MEMAFF: RETRIEVING AND GROUNDING AFFORDANCES FROM VISUAL MEMORY

Anonymous authors
Paper under double-blind review

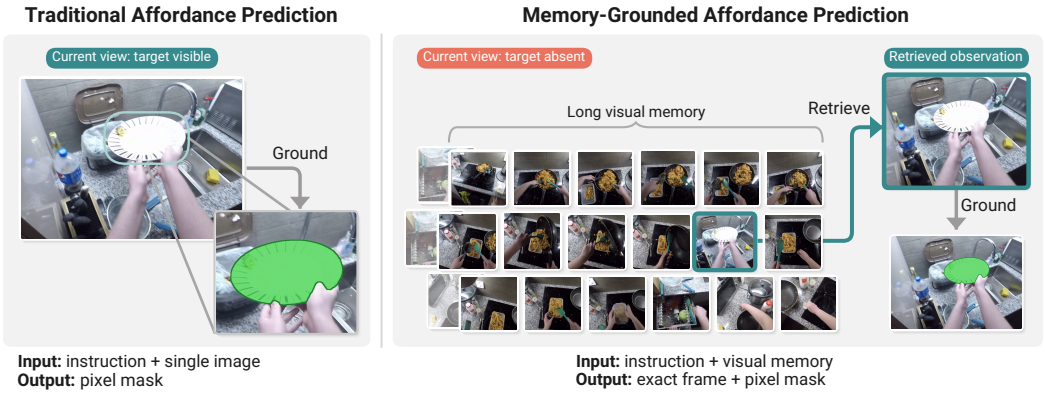

Figure 1: Traditional affordance prediction assumes that the task-relevant object is visible in the current observation. In memory-grounded affordance prediction, we allow the target to be absent from the current view; in this case, the agent must first retrieve the relevant observation from a long visual memory and then ground the actionable object at the pixel level.

ABSTRACT

Affordances bridge high-level task reasoning and low-level physical action by identifying what objects enable and where an agent can interact with them. However, existing affordance prediction methods typically assume that the task-relevant object is visible in the current image. This assumption breaks for embodied agents with persistent visual experience: an instruction may refer to an object seen much earlier, requiring the agent to search its visual memory before determining where to act. We formulate this setting as *Memory-Grounded Affordance Prediction* and introduce MEMAFF-BENCH, a benchmark with 3,628 tasks drawn from egocentric videos and indoor panoramas. Given an intent-style instruction and a sequence of past observations, an agent must retrieve the relevant observation and localize the target object at the pixel level. By systematically increasing the amount of visual memory, MEMAFF-BENCH reveals that existing systems struggle to identify task-relevant observations as memory grows, even when long-context models can technically process the full sequence. To address this challenge, we further develop MEMAFF, a training-free framework that retrieves promising observations, visually reranks them, and grounds the target object. MEMAFF matches or exceeds the frame accuracy of a $4\times$ larger Qwen3-VL-32B single-pass model while using one to two orders of magnitude fewer prompt tokens per query. Our analysis also identifies that reliable retrieval among many visually similar observations is still a central challenge, establishing memory-grounded affordance prediction as an important direction for embodied agents that must reason and act over long-term visual experience.

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.

1 INTRODUCTION

Affordances, the action possibilities that objects offer to an agent (Gibson, 1979), form a crucial bridge between visual perception and physical interaction, supporting robotic grasping, manipulation, and task planning (Hassanin et al., 2021; Apicella et al., 2025). Recent work has advanced weakly supervised part grounding (Luo et al., 2022; Li et al., 2023), 3D affordance understanding (Deng et al., 2021), multimodal reasoning and segmentation (Qian et al., 2024; Yu et al., 2025; Wang et al., 2025), agentic foundation-model pipelines (Zhang et al., 2025), and affordance prediction over panoramic indoor scenes (Zhang et al., 2026).

