Egocentric Daily Video Question Answering with Token-Efficient Storyboard Retrieval

Zihao Ding
Rutgers University
Piscataway, NJ, USA
zd75@rutgers.edu

Tun-Yuan Chang
National Tsing Hua University
Hsinchu, Taiwan
tunyuan@gapp.nthu.edu.tw

Cheng-Hsin Hsu
National Tsing Hua University
Hsinchu, Taiwan
chsu@cs.nthu.edu.tw

Yao Liu
Rutgers University
Piscataway, NJ, USA
yao.liu@rutgers.edu

Abstract

Always-on AI glasses enable continuous egocentric video recording that can support daily video question answering (QA). However, processing hours of personal footage poses significant challenges under the compute and token constraints of modern vision-language models (VLMs). In this paper, we present a token-aware retrieve-then-reason architecture for egocentric QA that indexes visual logs and retrieves a small set of relevant evidence for VLM inference. Using a two-tier evaluation, we first establish an intrinsic reasoning baseline for single video QA, and then measure the end-to-end behavior when retrieval must search a personal video archive. Across two datasets, we identify a critical "modality gap": unlike narrative web videos that rely on audio, egocentric QA depends heavily on visual evidence. Furthermore, we find that storyboards, by packing multiple frames in a chronologically arranged grid, provide a strong token-accuracy trade-off, achieving similar or better accuracy than frame-based baselines with up to 85% fewer prompt tokens. Tile-level retrieval further improves accuracy at a higher token cost. Our results highlight the importance of evidence packaging and retrieval policies for practical, resource-constrained daily QA.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; **Multimedia information systems**; • **Human-centered computing** → *Ubiquitous and mobile computing*.

Keywords

Egocentric video, Multi-video retrieval, Storyboard representations, Vision-Language Models, Token-efficient inference

NOSSDAV '26, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2534-0/2026/04
https://doi.org/10.1145/3798065.3798076

ACM Reference Format:
Zihao Ding, Tun-Yuan Chang, Cheng-Hsin Hsu, and Yao Liu. 2026. Egocentric Daily Video Question Answering with Token-Efficient Storyboard Retrieval. In *The 36th edition of the Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '26), April 04–08, 2026, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3798065.3798076

1 Introduction

Recent advances in wearable sensing have led to AI glasses equipped with cameras and microphones that enable continuous first-person capture [**?**]. Unlike smartphones, which require deliberate capture, such devices could support always-on egocentric sensing with minimal user intervention. Although continuous recording is not yet a standard consumer feature, it motivates egocentric daily video question answering (QA), where users ask natural language questions about recent experiences (e.g., "Where did I leave my keys?") and receive answers grounded in their visual history [**? ?**].

However, supporting daily video QA poses significant systems challenges. Egocentric video recording generates hours of continuous data per day, while modern vision-language models (VLMs) are constrained by limited context window length and token budget. In practice, these constraints motivate a **retrieve-then-reason** architecture where the system stores and indexes compact evidence and retrieves a small set of query-relevant evidence (e.g., visual frames, storyboards, audio transcripts) to fit within the VLM's context limits [**? ?**].

Furthermore, egocentric video data is intrinsically different from traditional narrative-driven content found in standard video benchmarks (e.g., Video-MME [**?**]). In narrative videos, significant information is conveyed through speech, allowing queries to be answered from audio alone. In contrast, daily egocentric video data is continuous, unedited, and often unnarrated, meaning the audio signal can be far less information-dense. As a result, methods optimized for narrative videos may fail in this domain. This necessitates a rigorous evaluation of how evidence representation and retrieval choices trade off answer quality for system cost when the system is forced to rely primarily on visual data.

In this paper, we present a token-aware architecture for egocentric QA that enables the "retrieve-then-reason" pipeline under strict token budgets. To optimize token efficiency within the VLM's limited context window, we use storyboard representations, packing temporal sequences into dense grids, rather than retrieving individual video clips. We evaluate this design using a two-tier protocol: first, we establish an intrinsic reasoning baseline of model capabilities using ground-truth evidence, and second, we measure end-to-end system performance when retrieval must efficiently locate evidence within a large personal multi-video archive.

