EGOSKILL: RETHINKING WHAT TO STORE AND HOW TO REASON IN LONG-FORM EGOCENTRIC VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form egocentric video question answering (VQA) requires reasoning over evidence that may be sparse, distributed across distant moments, and heterogeneous across question types. We introduce **EgoSkill**, a fully training-free framework that factorizes this problem into two complementary inference-time components: *what to store* and *how to execute*. For *what to store*, we propose **Dynamic Hierarchical Memory (DH-Mem)**, a query-agnostic representation of long-form egocentric video. DH-Mem represents local events as temporally grounded substeps and semantically composes related substeps, even when they are temporally non-contiguous, into higher-level procedural steps and a video-level goal. For *how to execute*, we introduce **Evidence Construction Skills (EC-Skills)**, reusable query-aware procedures that adapt evidence retrieval, visual processing, tool use, and verification to the requirements of each question type. Together, these components realize a store–execute factorization, with DH-Mem providing query-agnostic video context and EC-Skills constructing query-specific evidence for frozen MLLMs, without task-specific parameter updates. Across HD-EPIC, EgoSchema, and QAEgo4D, EgoSkill consistently improves the corresponding frozen MLLM baselines, demonstrating the effectiveness of factorizing long-form egocentric VQA into query-agnostic representation and question-dependent evidence construction.

1 INTRODUCTION

Long-form egocentric video (Grauman et al., 2022; Damen et al., 2020; Grauman et al., 2024) captures how people interact with objects, environments, and ongoing activities over extended periods, making it a useful substrate for embodied assistants and agents. Yet reasoning over such videos remains difficult for multimodal large language models (MLLMs) (Luo et al., 2026; Pei et al., 2026; Ye et al., 2025). Relevant evidence can be sparse and distributed across distant moments, and different questions may require very different evidence. These challenges raise two complementary questions for long-form egocentric video understanding: *what information should be preserved from the video, and how should that information be used to answer a particular query?*

We argue that answering the first question requires rethinking a common prior in video representation: temporal proximity implies semantic relatedness. At short temporal scales, temporal continuity provides a strong inductive bias: adjacent frames typically belong to the same motion, object interaction, or local action, making contiguous clips a natural unit of representation. At higher semantic levels, however, this prior becomes less reliable. Long-form egocentric activities can be interleaved, interrupted by unrelated events, and resumed later (Price et al., 2022; Peirone et al., 2025). Consequently, temporally distant observations may belong to the same procedural step, whereas temporally adjacent observations around an interruption may belong to different activities. Thus, while temporal continuity remains appropriate for representing local events, temporal containment need not define their higher-level organization.

Motivated by this observation, we introduce **Dynamic Hierarchical Memory (DH-Mem)** for *what to store*. DH-Mem is constructed before any future query is known and therefore remains query-agnostic, while preserving semantic relationships that may span large temporal gaps. It first

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.

represents the video as timestamped local substeps and then organizes them into higher-level steps according to semantic relatedness rather than temporal adjacency alone. This allows non-contiguous observations describing the same activity to be composed into a common procedural step while unrelated intervening events remain separate. The resulting steps are further consolidated into global activity or goal context, yielding a hierarchy that preserves temporal continuity locally while organizing long-range observations semantically.

A well-structured memory, however, does not determine how a particular question should be solved. This gives rise to the orthogonal problem of *how to execute*. Once a question is known, the appropriate inference strategy becomes inherently query-dependent. Some questions require broad procedural context, whereas others depend on a short temporal interval or fine-grained visual evidence. Others require specialized processing, such as text recognition, spatial reasoning, or candidate-wise verification. Applying the same inference procedure to such heterogeneous questions can therefore underutilize even a high-quality memory.

To address this second problem, we introduce **Evidence Construction Skills (EC-Skills)**: reusable, query-aware procedures that determine how evidence is constructed for a frozen MLLM. Depending on the question, a skill may retrieve relevant memory, adapt visual input, invoke an auxiliary tool, or perform candidate-wise verification. Importantly, these skills operate at the inference interface rather than modifying model parameters. The same query-agnostic memory can therefore be reused across questions, while the evidence presented to the model is adapted to the requirements of each question.

Based on this separation, we present **EgoSkill**, a framework that factorizes long-form egocentric reasoning into query-agnostic memory construction and query-aware execution. DH-Mem determines *what to store* by organizing temporally local observations into semantically coherent long-range structure, while EC-Skills determine *how to execute* by constructing task-appropriate evidence from memory, raw visual observations, and optional tools. This store–execute factorization allows EgoSkill to specialize inference behavior across heterogeneous questions without modifying the underlying MLLM. EgoSkill therefore operates entirely at inference time with frozen models and remains modular with respect to both the answer model and auxiliary tools. Across HD-EPIC (Perrett et al., 2025), EgoSchema (Mangalam et al., 2023), and QAEgo4D (Di & Xie, 2024), EgoSkill consistently improves the corresponding frozen MLLM baselines, demonstrating the value of separating persistent video representation from query-conditioned execution.

Our contributions are summarized as follows:

- We introduce **EgoSkill**, a store–execute factorization framework for long-form egocentric VQA that separates query-agnostic representation of persistent video context from query-aware procedures for constructing task-appropriate evidence at inference time.
- For *what to store*, we propose **DH-Mem**, which preserves temporal continuity for local observations while semantically composing potentially non-contiguous observations into coherent higher-level procedural steps.
- For *how to execute*, we introduce **EC-Skills**, reusable input-side procedures that adapt evidence retrieval, visual processing, tool use, and verification to the requirements of each question type, yielding substantial gains for frozen MLLMs across three egocentric VQA benchmarks.

2 Related Work

Egocentric Video Understanding. Egocentric video understanding is important for embodied agents, robotics, assistive systems, and AR/VR applications, as first-person video captures how people interact with objects, environments, and ongoing activities. Recent egocentric video QA benchmarks evaluate increasingly diverse capabilities, including long-range reasoning, episodic memory, procedural understanding, scene-text reasoning, cross-domain generalization, and streaming understanding (Mangalam et al., 2023; Di & Xie, 2024; Perrett et al., 2025; Zhou et al., 2025; Li et al., 2026c; Wang et al., 2026). Recent egocentric MLLM approaches improve first-person video understanding through data- and model-centric adaptation (Ye et al., 2025; Luo et al., 2026; Pei et al., 2026; Yang et al., 2025). In contrast, we pursue a fully training-free approach to long-form egocentric video QA through dynamic hierarchical memory and evidence construction skills.

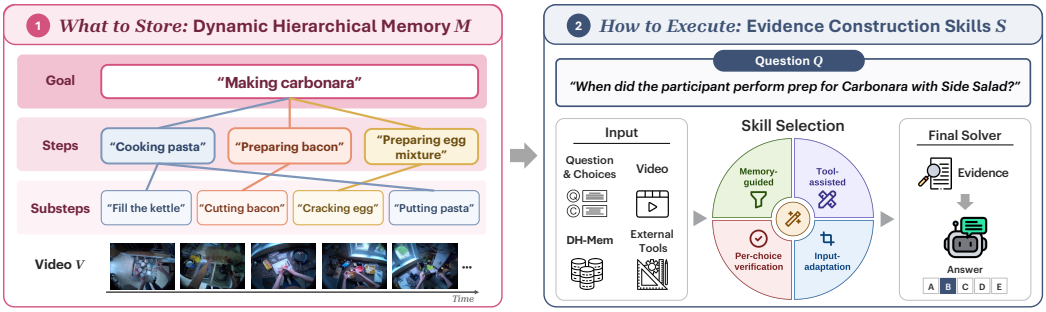

Figure 1: **Overview of EgoSkill.** EgoSkill factorizes long-form egocentric VQA into *what to store* and *how to execute*. Dynamic Hierarchical Memory (DH-Mem) builds a query-agnostic hierarchy of substeps, semantically composed steps, and a video-level goal. For each question, a selected Evidence Construction Skill (EC-Skill) constructs question-dependent evidence from the video, DH-Mem, and optional tools, which is then passed to a frozen MLLM solver to produce the answer.

Memory for Long-Video Understanding. Recent long-video understanding methods have explored diverse memory representations to preserve and organize information over extended temporal contexts, including episodic, semantic, visual, object-centric, hierarchical, and graph-structured memories (Fan et al., 2024; Huang et al., 2025; Choi et al., 2026; Yeo et al., 2026; Yang et al., 2025; Chu et al., 2025; Goletto et al., 2024). Among them, hierarchical representations are particularly relevant to human activities, which naturally exhibit structure across goals, steps, and local actions (Song et al., 2023; Islam et al., 2024; Kang et al., 2025; Peirone et al., 2025; Price et al., 2022). Building on these directions, we construct a query-agnostic hierarchical memory for long-form egocentric VQA. Rather than relying on temporal containment to form higher-level units, DH-Mem preserves temporal continuity for local substeps while semantically composing potentially non-contiguous observations into coherent procedural steps.

Inference-time Composition and Agent Skills. Prior work has explored compositional approaches to multimodal reasoning, combining pretrained models, external tools, retrieval, and agentic workflows at inference time (Zeng et al., 2022; Gupta & Kembhavi, 2023; Fan et al., 2024). More recently, agent skills have emerged as reusable packages of procedural knowledge that equip agents with task- or domain-specific workflows (Zhang et al., 2025a; Li et al., 2026b). Recent work further studies how such skills can guide inference across multimodal tasks (Li et al., 2026a). Following this perspective, we introduce EC-Skills for long-form egocentric VQA, which provide reusable procedures for question-dependent evidence construction from memory, video, and optional tools.

3 Method

3.1 Framework Overview

Problem Formulation. Given a long-form egocentric video V and a set of questions $\mathcal{Q}$, our goal is to predict an answer $\hat{a}$ for each question $q \in \mathcal{Q}$. We first construct a query-agnostic video memory,

$$M = \mathcal{B}(V), \tag{1}$$

where $\mathcal{B}$ denotes the memory builder.

