D^3-OBSERVE: A DIFFERENCE-AWARE, DELAYED-COMMITMENT, DUAL-ROUTE OBSERVATION AGENT FOR EFFICIENT VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video question answering still lacks a reliable way to identify the visual evidence that distinguishes competing answers under a limited observation budget. Relevance-based selection may capture the question topic yet omit the state, order, count, or relation needed to determine the answer. A separate bottleneck arises when textual intermediates omit visual details available in the original frames. To address these limitations, we present D^3-Observe, a training-free observation agent that uses the question and all answer options to select visual evidence before committing to an answer. It then passes the retained RGB frames directly to Vision-Language Models (VLMs) in a single call. Additionally, a label-free shared-observation route amortizes visual processing across multiple questions about the same video. Across five benchmarks spanning long and short videos (Video-MME, MLVU, EgoSchema, NExT-QA, and LVBench), D^3-Observe achieves 64.94% macro accuracy with an average of 10.21 frames sent per question. It improves upon uniform 32-frame inference with Qwen3.5-27B by 2.02 percentage points (95% CI [$+0.84$, $+3.21$]) while reducing sent frames by 68.1%. By holding frame inputs fixed, we further isolate the effect of the textual bottleneck and show that direct visual answering consistently outperforms text-mediated answering.

1 INTRODUCTION

Vision-Language Models (VLMs) now support multimodal reasoning across both images and videos (Alayrac et al., 2022; Li et al., 2023; Lin et al., 2024; Wang et al., 2024a; Li et al., 2024; Zhu et al., 2025). Yet VideoQA remains demanding because a system must locate sparse events, preserve temporal relations, and distinguish visually similar alternatives across short clips and videos lasting minutes or hours. Model capability alone does not resolve the cost of observing such videos. Sending a dense frame sequence to a foundation VLM incurs substantial visual-token and inference cost, whereas aggressive sequence reduction can remove the few observations that determine the answer. Efficient VideoQA is therefore an evidence-acquisition problem whose central criterion is whether limited observations preserve the visual differences between competing answer hypotheses.

The required evidence is sparse, heterogeneous, and conditioned on competing answers. Uniform sampling covers time broadly but spends capacity on redundancy and can miss brief events. Caption-based pipelines reduce direct visual input, yet add preprocessing cost and may omit the detail that separates plausible options (Zhang et al., 2024a; Wang et al., 2024b; 2025b). Two requirements remain unmet. The first is to preserve candidate-neutral evidence before answer formation, and the second is to reuse one visual observation when several questions share a video.

One line of work performs query-conditioned frame selection to reduce visual input. These methods improve diversity, temporal search, dynamic resolution, or list-wise selection (Yu et al., 2023; Liu et al., 2025; Sun et al., 2025; Zhang et al., 2025; Tang et al., 2025; Ye et al., 2025). However, question relevance is not equivalent to decision utility. Figure 1 illustrates this mismatch. Frames that visibly contain the queried object may still omit the state or temporal relation that separates the options, while a caption can preserve the topic and lose the decisive pixels. The relevant observation target

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: **From query relevance to difference coverage.** (a) Query–frame similarity and caption pipelines rank observations by overlap with the question. Frames that contain queried objects can receive high scores while omitting the order or state that separates the options, while verbalization can keep the topic and drop those pixel distinctions. (b) D^3-Observe derives candidate-neutral evidence axes from the question and all options before any answer is formed, then retains frames that cover those differences.

is therefore defined by the differences between the options, which motivates a shift from relevance ranking to candidate-neutral difference coverage.

Another line of work uses semantic models to inspect and revisit video evidence. VideoAgent (Wang et al., 2024b) and VideoTree (Wang et al., 2025b) reason over caption-derived representations, while VCA (Yang et al., 2025) uses iterative VLM analysis for exploration and answering. A.I.R. (Zou et al., 2026) first applies a CLIP-based coarse pass, followed by iterative VLM analysis for exploration and answering. Caption mediation can omit visual distinctions, whereas iterative inspection increases calls and makes later acquisition depend on earlier judgments. In our frozen five-benchmark aggregate, the A.I.R. route requires 38.424 amortized VLM calls per question (Table 9). These observations point to a shared gap: an efficient system must identify which evidence distinguishes the options before answering, while avoiding repeated visual processing when questions share a video. Candidate-neutral difference coverage addresses the first requirement, delayed commitment keeps answer formation after evidence acquisition, and shared observation addresses the second by amortizing visual encoding.

We introduce D^3-Observe, a unified video observation agent centered on candidate-neutral difference coverage. Rather than asking which frames match the question, it asks which observations can distinguish the competing answer hypotheses. The framework realizes this principle under two workload regimes. On the per-question route, the question and all options are processed symmetrically to derive observable evidence axes for action, state, count, order, boundary, or identity. A training-free, CLIP-free selector combines these axes with uniform temporal coverage and lightweight visual signals to allocate an adaptive frame budget. On eligible long videos with repeated questions, a label-free router instead groups several questions around one uniformly sampled raw-frame observation. This shared route does not construct a text index, and each question is still decided from the original pixels and its own options. Across both routes, delayed commitment prevents answer formation until observation allocation is complete, and no separate text-only LLM intervenes. Candidate-neutral difference coverage therefore defines the evidence target, delayed commitment preserves that evidence, and shared visual observation amortizes its acquisition cost. Across five benchmarks, D^3-Observe achieves 64.94% macro accuracy with 10.21 sent frames on average. It exceeds the 32-frame Qwen3.5-27B (Qwen Team, 2026) baseline by 2.02 percentage points while reducing the average number of sent frames by 68.1%.

Our contributions are threefold:

- We reformulate per-question frame acquisition as candidate-neutral difference coverage. A training-free, CLIP-free selector derives observable evidence axes symmetrically from all options and allocates frames to preserve their distinctions under a constrained visual budget.
- We introduce difference-preserving delayed commitment. Evidence acquisition is completed before answer formation, and the answering VLM operates directly on retained pixels without a text-only LLM intermediary.
- We extend this evidence-preservation principle with a label-free shared visual observation route for long multi-question videos. The router uses video duration and question multiplicity alone to answer several questions from one jointly encoded frame set, while all other inputs retain per-question difference-aware acquisition.

2 Related Work

VLMs for long-form video understanding. Modern VLMs acquire multimodal knowledge from large-scale pre-training (Alayrac et al., 2022; Li et al., 2023) and process images and videos within unified architectures (Lin et al., 2024; Wang et al., 2024a; Li et al., 2024; Zhu et al., 2025). Long-video systems extend their usable horizon through long-context transfer or activation memory (Zhang et al., 2024b; Balazevic et al., 2024). Other systems use captions, retrieved text, or learned memories to support video understanding (Zhang et al., 2024a; Luo et al., 2025; Ma et al., 2025; Yuan et al., 2025). Textual intermediates may omit the evidence that separates plausible answers. D^3-Observe treats preservation of these option-discriminative details as the central observation objective.

Query-conditioned frame selection. Frame selectors use CLIP-style alignment (Radford et al., 2021), multimodal models, or diversity criteria to optimize query relevance, temporal coverage, resolution, or redundancy (Yu et al., 2023; Liu et al., 2025; Sun et al., 2025; Zhang et al., 2025; Tang et al., 2025; Ye et al., 2025; Song et al., 2026; Li et al., 2026). Recent methods also route global and local sampling or derive scores from multimodal attention (Zhang et al., 2026; Wang et al., 2026). VideoZoomer learns temporal focus with reinforcement learning (Ding et al., 2026). Query similarity alone need not identify the observation that separates competing answers, particularly for changes, order, counting, comparison, and negation. Our candidate-neutral difference coverage instead derives observable evidence axes from all answer hypotheses. Uniform coverage and training-free signals then allocate the per-question frame budget.

VLM-guided and agentic evidence acquisition. A separate family uses semantic models to guide evidence acquisition, but its model chains differ. VideoAgent (Wang et al., 2024b) and VideoTree (Wang et al., 2025b) reason over caption-derived representations. Other methods learn selection policies or apply multimodal frame assessment (Yu et al., 2025; 2023; Hu et al., 2025). VCA (Yang et al., 2025) uses one VLM as both its reward model and exploration agent. A.I.R. (Zou et al., 2026) combines CLIP-based initialization with iterative analysis and uses the same VLM backbone for analysis and answering. Caption-mediated pipelines may lose option-discriminative details during verbalization. Iterative VLM inspection preserves multimodal access, but increases model calls and conditions later search on intermediate judgments. D^3-Observe instead completes candidate-neutral evidence acquisition before answer formation. Its answering VLM reads retained pixels directly, while shared visual observation amortizes raw-frame processing without a text index.