Overall, our work makes the following contributions:

- **A token-aware prototype:** We adapt a multi-video retrieval pipeline [**?**] to reason over full-day egocentric logs. Unlike generic QA baselines, our prototype system is explicitly designed to balance the precision of retrieved visual evidence against strict prompt-token budgets of modern VLMs.
- **Quantifying the "modality gap."** We identify a clear divergence between narrative and egocentric data. While web video benchmarks often rely on audio, our experiments show that daily egocentric QA depends critically on robust visual evidence.
- **The efficiency of storyboards.** We demonstrate that packing frames into dense *storyboards* yields a strong token–accuracy trade-off, matching or outperforming frame-based baselines on egocentric data at substantially lower prompt-token cost.

2 Background and Problem Formulation

2.1 Egocentric Daily Video QA

Egocentric sensing is becoming increasingly feasible with commercial AI glasses and wearable cameras that support first-person video capture (often paired with microphones) with minimal user intervention [**?**]. This enables a new interaction paradigm: *egocentric daily video question answering*, where a user asks natural-language questions about their own recent experiences, and the system answers by grounding the response in the captured visual record [**? ? ? ?**].

We study daily video QA as a multimedia system: the input is a growing personal archive of long-form, unedited, first-person recordings, and the system must bridge high-fidelity capture with resource-constrained inference. Let $\mathcal{V} = \{V_1, \ldots, V_N\}$ denote an archive of video segments captured over a day (or longer), and let q be a query asked after capture. The system returns an answer a (in our case, a multiple-choice label) grounded in evidence from $\mathcal{V}$. Unlike conventional VideoQA tasks that assume the relevant clip is known, daily QA must operate over an archive: it must first determine which parts of which videos contain evidence relevant to q before reliable reasoning is possible [**? ?**].

Throughout the paper, we use information modalities to refer to the types of signals that may support answering q, e.g., visual frames or storyboards, audio transcripts, and extracted text, such as optical character recognition (OCR).

2.2 Related Work

Egocentric daily QA is related to temporal grounding and moment retrieval, which localize a time interval in an untrimmed video given a text query. Earlier benchmarks (e.g., DiDeMo and Charades-STA/TALL) focused on aligning language with fine-grained temporal segments [**? ?**]. However, the daily setting introduces an additional axis: *multi-video ambiguity*. When an archive spans many recordings, a query may match multiple visually similar moments across different videos (e.g., repeated routines), and the system must identify both the correct video and the relevant moment(s). This motivates Video Corpus Moment Retrieval (VCMR), which explicitly studies the retrieval of moments from large video collections [**? ?**]. Egocentric benchmarks such as Ego4D's natural language queries further emphasize this corpus-scale structure in first-person data, where hours of activity contain only brief query-relevant events [**?**].

Unlike prior work, our work does not predict timestamped boundaries. Instead, daily QA is inherently a *retrieve-then-reason* problem over a personal video archive, where retrieval quality and evidence packaging directly shape downstream answerability.

2.3 Resource and Privacy Constraints

Egocentric daily QA is constrained by both model limitations and deployment realities.

Context vs. token budgets. Modern VLMs extend large language models (LLMs) with a visual front-end (e.g., an image/video encoder) so that the LLM can condition on non-text inputs such as images, frame sequences, or storyboard-like summaries. In practice, VLM inputs are still bounded by a finite context window and token budget, which limits how much visual evidence can be presented per query [**? ?**]. As a result, a daily QA system must decide *what to store* (offline representation) and *what to present at query time* (online evidence selection), trading off evidence coverage against inference cost.

Resource constraints on user devices. Continuous first-person capture is storage and compute intensive. Prior work on egocentric memory highlights that retaining raw video at scale is impractical, motivating on-device processing that converts streams into compact, queryable representations [**?**]. These constraints are especially relevant for wearables, where battery capacities, thermal limits, and local storage space restrict both (i) how much content can be kept and (ii) how much can be transmitted to an inference endpoint.

Privacy exposure in personal archives. Egocentric video often contains sensitive information about both the wearer and bystanders. Recent studies show that first-person recordings can leak identity and other private attributes, and that bystander privacy raises additional ethical and practical concerns [**?** **?**]. This motivates a privacy-first assumption common to personal data systems: raw video should remain under user control, and external model interaction should minimize exposure of identifiable content.

Overall, these constraints yield the central system design question studied in this paper: *How do evidence representation choices and retrieval policies trade off answer quality against token usage, storage footprint, and privacy exposure in multi-video egocentric daily QA?* We explore this question through a prototype system and evaluate these trade-offs under controlled settings.