Let $\mathcal{S}$ denote the set of evidence construction skills. For each question $q \in \mathcal{Q}$, a selected evidence construction skill $s \in \mathcal{S}$ constructs question-dependent evidence using video, memory, and optional auxiliary tools,

$$\mathcal{E}_q = s_q(V, M, \mathcal{T}), \tag{2}$$

where $\mathcal{T}$ denotes the set of available tools. The resulting evidence $\mathcal{E}_q$ may contain information retrieved from DH-Mem, visual evidence extracted from the original video, tool-derived evidence, or their combination.

Finally, the frozen solver MLLM F_{sol} predicts the answer from the question and the constructed evidence,

$$\hat{a} = F_{\text{sol}}\left(q, \mathcal{E}_q\right). \tag{3}$$

Store–Execute Factorization. This formulation separates long-form egocentric VQA into two complementary roles. Dynamic hierarchical memory (DH-Mem) implements *what to store*, determining which information from the video should be persistently represented in M. Evidence construction skills (EC-Skills) implement *how to execute*, determining how question-dependent evidence $\mathcal{E}_q$ should be constructed for the frozen answer model. We describe the two components in Sections 3.2 and 3.3.

3.2 Dynamic Hierarchical Memory

The memory builder $\mathcal{B}$ in Eq. 1 constructs Dynamic Hierarchical Memory (DH-Mem) M from the input video V. DH-Mem represents the video at three levels: timestamped local substeps, semantically composed steps, and a video-level goal. Its key design is to preserve temporal continuity at the local level while organizing higher-level observations according to semantic relatedness rather than temporal adjacency.

Substep Construction. We divide the video V into N non-overlapping fixed-length temporal chunks $\{v_i\}_{i=1}^{N}$. For each chunk v_i, a frozen MLLM F_{sub} generates a temporally grounded substep description h_i describing visible actions, manipulated objects, object states, and scene context. To maintain object identity and action continuity across neighboring chunks, the generation of h_i is conditioned on a short history of previously generated substeps:

$$h_i = F_{\text{sub}}\left(v_i; \{h_j\}_{j=\max(1,i-W)}^{i-1}\right), \tag{4}$$

where W denotes the history-window size.

Semantic Step Composition. We then compose the substep descriptions into higher-level steps based on semantic relatedness rather than temporal adjacency. Each substep description h_i is encoded using a frozen text embedding model f_{emb}. Given the first n substeps, where $1 \le n \le N$, we group their representations using average-linkage agglomerative clustering with cosine similarity:

$$\mathcal{C}^{(n)} = \{C_1^{(n)}, \ldots, C_{K_n}^{(n)}\} = \text{Cluster}\left(f_{\text{emb}}(h_1), \ldots, f_{\text{emb}}(h_n)\right). \tag{5}$$

Here, K_n denotes the number of induced steps, and $C_k^{(n)} \subseteq \{1, \ldots, n\}$ contains the indices of the substeps assigned to the k-th cluster. No temporal-contiguity constraint is imposed during clustering, allowing temporally distant substeps to be grouped into the same step when they are semantically related. As new substeps accumulate, clustering is recomputed over all available substeps. Thus, both the semantic grouping $\mathcal{C}^{(n)}$ and the number of steps K_n can evolve as the video progresses.

For each cluster $C_k^{(n)}$, a frozen MLLM F_{step} consolidates its constituent substeps into a higher-level step description,

$$m_k^{(n)} = F_{\text{step}}\left(\{h_i \mid i \in C_k^{(n)}\}\right). \tag{6}$$

As the semantic grouping evolves, the composition and descriptions of the induced steps may change, even when the total number of steps remains unchanged. This evolving semantic organization is what makes DH-Mem dynamic. Each step retains its constituent substeps and the timestamps of their source chunks.

Goal Consolidation. Finally, a frozen MLLM F_{goal} consolidates the current step descriptions into a video-level goal,

$$g^{(n)} = F_{\text{goal}}\left(m_1^{(n)}, \ldots, m_{K_n}^{(n)}\right). \tag{7}$$

After n substeps, the resulting hierarchy consists of the goal $g^{(n)}$, the induced steps $\{m_k^{(n)}\}_{k=1}^{K_n}$, and their constituent substeps. We denote this hierarchy as $M^{(n)}$. For the full video, $n = N$, and the final DH-Mem is $M = M^{(N)}$. An example memory is provided in Table 16 (Appendix).

3.3 Evidence Construction Skills

For each question $q \in \mathcal{Q}$, the selected evidence construction skill (EC-Skill) s constructs the question-dependent evidence $\mathcal{E}_q$ in Eq. 2. A skill is a reusable inference procedure that specifies how evidence

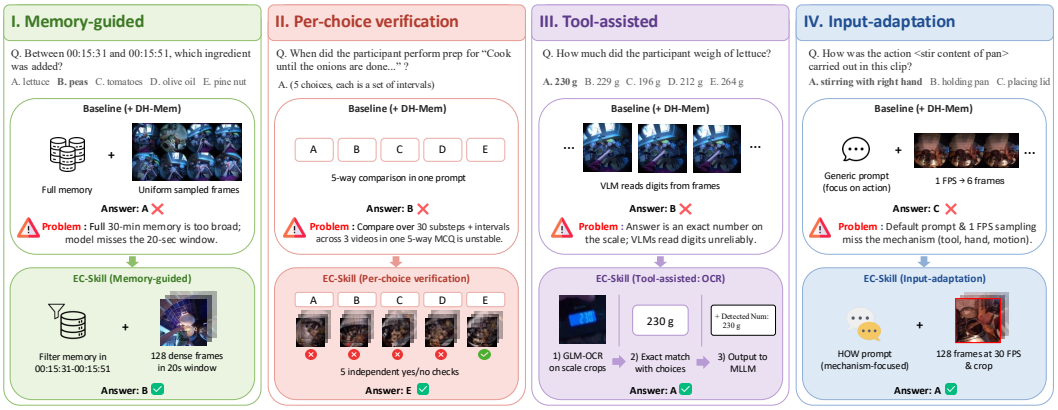

Figure 2: **Examples of evidence construction with four EC-Skill families.** All skills share one query-agnostic DH-Mem, which each skill uses as needed. The selected evidence construction skill constructs evidence from the memory, video frames, and optional external tools, and decides how the frozen MLLM produces the final answer.

should be obtained from the video V, DH-Mem M, and optional tools $\mathcal{T}$, and how it should be presented to the frozen solver MLLM F_{sol}. A skill need not use all available sources; depending on the question, it may rely on memory, return directly to the video, invoke an auxiliary tool, or combine multiple sources.

Four EC-Skill Families. We organize EC-Skills into four families according to their primary intervention on the inference process (Figure 2).

- **I. Memory-guided.** Memory-guided skills determine *which stored evidence should be used*. They select the relevant memory level, entity, or temporal region from DH-Mem and focus the visual input on the corresponding part of the video. This reduces irrelevant context when the information required by the question is already represented in memory.
- **II. Per-choice verification.** Per-choice verification determines *how candidate hypotheses should be evaluated*. When a direct comparison among complex answer choices is difficult, the skill decomposes the problem into candidate-wise verification. It constructs evidence separately for each candidate and invokes F_{sol} to assess whether the candidate is supported, after which the candidates are ranked by their verification scores.
- **III. Tool-assisted.** Tool-assisted skills determine *what additional evidence should be derived*. When the required information cannot be reliably obtained from the default video or memory input, the skill invokes an auxiliary tool and converts its output into evidence for F_{sol}. The resulting evidence may be textual, numeric, spatial, or visual, depending on the tool and question.
- **IV. Input-adaptation.** Input-adaptation skills determine *how existing visual evidence should be presented*. When the relevant information is already visible in the video, the skill changes the input configuration—such as frame sampling, temporal coverage, cropping, visual annotation, or additional prompting —so that the solver can attend to the aspect required by the question.

Procedure and Configuration. We separate the reusable procedure that defines an EC-Skill from its question-specific configuration. The procedure specifies the underlying inference operation, whereas the configuration determines how that operation is instantiated for a particular question. Thus, the same EC-Skill procedure can be reused across question types and benchmarks with different configurations. For example, a memory-guided skill may follow the same retrieval procedure while using different temporal or semantic cues depending on the question.

Skill Selection. Skill selection is implemented using predefined mappings from question types to EC-Skill procedures. Question types without an assigned skill are answered by the frozen solver using DH-Mem without additional skill-specific processing. The complete skill assignments and question-specific configurations are provided in Appendix E.

4 EXPERIMENTS

4.1 BENCHMARKS AND METRICS

We evaluate EgoSkill on three long-form egocentric video question answering (VQA) benchmarks: HD-EPIC (Perrett et al., 2025), EgoSchema (Mangalam et al., 2023), and QAEgo4D (Di & Xie, 2024). We report accuracy for all benchmarks; for HD-EPIC, we additionally report Macro-7, the average accuracy across its seven macro-categories.

HD-EPIC (Perrett et al., 2025) is an egocentric VQA benchmark with 26,550 questions across 30 subtasks, grouped into seven macro-categories: Recipe, Ingredient, Nutrition, Fine-grained Action, 3D Perception, Object Motion, and Gaze.

EgoSchema (Mangalam et al., 2023) is a long-range egocentric multiple-choice QA benchmark with over 5,000 human-curated question-answer pairs spanning diverse human activities. We use the official 500-question subset, referred to as the EgoSchema$_{\texttt{subset}}$.

QAEgo4D (Di & Xie, 2024) is a long egocentric video QA benchmark with temporally grounded questions. We use QAEgo4D$_{\texttt{test-mc}}$, a 500-question multiple-choice test split.

