.18734.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.

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, 2026a. URL https://arxiv.org/abs/2607.24447.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=1OQKqLFcbB.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling Adaptive, Iterative, and Reasoning-based Frame Selection For Video Question Answering, 2025. URL https://arxiv.org/abs/2510.04428v3.

Appendix

A Component Ablation with Comparable Selection Counts

The full-test-set ablation in Table 4 gives every method 32 evidence frames, but VEI supplies more of these through its own selections and requires fewer uniform additions. To examine whether its advantage persists when this difference is smaller, we filter the same 502 MLVU questions by the number of frames each method selects.

Selection rule. For each question, we count each method's valid, unique selections before uniform filling and cap this count at 32. We retain the question if the capped counts for Base FS and + EL are each within ± 3 frames of + EL + BP (VEI). Thus, questions where all three independently select at least 32 frames automatically qualify. The original answering protocol is unchanged: excess selections are subsampled and shortages are uniformly supplemented to obtain exactly 32 evidence frames. We apply the filter separately to Training-Free and Self-Distilled models; within each row, all three configurations are evaluated on the same retained questions.

Before capping, all three methods select exactly the same number of frames on 67 Training-Free and 65 Self-Distilled questions. Separately, all three select at least 32 frames on 60 and 66 questions, respectively. Both conditions satisfy the inclusion rule above.

Selector	Comparable questions	Base FS	+ EL	+ EL + BP (VEI)
Training-Free VEI	220 / 502	70.0	66.4	**74.6**
Self-Distilled VEI	218 / 502	78.9	78.0	**80.7**

Table 6: **VEI component ablation with comparable selection counts.** Accuracy (%) on shared questions within each row; every configuration uses a final 32-frame Evidence Set.

Targets need target-wise selection. Table 6 shows that VEI outperforms both alternatives even when their selected-frame counts are similar. Base FS remains competitive, while adding targets without target-wise selection lowers accuracy on these subsets. This suggests that a target list alone does not ensure useful evidence acquisition: a single global list may fail to turn the planned targets into corresponding observations. Coupling localization with budget planning makes these targets actionable, recovering the advantage over direct selection in both models. The result supports this coupling, rather than treating target listing as an independently effective step.

Beyond the VEI format. Although self-distillation trains on VEI traces, the distilled model also performs well with Base FS and + EL. This is consistent with learned preview-use skills carrying over to simpler selection prompts. The Base FS scores of 70.0% and 78.9% are suggestive, although the different subsets prevent a paired estimate of the distillation gain.

B Prompt Details

This appendix records the first-turn Video Evidence Indexing Prompt. Final answer generation follows Evidence Set construction in a separate turn. During self-distillation, the privileged block conditions the teacher's scoring of student-generated prefixes; it is removed at test time.

C Additional Notes on Cost Estimates

The cost estimates use measured token counts and a fixed effective token rate. They are approximate and depend on encoding, cache reuse, and serving assumptions.

The comparison in Table 7 uses the same 8B backbone family and token-rate assumption for both rows. Self-Distilled VEI uses a 50K context limit, compared with 256K for the full-context baseline.

Video-Evidence-Indexing Prompt

Input:
{Video Preview}
Question: {question}
Possible answer choices:
{options}

[Optional] privileged teacher information:
You are a teacher, and you are given the following information about the question.
The task type of this question is: {question type}.
The correct answer to this question is: **{answer text}**.
Please use this privileged information only to verify and refine a gold-standard reference solution, not to shortcut the reasoning. Answer Steps 1–3 exactly as specified, with analysis that derives the correct choice from visible evidence in the video. Finally, for each target, select a compact and diverse set of evidence frames that supports this analysis, maximizes question-relevant coverage, and minimizes redundancy, with at most {maximum frames per target} frames per target.

Step 1: Evidence Planning
Briefly answer: In order to locate the correct answer choice, what objects, scenes, actions, or keyframes should we focus on?

Step 2: Observation Targets
Based on Step 1, list a compact and diverse set of observation targets. Merge overlapping targets when appropriate and avoid redundancy.
Do not give more than {maximum number of targets} targets.