3 System Architecture

We propose a token-aware retrieve-then-reason architecture for egocentric daily video QA: given a video collection, the system converts videos into queryable *visual evidence units*, retrieves the most relevant subset, and conditions a VLM on retrieved evidence to produce a multiple-choice answer. Our goal is not to propose a new model but to enable controlled measurements of how evidence representations and retrieval strategies affect downstream QA quality and token cost.

3.1 System Overview and Scope

The system acts as middleware between egocentric capture and VLM inference (Figure 1): videos are processed on a user device for evidence preparation and indexing, and only a small set of retrieved evidence is forwarded to the VLM for QA, reducing both prompt-token cost and raw-data exposure. We use a retrieve-then-reason structure [**?**]. We consider model training, explicit timestamp localization, and closed-loop storage budget control to be out of scope for this work.

3.2 Evidence Representation and Retrieval

A central design choice is how a long video can be represented as a set of *evidence units* that can be indexed and retrieved.

We use three evidence formats: **Frames** (single images), **Storyboards** (3×3 sheets that pack frames), and optionally **Tiles** (i.e., individual cells of a storyboard sheet, used as fine-grained retrieval units). We evaluate two scopes: **Tier I (Target-Video)** provides evidence from the ground-truth video (no retrieval) using uniform frame sampling, while **Tier II (Archive Retrieval)** retrieves evidence from an index built using shot-driven keyframes. We use K to denote the retrieval depth and T to denote the final number of storyboard sheets passed to the VLM.

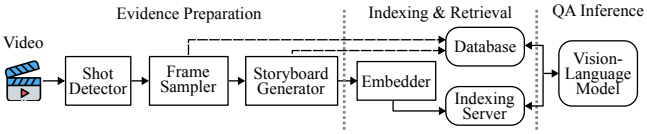

Figure 1: Proposed System Architecture

Frames. A frame evidence unit is a single image sampled from the video. For Tier I, we uniformly sample frames across the full duration under an image-count cap. For Tier II, we extract shot-driven keyframes to form the indexed candidate pool. At query time, the system retrieves the top-K frames by similarity between question and frame embeddings.

Storyboards and tiles. A storyboard sheet is a single image composed of a fixed grid of chronologically arranged frames (e.g., 3×3), where each cell is a tile (Figure 2). By packing frames, storyboards increase temporal information density per image token under a fixed prompt budget. Sheets are assembled from tier-specific sampled frames (uniform in Tier I, shot-driven in Tier II), preserving chronological order within each grid. Optionally, retrieval operates on tile embeddings for finer-grained matching, but the VLM always receives the full parent sheet. Multiple retrieved tiles mapping to the same sheet are deduplicated at the sheet level.

3.3 Evidence-Conditioned Answering

After retrieval, the system forms a VLM request consisting of the retrieved evidence images and the question text. The VLM is instructed to output a single multiple-choice label (A–D) in a strict format for automated evaluation. Optional free-form reasoning traces are also generated (Figure 2). We keep retrieval and inference separate: the VLM receives only images and text and re-encodes evidence internally. Each run logs the configuration, retrieved evidence identifiers, model output, and efficiency statistics (prompt tokens and end-to-end latency).

4 Evaluation Methodology

We evaluate (i) intrinsic VLM reasoning when evidence from the ground-truth video is provided directly, and (ii) end-to-end system behavior when evidence must be retrieved from a multi-video personal archive. We use a two-tier protocol introduced in §3: **Tier I (Target-Video)** as an intrinsic probe and **Tier II (Archive Retrieval)** as a system probe.

4.1 Datasets and Experimental Tiers

We conduct experiments on two datasets: (1) *Video-MME* [**?**] (N=49 videos, 147 questions), representing generic narrative-driven web videos, and (2) an *Egocentric Pilot* from Ego4D [**?**] (N=10 videos, 53 questions), consisting of continuous first-person activity logs (6-36 min/video).

Tier I (Target-Video): Intrinsic baseline. We provide evidence directly from the ground-truth video using uniform sampling for unbiased temporal coverage. We evaluate: **C0** (text-only), **C1** (audio-only; Video-MME only), and visual-only configurations (F13, S13) from Table 1. C0 and C1 are used in Tier I only.

Tier II (Archive Retrieval): End-to-end system probe. We build a shared index over the entire archive using shot-driven keyframes and retrieve relevant evidence given only the question text. Unless otherwise stated, retrieval is visual-only using F13, S13, and S13-T (Table 1); S23 is included in the efficiency analysis (§5.3).