4.2 IMPLEMENTATION DETAILS

Backbones and memory builder. We evaluate several MLLM backbones as F_{sol}, including Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), Qwen3.5 (Qwen Team, 2026), and GLM-4.6V (Hong et al., 2025). All solver backbones are kept frozen and are not fine-tuned on any evaluated benchmark.

Memory construction. Across HD-EPIC, EgoSchema, and QAEgo4D, we build DH-Mem with the same code, prompts, and hyperparameters. Videos are divided into 30-second chunks $\{v_i\}_{i=1}^{N}$ and sampled at 1 FPS. We instantiate F_{sub}, F_{step}, and F_{goal} with Qwen3-VL-8B. Substeps h_i are generated using greedy decoding, with the history-window size in Eq. 4 set to $W{=}10$. Substeps are embedded with Qwen3-Embedding-0.6B (Zhang et al., 2025b) and grouped using average-linkage agglomerative clustering over cosine distances, with similarity threshold $\tau{=}0.68$ (equivalently, a distance threshold of $1 - \tau = 0.32$). The resulting clusters are consolidated by F_{step} into steps, which are subsequently consolidated by F_{goal} into a video-level goal.

Auxiliary tools. For tool-assisted EC-Skills, the auxiliary tool set $\mathcal{T}$ is used only to construct input-side evidence. We use Depth Anything 3 (Lin et al., 2025) for 3D reconstruction, SAM3 (Carion et al., 2026) and Grounding DINO (Liu et al., 2024) for object detection and segmentation, GLM-OCR (Duan et al., 2026) for text and measurement reading, the USDA FoodData Central database (U.S. Department of Agriculture, Agricultural Research Service, 2019) for nutrition lookup, and Qwen3-VL-32B (Bai et al., 2025a) as a narration generator. The resulting tool outputs are converted into prompts, rendered images, structured text, or answer candidates and incorporated into the question-specific evidence $\mathcal{E}_q$ provided to the frozen solver F_{sol}. Appendix C provides prompts, hyperparameters, and additional implementation details.

4.3 RESULTS

Results on HD-EPIC VQA. Table 1 compares EgoSkill with baselines reported in HD-EPIC, recent egocentric video understanding systems, and additional frozen MLLM baselines. HD-EPIC remains challenging for general-purpose video-language models: Gemini Pro achieves 37.6, and even Qwen3-VL-235B reaches only 43.2, far below the sample human baseline of 90.3. Recent egocentric-oriented methods improve this range by introducing additional structure or task-specific training. From Pixels to Graphs (Taluzzi et al., 2025), built on Gemini 2.0 Flash (Google), uses structured scene and knowledge graphs and reports 44.2. OpenMMEgo (Luo et al., 2026) and Egocentric Co-Pilot (Yang et al., 2026), both based on Qwen2.5-VL-7B fine-tuning, report 42.4 and 46.2, respectively.

In contrast, EgoSkill keeps the solver MLLM frozen and improves performance through DH-Mem and EC-Skills. Using Qwen2.5-VL-7B, EgoSkill reaches 46.0, remaining competitive with recent fine-tuned egocentric systems without updating the solver parameters. With stronger frozen backbones, performance further increases to 56.1 with Qwen3-VL-8B and 59.6 with Qwen3.5-27B, exceeding the

Table 1: **HD-EPIC VQA (Perrett et al., 2025) results.** We compare EgoSkill with baselines reported in HD-EPIC, recent egocentric systems evaluated on HD-EPIC, and additional frozen MLLM baselines. Rows marked with "+ EgoSkill" add DH-Mem and EC-Skill to the same frozen backbone, without finetuning the backbone model.

Model	Recipe	Ingredient	Nutrition	Action	3D	Motion	Gaze	Avg.
Sample Human Baseline	96.7	96.7	85.0	92.5	93.8	92.7	75.0	90.3
VideoLlama2 (Cheng et al., 2024)	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4
LongVA (Zhang et al., 2024a)	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3
LLaVA-Video (Zhang et al., 2024b)	36.3	33.5	38.7	43.0	27.3	18.9	29.3	32.4
Gemini Pro (Team et al., 2024)	60.5	46.2	34.7	39.6	32.5	20.8	28.7	37.6
Qwen3-VL-32B (Bai et al., 2025a)	49.9	35.5	40.3	49.6	37.3	30.1	34.6	39.6
Qwen3-VL-235B (Bai et al., 2025a)	55.0	38.5	47.0	53.4	35.8	35.7	36.8	43.2
GLM-4.6V (Hong et al., 2025)	47.5	40.2	40.0	44.3	32.8	32.0	33.8	38.7
From Pixel to Graphs (Taluzzi et al., 2025)	64.6	54.7	38.0	48.3	42.6	30.2	31.2	44.2
OpenMMEgo (Luo et al., 2026)	-	-	-	-	-	-	-	42.4
Egocentric Co-Pilot (Yang et al., 2026)	-	-	-	-	-	-	-	46.2
Qwen2.5-VL-7B (Bai et al., 2025b)	35.1	34.5	37.3	41.8	29.7	30.8	34.0	34.8
+ EgoSkill	53.9	52.8	51.3	48.3	39.0	41.0	35.5	46.0
Qwen3-VL-8B (Bai et al., 2025a)	48.0	33.8	30.7	45.1	35.5	38.5	32.5	37.7
+ EgoSkill	77.6	60.5	56.3	52.4	50.2	49.1	46.3	56.1
Qwen3.5-27B (Qwen Team, 2026)	52.6	43.0	38.0	53.2	36.4	29.6	38.9	41.7
+ EgoSkill	85.5	69.2	61.0	63.7	46.7	37.3	53.6	59.6

strongest prior egocentric-oriented result in the table by 9.9 and 13.4 points, respectively. Across the three frozen backbones, EgoSkill improves over direct video prompting by 11.2 points for Qwen2.5-VL-7B, 18.4 points for Qwen3-VL-8B, and 17.9 points for Qwen3.5-27B. These consistent gains show that DH-Mem and EC-Skills can substantially improve frozen MLLMs without task-specific parameter updates.

Interleaved procedural structure in HD-EPIC. To examine whether the non-contiguous structure motivating DH-Mem is prevalent in real egocentric activities, we analyze the recipe-level procedural annotations in HD-EPIC. Among 116 videos with such annotations, 83 (71.6%) contain at least one interleaved step, where the same procedural step reappears after another activity intervenes. This supports the premise behind DH-Mem: temporal continuity is a useful local prior, but need not determine higher-level procedural organization. Figure 3 shows a representative case in which DH-Mem semantically groups temporally separated local observations into a shared higher-level step.

Results on other egocentric VQA benchmarks. Table 2 evaluates EgoSkill on $\text{EgoSchema}_{\texttt{subset}}$ and $\text{QAEgo4D}_{\texttt{test-mc}}$ across multiple Qwen backbones. We additionally compare with EgoThinker (Pei et al., 2026), an egocentric model fine-tuned from Qwen2-VL-7B. With the same Qwen2-VL-7B backbone, EgoSkill achieves 70.4 on $\text{EgoSchema}_{\texttt{subset}}$ and 57.2 on $\text{QAEgo4D}_{\texttt{test-mc}}$, outperforming EgoThinker by 3.8 and 2.0 points, respectively. Using Qwen3-VL-8B, the strongest backbone evaluated here, EgoSkill further reaches 77.6 and 62.2, respectively. These results demonstrate that the gains from DH-Mem and EC-Skills are not limited to HD-EPIC, but also extend to egocentric VQA benchmarks with different question distributions.

4.4 Ablation Study

We conduct ablation studies to examine the two main components of EgoSkill: DH-Mem and EC-Skills. Table 3 isolates the effect of memory organization, while Table 4 analyzes the cumulative effect of adding DH-Mem and EC-Skills.

Memory format ablation. Table 3 compares different memory formats on HD-EPIC using Qwen3-VL-8B. The +30 sec setting uses only the 30-second substeps, ordered chronologically without higher-level organization, whereas DH-Mem further organizes these substeps into semantically coherent procedural steps and global goal context.

The flat substep memory improves over the baseline, showing that temporally ordered local descriptions already provide useful long-range context. DH-Mem further improves overall performance,

Table 2: **EgoSchema$_{\text{subset}}$ (Mangalam et al., 2023) and QAEgo4D$_{\text{test-mc}}$ (Di & Xie, 2024) results.** Rows marked with "+ EgoSkill" add DH-Mem and EC-Skill to the same frozen backbone, without finetuning the backbone model.

Model	EgoSchema$_{\text{subset}}$	QAEgo4D$_{\text{test-mc}}$
EgoThinker	66.6	55.2
Qwen2-VL-7B	62.2	56.6
+ EgoSkill	70.4	57.2
Qwen2.5-VL-3B	58.0	50.4
+ EgoSkill	69.4	53.2
Qwen2.5-VL-7B	68.4	55.2
+ EgoSkill	73.8	58.0
Qwen3-VL-8B	72.0	56.6
+ EgoSkill	77.6	62.2

Table 3: **Memory format ablation on HD-EPIC using Qwen3-VL-8B.** "+30 sec" uses only chronologically ordered 30-second substeps, while "+DH-Mem" adds the full goal–step–substep hierarchy.

Category	Baseline	+30 sec	+DH-Mem
Recipe	48.0	69.5	**71.0**
Ingredient	33.8	43.0	**44.8**
Nutrition	30.7	35.3	**39.3**
Action	45.1	45.0	**48.3**
3D	35.5	**41.7**	40.9
Motion	38.5	37.0	**39.0**
Gaze	32.5	32.2	**40.7**
Avg.	37.7	43.4	**46.3**

Table 4: **Component analysis of EgoSkill.** Results show the cumulative effect of adding DH-Mem and EC-Skills across three egocentric video QA benchmarks.