- Target 1: short description.
- Target 2: short description.
- Target 3: short description.
- ...

Step 3: Evidence Indexing
For each observation target, index at most {maximum frames per target} evidence frames from the Video Preview.
Output each indexed set as a Python list of absolute timestamps in seconds.
Example format:

- Evidence Set 1: [12, 18, 23]
- Evidence Set 2: [45]
- Evidence Set 3: [61, 64]
- ...

Rules:

- Each indexed set must be a Python list.
- Each entry must be an integer timestamp in seconds.
- Each list must contain at most {maximum frames per target} values.
- The indexed frames should be compact, diverse, and relevant to the target.
- Avoid repeated timestamps or near-duplicate frames unless they are needed for temporal verification.

Figure 6: **Video Evidence Indexing Prompt.** The normal indexer receives the Video Preview, question, and answer choices. During self-distillation, the teacher additionally receives privileged information, including the question type and correct answer.

The dollar values are approximate token-based estimates, rather than costs derived directly from these configured limits.

Method	Video Processed (#Frames/Token)	Average Accuracy All Benchmarks	Estimated Cost All Benchmarks
Qwen3-VL-8B	$2048_{\text{Max}}/640_{\text{Tok}}$	65.4	$\sim$ \$87.8
Self-Distilled VEI	$1\text{fps}/24_{\text{Tok}} + 32/128_{\text{Tok}}$	66.9	$\sim$ \$19.1

Table 7: **Long-video accuracy and evaluation cost.** The estimate covers 4,751 test questions across the three datasets.

D Additional Discussion: Evidence Indexing

Long-video Video Evidence Indexing differs from standard keyframe sampling. In many QA examples, evidence may be spread across multiple moments: an object appears early, an action happens later, and the answer depends on comparing them. VEI therefore treats the preview-based indexing phase as Evidence Set construction, not isolated keyframe selection. The target list helps distribute the high-resolution frame budget across different evidence needs, while privileged self-distillation helps train this missing policy skill from student-generated rollouts.

E Qualitative Evidence and Failure Cases

In the example in Figure 7, CLIP retrieves multiple similar views of the animated tree, whereas VEI covers more stages of the decoration sequence. This illustrates how target-wise selection can gather complementary observations for this question.

Question	Outcome	Diagnosis
61	Hit, wrong	The selected evidence reaches the annotated event but misses its later outcome.
65	Hit, wrong	The interaction is localized, but the answer misses the semantic detail that no bribe occurred.
73	Miss, wrong	Selection concentrates on unrelated moments and misses the annotated event.
78	Miss, correct	The answer is correct despite missing the annotated window; the preview or alternative evidence may provide support.

Table 8: **Representative LVBench cases.** These cases distinguish localization, incomplete event coverage, and downstream reasoning, and show that a single annotated window need not exhaust all useful evidence.

F Additional Evidence Set Ablation

Table 9 compares Training-Free VEI with a variant that removes the selected Evidence Set under the same configured context limits. With the preview fixed, this ablation measures the benefit of adding the selected high-resolution evidence.

The results show that the Evidence Set contributes consistently across all budgets. Removing it reduces the average score by 11.1, 8.9, 5.2, and 5.0 points under 15K, 20K, 30K, and 50K context limits, respectively. The drop is largest in the lowest-budget setting, where the preview is useful for coarse temporal coverage but lacks enough visual detail for final answer verification. As the context budget increases, the preview itself becomes more informative, so the relative dependence on the Evidence Set decreases. However, the performance gap remains clear even at 50K, showing that VEI is not only a low-resolution preview method. Its benefit comes from the combination of dense cheap coverage and targeted high-resolution evidence reinjection.