Benchmarking rationale. We use multiple-choice QA for deterministic grading and reproducible comparison across configurations. Figure 2 confirms the model generates grounded reasoning traces before selecting an answer, validating multiple-choice accuracy as a proxy for visual reasoning capability [**?**].

Controlled single-video comparisons. To separate the effect of evidence packaging from per-frame resolution differences, §5.3 reports two controlled single-video comparisons on Video-MME (Table 4): (i) a *matched-accuracy* setting, which compares frames and storyboards at matched accuracy to measure token savings, and (ii) a *matched per-tile resolution* setting, which aligns per-frame and per-tile resolutions and token budgets across the two representations. These show that storyboard gains in the main results are not simply due to resolution differences.

4.2 System Configuration

Model Serving. We use *Qwen3-VL-8B-Instruct* [**?**] as the reasoner, served via vLLM [**?**] with `float16` precision. Decoding is deterministic (temperature = 0) with strict format enforcement. We set a maximum context length of 65,536 tokens to accommodate the settings with the largest number of tokens without truncation.

Retrieval & Indexing. Visual evidence is embedded using *OpenCLIP* [**?**] and indexed via *FAISS* [**?**]. Configurations are summarized in Table 1.

4.3 Metrics

We report three metrics: (i) *QA Accuracy (Acc.)*: fraction of questions answered correctly, (ii) *Retrieval Precision (Prec.)*: fraction of retrieved evidence units that originate from the ground-truth video (reported for Tier II), and (iii) *Token Cost (Toks.)*: average prompt tokens (text + image) per query as reported by vLLM, excluding output tokens.

Token accounting. Non-visual prompt content (question + options) contributes ≈2k tokens regardless of configuration. Since visual configurations in Table 5 range from 3.9k to 27.2k tokens, visual inputs account for the majority of prompt cost even in the most compact setting. Token cost therefore serves as a good proxy for inference cost, reflecting the efficiency impact of evidence representation and retrieval choices.

Sample storyboard (S13: 3×3, 1280px) from Ego4D Pilot Video 5014

(a) Open-Ended Mode *(Baseline style [**?**])*
Q: Where did I drop the drilling machine and the ruler?
GT: You dropped the drilling machine on the ground and dropped the ruler on the floor.
Model Output: "on the ground"
Evaluation: ✓ *Semantically correct (ground ≈ floor). However, the mismatch in length and phrasing makes automated scoring difficult, motivating our use of structured multiple-choice.*

(b) Reasoning-Guided Selection *(Our Protocol)*
Q: Where are the red pencil and folding ruler when the person picks up the angle grinder?
Options: A. Scaffolding plank; **B. Plastic dust sheet**; C. Electrical backbox; D. Behind metal track.
Model Reasoning: "...visible on the white plastic dust sheet on the ground, near the base of the wall..."
Model Selection: (B) ✓
Evaluation: ✓ *Deterministic grading with verifiable grounding trace.*

Figure 2: Qualitative capability comparison: (a) our system matches the descriptive capability of open-ended baselines [?**], and (b) for systematic evaluation, we utilize reasoning-guided multiple-choice. The model's internal trace (e.g., identifying the "white plastic dust sheet") confirms it is not guessing, but correctly interpreting the retrieved storyboard.**

5 Results

In this section, we report our findings in three parts. First, we isolate the effect of input modality to establish a baseline (§5.1). Second, we evaluate the viability of multi-video retrieval over a personal archive (§5.2). Finally, we analyze the critical systems trade-off between representation efficiency and accuracy (§5.3).

Table 1: Configurations used in evaluation.

Config	Modality	Representation	Notes
C0	Text-only	-	Question and answer options only
C1	Audio-only	ASR transcript	Whisper-generated audio transcript
F13	Visual	Discrete frames	Individual video frames (1280px)
S13	Visual	Storyboard	3×3 grid, 1280px resolution
S23	Visual	Storyboard	3×3 grid, 1920px resolution
S13-T	Visual	Storyboard	S13 with tile-level search, sheet-level VLM

Table 2: Target-Video Baseline (Tier I). Video-MME is audio-dominant (C1), while egocentric QA benefits primarily from visual evidence. In our default visual-only setup, storyboard packing (S13) improves over uniform frames (F13).