Model	Baseline	+DH-Mem	+EC-Skills
HD-EPIC			
Qwen2.5-VL-7B	34.8	37.2	**46.0**
Qwen3-VL-8B	37.7	46.3	**56.1**
Qwen3.5-27B	41.7	48.5	**59.6**
EgoSchema$_{\text{subset}}$			
Qwen2-VL-7B	62.2	68.6	**70.4**
Qwen2.5-VL-3B	58.0	67.2	**69.4**
Qwen2.5-VL-7B	68.4	70.2	**73.8**
Qwen3-VL-8B	72.0	73.4	**77.6**
QAEgo4D$_{\text{test-mc}}$			
Qwen2-VL-7B	56.6	**59.4**	57.2
Qwen2.5-VL-3B	50.4	**54.0**	53.2
Qwen2.5-VL-7B	55.2	57.4	**58.0**
Qwen3-VL-8B	56.6	60.6	**62.2**

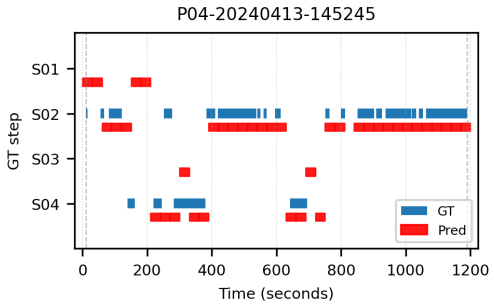

Figure 3: **Example of DH-Mem step construction on HD-EPIC video.** Blue bars denote the reference temporal intervals of each ground-truth step, and red bars denote the substeps from the predicted semantic cluster corresponding to each reference step. This shows how DH-Mem can group semantically related but temporally disjoint observations into a single step-level memory, rather than enforcing temporal contiguity.

indicating that its benefit comes not only from retaining local substep descriptions, but also from semantically organizing them into higher-level procedural structure.

Component analysis. Table 4 analyzes the cumulative effect of DH-Mem and EC-Skills across HD-EPIC, EgoSchema$_{\text{subset}}$, and QAEgo4D$_{\text{test-mc}}$. The baseline directly answers from sampled video frames, the memory-augmented setting additionally provides DH-Mem, and the full EgoSkill framework further applies EC-Skills to construct question-dependent evidence. Adding DH-Mem improves over the baseline in every evaluated setting, showing that query-agnostic hierarchical memory provides useful context across benchmarks and backbone models.

EC-Skills provide substantial additional gains on HD-EPIC, where different question types require heterogeneous forms of evidence construction, including temporal filtering, candidate-wise verification, auxiliary tools, and visual input adaptation. Similarly, on EgoSchema$_{\text{subset}}$, EC-Skills further improve performance over the DH-Mem setting across all evaluated backbones.

On QAEgo4D$_{\text{test-mc}}$, the effect is less uniform: EC-Skills improve the two larger backbones, while the DH-Mem setting performs slightly better for Qwen2-VL-7B and Qwen2.5-VL-3B. Our main experiments use a fixed EC-Skill policy established with Qwen3-VL-8B and transfer it unchanged across backbones. Additional per-backbone analysis in Appendix D shows that adapting the EC-Skill assignments and configurations recovers these drops, indicating that the preferred EC-Skill policy can depend on the solver backbone.

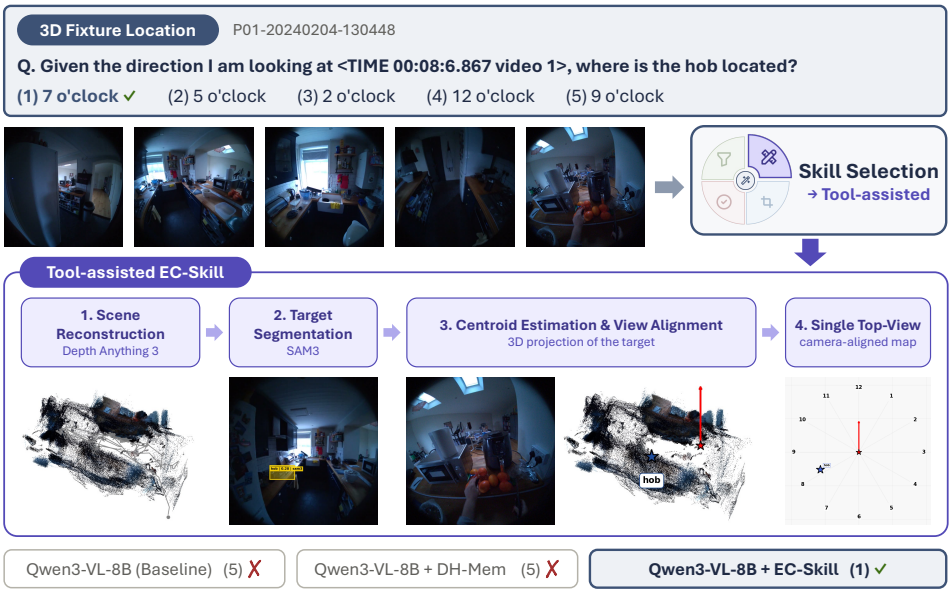

Figure 4: **Qualitative results of EC-Skill on HD-EPIC (Perrett et al., 2025).** For 3D fixture localization question, the skill renders a camera-aligned top-view map with the target fixture localized, providing structured spatial evidence for clock-direction prediction.

Overall, DH-Mem and EC-Skills play complementary roles. DH-Mem provides reusable, query-agnostic context for long-form video, whereas EC-Skills determine how evidence should be selected, transformed, augmented, or verified for a given question type. The consistent gains from DH-Mem, together with the additional improvements from EC-Skills in most settings, support the complementary roles of structured memory and question-dependent evidence construction.

4.5 Qualitative Analysis

Figure 4 illustrates a tool-assisted EC-Skill for a 3D fixture localization question. Given a query asking for the clock direction of the hob relative to the camera-facing direction, the EC-Skill first reconstructs the scene and estimates camera poses using Depth Anything 3, then localizes the target fixture with SAM3. The segmented target is projected into the reconstructed 3D space to estimate its spatial position relative to the camera, from which a camera-aligned top-view map is rendered. This single top-view image is provided to the frozen MLLM solver as task-aligned visual evidence for predicting the clock direction. The example shows how an EC-Skill can transform raw egocentric observations into a compact spatial representation through an explicit evidence-construction procedure.

5 Conclusion

We presented EgoSkill, a fully training-free store–execute framework for long-form egocentric VQA. EgoSkill separates query-agnostic video representation from query-aware evidence construction: DH-Mem preserves temporal continuity for local observations while semantically composing potentially non-contiguous observations into higher-level procedural structure, and EC-Skills provide reusable procedures that adapt evidence retrieval, visual processing, tool use, and verification to the requirements of each question. Across HD-EPIC, EgoSchema, and QAEgo4D, EgoSkill consistently improves frozen MLLM baselines, while our ablations show complementary benefits from structured memory and question-dependent evidence construction. These results highlight the value of explicitly separating what to store from how to execute when reasoning over long-form egocentric video with frozen MLLMs.

AI USE STATEMENT

We utilized GPT to refine the manuscript's wording and flow, and Claude for minor code revisions. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final content and results.

REPRODUCIBILITY STATEMENT

We provide the implementation details required to reproduce our experiments in the appendix, including model and decoding configurations, memory construction hyperparameters, computational resources, and prompt templates. We additionally document the benchmark-specific protocol definitions, EC-Skill procedures, and protocol-to-skill assignments and configurations used to obtain the reported results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024.

Yeeun Choi, Youngbeom Yoo, Joon-Young Lee, Hyolim Kang, and Seon Joo Kim. Keep it simple: Multi-key episodic memory retrieval for ultra-long video understanding. *arXiv preprint arXiv:2608.07663*, 2026.

Meng Chu, Yicong Li, and Tat-Seng Chua. Graphvideoagent: Enhancing long-form video understanding with entity relation graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4639–4648, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, 2024.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, et al. Glm-ocr technical report. *arXiv preprint arXiv:2603.10910*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*, pp. 92–110. Springer, 2024.

Google. Gemini 2.0 flash. URL https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2023.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, et al. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24169–24179. IEEE, 2025.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18198–18208, 2024.

Hyolim Kang, Yunsu Park, Youngbeom Yoo, Yeeun Choi, and Seon Joo Kim. Open-ended hierarchical streaming video understanding with vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20715–20725, 2025.

Haoyuan Li, Zhengdong Hu, Jun Wang, Hehe Fan, and Yi Yang. Skill-3d: Evolving scene-aware skills for agentic 3d spatial reasoning. *arXiv preprint arXiv:2606.07436*, 2026a.

Xiangyi Li, Yimin Liu, Wenbo Chen, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, et al. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026b.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6592–6600, 2026c.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Hao Luo, Zihao Yue, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Deheng Ye, and Zongqing Lu. Openmmego: Enhancing egocentric understanding for lmms with open weights and data. *Advances in Neural Information Processing Systems*, 38:25749–25781, 2026.

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*, volume 36, 2023.

Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Jiangmiao Pang, and Yu Qiao. Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. *Advances in Neural Information Processing Systems*, 38:44140–44168, 2026.

Simone Alberto Peirone, Francesca Pistilli, and Giuseppe Averta. Hiero: understanding the hierarchy of human behavior enhances reasoning on egocentric videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19862–19871. IEEE, 2025.

Toby Perrett, Will Price, Jorge Cartucho, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23901–23913, 2025.

Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13760–13769. IEEE, 2022.

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

Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in neural information processing systems*, 36:38863–38886, 2023.