3 Method

3.1 Overview

Let a video be V with duration T, and let a multiple-choice question be Q with an ordered option set $O = \{O_1, \ldots, O_m\}$. The system returns one option using a small number of RGB frames sent to a VLM. Sent-frame cost is measured at the model-call boundary. A frame decoded locally but omitted from every call incurs no cost, while a repeated transmission is counted each time. The objective is to construct a visual context that preserves evidence capable of separating the competing hypotheses in O before answer formation.

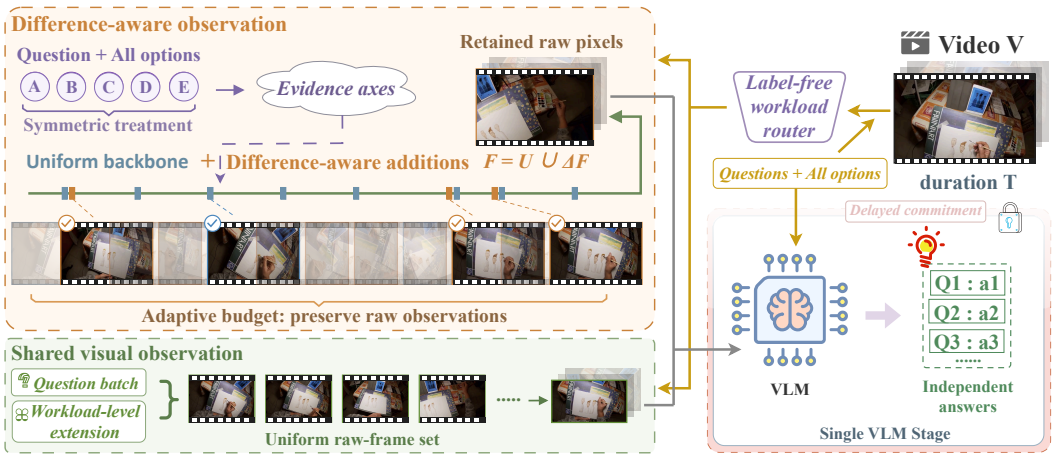

Figure 2: **Overview of D^3-Observe.** A label-free router selects shared observation for videos of at least 300 s with at least three questions, while other inputs use per-question acquisition. The individual route derives a candidate-neutral difference sketch, keeps a uniform 8-frame backbone, and adds training-free frames under an adaptive budget. The shared route encodes one raw 32-frame observation for each batch of 3–5 questions. Both routes allocate visual evidence before the answering VLM is called.

D^3-Observe separates observation from decision and then chooses between two workload routes. As shown in Figure 2, the system moves from the workload gate to the final visual context. The per-question branch builds an evidence sketch and adds frames to a uniform backbone, whereas the shared branch reuses one raw observation across a question batch. For the per-question route, a rule-first planner converts (Q, O) into a small set of candidate-neutral evidence axes, such as action, state, count, order, boundary, identity, or global activity. A training-free selector combines a uniform temporal backbone with axis-conditioned visual coverage, novelty, motion, detail, and region cues. It assembles the complete frame context before the answering VLM is called. For the shared route, a long video with several questions uses a label-free router that sends one shared raw observation to a joint VLM call, and each question keeps its own options and receives an independent answer. Thus difference coverage defines what to observe, delayed commitment defines when to decide, and shared visual observation defines how to amortize the observation cost.

3.2 Workload Routing

The router groups inputs by their resolved video path and uses no dataset name, question type, label, gold answer, subtitle, or model prediction. A group is eligible for the shared route only when $T \geq 300$ s and it contains at least three questions. Eligible groups are partitioned into balanced batches of at most five questions, and all remaining questions use the individual route. This deterministic gate is deliberately workload-level rather than answer-level, it cannot make a route decision from a task label or from a preliminary prediction. The shared route samples one chronological 32-frame raw context per batch and invokes the answering VLM once. The cost attributed to each question is the actual call cost divided by the batch size, namely $32/|\mathcal{B}|$ sent frames for a batch $\mathcal{B}$. No difference sketch or text index is constructed on this route, and sharing is limited to the raw visual observation.

Algorithm 1 summarizes this workload-level procedure. It applies the shared route only after the deterministic duration and question-count gate is met, and otherwise performs candidate-neutral selection before the answering VLM call.

3.3 Candidate-Neutral Observation

Evidence axes. Query relevance alone does not determine decision utility. A frame may contain the queried object yet leave its state, order, count, or relation to another object unresolved. We there-

Algorithm 1 D^3-Observe inference

Require: Workload $\mathcal{W}$, thresholds $\tau_T = 300\,\text{s}$ and $\tau_Q = 3$
$\mathcal{G} \leftarrow \text{GROUPBYVIDEO}(\mathcal{W})$
for each $(V, Q_V) \in \mathcal{G}$ **do**
  if $T_V \geq \tau_T$ and $|Q_V| \geq \tau_Q$ **then**
    for each $\mathcal{B} \in \text{BALANCEDBATCHES}(Q_V, 5)$ **do**
      $F \leftarrow \text{UNIFORMSAMPLE}(V, 32)$
      $\hat{\mathbf{y}}_{\mathcal{B}} \leftarrow \text{VLM}(F, \mathcal{B})$
    end for
  else
    for each $(Q, O) \in Q_V$ **do**
      $A \leftarrow \text{EVIDENCEAXES}(Q, O)$, $U \leftarrow \text{UNIFORMBACKBONE}(V, 8)$
      $B \leftarrow \text{ADAPTIVEBUDGET}(A, V)$
      $F \leftarrow \text{DIFFERENCECOVERAGE}(V, A, U, B)$
      $\hat{y} \leftarrow \text{VLM}(F, Q, O, A)$
    end for
  end if
end for

fore derive an evidence sketch from the question and all options jointly, without ranking or selecting an option. Each axis a_j is represented by an evidence type, a requirement, a preferred temporal region, and an option-value signature. Rule-first patterns cover temporal order and transition, global activity, causal relations, source or claim verification, identity and attributes, quantitative state, and local actions. If no high-precision pattern applies, the planner uses a generic visual-difference axis. Option text is used only to expose distinctions (for example, an explicitly ordered set of alternatives), and it never supplies a gold answer. The sketch is consequently a request for observable evidence, not a predicted answer.

One-shot selection. The selector decodes a local candidate pool of up to 64 timestamps. It first retains a complete, uniformly spaced 8-frame temporal backbone U, then ranks the remaining candidates with the axis-conditioned score defined as:

$$s(f) = w_c C(f) + w_n N(f) + w_m M(f) + w_d D(f) + w_e E(f) + w_r R(f), \tag{1}$$

where $w_c, w_n, w_m, w_d, w_e, w_r$ are fixed cue weights, and C measures temporal coverage, N visual novelty, M frame-to-frame change, D local detail, E image entropy, and R affinity to the axes' preferred regions. Equation 1 combines these local cues without introducing a learned selector. These features are computed locally from decoded RGB frames with OpenCV/Pillow, and no learned selector or CLIP model is used. A discrete adaptive evidence budget chooses $B \in \{8, 12, 16, 24, 32\}$ from the evidence scope and marginal coverage gain. The selected context is defined as:

$$F = U \cup \Delta F, \tag{2}$$

where ΔF contains the highest-scoring residual frames subject to the selected budget and coverage-preserving safeguards. Equation 2 makes explicit that the uniform backbone is always retained and that adaptive selection contributes only the residual set. The selector never observes a gold label, a baseline answer, or a previous answer-model output.

3.4 DELAYED COMMITMENT AND SHARED OBSERVATION

After F is fixed, the answering VLM receives the question, all options, the candidate-neutral evidence sketch, and the chronological raw frames in F. It compares every option independently and returns a structured option letter with supporting frame references when available. There is no answer-bearing scout call before selection, no text-only LLM between observation and decision, and no iterative relevance judgment that steers later acquisition. Figure 3 summarizes the two observation paths. This ordering gives delayed commitment its operational meaning. Evidence allocation finishes before answer formation, and the final decision still comes from the original pixels rather than from a caption or a text index.