Most existing methods nevertheless assume that the task-relevant object is already visible in the input image (Qian et al., 2024; Wang et al., 2025). This assumption breaks for embodied agents with persistent visual experience, as illustrated in Figure 1. A robot asked to "get me something to tighten a screw" may have last seen the screwdriver minutes earlier in another room. Before grounding an actionable region, it must search its visual history and recover the relevant observation. We call this setting *Memory-Grounded Affordance Prediction*: given a goal expressed in natural language and a sequence of past visual observations, retrieve the observation containing the task-relevant object and localize that object at the pixel level. The output is an object mask in a past observation; navigation and action execution are outside the evaluated task.

Current benchmarks do not measure this combination. Single-image affordance datasets assume away retrieval, whether they study part-level grounding (Myers et al., 2015; Luo et al., 2022), 3D understanding (Deng et al., 2021), or reasoning-based segmentation (Wu et al., 2025; Wang et al., 2025). Long-video benchmarks search a history, but typically return a timestamp, clip, or textual answer (Grauman et al., 2022; Mangalam et al., 2023), rather than the pixels on which an agent should act.

We introduce MEMAFF-BENCH to close this gap. Its evaluation set contains 3,628 tasks from three real-image sources: EPIC-KITCHENS with VISOR annotations (Damen et al., 2022; Darkhalil et al., 2022), long-form Ego4D video (Grauman et al., 2022), and PAP indoor panoramas (Zhang et al., 2026). We pair each intent-style instruction with a target frame and object mask. Two design choices make the measurement meaningful. First, *object-level uniqueness*: each task is filtered so that a single object in the source memory is intended to satisfy the instruction, and a sampled human audit checks the target and its mask. Second, a *controlled memory-size protocol*: for every task retained at a given scale, the instruction, target, and ground truth remain fixed while distractors increase from 120 to as many as 960 frames. We score both stages together using strict end-to-end mIoU, which is nonzero only when the exact annotated frame is selected.

Our evaluation on MEMAFF-BENCH shows that increasing context length alone does not solve the task. Single-pass systems are costly and often exceed their configured context limits; even when a 32B model can process all 960 frames, it retrieves the correct frame only 6–21% of the time while reading up to 2.5×10^5 prompt tokens per query. The central challenge is finding the task-relevant observation among many distractors, not merely fitting them into the context.

Motivated by this finding, we develop MEMAFF, a training-free retrieve-then-rerank reference system. MEMAFF retrieves a small set of promising observations, visually reranks them, and grounds the target object. It matches or outperforms the 32B single-pass model in frame accuracy on all three sources while using 14–116$\times$ fewer prompt tokens per query. However, its end-to-end performance still drops substantially as memory grows: a larger same-family reranker does not help, most of the lost score at large memories comes from targets that retrieval never ranks, and a learned selector transfers poorly across sources (Appendix B), identifying robust retrieval among visually similar observations as the main bottleneck.

In summary, this work makes four main contributions:

- We formulate *Memory-Grounded Affordance Prediction*, coupling retrieval from visual memory with pixel-level affordance localization and scoring the two stages jointly.
- We compile MEMAFF-BENCH, with 3,628 evaluable tasks over three real-image sources, object-level uniqueness filtering, human auditing, and controlled memory sizes from 120 to 960 frames.

- We systematically evaluate single-pass and retrieval-based systems, showing both hard resource limits and severe selection failures as memory grows; the best strict E2E score among the main systems is 0.144 at 120 frames and 0.070 at 960.
- We provide MEMAFF, a training-free retrieve-then-rerank reference system that is better than or comparable with a $4\times$ larger single-pass model at 14–$116\times$ fewer prompt tokens per query, and use ablations to locate the remaining bottleneck in candidate retrieval.