Condition	Video-MME	Ego4D Pilot
Text-only (C0)	25.2%	27.5%
Audio-only (C1)	**57.8%**	29.4%
Visual-only, uniform frames (F13)	48.8%	52.9%
Visual-only, storyboards (S13)	53.7%	**58.8%**

5.1 The Modality Gap

We begin with the **Target-Video (Tier I)** capability probe to characterize the "information density" of different modalities. In this setting, the ground-truth video is provided directly to the VLM (no retrieval), establishing an intrinsic reasoning baseline.

Observation 1: Video-MME is audio-dominant while Egocentric is visual-first. Table 2 compares performance across text-only (C0), audio-only (C1), and visual-only conditions (Tier I: uniform F13 vs. storyboard S13). On Video-MME, the audio-only condition achieves **57.8%** accuracy, outperforming both the uniform visual baseline (48.8%) and the storyboard condition (53.7%). This confirms that generic video benchmarks often reward "listening" over "watching," as narration provides the primary semantic signal.

In contrast, the Egocentric Pilot dataset shows no benefit from audio. The audio-only performance (29.4%) shows only a marginal change over the blind text-only baseline (27.5%), indicating no clear benefit from audio in this pilot. On the other hand, visual evidence is essential: even simple uniform sampling achieves 52.9%, and storyboard packing drives accuracy further to **58.8%**.

This empirical gap confirms that egocentric daily QA is a distinct systems challenge: questions are grounded in what the user saw, requiring high-quality visual retrieval rather than speech transcripts.

Table 3: Multi-Video Archive Retrieval (Tier II). Visual-only retrieval is noisy on Video-MME (low Prec.) but reliable on Ego4D. F13 retrieves top-K frames; S13 feeds top-T sheets; S13-T retrieves K tiles and feeds T deduplicated parent sheets.

Config	Video-MME		Ego4D Pilot	
	Acc.	Prec.	Acc.	Prec.
F13 ($K = 20$)	55.0%	16.6%	58.8%	65.2%
S13 ($T = 3$)	41.7%	21.1%	**66.7%**	62.8%
S13-T ($K = 40$, $T = 10$)	50.0%	21.0%	**70.6%**	64.1%

5.2 Viability of Multi-Video Archive Retrieval

We next evaluate the **Archive-Retrieval (Tier II)** pipeline, where the system must identify relevant visual evidence from a multi-video archive before answering. A central challenge is retrieval noise when the ground-truth video source is unknown.

Observation 2: Visual-only retrieval underperforms on narrative web videos but succeeds on egocentric video logs. Table 3 reports the QA accuracy (Acc.) and Retrieval Precision (Prec.), defined as the fraction of retrieved evidence units within the final token budget that originate from the correct source video. On Video-MME, the *visual-only* retriever is less reliable (Prec. ≈16%-21%) as many questions depend on narration rather than distinctive visual cues. As a result, end-to-end accuracy falls below the single-video baseline for storyboard configurations.

In contrast, on the Ego4D pilot, retrieval is substantially more reliable. Retrieval Precision increases to **62%-64%** for storyboard-based configurations, indicating that first-person visual cues (objects, environments) are distinct enough for reliable indexing. As a result, end-to-end accuracy reaches **66.7%-70.6%**, surpassing the Target-Video baseline. This improvement likely stems from (i) shot-driven keyframe extraction in Tier II, which selects more informative visual evidence than the uniform sampling used in Tier I, and (ii) the denoising effect of retrieval in Tier II, which filters out irrelevant segments that would otherwise distract the VLM during full-video inference.

5.3 Efficiency and Performance Trade-offs

Finally, we analyze how visual representation choices affect system performance. A practical system must maximize accuracy while minimizing token usage, which serves as a proxy for inference cost and latency.

Observation 3: Storyboards provide a strong token-accuracy trade-off in our default settings. A key challenge in comparing storyboards to frames is that grid packing simultaneously reduces per-frame resolution and expands temporal coverage. Table 4 addresses this with two

Table 4: Controlled visual-only single-video comparisons (Video-MME). We compare Frames and Storyboards under two settings. Top: matched-accuracy comparison. Bottom: matched-resolution comparison. For storyboards, T denotes the number of storyboard sheets passed to the VLM (full S13 has 52 sheets per video). Img and Latency report average payload size (MB) and end-to-end time (s).