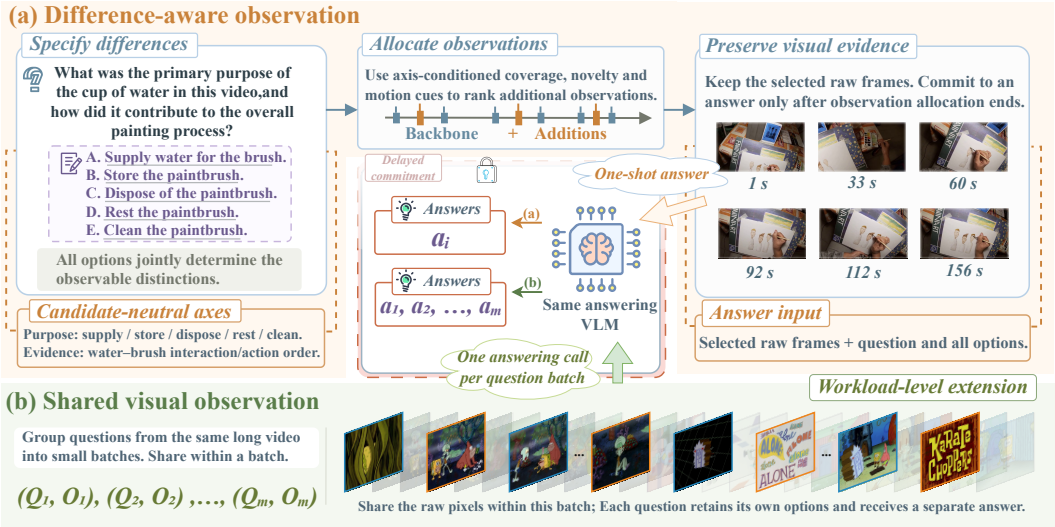

Figure 3: **Difference-preserving observation allocation.** (a) A 64-frame pool is reduced to one visual context before the first VLM call. Eight uniform frames preserve temporal coverage, and additional frames are prioritized by candidate-neutral evidence axes and local visual cues. (b) Eligible long videos share one raw 32-frame observation across several questions. Each question retains its own options and is answered from the original pixels.

The shared route addresses a different bottleneck, namely repeated questions about one long video. For each eligible batch $\mathcal{B}$, the runner uniformly samples and encodes one chronological 32-frame raw observation, then supplies that same visual context together with the full question-and-option record for every member of $\mathcal{B}$. Figure 3a illustrates the per-question interface, whereas Figure 3b illustrates the shared interface. The VLM is instructed to solve each question independently, so an answer for one question cannot be copied to another and each question retains its own candidate set. The observation is shared only at the pixel/encoding level, and it is never converted into a reusable text index. This route amortizes visual encoding across the batch. Candidate-neutral difference selection is applied on the individual route, where each question receives its own evidence sketch.

All frame selection is training-free and performed before the first answering call on the individual route. The candidate pool, uniform backbone, budget tiers, router thresholds, and batch limit are fixed by the released configuration and independent of dataset labels. For a shared batch $\mathcal{B}$, the amortized sent-frame cost per question is defined as:

$$C_{\text{sent}}(\mathcal{B}) = \frac{|F_{\mathcal{B}}|}{|\mathcal{B}|}. \tag{3}$$

where $F_{\mathcal{B}}$ is the shared RGB frame set sent to the VLM and $|\mathcal{B}|$ is the number of questions answered from that observation. Equation 3 charges the shared visual context once and then distributes that charge across the questions that receive the observation. The implementation records selected timestamps, axis types, route decisions, model-call counts, and sent-frame accounting for every question. Accuracy is reported per dataset, with route-specific strengths and failure boundaries examined in the experiments.

4 Experiments

4.1 Experimental setup

We evaluate on the official evaluation portions of Video-MME (long split), MLVU, EgoSchema, NExT-QA, and LVBench (Fu et al., 2025; Zhou et al., 2025; Mangalam et al., 2023; Xiao et al., 2021; Wang et al., 2025a). The released evaluation sizes are 900, 2174, 500, 4996, and 1549 questions, respectively. The controlled Qwen and D^3-Observe intersections contain 898, 2146, 499, 4936,

Table 1: Experimental Results on LVBench. ER, EU, KIR, TG, Rea, and Sum denote entity recognition, event understanding, key-information retrieval, temporal grounding, reasoning, and summarization. Accuracy is reported in percent. * denotes reported results.

Method	Frames	ER	EU	KIR	TG	Rea	Sum	Overall
Qwen3.5-27B (Qwen Team, 2026)	32.00	38.80	40.90	36.00	32.40	41.90	35.10	39.48
GPT-4o* (Hurst et al., 2024)	64.00	35.90	30.80	35.50	28.30	33.50	34.50	34.70
GLM4V-Plus* (Hong et al., 2024)	20.00	39.90	35.80	34.80	37.70	40.00	32.80	38.30
InternVL2-40B* (Chen et al., 2024)	16.00	37.40	39.70	**43.40**	31.40	42.50	**41.40**	39.80
TimeChat* (Ren et al., 2024)	> 96	21.90	21.70	25.90	22.70	25.00	24.10	22.30
LLaVA-NeXT* (Zhang et al., 2024c)	32.00	30.10	31.20	34.10	31.40	35.00	27.60	32.20
VideoAgent* (Wang et al., 2024b)	25.50	28.00	30.30	28.00	29.30	28.00	36.40	29.30
VideoTree* (Wang et al., 2025b)	103.20	30.30	25.10	26.50	27.70	31.90	25.50	28.80
VCA* (Yang et al., 2025)	20.00	**43.70**	40.70	37.80	**38.00**	46.20	27.30	41.30
A.I.R. (Zou et al., 2026)	31.88	42.69	38.22	41.66	37.91	46.80	28.42	41.63
D^3-Observe	7.26	40.90	**42.30**	42.30	30.10	**52.50**	40.40	**42.39**

Table 2: Experimental Results on Video-MME. Accuracy is reported in percent. For our runs, frames denote the average number sent at the model-call boundary. * denotes reported results.

Method	Frames	Accuracy
Qwen3.5-27B (Qwen Team, 2026)	8.00	52.00
Qwen3.5-27B	32.00	60.02
VideoAgent* (Wang et al., 2024b)	24.60	46.40
VideoTree* (Wang et al., 2025b)	98.00	53.10
VCA* (Yang et al., 2025)	18.10	56.30
A.I.R. (Zou et al., 2026)	31.62	52.89
D^3-Observe	10.69	**60.24**

Table 3: Experimental Results on MLVU. Accuracy is reported in percent. For our runs, frames denote the average number sent at the model-call boundary. * denotes reported results.

Method	Frames	Acc.
Qwen3.5-27B (Qwen Team, 2026)	8.00	58.57
Qwen3.5-27B	32.00	67.94
BOLT* (Liu et al., 2025)	32.00	66.80
Q-Frame* (Zhang et al., 2025)	8.00	54.40
Frame-Voyager* (Yu et al., 2025)	8.00	49.80
A.I.R. (Zou et al., 2026)	28.03	**70.09**
D^3-Observe	8.83	68.36

and 1512 questions. The smaller effective counts result from missing or invalid model outputs in the frozen evaluation logs, rather than from changing a benchmark split. The primary backbone is the deployed Qwen3.5-27B (Qwen Team, 2026) model. Its provider identifier is qwen3.5-27b. It runs with temperature zero and thinking disabled. The two Qwen settings, the GLM-4.6V (Z.ai, 2025) validation, the A.I.R. (Zou et al., 2026) runs, and D^3-Observe use the same frozen benchmark versions and question lists. The controlled baselines use uniform midpoint sampling with the same answer prompt and parser as the final VLM call. We report 8-frame and 32-frame contexts for this backbone when both settings are used.

Accuracy is computed independently for each dataset. Sent-frame cost is measured at the VLM-call boundary. Locally decoded but unsent frames are not counted, repeated transmissions are counted, and a shared batch is charged by its actual transmitted frames divided by the number of answered questions. The efficiency summary uses five metrics. Frames count transmitted RGB frames, Calls count VLM invocations per question after shared-batch amortization, and Tokens denote the recorded model-token count per question. Wall-clock is elapsed experiment time in seconds divided by the evaluated questions, rather than single-request latency. The overall frame count is sample weighted. For uncertainty, we resample videos rather than individual questions and report percentile intervals from 10,000 bootstrap replicates. This preserves within-video dependence for benchmarks containing several questions per video.