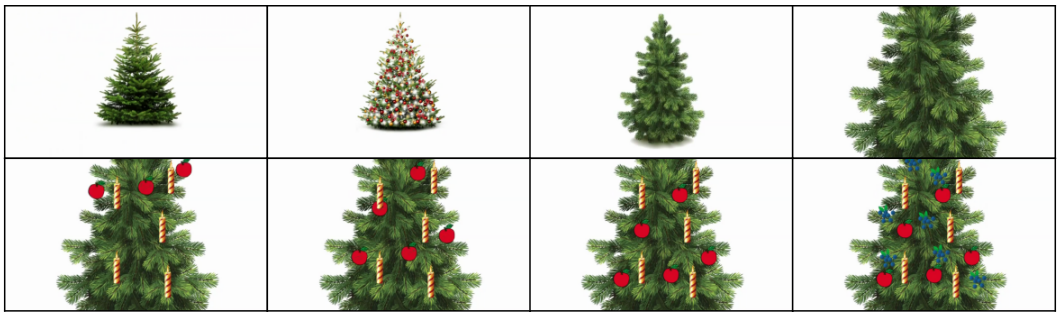
[Question] When demonstrating the Germany modern Christmas tree is initially decorated with apples, candles and berries, which kind of the decoration has the largest number?
[Choices] A. Apples. B. Candles. C. Berries. D. Same number.
[Correct Answer] C. Berries

Uniform Sampling



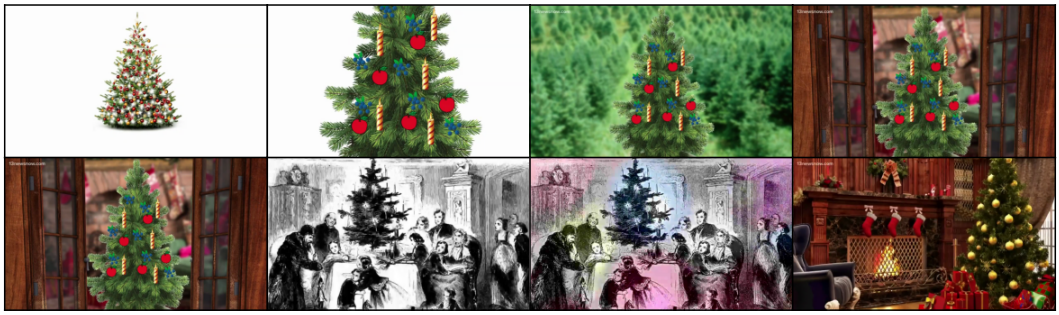
Answer: D. The three kinds are of the same number. **[Incorrect]**

Similarity-based Frame Ranking with Pre-trained CLIP



Answer: A. Apples. **[Incorrect]**

Video Evidence Indexing (VEI)



Answer: C. Berries. **[Correct]**

Figure 7: **EvSet Case Analysis.** Uniform sampling misses the close-up needed to compare decorations. CLIP retrieves repetitive tree views, while VEI includes detailed views of the completed decoration and answers correctly.

Method	Context Limit	VideoMME w/o sub.				LVBench	MLVU$_{MCQ}$	Average Score (%)
		Short	Medium	Long	Overall	Train	Test	
Use 5.9% Context Width Across All Benchmarks								
Training-Free VEI	15K	70.4	53.9	51.4	58.6	36.7	46.2	47.2
w/o Evidence Set		45.1	43.1	45.0	44.4	30.9	32.9	36.1 (−11.1)
Use 7.8% Context Width Across All Benchmarks								
Training-Free VEI	20K	73.2	64.1	55.9	64.4	46.1	61.2	57.2
w/o Evidence Set		59.9	55.9	55.7	57.1	40.7	47.0	48.3 (−8.9)
Use 11.7% Context Width Across All Benchmarks								
Training-Free VEI	30K	75.6	67.2	59.0	67.3	51.6	63.5	60.8
w/o Evidence Set		68.1	62.4	59.6	63.4	47.4	56.0	55.6 (−5.2)
Use 19.5% Context Width Across All Benchmarks								
Training-Free VEI	50K	77.7	70.3	63.6	70.5	53.5	66.7	63.6
w/o Evidence Set		72.0	66.1	60.2	66.1	50.8	59.0	58.6 (−5.0)

Table 9: **Effect of Evidence Set access under matched configured context limits.** For each low-resolution preview budget, we compare training-free VEI with the same model after removing access to the selected high-resolution Evidence Set. Blue values report the absolute average-score drop caused by removing Evidence Set frames.