2 RELATED WORK

Affordance prediction and grounding. Early affordance methods regress heatmaps or masks for an affordance category, often learning from human–object interactions or exocentric demonstrations (Luo et al., 2022; Li et al., 2023); related formulations address 3D objects and tool parts (Deng et al., 2021; Myers et al., 2015). More recent work combines multimodal language models with pixel-level grounding (Qian et al., 2024; Yu et al., 2025; Wang et al., 2025; Wu et al., 2025), coordinates foundation models in an agentic pipeline (Zhang et al., 2025), or expands prediction to panoramic scenes (Zhang et al., 2026). Despite their differences, these methods generally assume that the relevant object is already present in a given image. We instead require the system to find that image within a long memory before grounding the object. Because several objects may satisfy the same functional instruction, our benchmark filters tasks for object-level uniqueness.

Egocentric video understanding and embodied memory. Our setting relates to egocentric episodic-memory tasks, which ask when and where an object was seen (Grauman et al., 2022), and to long-form video reasoning benchmarks (Mangalam et al., 2023). Embodied agents also maintain scene or episodic memories for navigation and planning (Fang et al., 2019). These settings usually output timestamps, clips, or text. Our setting additionally requires pixel-level grounding in the retrieved observation, connecting "what was seen where" to "where the agent can act".

Long-context and memory-augmented multimodal reasoning. Long-context video models can consume many frames in a single request (Zhang et al., 2024a;b; Bai et al., 2025). Strong needle-style retrieval results, however, often use a visually distinctive probe inserted into unrelated content. A real memory of one kitchen or apartment instead contains many near-duplicate views and repeated objects. Consistent with known long-context retrieval difficulties (Liu et al., 2024a; Xiao et al., 2024), we find that a larger context window removes overflow but not selection errors. An alternative is external memory that is written once and queried on demand, as in retrieval-augmented generation (Lewis et al., 2020) and memory-centric agents (Park et al., 2023; Packer et al., 2023). MEMAFF follows this decoupled design while specializing the stored representations and queries for affordance retrieval.

3 THE MEMAFF-BENCH BENCHMARK

MEMAFF-BENCH contains 3,628 evaluation tasks from three complementary real-image sources, spanning egocentric videos and indoor panoramas, with visual memories scaled from 120 to as many as 960 frames (Figure 2).

3.1 TASK FORMULATION

A task consists of a memory $\mathcal{M} = \{f_1, \ldots, f_N\}$, an intent-style instruction q that expresses a goal without naming the target object, and ground truth $(f^\star, m^\star)$: the annotated target frame and object mask. A system must retrieve $f^\star$ and segment the object in that frame. Each task is filtered so that a single object in the source memory is intended to satisfy q (*object-level uniqueness*; Section 3.3); residual ambiguity at larger memories is discussed in Section 6.

Evaluation. We evaluate the complete task using *strict end-to-end mIoU* (E2E): a prediction receives the mask IoU only when the system retrieves the correct frame, and otherwise receives zero. Averaged over tasks, E2E therefore equals frame accuracy (Acc@1) times the mean mask IoU on correctly retrieved frames; we report both factors as diagnostics. The metric measures recovery of the annotation: it does not credit another view of the target or another suitable object.