Tables compare uniform Qwen3.5-27B, A.I.R., published methods, and D^3-Observe. Our Qwen runs and method share the answer prompt, parser, and evaluation protocol. Reported results retain source-specific frame counts and require matching metrics and benchmark definitions. Results marked with an asterisk are taken from published papers.

4.2 Main results

The main comparison is reported in Table 1, which gives the capability-level LVBench view. Table 2 and Table 3 provide the longer-video aggregates, while Table 4 and Table 5 test transfer to shorter clips. Relative to the uniform 32-frame Qwen setting, D^3-Observe gains 0.22, 0.42, 6.01, 0.55, and 2.91 percentage points on Video-MME, MLVU, EgoSchema, NExT-QA, and LVBench, respec-

Table 4: Experimental Results on EgoSchema. Accuracy is reported in percent. For our runs, frames denote the average number sent at the model-call boundary. * denotes reported results.

Method	Frames	Acc.
Qwen3.5-27B (Qwen Team, 2026)	8.00	62.53
Qwen3.5-27B	32.00	67.74
GPT-4o* (Hurst et al., 2024)	8.00	69.00
GPT-4o*	32.00	70.40
VideoAgent* (Wang et al., 2024b)	10.70	60.80
VideoTree* (Wang et al., 2025b)	69.50	67.00
LVNet* (Park et al., 2024)	12.00	68.20
VCA* (Yang et al., 2025)	7.20	73.60
A.I.R. (Zou et al., 2026)	32.00	71.00
D^3-Observe	9.29	**73.75**

Table 5: Experimental Results on NExT-QA. Accuracy is reported in percent. For our runs, frames denote the average number sent at the model-call boundary. * denotes reported results.

Method	Frames	Acc.
Qwen3.5-27B (Qwen Team, 2026)	8.00	74.57
Qwen3.5-27B	32.00	79.40
LLoVi* (Zhang et al., 2024a)	90.00	67.70
LVNet* (Park et al., 2024)	12.00	72.90
VideoAgent* (Wang et al., 2024b)	8.20	71.30
VideoTree* (Wang et al., 2025b)	12.60	75.60
Hu et al. (2025)*	32.00	78.40
LangRepo* (Kahatapitiya et al., 2025)	180.00	60.90
A.I.R. (Zou et al., 2026)	22.79	**81.85**
D^3-Observe	11.71	79.94

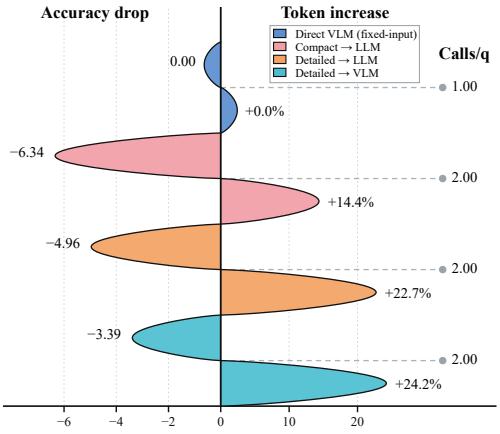

Figure 4: Fixed-input visual-to-text comparison on MLVU with Qwen3.5-27B. The same selected frames are supplied to each decision interface, which isolates the effect of the decision path.

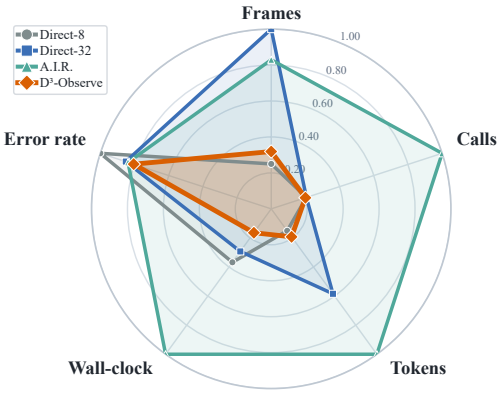

Figure 5: Efficiency comparison across the five-benchmark aggregate. Radar axes show transmitted frames, amortized VLM calls, recorded tokens, wall-clock time, and error rate. The two reference runs use Qwen3.5-27B with 8 and 32 frames. Lower values preferred.

tively, while using fewer sent frames on every dataset. The LVBench breakdown further shows that the notable gains occur in reasoning and summarization. Against A.I.R., D^3-Observe is more accurate on Video-MME and LVBench with a smaller frame budget, while the A.I.R. rows remain higher on MLVU and NExT-QA at substantially larger budgets. On those datasets, D^3-Observe offers a more efficient operating point while retaining highly competitive or even superior accuracy.

4.3 Component Analysis

Direct visual answering preserves decision-relevant evidence. To examine whether textual intermediates discard decision-critical visual details under a fixed visual input budget, we conduct an interface comparison on MLVU. Figure 4 and Table 6 present the comparison with Qwen3.5-27B (Qwen Team, 2026). LLM-mediated arms use DeepSeek-V4-Pro (DeepSeek-AI, 2026). Both displays hold the selected RGB frames fixed for each question and vary only the decision interface. Every arm receives the same mean input of 9.36 frames per question. Direct visual answering reaches 61.33%, whereas the visual-to-text interfaces reach 54.98%–57.94% and require a second model call. The same-VLM text control also remains below the direct visual path, showing that the loss is not explained by replacing the answering backbone alone. Repeating the fixed-input protocol across all five benchmarks preserves the same point-estimate ordering.

Table 6: Fixed-input visual-to-text comparison on MLVU with Qwen3.5-27B. All arms use the same selected-frame input, while only the decision interface changes.

Decision path	Accuracy	Frames	Calls	Tokens
Direct VLM	**61.33**	9.36	**1.00**	**9668.0**
Compact → LLM	54.98	9.36	2.00	11057.0
Detailed → LLM	56.37	9.36	2.00	11864.7
Detailed → VLM	57.94	9.36	2.00	12009.5

Table 7: Ablation of observation mechanisms on MLVU. Accuracy is reported in percent, and frames denote the average number sent at the model-call boundary.

Configuration	Accuracy	Frames
Full method	**68.40**	**8.84**
w/o difference coverage	64.18	8.84
w/o shared observation	61.76	9.37
w/o both mechanisms	61.81	9.37

Table 8: Cross-backbone validation using GLM-4.6V as the answering VLM. Accuracy is reported in percent.

Dataset	Model	Frames	Accuracy
Video-MME	GLM-4.6V	8	47.38
	GLM-4.6V	32	51.94
	D^3-Observe	10.68	**54.10**
EgoSchema	GLM-4.6V	8	56.20
	GLM-4.6V	32	62.37
	D^3-Observe	9.30	**63.58**
LVBench	GLM-4.6V	8	33.73
	GLM-4.6V	32	**39.02**
	D^3-Observe	7.26	38.56

Table 9: Efficiency comparison across the five-benchmark aggregate. Frames are transmitted RGB frames, calls are VLM invocations after shared-batch amortization, tokens are recorded model tokens, and wall-clock (s) is elapsed experiment time divided by evaluated questions rather than single-request latency. Error rate is 100 minus macro accuracy.

Method	Frames	Calls	Tokens	Wall-clock	Error rate (%)
Qwen3.5-27B	**8.00**	1.000	**4.64k**	22.39	43.57
Qwen3.5-27B	32.00	1.000	17.96k	17.83	37.08
A.I.R.	26.55	38.424	30.75k	60.90	36.51
D^3-Observe	10.21	**0.734**	5.92k	**10.01**	**35.06**

The two observation mechanisms provide complementary gains. Table 7 isolates difference coverage and shared observation. In this separate component-ablation evaluation, the full method reaches 68.40% with 8.84 sent frames over its $N = 2155$ common valid-question set. Removing difference coverage lowers accuracy while keeping the frame budget matched. Removing shared observation lowers accuracy and increases the amortized frame cost because each question acquires its own context. Appendix Table 12 reports paired module effects and video-level intervals.

The observation policy transfers across VLM backbones. Table 8 keeps the D^3-Observe architecture fixed and replaces only the Qwen3.5-27B answering backbone with GLM-4.6V. With identical frame budgets and parsing, the policy improves over both uniform baselines on Video-MME and EgoSchema, while LVBench remains harder than the 32-frame control.