Agnese Taluzzi, Davide Gesualdi, Riccardo Santambrogio, Chiara Plizzari, Francesca Palermo, Simone Mentasti, and Matteo Matteucci. From pixels to graphs: using scene and knowledge graphs for hd-epic vqa challenge, 2025. URL https://arxiv.org/abs/2506.08553.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

U.S. Department of Agriculture, Agricultural Research Service. FoodData Central, 2019. URL https://fdc.nal.usda.gov/.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yifei Wang, Zhenkai Li, Tianwen Qian, Huanran Zheng, Zheng Wang, Yuqian Fu, and Xiaoling Wang. Streameqa: Towards streaming video understanding for embodied scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9422–9432, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Sicheng Yang, Yukai Huang, Weitong Cai, Shitong Sun, Fengyi Fang, You He, Yiqiao Xie, Jiankang Deng, Hang Zhang, Jifei Song, et al. Egocentric co-pilot: Web-native smart-glasses agents for assistive egocentric ai. In *Proceedings of the ACM Web Conference 2026*, pp. 8862–8873, 2026.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mmego: Towards building egocentric multimodal llms for video qa. In *International Conference on Learning Representations*, volume 2025, pp. 71705–71723, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Barry Zhang, Keith Lazuka, and Mahesh Murag. Equipping agents for the real world with agent skills. *Anthropic Engineering Blog*, 2025a.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b. URL https://arxiv.org/abs/2410.02713.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3363–3373. IEEE, 2025.

APPENDIX

A LICENSE

Table 5: **Summary of licenses for key artifacts.** We strictly adhere to the licensing terms and usage policies of all external datasets, models, and tools utilized in our framework.

Artifact	Type	License
HD-EPIC	Dataset	CC BY-NC 4.0
EgoSchema	Dataset	MIT License
QAEgo4D	Dataset	MIT License
Qwen-VL Series	MLLM	Apache License 2.0
GLM-4.6V	MLLM	MIT License
GLM-OCR	Tool	MIT License
SAM3	Tool	Meta SAM License
Grounding DINO	Tool	Apache License 2.0
Depth Anything 3	Tool	Apache License 2.0

B BENCHMARK DETAILS

We evaluate on three egocentric multiple-choice VideoQA benchmarks, summarized in Table 6. The benchmarks differ in video duration, question design, and required reasoning skills, providing complementary evaluation settings for long-form egocentric video understanding.

Table 6: **Summary of the evaluation benchmarks.** MC denotes multiple-choice VideoQA.

Benchmark	Duration	#Videos	#QA	Type
HD-EPIC	15.9 ($\pm$ 14.5) min	156	26,550	MC
QAEgo4D$_{\texttt{test-mc}}$	8.3 min	148	500	MC
EgoSchema$_{\texttt{subset}}$	3 min	500	500	MC

HD-EPIC (Perrett et al., 2025) is a highly detailed egocentric VQA benchmark collected from unscripted kitchen activities. We evaluate on 156 videos with 26,550 multiple-choice questions. The benchmark contains 30 subtasks grouped into seven macro-categories: Recipe, Ingredient, Nutrition, Fine-grained Action, 3D Perception, Object Motion, and Gaze, with 8, 6, 3, 4, 4, 3, and 2 subtasks, respectively. It emphasizes detailed human-object interaction, recipe understanding, object motion, 3D perception, and gaze-based reasoning.

EgoSchema (Mangalam et al., 2023) evaluates long-range video-language understanding over three-minute Ego4D clips. We use the official 500-question subset, denoted EgoSchema$_{\texttt{subset}}$. Each question has five answer choices and is designed to require understanding information distributed across an extended temporal context.

QAEgo4D (Di & Xie, 2024) focuses on question answering in long egocentric videos with temporal grounding. We use QAEgo4D$_{\texttt{test-mc}}$, a 500-question multiple-choice test split over 148 videos. This benchmark evaluates whether a model can answer episodic questions that often require locating relevant temporal evidence in long first-person videos.

C ADDITIONAL IMPLEMENTATION DETAILS

For HD-EPIC, each question is paired with up to 128 frames decoded at 1 FPS from the native 1408×1408 / 30 FPS source, with frames center-cropped at 1.0 (no crop) and resized to 512×512.

For EgoSchema$_{\texttt{subset}}$, we uniformly sample 32 frames per question across the full clip; frames use the lmms-eval default Qwen-VL processor settings.

For QAEgo4D$_{\texttt{test-mc}}$, we uniformly sample 32 frames per question, with the same processor settings as EgoSchema.

Table 7: **HD-EPIC VQA full subtask results.** 30 subtask results across all backbones and inference settings (% Acc.).

	Recipe Recognition	Multi-Recipe Recognition	Multi-Step Localisation	Step Localisation	Prep Localisation	Step Recognition	Rough Step Localisation	Following Activity Recognition	Ingredient Retrieval	Ingredient Weight	Ingredients Order	Ingredient Adding Localisation	Ingredient Recognition	Exact Ingredient Recognition	Image Nutrition Estimation	Nutrition Change	Video Nutrition Estimation	Action Recognition	How Recognition	Why Recognition	Action Localisation	Fixture Location	Object Location	Object Contents Retrieval	Fixture Interaction Counting	Object Movement Itinerary	Object Movement Counting	Stationary Object Localisation	Gaze Estimation	Interaction Anticipation
Qwen2.5-VL-7B																													
Baseline	34.0	70.0	32.0	40.0	20.0	15.0	22.0	48.0	27.0	34.0	38.0	38.0	40.0	30.0	32.0	16.0	64.0	57.7	38.6	50.6	20.4	27.4	31.0	29.0	31.3	25.4	40.0	27.0	41.8	26.2
+ DH-Mem	12.0	68.0	52.0	46.0	33.0	42.0	51.0	64.0	45.0	40.0	50.0	42.0	26.0	30.0	26.0	12.0	64.0	60.2	41.0	55.4	21.6	24.8	36.6	47.5	29.3	23.2	39.5	27.0	39.8	25.6
+ EgoSkill	12.0	68.0	52.0	46.0	68.0	70.0	51.0	64.0	75.0	78.0	50.0	54.0	26.0	34.0	26.0	**64.0**	64.0	60.2	45.6	61.2	26.2	30.0	37.0	47.5	41.7	37.0	39.5	**46.5**	45.4	25.6
Qwen3-VL-8B																													
Baseline	50.0	68.0	60.0	36.0	30.0	29.0	47.0	64.0	24.0	38.0	52.0	37.0	28.0	24.0	24.0	14.0	54.0	60.9	44.2	51.6	23.6	26.0	31.2	55.0	29.7	33.0	53.5	29.0	42.6	22.4
+ 30 sec	50.0	72.0	92.0	68.0	65.0	61.0	76.0	72.0	61.0	36.0	52.0	49.0	40.0	20.0	24.0	18.0	64.0	60.8	43.6	51.8	23.8	25.2	42.6	**63.5**	35.3	36.4	46.0	28.5	42.1	22.3
+ DH-Mem	62.0	72.0	86.0	72.0	66.0	64.0	74.0	72.0	69.0	36.0	54.0	50.0	40.0	20.0	28.0	20.0	70.0	63.8	44.2	59.2	25.9	21.4	46.4	60.5	35.3	38.0	48.5	30.5	45.6	35.8
+ EgoSkill	62.0	78.0	86.0	76.0	81.0	87.0	79.0	72.0	87.0	78.0	54.0	70.0	40.0	34.0	37.0	62.0	70.0	63.8	52.8	64.6	28.2	40.8	50.8	60.5	**48.7**	**52.4**	48.5	46.5	56.8	35.8
Qwen3-VL-32B																													
Baseline	40.0	76.0	84.0	32.0	32.0	30.0	31.0	74.0	24.0	28.0	56.0	37.0	42.0	26.0	23.0	28.0	70.0	66.4	51.0	57.8	23.3	30.2	41.6	41.5	36.0	20.4	42.5	27.5	46.3	22.9
+ DH-Mem	72.3	72.0	98.0	80.0	81.0	74.0	77.0	76.0	70.0	28.0	60.0	55.0	54.0	22.0	28.0	18.0	68.0	70.1	48.2	60.6	26.0	26.4	53.4	51.5	40.3	22.2	49.0	31.5	46.4	32.3
Qwen3.5-27B																													
Baseline	42.0	68.0	82.0	46.0	40.0	42.0	35.0	66.0	35.0	46.0	60.0	47.0	42.0	28.0	28.0	12.0	74.0	69.5	55.2	60.6	27.5	32.2	39.2	45.0	29.0	16.8	46.5	25.5	45.7	32.1
+ DH-Mem	**74.0**	80.0	**100.0**	**86.0**	79.0	75.0	84.0	72.0	78.0	38.0	60.0	59.0	**64.0**	28.0	24.0	18.0	**80.0**	71.6	55.6	62.8	28.8	33.2	47.6	46.5	34.7	23.2	23.5	30.5	49.6	35.1
+ EgoSkill	74.0	**86.0**	100.0	86.0	**85.0**	**91.0**	**90.0**	72.0	**91.0**	**80.0**	60.0	**80.0**	64.0	**40.0**	**43.0**	60.0	80.0	**80.8**	**65.6**	**70.6**	**37.6**	**41.4**	**54.4**	46.5	44.3	42.0	23.5	46.5	**65.5**	**41.6**
Qwen3-VL-235B																													
Baseline	56.0	74.0	78.0	48.0	44.0	27.0	35.0	**78.0**	29.0	40.0	56.0	36.0	42.0	28.0	29.0	44.0	68.0	70.5	55.0	63.6	24.6	29.2	33.0	46.5	34.7	19.2	52.0	36.0	46.4	27.1
+ DH-Mem	58.0	76.0	98.0	82.0	82.0	74.0	80.0	76.0	74.0	40.0	**64.0**	47.0	56.0	8.0	18.0	22.0	74.0	71.4	54.2	67.0	27.9	28.4	44.6	51.0	36.0	22.8	**54.0**	37.5	61.9	31.2
GLM-4.6V																													
Baseline	42.0	66.0	78.0	44.0	32.0	22.0	18.0	78.0	25.0	44.0	54.0	40.0	44.0	34.0	24.0	24.0	72.0	59.4	39.0	56.0	23.0	23.2	27.0	52.0	29.0	18.6	45.5	32.0	45.2	22.3
+ DH-Mem	66.0	74.0	98.0	64.0	71.0	70.0	70.0	68.0	70.0	40.0	64.0	47.0	50.0	24.0	29.0	22.0	78.0	62.5	43.4	59.6	25.1	18.8	41.2	51.5	33.0	27.0	48.5	36.5	45.4	27.2

Decoding and reproducibility. All MLLM calls use greedy decoding (`do_sample`=False, `temperature`=None) in `bfloat16` with `flash_attention_2` (SDPA fallback if unavailable). Answering calls emit a single token (`max_new_tokens`=2, letter A–E); Memory-building calls allocate up to 2048 new tokens per substep, step, goal generation. On HD-EPIC we fix the random seed to 2025 throughout and cache decoded frames on disk so that re-running with the same cache yields byte-identical predictions; $\text{QAEgo4D}_{\texttt{test-mc}}$ and $\text{EgoSchema}_{\texttt{subset}}$ are run through lmms-eval with greedy decoding and the framework's default fixed seeds.