Table 10: Training Configuration for Privileged Self-Distillation

Parameter	**SD**
Learning Rate	1×10^{-4}
Effective Batch Size	8
LoRA Rank (r)	64
LoRA Alpha (α)	128
LoRA Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Max Completion Length	1024
Number of Generations per Prompt	1
Sampling Temperature	1.0
KL Coefficient (β)	–
Training Steps	1500

G Accuracy by Question Type

Tables 11–13 compare Uniform FS, Training-Free VEI, and Self-Distilled VEI across each dataset's question types and all four preview resolutions. Each comparison uses the same preview and 32-frame evidence budget. Bold marks the highest displayed accuracy within each question type and preview resolution, including ties. Overall results appear in Table 1.

VideoMME. At one token per frame, self-distillation improves over Training-Free VEI on 10 of 12 question types, with the largest gain in temporal perception. At 24 tokens, gains remain clear in action reasoning and OCR, while counting and several perception categories favor Training-Free VEI. Training therefore changes the capability profile rather than improving every question type equally.

Question type	1 token/frame			6 tokens/frame			12 tokens/frame			24 tokens/frame		
	U	TF	SD	U	TF	SD	U	TF	SD	U	TF	SD
Action Reasoning	56.1	50.5	**61.1**	57.9	55.8	**62.8**	58.3	58.6	**64.6**	60.0	61.8	**67.7**
Action Recognition	57.8	56.2	**63.3**	62.0	63.9	**65.5**	65.2	64.9	**66.1**	65.2	67.1	**69.5**
Attribute Perception	77.9	70.3	**79.3**	**81.1**	75.2	80.6	82.0	79.3	**82.4**	78.8	**83.3**	82.4
Counting Problem	36.2	**39.2**	38.1	39.9	**43.7**	41.4	42.9	**48.1**	44.4	41.4	**48.1**	42.5
Information Synopsis	**82.0**	76.5	79.9	**82.0**	77.7	81.1	**83.9**	82.4	82.0	84.8	83.9	**85.9**
Object Reasoning	**61.2**	58.2	**61.2**	**64.5**	63.4	63.2	64.8	65.2	**66.5**	66.7	**68.9**	67.8
Object Recognition	**67.0**	64.1	**67.0**	68.6	**70.9**	69.8	69.8	70.9	**71.2**	70.6	75.7	**75.8**
OCR Problems	**71.9**	61.2	68.4	71.9	70.5	**74.1**	72.7	73.4	**75.5**	72.7	79.1	**83.3**
Spatial Perception	**66.7**	59.3	63.0	**70.4**	59.3	64.8	**70.4**	68.5	66.7	64.8	**66.7**	65.8
Spatial Reasoning	**82.1**	75.0	75.0	**85.7**	78.6	80.4	**83.9**	78.6	75.0	80.4	**82.1**	81.6
Temporal Perception	65.5	50.9	**72.7**	76.4	70.9	**78.2**	76.4	**78.2**	74.6	78.2	**81.8**	79.4
Temporal Reasoning	48.6	42.9	**52.0**	50.9	52.5	**53.7**	57.1	57.6	**61.0**	59.9	65.0	**67.1**

Table 11: **Question-type accuracy on VideoMME.** Scores (%) compare Uniform FS (U), Training-Free VEI (TF), and Self-Distilled VEI (SD) at each preview resolution.

LVBench. Training-Free VEI exceeds uniform sampling on temporal grounding at every preview resolution. Self-distillation further improves this category across all four resolutions and, at 24 tokens, improves all six question types over Training-Free VEI. The gains extend beyond temporal grounding to information retrieval and reasoning.