Settings (no-audio)	Accuracy	Tokens	Img (MB)	Latency (s)
Matched-accuracy (46.9%)				
Frames (1280×704, 25 frames per video, tok≈25k)	46.9%	24.9k	2.25	3.94
Storyboards (3×3 @1280px, T=13, tok≈13k)	46.9%	12.8k	2.08	2.67
Matched per-tile resolution (426×234)				
Frames (426×234, 300 frames per video, tok≈40k)	55.8%	40.5k	5.36	7.79
Storyboards (3×3 @1280px, T=39, tok≈40k)	55.1%	40.1k	6.23	5.85

controlled single-video comparisons on Video-MME. (i) In the matched-accuracy setting, we select configurations that achieve matched accuracy (46.9%). "Frames" use 25 full-resolution frames (1280×704), while "Storyboards" use 13 sheets (3×3 @1280px). At the same accuracy level, storyboards halve token usage (12.8k vs. 24.9k), showing the efficiency benefit of storyboard packing. (ii) In the matched per-tile resolution setting, we downsample frames to the storyboard per-tile resolution (426×234) and choose the number of frames to match storyboard accuracy. Under this control, frames and storyboards achieve similar accuracy (55.8% vs. 55.1%) with similar token usage (≈40k).

With resolution effects controlled, Table 5 shows efficiency-accuracy trade-offs under archive retrieval (Ego4D). The individual frames approach (IDs 1-2) scales poorly: doubling retrieval depth K from 5 to 10 improves accuracy (51.0%→56.9%) but nearly doubles token cost to ≈27.2k. In contrast, the S13 storyboard setting (ID 3) achieves **66.7%** accuracy with only ≈3.9k tokens, reflecting higher temporal information density per token from storyboard packing.

Observation 4: System knobs trade accuracy for cost. Increasing storyboard resolution from 1280px (S13) to 1920px (S23) increases tokens (3.9k → 14.3k) without improving accuracy (66.7% → 62.7%), suggesting temporal coverage matters more than extra pixels. Tile-level retrieval (S13-T) yields the highest accuracy (70.6%) at 12.2k tokens, providing a tunable efficiency-accuracy knob for deployment.

Table 5: Storyboards (3×3 grid) offer a strong efficiency-accuracy trade-off. F13 denotes frame-based evidence, K denotes retrieval depth, and T is the number of storyboard sheets passed to the VLM. Toks. reports average prompt tokens per query (text + image tokens). For S13-T, retrieval ranks K tiles and the VLM receives the top-T deduplicated parent storyboard sheets.

ID	Config	Spec	Acc.	Toks.	Notes
1	F13	$K = 5$	51.0%	13.4k	Baseline
2	F13	$K = 10$	56.9%	27.2k	Baseline
3	**S13**	**1280px,** $T = 3$	**66.7%**	**3.9k**	**Best Eff.**
4	S23	1920px, $T = 5$	62.7%	14.3k	High-Res
5	S13-T	1280px, $K = 40$ $T = 10$	**70.6%**	12.2k	Max Acc.

6 Conclusion and Future Work

In this paper, we presented a token-aware retrieve-then-reason architecture for egocentric daily video QA that bridges always-on recording with the prompt-token limits of modern VLMs. Through a two-tier evaluation, we identified a clear modality gap: narrative web benchmarks (e.g., Video-MME) can often be solved from audio alone, whereas egocentric daily QA is fundamentally visual-first and depends on reliable visual evidence. Across controlled configurations, we find that storyboard packaging achieves a strong token-accuracy trade-off, reaching similar or better accuracy to frame-based inputs at substantially lower token usage. Tile-level retrieval further improves accuracy at a higher cost. Controlled single-video comparisons confirm that these trends are not artifacts of resolution differences, highlighting evidence packaging and retrieval policy as primary levers for practical daily QA.

Future work will incorporate richer on-device evidence signals (e.g., OCR and lightweight metadata) while preserving user-controlled privacy boundaries, develop adaptive retrieval policies that adjust the efficiency-coverage trade-off based on query difficulty and resource constraints, and extend evaluation to finer-grained spatial queries and additional VLM backends. We also aim to separate network transfer time from model inference time in end-to-end measurements.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. The work of Zihao Ding and Yao Liu was partially supported by NSF under grant CNS-2200048. Cheng-Hsin Hsu was partially supported by MOST (Taiwan) under grant 114-2918-I-007-004.