Model size and Budget. All experiments run on a single node with 8×NVIDIA H200 (141 GB) using HuggingFace's `device_map="auto"` for the Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3-VL-32B, Qwen3.5-27B backbones and FSDP-style sharding via `accelerate launch` for the Qwen3-VL-235B-A22B-Instruct and GLM-4.6V(106B). For tool-assisted skills, we additionally integrated several specialized auxiliary models, including SAM 3 (848 M), DA3-streaming (120 M), GLM-OCR (0.9 B), and Grounding DINO-B ($\sim$ 240 M). In total, our experiments required approximately 3,600 GPU hours, including dynamic hierarchical memory construction, benchmark evaluation, and skill discovery.

Prompt details. For reproducibility, we provide the prompt templates used in our pipeline. Tables 12–14 show the prompts for dynamic hierarchical memory construction, while Table 15 shows the prompt for narration generation.

D Additional Analysis

Per-subtask analysis on HD-EPIC VQA. Table 7 reports per-subtask accuracy across all backbones and inference settings. The overall gains across diverse question categories show that DH-Mem and EC-Skills improve frozen MLLMs across heterogeneous question types in long-form egocentric VQA.

Backbone-specific EC-Skill policies. For the main experiments, we establish a fixed EC-Skill policy with Qwen3-VL-8B and transfer it unchanged to the other solver backbones. This setting evaluates the cross-backbone transferability of the same question-type-to-skill assignments and associated configurations. Table 8 additionally reports a diagnostic analysis in which the EC-Skill policy is adapted to each backbone. In some protocols, the preferred EC-Skill procedure itself differs across backbones, while in others only its configuration changes. The resulting improvements recover the small drops observed for the smaller backbones on QAEgo4D and yield additional gains in several settings. These results suggest that EC-Skill procedures are reusable across solver backbones, while their preferred assignment and configuration can depend on the backbone model.

Table 8: **Backbone-specific EC-Skill policy analysis on EgoSchema$_{\texttt{subset}}$ and QAEgo4D$_{\texttt{test-mc}}$.** The main experiments transfer a fixed EC-Skill policy established with Qwen3-VL-8B across solver backbones. Here, we additionally report a diagnostic analysis using backbone-specific skill assignments and configurations.

Model	EgoSchema$_{\texttt{subset}}$	QAEgo4D$_{\texttt{test-mc}}$
Qwen2-VL-7B	62.2	56.6
+ EgoSkill	72.4	61.0
Qwen2.5-VL-3B	58.0	50.4
+ EgoSkill	69.8	55.4
Qwen2.5-VL-7B	68.4	55.2
+ EgoSkill	73.8	58.6
Qwen3-VL-8B	72.0	56.6
+ EgoSkill	77.6	62.2

Qualitative analysis of DH-Mem. Figure 5 illustrates how DH-Mem supports recipe-step localization. A flat 30-second memory retains local observations as chronologically ordered chunks without explicitly organizing them by higher-level procedural meaning. In contrast, DH-Mem semantically composes related substeps into higher-level procedural steps, making the context relevant to the queried activity more explicit. In this example, this structured representation enables the frozen MLLM to identify the correct temporal interval, whereas both direct video prompting and the flat substep memory fail.

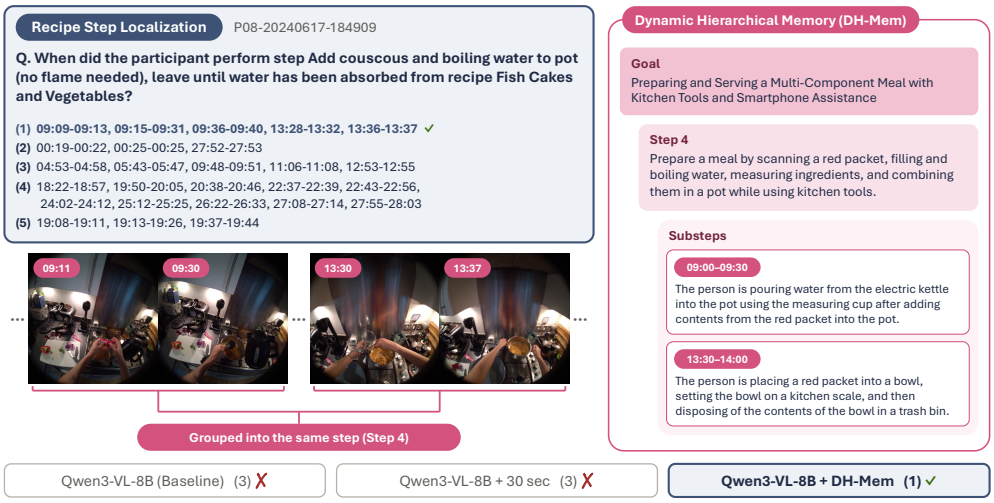

Figure 5: **Qualitative results of DH-Mem on HD-EPIC (Perrett et al., 2025).** DH-Mem groups semantically related local observations into higher-level procedural steps, making the context relevant to the queried recipe step explicit. This structured memory leads to the correct localization, whereas direct video prompting and a flat 30-second memory produce incorrect predictions.

E Evidence Construction Skills Details

This appendix provides the concrete EC-Skill procedures and configurations used in our experiments. We first describe how question types, which we refer to as *protocols*, are defined for each benchmark in §E.1. We then detail the four EC-Skill families and their concrete experimental instantiations in §E.2. Finally, we provide the protocol-to-skill assignments and associated configurations used in the main experiments.

E.1 Protocol definition

HD-EPIC provides a subtask label for each question in its released annotations. We therefore use the provided subtask directly as the question protocol and assign each protocol to its predefined EC-Skill procedure.

EgoSchema$_{\texttt{subset}}$ and QAEgo4D$_{\texttt{test-mc}}$ provide question text but no explicit question-type labels. For these benchmarks, we assign each question to a protocol using an ordered set of regular-expression rules applied to the question stem, with the first matching rule determining the protocol. Tables 9 and 10 summarize the resulting protocols, representative question forms, and the number of questions assigned to each protocol in the 500-question evaluation sets.

Table 9: QAEgo4D$_{\texttt{test-mc}}$ 6 protocol distribution.

Protocol	N	Question shape
Q1 loc_recall	36	"where did I put / place / leave X?"
Q2 loc_prior	143	"where was X before / when / after . . . ?"
Q3 obj_action	184	"what / which did I *verb*?"
Q4 attribute	58	"what color / size / shape / kind / material was X?"
Q5 count	61	"how many / how much X?"
Q6 social	18	"who . . . ?"

Table 10: EgoSchema$_{\texttt{subset}}$ 8 protocol distribution.

Protocol	N	Question shape
P1 goal	194	"what is the overall goal / primary objective?"
P2 steps	79	"what are the key recurring steps / phases?"
P3 why	51	"why does the participant do this?"
P4 method	149	"how does the participant perform / what method?"
P5 tool	12	"what is the primary tool / object?"
P6 moment	8	"most significant moment / in what order?"
P7 anticip.	2	"what will happen next?"
P8 spatial	5	"where does the activity happen?"

E.2 EC-Skill Families and Configurations

We organize the EC-Skills used in our experiments into the four families introduced in §3.3. Each subsection below describes the underlying evidence-construction procedure together with the concrete configurations used in our experiments. The same EC-Skill procedure can be reused across benchmarks and question protocols, while its prompting, temporal cues, sampling strategy, or other configuration details may vary according to the question. We indicate the benchmarks for which each configuration is used.

Memory-guided skills (Family I). These skills shape *how the memory is presented* to the answer model: either by adding a focusing directive on top of the full memory, or by filtering the memory to a relevant subset before the final call.

- **Memory-hint recognition** — the lowest-cost variant. The full DH-Mem is rendered as text and a single protocol-specific instruction is inserted between the memory and the question, telling the model which memory field (goal, steps, substeps) is most relevant. *Instances:* HD-EPIC `recipe_step_recognition`, `recipe_multi_recipe_recognition`, `ingredient_exact_recognition`; EgoSchema: `P2--P8`. Variation is in the hint sentence only; the mechanism is shared.
- **Entity-anchored memory window** — parses the question's noun phrase, expands it with a synonym pool, filters memory substeps to those mentioning the entity, and redirects the frame budget into the time window around those mentions. *Instances:* QAEgo4D `Q2`. Variation is in window placement (post / pre / union) and frame budget.

Table 11: **Family × benchmark assignment.** The four inference-skill families are dispatched per (benchmark, protocol).