D^3-Observe provides a more efficient operating point. Figure 5 and Table 9 summarize the five-benchmark efficiency audit. D^3-Observe has the highest macro accuracy and lowest per-question wall-clock time, using fewer frames, calls, and tokens than the 32-frame Qwen baseline and A.I.R.

The appendix records denominators, costs, intervals, paired effects, and prompt contracts used for these comparisons.

5 Conclusion

We presented D^3-Observe, a training-free observation policy for VideoQA that targets differences between answer hypotheses rather than question relevance. Candidate-neutral evidence axes guide adaptive selection, delayed commitment defers decisions to retained RGB frames, and shared observation amortizes repeated questions on eligible long videos. Across five long- and short-video benchmarks, the method reached 64.94% macro accuracy with 10.21 sent frames on average while reducing the fixed-input visual-to-text bottleneck. Direct visual answering better preserves decision-critical pixels than text-mediated interfaces, while capability scores vary by question type. Shared observation activates only when video duration and question count meet its workload gate. These results motivate richer temporal evidence models while preserving direct visual answering.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing. We reviewed all AI-assisted text and take responsibility for the final content of this work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J. Henaff. Memory consolidation enables long-context video understanding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2527–2542, 2024. URL https://proceedings.mlr.press/v235/balazevic24a.html.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. URL https://arxiv.org/abs/2404.16821.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report, 2026. URL https://arxiv.org/abs/2606.19348.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. VideoZoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ARHCFvgx6G.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. URL https://arxiv.org/abs/2408.16500.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13702–13712, 2025. doi: 10.1109/CVPR52734.2025.01279.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5627–5646, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.294. URL https://aclanthology.org/2025.findings-acl.294/.

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. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. MaxInfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7198–7207, 2026. doi: 10.1109/WACV61042.2026.00695.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3318–3327, 2025. doi: 10.1109/CVPR52734.2025.00315.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, volume 38, pp. 186208–186233, 2025. doi: 10.52202/085713-5602.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Official citation for the Qwen3.5-27B model card; model card accessed September 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Baiyang Song, Jun Peng, Yuxin Zhang, Guangyao Chen, Feidiao Yang, and Jianyuan Guo. KTV: Keyframes and key tokens selection for efficient training-free video LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):9060–9068, 2026. doi: 10.1609/aaai.v40i11.37862.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP^3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24090–24101, 2025. doi: 10.1109/ICCV51701.2025.02233.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025. doi: 10.1109/CVPR52734.2025.02711.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. URL https://arxiv.org/abs/2409.12191.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a. doi: 10.1109/ICCV51701.2025.02131.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76, 2024b. doi: 10.1007/978-3-031-72989-8_4.

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. *arXiv preprint arXiv:2607.15689*, 2026. URL https://arxiv.org/abs/2607.15689.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025b. doi: 10.1109/CVPR52734.2025.00311.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20168–20179, 2025. doi: 10.1109/ICCV51701.2025.01876.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_Re-thinking_Temporal_Search_for_Long-Form_Video_Understanding_CVPR_2025_paper.html.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76749–76771, 2023. doi: 10.52202/075280-3354.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LNL7zKvm7e.

Huaying Yuan, Zheng Liu, Minghao Qin, Hongjin Qian, Yan Shu, Zhicheng Dou, Ji-Rong Wen, and Nicu Sebe. Memory-enhanced retrieval augmentation for long video understanding. *arXiv preprint arXiv:2503.09149*, 2025. URL https://arxiv.org/abs/2503.09149.

Z.ai. GLM-4.6V. Model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V. Accessed September 2026.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024a. doi: 10.18653/v1/2024.emnlp-main.1209.

Haiyue Zhang, Yi Bin, Xun Jiang, Zeyu Ma, Duo Peng, Guoqing Wang, Yang Yang, and Heng Tao Shen. VisualRouter: Query-grounded visual sampling for long video understanding. *arXiv preprint arXiv:2607.28463*, 2026. URL https://arxiv.org/abs/2607.28463.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Q-Frame_Query-aware_Frame_Selection_and_Multi-Resolution_Adaptation_for_Video-LLMs_ICCV_2025_paper.html.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024c. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

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. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SZVpOKw0YD.

A APPENDIX

This appendix is organized into six parts. Appendix A defines the sent-frame and call accounting and reports the uncertainty summaries. Appendix B expands the component analyses that support the main claims. Appendix C gives the complete inference procedure, Appendix D records the frozen prompt contracts, Appendix E provides category-level benchmark breakdowns, and Appendix F presents qualitative route traces for representative questions. The sections use the same evaluation conventions as the main text and provide additional evidence without changing the reported benchmark results.

A.1 INFERENCE-COST ACCOUNTING

This subsection defines the accounting convention used by the efficiency table and separates transmitted-frame cost from local decoding. It also records how shared observations are charged to individual questions.

The main text efficiency display reports the aggregate operating point. Table 10 resolves that result by dataset and makes the route assignment explicit. A shared batch is charged by its transmitted frames divided by the number of answered questions, whereas an individual route is charged by its complete visual context. The reduction column uses the uniform 32-frame Qwen3.5-27B (Qwen Team, 2026) run as the reference. Calls are reported after allocating one shared VLM call across the questions in its batch. For dataset d, the reported reduction is:

$$R_d = 100\left(1 - \frac{\bar{F}_d}{32}\right), \tag{4}$$

where $\bar{F}_d$ is the sample-weighted number of RGB frames transmitted per question after route-specific accounting. This definition counts a shared raw observation once and distributes it over the questions in its batch, matching the call-boundary convention used throughout the experiments. Equation 4 therefore reports a cost reduction relative to the fixed 32-frame reference, not an accuracy difference.

Table 10: **Inference-cost accounting by benchmark.** Sent frames are counted at the model-call boundary. Reduction is relative to uniform 32-frame Qwen3.5-27B inference, and calls are amortized over questions in each shared batch. The overall row is sample weighted.

Dataset	N	Execution route	Sent frames	Reduction vs. 32 frames	Calls
Video-MME	898	Shared batches of 3	10.69	66.59%	0.333
MLVU	2146	Mixed shared and per-question	8.83	72.41%	0.583
EgoSchema	499	Per-question difference route	9.29	70.97%	1.000
NExT-QA	4936	Per-question difference route	11.71	63.41%	1.000
LVBench	1512	Shared batches of 3 to 5	7.26	77.31%	0.227
Overall	9991	Sample-weighted	**10.21**	**68.11%**	**0.734**

The per-dataset rows show how the same accounting convention behaves under different workload mixes. The lowest sent-frame counts occur on the shared routes, where one raw observation is distributed across several questions, while the overall row preserves the sample-weighted aggregate reported in the main text.

Table 11 reports the video-level bootstrap intervals used to quantify uncertainty for the five benchmark comparisons. The macro interval remains entirely above the uniform 32-frame reference, while the per-dataset intervals show where the aggregate gain is concentrated. Resampling videos rather than questions keeps these intervals aligned with the repeated-question structure of the benchmarks.

B EXTENDED COMPONENT ANALYSES

Table 12 gives paired effects for the two observation mechanisms, including video-level confidence intervals and fixed/break counts. Both mechanisms contribute positively on MLVU: difference cov-

Table 11: **Bootstrap confidence intervals.** Intervals use 10,000 bootstrap resamples with video as the resampling unit. Delta denotes D^3-Observe minus the 32-frame Qwen3.5-27B baseline in percentage points, and the bootstrap seed is 42.

Dataset	Questions	Videos	D^3-Observe [95% CI]	Qwen3.5-27B (32 frames) [95% CI]	Delta [95% CI]
Video-MME	898	300	60.24 [56.67, 63.79]	60.02 [56.70, 63.17]	$+0.22$ [-3.33, $+3.78$]
MLVU	2146	1108	68.36 [66.24, 70.47]	67.94 [65.89, 69.98]	$+0.42$ [-1.52, $+2.38$]
EgoSchema	499	499	73.75 [69.94, 77.56]	67.74 [63.53, 71.74]	$+6.01$ [$+2.61$, $+9.22$]
NExT-QA	4936	563	79.94 [78.75, 81.11]	79.40 [78.09, 80.66]	$+0.55$ [-0.36, $+1.43$]
LVBench	1512	101	42.39 [39.78, 45.03]	39.48 [36.67, 42.36]	$+2.91$ [$+0.27$, $+5.56$]
Macro	Five datasets		**64.94 [63.65, 66.20]**	62.92 [61.62, 64.17]	$\mathbf{+2.02}$ [$\mathbf{+0.84}$, $\mathbf{+3.21}$]

erage improves decision quality at a matched frame budget, while shared observation provides an additional gain and reduces the amortized visual cost of repeated questions.