Question type	1 token/frame			6 tokens/frame			12 tokens/frame			24 tokens/frame		
	U	TF	SD	U	TF	SD	U	TF	SD	U	TF	SD
entity recognition	42.7	35.3	**42.8**	44.0	47.0	**48.6**	48.9	52.6	**54.1**	53.8	55.4	**57.8**
event understanding	38.2	35.1	**45.0**	39.9	45.0	**49.3**	42.2	48.2	**50.5**	47.1	49.9	**55.0**
key information retrieval	39.5	37.5	**47.4**	45.0	46.7	**49.8**	48.8	54.6	**55.7**	55.0	55.3	**62.1**
reasoning	35.8	34.3	**42.8**	36.8	36.8	**51.2**	42.8	46.3	**50.8**	44.8	44.8	**52.7**
summarization	31.0	**43.1**	32.8	34.5	**46.6**	44.8	39.7	**46.6**	**46.6**	39.7	44.8	**46.5**
temporal grounding	37.3	56.4	**60.0**	40.9	51.8	**58.2**	45.9	56.4	**59.6**	48.6	56.4	**59.9**

Table 12: **Question-type accuracy on LVBench.** Scores (%) compare Uniform FS (U), Training-Free VEI (TF), and Self-Distilled VEI (SD) at each preview resolution.

MLVU. At one token per frame, self-distillation improves over Training-Free VEI across all nine question types. At 24 tokens, the largest gains occur in *ego* and *count*, while gains are smaller or mixed in the other categories. Together, the three datasets show that training helps across varied question types under extreme preview compression, with more task-dependent gains at higher resolution.

Question type	1 token/frame			6 tokens/frame			12 tokens/frame			24 tokens/frame		
	U	TF	SD	U	TF	SD	U	TF	SD	U	TF	SD
anomaly_reco	48.7	46.2	**51.3**	53.9	48.7	**56.4**	53.9	**56.4**	**56.4**	51.3	53.9	**56.4**
count	26.7	11.7	**30.0**	30.0	21.7	**53.3**	31.7	25.0	**56.7**	30.0	35.0	**55.0**
ego	56.6	41.5	**60.4**	54.7	58.5	**77.4**	58.5	58.5	**79.3**	60.4	56.6	**79.3**
needleQA	50.0	50.0	**60.0**	58.3	**80.0**	78.3	61.7	81.7	**83.3**	71.7	**85.0**	83.3
order	42.9	30.0	**62.9**	55.7	65.7	**77.1**	61.4	68.6	**84.3**	60.0	75.7	**82.9**
plotQA	42.0	50.0	**56.0**	48.0	50.0	**66.0**	52.0	**62.0**	**62.0**	60.0	68.0	**70.0**
sportsQA	41.7	41.7	**47.2**	41.7	**50.0**	38.9	47.2	**50.0**	47.2	44.4	**50.0**	47.2
topic_reasoning	**85.7**	79.1	81.3	**84.6**	**84.6**	80.2	84.6	**86.8**	84.6	83.5	**85.7**	84.6
tutorialQA	53.5	51.2	**60.5**	44.2	**69.8**	65.1	58.1	60.5	**74.4**	58.1	**67.4**	**67.4**

Table 13: **Question-type accuracy on MLVU.** Scores (%) compare Uniform FS (U), Training-Free VEI (TF), and Self-Distilled VEI (SD) at each preview resolution.

H LIMITATIONS

Our evaluation focuses on a small set of 8B VLMs, and results may vary with model scale and pretraining. We report single-run evaluations without estimates of training-seed variability. Broader adaptive-selector comparisons would extend our uniform, CLIP-retrieval, and nested-VEI controls.

Our CLIP† baseline uses a publicly pretrained CLIP model to embed the question and each candidate frame independently. We rank frames by question–image embedding similarity and select the top 32, retaining the same preview and high-resolution evidence budget as Training-Free VEI.

The localization analysis measures temporal evidence recovery; direct measurements of target coverage and visual redundancy remain outside this study. The distilled checkpoint performs both indexing and answering, so the QA results characterize the complete pipeline without separately quantifying changes in answering behavior.

Very low-resolution previews can obscure evidence needed for indexing, while a fixed-size Evidence Set may under-cover extended events. VEI processes completed videos and does not address causal online streaming. Our efficiency analysis focuses on context capacity; end-to-end latency and memory across both inference calls remain to be characterized.