Benchmark	I. Mem.-guided	II. Per-choice	III. Tool-assisted	IV. Input-adapt.
EgoSchema	P2–P8	P1	—	—
QAEgo4D	Q2	Q3 , Q6	—	—
HD-EPIC	`step_recog`, `multi_recipe`, `ing_exact`, `ing_retrieval`	`prep_loc`, `rough_step_loc`, `step_loc`, `ing_add_loc`, `action_loc`	**OCR**: `ing_weight`. **USDA LUT**: `nut_change`, `nut_img`. **SAM3+DA3**: `3d_fix_loc`, `3d_obj_loc`, `3d_fix_cnt`. **32B narr.**: `motion_itin`, `motion_stat`.	`how`, `why`, `gaze_est`, `gaze_antic`

- **Timestamp-anchored memory window** — the same mechanism as "Entity-anchored memory window" but the anchor is a `<TIME>` tag parsed from the question rather than an entity. Memory substeps overlapping the tagged interval (± pad) are kept, and the frame budget is concentrated there. *Instances:* HD-EPIC `ingredient_retrieval` (128 frames dense in the window).

Per-choice verification skills (Family II). A single template that decomposes an n-way MCQ into n independent yes/no calls and reads first-token logits to pick the choice with the strongest positive evidence.

- **Per-choice yes/no verification** — for each candidate answer c, we build a prompt "*Is c the correct answer to question Q?*", forward once, and compute $\mathrm{score}(c) = \mathrm{logit}(\mathrm{YES}) - \mathrm{logit}(\mathrm{NO})$ on the first-token logits (averaged across `{YES,Yes,yes}` and `{NO,No,no}` BPE variants). The predicted letter is $\arg\max_c \mathrm{score}(c)$. *Instances:* HD-EPIC `recipe_prep_localization`, `recipe_rough_step_localization`, `recipe_step_localization`, `ingredient_adding_localization`, `fine_grained_action_localization`; QAEgo4D `Q3,Q6`; EgoSchema `P1`. Variation is in the choice anchor shape (interval / noun / essay).

Tool-assisted skills (Family III). These skills invoke an external tool before the final MLLM call to construct evidence that frozen prompting does not extract reliably. The tool output is fed back to the model as structured text, image crop, rendered scene, or auxiliary narration; the model then picks the multiple-choice letter.

- **Digit OCR with exact-match override** — GLM-OCR is run on dense frames with two crop regions over the digital-scale display; if a detected integer exactly matches one of the choice values, that letter is emitted (OCR overrides the VLM); otherwise the baseline MLLM is called as fallback. *Instance:* HD-EPIC `ingredient_weight`.
- **USDA nutrient look-up table** — the VLM is asked only to name the ingredient (or amount in grams); a deterministic solver then fits the 4-vector (kcal, carb, fat, protein) of each choice against a 277-entry USDA LUT entry indexed by the VLM's ingredient string, and picks the minimum-residual choice. *Instances:* HD-EPIC `nutrition_change` (4D residual against a VLM-anchored LUT row), `nutrition_image_estimation` (LUT × VLM-estimated grams).
- **3D-geometry / per-frame detection stack** — SAM3 segmentation + Depth-Anything-3 reconstruction + Grounding-DINO detection trackers produce a structured artifact such as a bounding box, top-view PNG, or detection track; the VLM then reads this artifact and selects the letter. *Instances:* HD-EPIC `3d_fixture_location` (SAM3 + DA3 + topview render), `3d_object_location` (SAM3 bbox crop + before/after composite frame), `3d_fixture_interaction_counting` (two-pass: bbox → class noun → narration filter → letter call).
- **Long-clip atomic-narration generator** — a Qwen3-VL-32B-Instruct narration generator is run over 30-second sliding clips of each video and emits one high-density atomic-action

sentence per clip. The dense narration stream is used as the memory source for downstream questions whose answers depend on micro-motions that the coarser 8B substep captions in DH-Mem routinely lose ("slides 5 cm to the left and back, twice"). *Instances:* HD-EPIC `object_motion_itinerary`, `object_motion_stationary_localization`.

HD-EPIC Family-III total: 8 subtasks. Family III is HD-EPIC-only because only HD-EPIC asks questions whose answers are extra-modal (exact weight, calorie change, 3D direction, per-frame counts, atomic object motion); QAEgo4D-MC and EgoSchema-subset are pure MCQ over scene context and do not require tool extraction.

Input-adaptation skills (Family IV). A single MLLM call whose prompt and / or visual shape is tuned for a specific question type.

- **Subtask-tuned prompt directive** — a handcrafted instruction block names the axes the model should attend to, with no other change. *Instances:* HD-EPIC `fine_grained_action_recognition` (verb-noun extraction), `fine_grained_how_recognition` (how-axis: focus on hand motion, tool usage, body posture, contact points), `fine_grained_why_recognition` (why-axis: focus on downstream goal / current task context).
- **Visual framing tweak** — the frame budget and / or crop are reshaped to put the answer-relevant pixels in view. No choice decomposition, no tool. *Instances:* HD-EPIC `gaze_estimation` (center crop, dual-view ego frame), `gaze_interaction_anticipation` (clip-tail frames with `-no_memory` to block future leakage).

Limitations and Future Work

Our current EC-Skills are human-designed, and skill selection relies on predefined mappings from question types to skill procedures. While this design enables controlled and interpretable evidence construction, it requires manual specification of both the available skills and their assignments. An important direction for future work is to enable agents to automatically discover, generate, or compose EC-Skills from task requirements, allowing the evidence-construction policy to adapt beyond a fixed human-designed skill set.

Table 12: **DH-Mem substep prompt.**

```
SYSTEM
    You are an expert AI assistant for fine-grained video understanding. Your goal is to analyze visual
      evidence objectively and provide precise action descriptions.

USER
    You are an annotation model for fine-grained video understanding.

    INPUTS
    - CURRENT FRAMES: Visual evidence provided in chronological order.
    - TEMPORAL CONTEXT: The timestamp range (Start to End) and frame count provided in the user prompt.
    - PREVIOUS ACTION HISTORY: A list of immediately preceding actions. Use this to maintain consistency in
      object identity and action flow.

    TASK
    First, analyze the visual evidence to resolve object ambiguity. Then, generate a single, objective
      sentence describing WHAT the person is doing.

    ***[TEMPORAL & LOGICAL REASONING GUIDELINES]***
    1. ANALYZE STATE TRANSITIONS & SEQUENCE (Sequential Completeness):
    - **Track the Full Sequence:** Do not judge the action solely by the first or final frame. If the video
      segment involves multiple distinct interactions (e.g., retrieving Object A, then retrieving Object B),
      the final description must capture the **cumulative sequence of actions**, not just the final state.
    - **No Omission:** Do not omit earlier distinct objects (e.g., items retrieved before the final
      placement) simply because the user interacts with a different object at the very end.
    - **State Definition:** Describe the **active process** (e.g., "placing," "pouring," "stirring") rather
      than the inferred result (e.g., "done," "finished"). Ensure the narrative includes the initiation of
      the action if it occurred within the start timestamps.

    2. FUNCTIONAL AFFORDANCE:
    - If a heavy object (like a pan or container) is placed ON TOP of another object, the bottom object must
      be a sturdy base.
```

```
3. HISTORICAL CONSISTENCY (Object Permanence & Identity):
- **Maintain Identity across State Changes:** Refer to the "PREVIOUS ACTION HISTORY". If a specific
  object's identity was established in previous steps, **persist that specific label** even if its
  visual appearance changes due to processing (e.g., slicing, cutting, or mixing). Do not downgrade the
  label to a generic visual descriptor (e.g., use the specific ingredient name if known) unless the
  original identity is irretrievably lost.
- **Prioritize History for Identity:** Trust the visual evidence for the current state (e.g., "cooked,"
  "sliced"), but trust the history for the identity. **Once an object represents a specific class,
  maintain that label even if its appearance changes. Do not rename it based on visual similarity alone;
  only update the identity if a clear object exchange occurs or the visual evidence unambiguously
  disproves the initial classification.**
- **Multi-Object Integration:** If the Visual Analysis identifies interactions with **multiple distinct
  objects** (e.g., a tool, a container, and an ingredient), the ###FINAL sentence must explicitly
  integrate **all primary objects** involved to provide a complete context.
- **Identity Lock Under Continuous Contact (Anti-Drift):** If the person maintains continuous physical
  contact with an object across adjacent frames (or from the immediately preceding action), you must
  keep the same identity label for that object. You may rename/reclassify it **only** if the CURRENT
  FRAMES provide explicit visual evidence of an object swap (e.g., the prior object is put down and a
  different object is picked up) or unambiguous re-identification cues (e.g., clearly readable
  text/logo, clearly different form factor) that directly contradict the prior label.
- **No Content/Category Guessing From Context:** Do not infer an object's contents, ingredient type, or
  functional category from contextual expectations alone (e.g., typical cooking patterns). If identity
  is uncertain and not established in history, use neutral visual descriptors (color/material/shape)
  rather than a speculative category.

4. NO UNVERIFIED STATE CLAIMS (Critical):
- Do **NOT** infer cooking progress or completion states (e.g., "cooked", "boiled", "done", "ready", "
  served") unless there is direct visual evidence in the CURRENT FRAMES.

5. HANDLING AMBIGUITY:
- **Specific Naming:** If an object's identity is known from the History or is visually unambiguous, use
  its **specific name** (e.g., "Parmesan cheese," "Screwdriver").
- **Visual Attribute:** Only when the class is completely unknown and not in the History, describe its
  **visual attributes** (e.g., "a black device") rather than guessing.

OUTPUT FORMAT
You must output exactly two blocks.

Block 1: ## Visual Analysis
- **Active Objects List:** First, list ALL distinct objects the person interacted with or held during
  this entire sequence (e.g., [First Object, Second Object]).
- **Sequential Flow:** Briefly describe the order of interactions (e.g., Interaction with Object A ->
  Interaction with Object B -> Combined Action).
- **Consistency Check:** Explicitly confirm if these objects match any in the PREVIOUS HISTORY.
- **Reasoning:** Explain *why* you identified these objects as X, referring to visual cues.

Block 2: ## Substep Description
- START THE FINAL SENTENCE WITH "###FINAL:"
- **Consolidation Rule:** Your sentence MUST include ALL objects listed in the "Active Objects List"
  above. Do not omit an object just because the interaction happened at the start of the clip.
- Provide a single, complete sentence in the present progressive tense.
- Do not use brackets like < > anymore, just use the marker.
- Example: ###FINAL: The person is [Action_Verb] [Object_A] and [Object_B] [Location] after
  [Previous_Action] them.
```