Table 12: **Paired module effects on MLVU.** Effects compare the full configuration with the corresponding ablation, while intervals use video-level bootstrap resampling.

Component	Effect	95% CI	FIX/BREAK
Difference coverage	$+4.22$	[$+3.08$, $+5.36$]	126/35
Shared observation	$+6.64$	[$+4.96$, $+8.34$]	236/93

The paired contrasts in Table 13 quantify the accuracy changes induced by textual intermediates under the fixed-input interface study. DeepSeek denotes DeepSeek-V4-Pro (DeepSeek-AI, 2026).

Table 13: **Paired effects in the visual-to-text bottleneck ablation.** Delta is the first arm minus the second arm. The p values use a two-sided exact binomial test on discordant question pairs.

Contrast	Delta	95% CI	FIX/BREAK	p
Compact with DeepSeek vs. end-to-end VLM	-6.34	[-8.26, -4.35]	162/295	5.04×10^{-10}
Detailed with DeepSeek vs. end-to-end VLM	-4.96	[-6.83, -3.06]	155/259	3.62×10^{-7}
Detailed with same VLM vs. end-to-end VLM	-3.39	[-5.04, -1.69]	123/194	7.92×10^{-5}
Detailed vs. compact, both with DeepSeek	$+1.38$	[-0.33, $+3.06$]	187/158	0.132
DeepSeek vs. same VLM, both using detailed text	-1.57	[-3.36, $+0.18$]	158/191	0.087

The paired table shows that each text-mediated route loses accuracy relative to direct visual answering under the same selected-frame input. The largest negative effects occur when the compact or detailed representation is passed to a separate language model, whereas the same-VLM text route narrows but does not remove the gap. These paired effects motivate the cross-benchmark breakdown that follows.

Scope and accounting of the fixed-input study. This study examines whether an intermediate textual representation can discard visual evidence that remains available in the original frames. After visual observation has been fixed, the direct VLM path, the compact-text path with an LLM, the detailed-text path with an LLM, and the detailed-text path with the same answering VLM receive the identical selected RGB frames. The frame set may differ between datasets, but it does not differ between interfaces within a dataset. Accuracy is computed on the common set of questions for which all four interfaces return a completed and validly parsed answer. Frames count transmissions at the model-call boundary, calls count the required model invocations after any shared accounting, and tokens record the corresponding input and output usage. The paired intervals resample videos rather than individual questions to preserve within-video dependence. The adaptive selector, shared-observation route, and workload router are not applied in this control study, so its values isolate the decision interface and should not be interpreted as full D^3-Observe results.

The common sets contain 824, 2097, 464, 4770, and 1494 questions on Video-MME, MLVU, EgoSchema, NExT-QA, and LVBench, respectively. They are drawn from the same evaluation records used by the corresponding main experiments after removing questions that fail the frozen observation validation, while the benchmark splits remain unchanged. These counts are smaller than the main strict intersections because this study additionally requires a common valid output from all four interfaces. Every retained question has one such output, and the audit records no unresolved model errors, frame-count mismatches, or answer leakage. The four paths use Qwen3.5-27B (Qwen Team, 2026) for visual observation, while the two LLM-mediated paths use DeepSeek-V4-Pro (deepseek-v4-pro) under the released prompt contract. Compact descriptions are capped at 800 characters, detailed descriptions retain the longer evidence ledger, and the final text-answering stage is limited to 32 tokens. These controls keep the comparison focused on the representation passed to the decision stage.

Applying this protocol across all five benchmarks gives the same ordering of point estimates. Table 14 reports the paired effects and their uncertainty, while the point estimates for the four decision paths are given in Table 15. The intervals exclude zero for all three text-mediated paths on Video-MME, MLVU, and NExT-QA, whereas they are wider on EgoSchema and LVBench. Overall, 11 of the 15 paired intervals exclude zero. Detailed text improves over compact text on every benchmark by point estimate. The detailed path nevertheless remains below the direct visual path on all five datasets, and the same-VLM text control is also below direct visual answering at every point estimate.

Table 14: **Paired cross-benchmark effects of text-mediated decision interfaces.** Delta is text-mediated accuracy minus direct visual accuracy in percentage points. Intervals use 10,000 video-level bootstrap resamples with seed 42, preserving the dependence among questions from the same video.

Dataset	Text-mediated decision path	Delta	95% CI
Video-MME	Compact → DeepSeek	-6.31	$[-9.66, -3.04]$
	Detailed → DeepSeek	-4.98	$[-8.36, -1.62]$
	Detailed → same VLM	-4.13	$[-7.24, -0.98]$
MLVU	Compact → DeepSeek	-6.34	$[-8.26, -4.35]$
	Detailed → DeepSeek	-4.96	$[-6.83, -3.06]$
	Detailed → same VLM	-3.39	$[-5.04, -1.69]$
EgoSchema	Compact → DeepSeek	-2.16	$[-6.03, +1.72]$
	Detailed → DeepSeek	-0.65	$[-4.09, +2.80]$
	Detailed → same VLM	-0.43	$[-3.88, +2.80]$
NExT-QA	Compact → DeepSeek	-5.41	$[-6.53, -4.24]$
	Detailed → DeepSeek	-2.12	$[-3.22, -1.03]$
	Detailed → same VLM	-1.34	$[-2.31, -0.35]$
LVBench	Compact → DeepSeek	-6.16	$[-8.74, -3.60]$
	Detailed → DeepSeek	-3.75	$[-6.34, -1.13]$
	Detailed → same VLM	-1.74	$[-3.99, +0.41]$

Table 15 gives the corresponding fixed-input accuracies. The direct visual path has the highest point estimate in all 15 pairwise comparisons against the three text-mediated paths. This ordering, together with the paired intervals in Table 14, shows that the observed gap is not explained by replacing the answering backbone alone.

Accuracy and cost move in the same direction. The fixed-input comparison does not trade accuracy for a cheaper decision stage. In the sample-weighted aggregate, direct visual answering uses 1.002 calls and 6456.5 recorded tokens per question, whereas the compact, detailed, and same-VLM text routes use 2.000 calls and 7922.5, 8788.7, and 8948.4 tokens, respectively. Thus every text-mediated route adds a model call and increases token usage by 22.7%, 36.1%, or 38.6% relative to the direct path, while remaining less accurate. These values are descriptive costs for the fixed-input interface study and are separate from the full-pipeline cost accounting reported in Appendix A.

Table 15: **Fixed-input decision-interface comparison across five benchmarks.** Within each dataset, the four decision paths receive the same selected frames and only the interface changes. Accuracy is reported in percent. N is the common set of questions with a completed, validly parsed output for every interface. These rows evaluate the interface effect under fixed visual input and do not represent the complete D^3-Observe pipeline.

Dataset	N	Direct VLM	Compact → LLM	Detailed → LLM	Detailed → VLM
Video-MME	824	**54.13**	47.82	49.15	50.00
MLVU	2097	**61.33**	54.98	56.37	57.94
EgoSchema	464	**74.78**	72.63	74.14	74.35
NExT-QA	4770	**80.02**	74.61	77.90	78.68
LVBench	1494	**38.69**	32.53	34.94	36.95

Algorithm 2 Complete D^3-Observe inference and sent-frame accounting

Require: $\mathcal{W} = \{(V_i, Q_i, O_i)\}_{i=1}^{n}$, $\tau_T = 300\,\mathrm{s}$, $\tau_Q = 3$, $K_{\max} = 5$
Ensure: Predictions $\hat{\mathcal{Y}}$ and sent-frame cost C