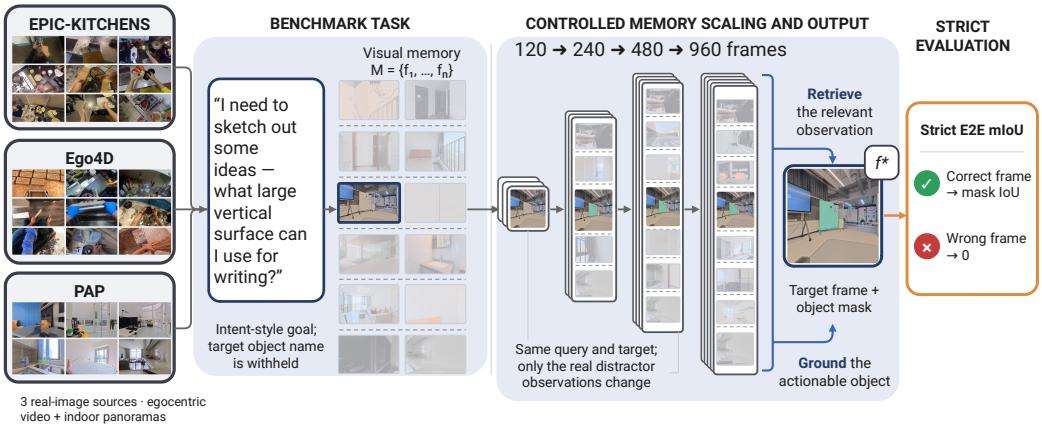

Figure 2: Overview of MEMAFF-BENCH. Tasks are constructed from three real-image sources and pair an intent-style instruction with a visual memory, a target frame, and an object mask. Memory size is increased from 120 to 960 frames with more real distractor observations while keeping the query and target fixed. Strict end-to-end mIoU awards mask IoU only when the exact target frame is retrieved.

Controlled memory sizes. We construct memories at $N \in \{120, 240, 480, 960\}$ frames. For every task retained at a given scale, the instruction, target frame, and mask are unchanged; only the distractor frames change. Larger memories sample the same video more densely for EPIC and Ego4D, or append additional real rooms for PAP, never synthetic images. PAP uses the 585 tasks shared by the three larger sizes (600 at 120 frames), and an Ego4D task is removed at a size its video can no longer fill. Because distractor composition changes along with count and the task sets shrink slightly with scale, Table 1 and all result tables report the denominators at each size.

3.2 SOURCES AND CONSTRUCTION

EPIC-KITCHENS (EPIC). From EPIC-KITCHENS-100 videos and VISOR object annotations (Damen et al., 2022; Darkhalil et al., 2022), we construct 2,621 tasks over 92 kitchen videos. Target localization originates from the human VISOR annotation; SAM 2 redraws the boundary so that the mask format matches Ego4D. Instructions are generated from the annotated verb and noun, rewritten into intent style, and uniqueness-checked. Of these tasks, 2,589 enter the controlled evaluation; 32 are excluded because their target frame is absent from the extracted full-video frame pool.

Ego4D. From Ego4D (Grauman et al., 2022) we build 731 tasks, 452 of which come from the 119 full-length videos available to us for dense sampling. In each memory we exclude every non-target frame whose Qwen3-VL-8B frame inventory (the inventory that MEMAFF's text arm also retrieves over; Section 4.1) names the target object; 13 tasks whose videos cannot then fill 120 frames are dropped, leaving 439. Target frames come from source annotations; masks are generated using LocateAnything-3B (Wang et al., 2026) followed by SAM 2 (Ravi et al., 2025). Dense memories from a single video make Ego4D the most temporally redundant source.

PAP. From PAP-12K indoor panoramas (Zhang et al., 2026), we render eight perspective views per room and concatenate 15 rooms into a 120-frame memory, yielding 600 tasks over 155 episodes. Larger memories append additional real rooms. Instructions are provided by the source benchmark and are already intent-style; target localization is the source panorama annotation projected into the rendered view. Only the ordering of rooms is composed: every frame remains real imagery, but the room sequence need not correspond to an agent's trajectory.

3.3 QUALITY CONTROL

Quality control has two stages. First, an LLM judge examines the candidate objects in each source memory (for Ego4D, the original sparse memory) and retains a task only when it judges exactly one

Table 1: MEMAFF-BENCH statistics. "Retained": tasks passing the construction filters (a sample was human-audited); "evaluation": tasks entering the 120-frame benchmark. GT: origin of target localization; EPIC and Ego4D boundaries are then refined with SAM 2.