Table 13: **DH-Mem step prompt.**

```
SYSTEM
    You are an expert AI assistant specialized in Hierarchical Video Understanding.

USER
    Your task is to summarize a sequence of fine-grained "Substeps" (atomic actions) into a single,
      high-level "Step" description (goal-oriented action).

    ### INPUT DATA
    You will be provided with a list of timestamped substeps.
    - These substeps represent a continuous temporal segment belonging to a single higher-level activity.

    ### GOAL
    Synthesize the substeps into ONE concise, imperative sentence that describes the overarching activity
      based strictly on observed facts.

    ### GUIDELINES (Strict Adherence Required)
    1. **Abstraction Level**:
    - Aggregate micro-movements (e.g., "grasp handle," "turn knob") into a functional action (e.g., "Turn on
      the stove").
    - Focus on **WHAT** is being done, rather than **HOW** it feels or looks.

    2. **Temporal Aggregation**:
    - Merge sequential actions logically using connectors like "and" or "then".
    - If preparatory actions (e.g., picking up a tool) lead to a main action, prioritize the main action.

    3. **Strict Groundedness (Fact-Based Only)**:
    - **NO Hallucination of States**: Do not describe the final outcome or texture (e.g., "until crispy,"
      "smoothly," "tightly") unless explicitly stated in the substeps.
    - **NO External Context**: Do not refer to unseen instructions (e.g., "according to the manual," "as per
      the recipe").
    - Only include objects and actions that are explicitly present in the input.

    4. **Generalization**:
    - This logic must apply to any domain (technical, household, etc.).
    - Treat specific items simply as objects being manipulated.

    ### EXAMPLES

    **Example 1 (Technical Domain)**
    substeps:
    - Pick up the screwdriver.
    - Align the screwdriver with the screw head.
    - Turn the screwdriver clockwise.
    Output:
    Fasten the screw using a screwdriver.

    **Example 2 (Household Domain)**
    substeps:
    - Slice the bacon into strips using a knife.
    - Put the sliced bacon into a heated pan.
    - Stir the contents.
    Output:
    Slice the bacon and cook it in a pan.

    ### OUTPUT FORMAT
    - Output exactly ONE imperative sentence.
    - Do not use quotes, numbering, or introductory text.
```

Table 14: **DH-Mem goal prompt.**

```
SYSTEM
    You are an expert AI assistant specialized in High-Level Activity Summarization and Goal Inference.

USER
    Your task is to infer the high-level **GOAL** of the activity based on the provided sequence of
      **STEPS**.

    ### INPUT DATA
    You will be provided with a chronological list of "Steps" (actions) that occurred in a video.

    ### GOAL
    Synthesize the steps into a concise Title Phrase that describes what the person is ultimately achieving.

    ### GUIDELINES (Strict Adherence Required)
    1. **Holistic Inference (Parts to Whole)**:
    - Identify the final product or system being created, modified, or repaired.
    - If multiple steps describe preparing separate components (e.g., base, topping, sub-assembly), combine
      them into a **specific name for the final dish or task** rather than listing them separately.
    - **Logic**: If the combination of ingredients or parts points to a well-known result (e.g., specific
      recipe, specific repair), use that specific name.

    2. **Handling Multi-Tasking vs. Single-Tasking**:
    - **Single Task**: If all steps logically contribute to one final outcome, output a single goal (e.g.,
      "Making Carbonara").
    - **Multi-Task**: Use "**and**" only if the activities describe clearly distinct, independent products
      that do not share a common purpose (e.g., "Making Pasta and Preparing a Salad").

    3. **Strict Groundedness**:
    - Base the inference strictly on the observed objects and actions.
    - Do NOT use subjective adjectives (e.g., "delicious," "perfect") or assume unseen external contexts.

    4. **Style & Format**:
    - Start with a Gerund (Verbing) or a Specific Noun Phrase.
    - Use Title Case.
    - 10–15 words total; no period at the end.
    - NO sentences, NO numbering, NO explanations.

    ### EXAMPLES

    **Example 1 (Integrating Components - Culinary)**
    steps:
    1. Form ground beef into patties and grill them.
    2. Slice tomatoes, onions, and lettuce.
    3. Slice buns and toast them on the grill.
    4. Assemble the meat and vegetables between the buns.
    Output:
    Making Hamburgers

    **Example 2 (Distinct Multi-Task)**
    steps:
    1. Season the chicken and roast it in the oven.
    2. Wash lettuce, chop tomatoes, and mix with dressing.
    Output:
    Roasting Chicken and Preparing Salad

    **Example 3 (Technical/Assembly Domain)**
    steps:
    1. Unscrew the back panel of the laptop.
    2. Disconnect the old battery connector.
    3. Insert the new battery and connect it.
    4. Screw the back panel back in place.
    Output:
    Replacing a Laptop Battery

    ### OUTPUT FORMAT
    - Output exactly ONE line containing the goal title(s).
```

Table 15: **Atomic-action narration extractor prompt**. A Qwen3-VL-32B model is run once per 30-second DH-Mem window with the SYSTEM and USER messages below; the resulting per-video JSON corpus is consumed by the two HD-EPIC motion skills (`object_motion_itinerary` and `stationary_object_localization`) as structured text at inference time.

```
SYSTEM
    You are an expert at decomposing egocentric kitchen video clips into atomic actions.
    An atomic action is a single short purposeful movement (0.3-5 seconds) performed by
    one hand or both hands together.

    ### SENTENCE STRUCTURE (build each narration from these slots; do not invent scenes you cannot see)
    <Verb> the <object> [<adjective/location qualifier>] [with the <left|right|both> hand[s]] [to <intent>].

    Examples of the slot pattern (DO NOT copy the example wording -- describe what is
    actually happening in the clip):
      - "<Verb1> the <object1> with the <hand>."
      - "<Verb2> the <object2> from the <location> with the <hand> to <intent>."
      - "<Verb3> the <object3> by holding the <object3-part> with the <hand>."

    ### STRICT RULES
    1. Describe ONLY what is visibly happening in THIS clip; do not generalize or invent
       actions from the system prompt.
    2. Imperative form, starting with a verb in present tense.
    3. Each action is a SINGLE movement; do not summarize multiple movements into one --
       split a multi-step interaction (e.g. open-reach-pickup-close) into separate atomic
       narrations.
    4. List EVERY distinct movement you see, including brief sub-second ones and ones that
       overlap with other movements. No upper limit.
    5. Order = chronological (no start/end timestamps required).
    6. DO NOT repeat the same action narration twice.
    7. Use lowercase verbs and nouns in the structured fields.

    ### OUTPUT FORMAT
    OUTPUT strict JSON only (no prose, no markdown fences, no explanation):
    {"actions":[
      {"narration":"<imperative sentence>",
       "verbs":["<verb>", ...],
       "nouns":["<noun phrase>", ...],
       "hands":["<left hand|right hand|both hands|none>"]}
    ]}

USER
    Look at the {win_dur}-second clip below (frames sampled at 2 fps in chronological
    order). List EVERY distinct atomic movement you observe, in the order they happen, no
    matter how brief. Split multi-step interactions (open-reach-pickup-close, etc.) into
    separate atomic narrations. Do not cap the count -- produce as many as the clip
    genuinely contains. Output JSON only.
```

Table 16: **Example of Dynamic Hierarchical Memory.**

```
Goal: Preparing Coffee and Orange Juice and Cleaning Kitchen Appliances
S01 Pour milk into a coffee machine and a mug, stir the coffee, dispose of waste, and rinse a container.
  [0–30s] The person is pouring milk from a bottle into a coffee machine after placing a mug on the counter.
  [120–150s] The person is placing a plastic bag into a food waste bin, adjusting a boiler, and pouring milk
      from a container into a mug while stirring it with a wooden spoon.
  [150–180s] The person is stirring coffee in a mug, disposing of an item in a trash bin, and rinsing a
      metal container in the kitchen sink.
S02 Juice oranges using a juicer and pour the juice into a glass, then clean the juicer and food processor.
  [30–60s] The person is cutting a bag of oranges with scissors after retrieving a milk bottle from the
      refrigerator and using a coffee machine.
  [90–120s] The person is placing a juicer on the counter, preparing to juice an orange on a cutting board
      with a knife, and simultaneously retrieving a kitchen towel from a drawer.
  [180–210s] The person is squeezing an orange into a juicer on a cutting board after slicing it with a
      knife.
  [210–240s] The person is squeezing an orange into a juicer on a cutting board after placing the orange
      peel into a trash bin.
  [240–270s] The person is squeezing an orange into a juicer on a cutting board after slicing it with a
      knife.
  [270–300s] The person is squeezing an orange into a juicer on a cutting board.
  [300–330s] The person is washing the juicer's central component in the kitchen sink with a sponge after
      juicing an orange and placing it on a dish rack to dry.
  [330–360s] The person is pouring orange juice from the food processor into a glass after washing the
      juicer's central component and placing it on the dish rack.
  [360–390s] The person is pouring orange juice from the food processor into a glass, then washing the food
      processor's central component in the kitchen sink with a sponge and placing it on the dish rack to dry.
S03 Plug in the food processor, open the dishwasher, retrieve a black container, and return it to the
      dishwasher.
  [60–90s] The person is plugging in a food processor, opening a dishwasher, retrieving a black container
      from the dishwasher rack, and placing it back inside the dishwasher.
S04 Wipe the stove with a kitchen towel.
  [390–396.267s] The person is wiping the stove with a kitchen towel after retrieving it from under the
      stove.
```