$\hat{\mathcal{Y}} \leftarrow \emptyset, C \leftarrow 0$
$\mathcal{G} \leftarrow$ GROUPBYRESOLVEDVIDEOPATH($\mathcal{W}$)
for each $(V, Q_V) \in \mathcal{G}$ **do**
  if $T_V \geq \tau_T$ and $|Q_V| \geq \tau_Q$ **then**
    for each $\mathcal{B} \in$ BALANCEDBATCHES($Q_V, K_{\max}$) **do**
      $F_{\mathcal{B}} \leftarrow$ UNIFORMSAMPLE($V, 32$)
      $\hat{\mathbf{y}}_{\mathcal{B}} \leftarrow \mathrm{VLM}(F_{\mathcal{B}}, \{(Q, O) : (Q, O) \in \mathcal{B}\})$
      Append one independently parsed answer per question
      $C \leftarrow C + |F_{\mathcal{B}}|$
    end for
  else
    for each $(Q, O) \in Q_V$ **do**
      $A \leftarrow$ RULEFIRSTEVIDENCEAXES(Q, O)
      $P \leftarrow$ DECODECANDIDATEPOOL($V, 64$)
      $U \leftarrow$ UNIFORMBACKBONE($P, 8$)
      for each $f \in P \setminus U$ **do**
        $s(f) \leftarrow w_c C_f + w_n N_f + w_m M_f + w_d D_f + w_e E_f + w_r R_f$
      end for
      $B \leftarrow$ ADAPTIVEEVIDENCEBUDGET($A, P, \{8, 12, 16, 24, 32\}$)
      $\Delta F \leftarrow$ COVERAGECONSTRAINEDTOP($P \setminus U, s, B - |U|$)
      $F \leftarrow$ CHRONOLOGICALSORT($U \cup \Delta F$)
      $\hat{y} \leftarrow \mathrm{VLM}(F, Q, O, A)$
      Append the structured or format-recovered answer
      $C \leftarrow C + |F|$
    end for
  end if
end for
return $\hat{\mathcal{Y}}, C$

C COMPLETE INFERENCE PROCEDURE

Algorithm 2 expands the compact procedure in Algorithm 1. The local candidate pool is decoded without incurring sent-frame cost. Both routes fix their complete visual context before the answering VLM is invoked. The shared route reuses raw pixels but does not construct a difference sketch, and neither route inserts a text-only LLM. For an individual question, the selected context is the union of the uniform backbone and the coverage-constrained residual set. For a shared batch, the same raw context is sent once and its cost is distributed according to Equation 3. The pseudocode therefore makes explicit which operations affect model-call accounting and which local decoding steps remain outside the sent-frame budget.

D PROMPT AND INTERFACE CONTRACT

The experiments use frozen prompt contracts rather than dataset-specific answer instructions. This appendix records the model-visible interface so that the observation mechanism is separable from the final decision format.

Reasoning Prompt (Per-question route)

System Prompt. You are a visual question answering model. You receive one question, its complete ordered option set, a candidate-neutral evidence sketch, and a chronological set of RGB frames. The evidence sketch describes what must be checked, not which option is correct. Inspect the original pixels and compare every option independently. Preserve temporal order, action, state, count, identity, and relations between entities. If an event is not visible, treat it as unknown rather than false. Do not infer an answer from the axis name, the frame timestamp, or the position of an option.
User. Question: `<question>`
Options: `<option_1>`, `<option_2>`, ..., `<option_m>`
Evidence sketch: `<axis_type>, <requirement>, <temporal_region>`
Frames: `<frame_1>, <frame_2>, ..., <frame_n>`
Output format. Return exactly one JSON object with fields `answer`, `evidence`, and `frames`. The first field contains one option letter and the last field contains the valid frame indices.
Do not return a ranking, a second answer, hidden assumptions, or a claim that cannot be verified from the supplied frames.

Shared-observation Prompt (Shared route)

System Prompt. You are a visual question answering model working on a shared raw observation. The frame set belongs to one video and is supplied once for a batch of questions. Solve each question independently with its own ordered options. The answer to one question must not be copied into another question and must not be used as evidence for another question. Use only the original pixels and the question records. Dataset labels, answer keys, annotated evidence times, captions, and text indexes are unavailable.
User. Shared frames: `<frame_1>, ..., <frame_32>`
Question records: `<question_id, question, ordered options>` for each batch member.
Output format. Return exactly one answer entry for every question identifier, preserving the input order. Each entry must contain one option letter and a short reason grounded in the shared frames. Do not merge questions, omit an identifier, output a global answer, or introduce information from any external text source.

Representative input. Figure 3a shows an illustrative per-question contract, while Figure 3b shows the shared contract on a question batch. These examples illustrate the two interfaces only, and all five benchmarks use the same route predicates, option-symmetric evidence rules, and output parser.

E DETAILED BENCHMARK RESULTS

This section reports category-level breakdowns, audit results, and additional experimental analyses that are not repeated in the main text. Our runs use the same model-call boundary and frozen evaluation records as the main experiments, so the detailed views can be read as decompositions of the main results rather than as separate evaluations.

The first breakdown asks whether the aggregate Video-MME result is driven by a single domain or is distributed across different types of video content. Table 16 compares the two uniform Qwen settings, published agentic baselines, A.I.R., and D^3-Observe under the same domain headings used in the benchmark. VideoAgent, VideoTree, and VCA results are from Yang et al. (2025), with source-reported frame counts, while other frame counts use model-call-boundary accounting.

Table 16: **Video-MME results by domain.** Accuracy is reported in percent. * denotes reported results.

Method	Frames	Artistic	Film/TV	Knowledge	Life	Multi.	Sports	Avg.
Qwen3.5-27B (Qwen Team, 2026)	8.00	51.67	45.00	63.70	46.67	33.33	47.97	52.00
Qwen3.5-27B	32.00	57.50	58.33	**71.85**	46.19	**63.33**	60.81	60.02
VideoAgent* (Wang et al., 2024b)	24.60	47.50	42.50	52.20	44.70	36.60	42.70	46.40
VideoTree* (Wang et al., 2025b)	98.00	51.60	52.50	60.70	49.50	40.00	48.60	53.10
VCA* (Yang et al., 2025)	18.10	**65.80**	55.00	56.90	51.90	46.70	59.30	56.30
A.I.R. (Zou et al., 2026)	31.62	54.23	61.25	57.58	47.06	52.22	44.99	52.89
D^3-Observe	10.69	56.67	**65.00**	62.59	**55.24**	43.33	**65.54**	**60.24**

The domain results are distributed rather than concentrated in one category. D^3-Observe reaches the highest average accuracy with fewer frames than the 32-frame Qwen setting and all agentic baselines, while the 8-frame Qwen setting uses slightly fewer frames but is less accurate. Its largest margins appear on Film/TV, Life, and Sports, while VCA remains strongest on Artistic. This pattern supports the main claim that the observation policy improves decision-relevant coverage across heterogeneous content rather than only on one domain.

The second breakdown tests whether the MLVU aggregate is consistent across its task families. Table 17 reports the task-level accuracies alongside the same frame accounting used in the main tables.

Table 17: **MLVU task-level results.** Accuracy is reported in percent, and frames denote the average number sent at the model-call boundary.

Method	Frames	PQA	NQA	AC	AO	ER	AR	TR	Avg.
Qwen3.5-27B (Qwen Team, 2026)	8.00	61.74	62.78	15.20	39.84	59.43	68.00	**91.02**	58.57
Qwen3.5-27B	32.00	77.08	73.30	20.10	55.47	**67.14**	**73.50**	89.06	67.94
A.I.R. (Zou et al., 2026)	28.03	**79.93**	85.35	**36.89**	**66.80**	65.90	72.50	83.27	**70.09**
D^3-Observe	8.83	71.59	**86.08**	21.08	52.73	66.29	72.00	90.62	68.36

D^3-Observe achieves 68.36% accuracy with 8.83 sent frames per question, improving over the 32-frame Qwen baseline by 0.42 percentage points while reducing frame transmissions by 72.41%. It also uses 68.50% fewer sent frames than A.I.R. The task breakdown shows complementary strengths: D^3-Observe leads NQA; A.I.R. leads PQA, AC, and AO; the 32-frame Qwen baseline leads ER and AR; and the 8-frame baseline leads TR.

Finally, Table 18 resolves the short-video transfer result by question type. It shows where each operating point gains or loses accuracy without introducing a second frame-accounting convention.

Table 18: **NExT-QA results by question type.** Accuracy is reported in percent, and frames are average transmissions at the model-call boundary.

Method	Frames	CH	CW	DC	DL	DO	TC	TN	TP	Avg.
Qwen3.5-27B (Qwen Team, 2026)	8.00	79.67	73.87	61.14	88.32	85.81	77.25	64.89	67.92	74.57
Qwen3.5-27B	32.00	84.72	78.86	70.29	91.07	**88.45**	81.37	71.12	58.49	79.40
A.I.R. (Zou et al., 2026)	22.79	**84.76**	**82.41**	**75.35**	91.09	87.14	**83.90**	**73.85**	**73.18**	**81.85**
D^3-Observe	11.71	82.94	80.07	69.14	**92.44**	86.80	80.15	73.50	69.81	79.94

The category results separate the strengths of the compared methods. A.I.R. reaches the highest average in this breakdown, but uses 22.79 sent frames compared with 11.71 for D^3-Observe. Across all methods, D^3-Observe leads DL, the 32-frame baseline leads DO, and A.I.R. leads the remaining categories. The detailed results therefore show a clear accuracy–cost trade-off on short-video transfer, with D^3-Observe retaining competitive accuracy at a much smaller visual budget.

F Qualitative Route Traces

This section illustrates how the two workload routes turn a question and its answer options into a visual context. The examples are qualitative traces rather than additional benchmark measurements.

Each trace shows the question, the candidate options, the initial RGB input, the route decision, the evidence contract, the retained frames, and the final answer interface. The examples cover both per-question and shared observation, and include temporal order, object identity, appearance, location, and surveillance-style activity.

Figures 6 and 7 show per-question cases. Their route traces derive option-symmetric evidence requirements before the answering VLM is called, then retain a small set of frames that exposes the relevant object, action, or temporal relation. The displayed frames are the actual visual inputs used by the final decision interface.

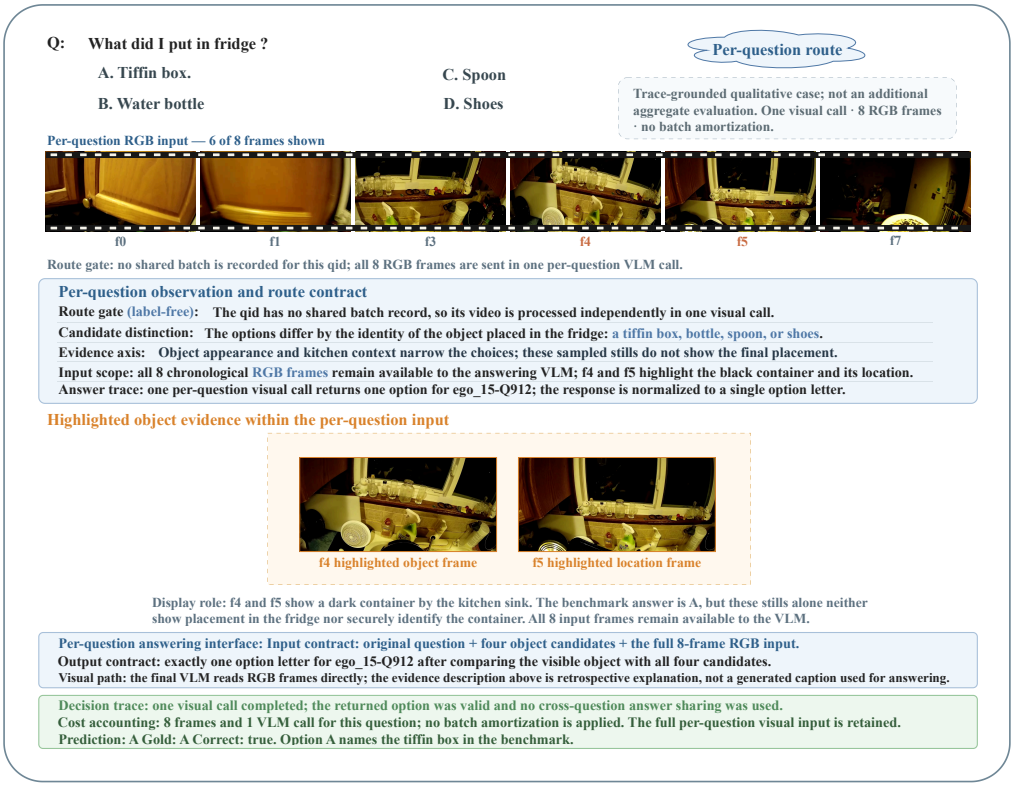

Figure 6: Per-question route on an EgoSchema object-identity question. The displayed frames show the sampled kitchen context and the model's answer trace.



Figure 7: Per-question route on a temporal-order question. The displayed frames illustrate the sampled context and answer trace for the candidate event orders.

Figures 8, 9, and 10 show shared-observation cases. A single raw frame context is sent for the question batch, while each question preserves its own options and answer record. The traces make the amortization boundary visible: sharing occurs at the visual observation level, not by copying one question's answer into another.



Figure 8: Shared-observation route for an appearance question. The same raw observation is reused across the batch, while each question retains its own options and answer record.



Figure 9: Shared-observation route for a location question. The shared context and answer trace are shown alongside illustrative frames from the video.

The remaining traces demonstrate that the same interface applies to activity recognition and phone-call context. Figure 10 shows a shared route whose options concern a person's location and interaction, while Figure 11 shows a per-question route that distinguishes irregular activity from ordinary store movement. Together, these cases show that the route choice changes the observation schedule, not the requirement that the final answer be grounded in the retained pixels.



Figure 10: Shared-observation route for a phone-call question. The retained frames preserve both the shoreline context and the person's interaction.

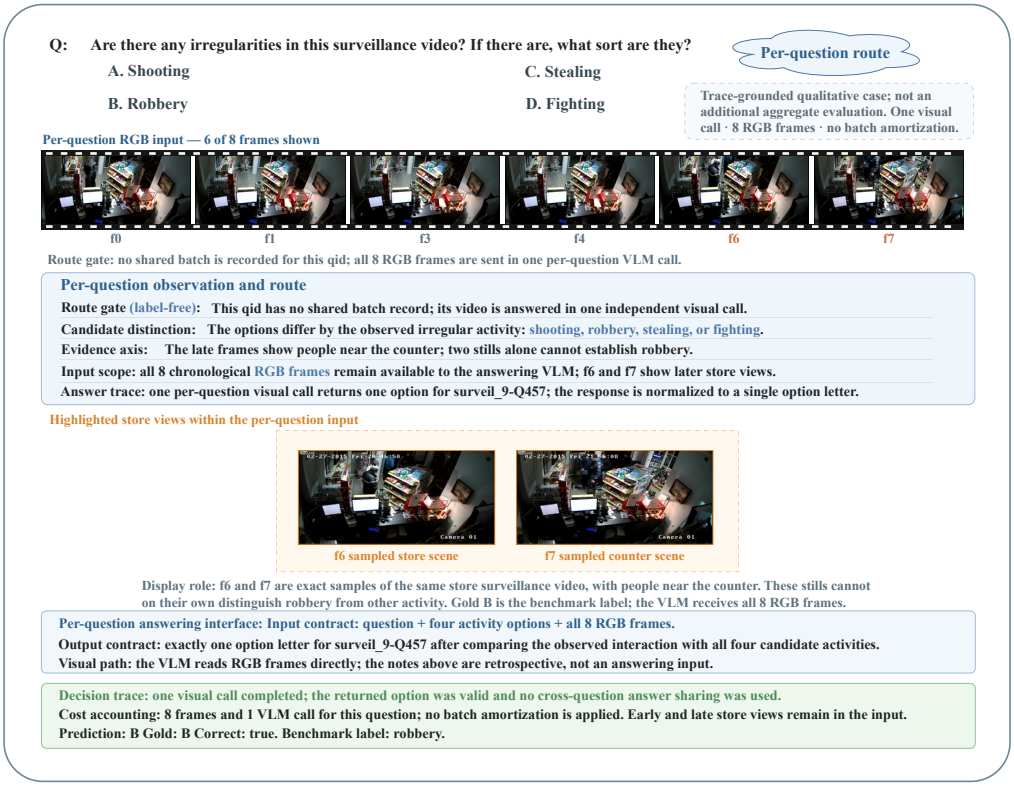

Figure 11: Per-question route for surveillance activity recognition. The displayed frames show the sampled store context and the model's answer trace.

Taken together, Figures 6–11 show that the two routes differ in when visual evidence is acquired and amortized, not in the standard used to justify the final answer. Per-question traces retain frames that separate competing states or event orders, whereas shared traces reuse one raw observation while keeping the question records independent. Across both routes, the answer is formed only after the retained RGB evidence has been exposed to the answering VLM.