Source	Retained	Evaluation	N=120	N=240	N=480	N=960	GT origin
EPIC	2,621	2,589	2,589	2,589	2,589	2,588	VISOR human annotation
Ego4D	731	439	439	433	415	370	keyframe + detector
PAP	600	600	600	585	585	585	panorama annotation
Total	3,952	3,628	3,628	3,607	3,589	3,543	

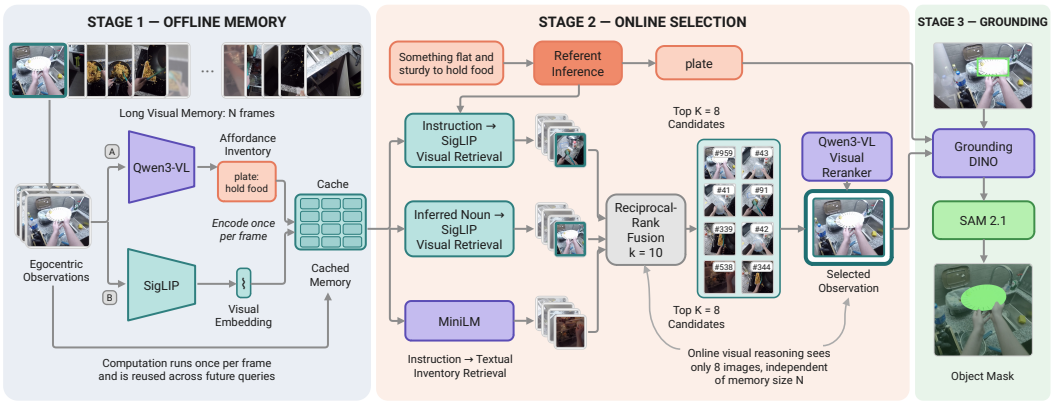

Figure 3: Overview of MEMAFF. During offline memory construction, each observation is encoded once into a textual affordance inventory and a visual embedding. At query time, referent inference and three complementary retrieval paths produce a fused shortlist of eight candidates, which a vision-language model reranks. Grounding-DINO and SAM 2.1 then produce the final object mask on the selected observation.

object suitable; the larger memories are not re-checked for any source (Section 6). The check is object-level because the same object may appear in many nearby frames.

We then manually audit random samples from all three data sources, checking whether the intended target is correct and its localization is accurate. Overall, 109 of 119 audited tasks pass these checks (91.6%; EPIC 27/30, Ego4D 53/60, PAP 29/29). Automatic filtering and this sampled audit are complementary checks on task validity and annotation quality, not a guarantee that every task has a unique valid answer.

4 MEMAFF: A TRAINING-FREE REFERENCE SYSTEM

MEMAFF establishes what can be achieved on MEMAFF-BENCH without task-specific training, using only frozen public components (Figure 3). Unlike one-shot or single-image systems, which re-process all N frames for every instruction, MEMAFF separates the cost: (i) each frame is encoded once into cached visual and textual representations; (ii) online candidate retrieval scans only those cached representations; and (iii) online visual reasoning is restricted to K candidates, so its visual cost is constant in N and the retrieval stage exposes the remaining scaling bottleneck. Memory construction still scales with N; it is paid once, amortized over queries, and excluded from the per-query token counts below.

4.1 OFFLINE MEMORY

For each observed keyframe, MEMAFF caches two instruction-agnostic representations. First, Qwen3-VL-8B-Instruct (Bai et al., 2025) produces a compact *affordance inventory*: movable objects in the frame and tasks they can support (e.g., "`screwdriver: tighten screws`"). Second, SigLIP-large-384 (Zhai et al., 2023) produces a visual embedding. Both are computed once

per frame and reused across future queries; the inventories are model-generated descriptions, not verified annotations.

4.2 Online Selection: Reasoning Before Retrieval

At query time, MemAff reduces the memory to $K = 8$ candidates using cached representations and then visually reranks that shortlist.

Inferring the referent. An intent-style instruction names a goal rather than an object, so a contrastive encoder would have to bridge "something to tighten a screw" directly to a screwdriver image. A short text-only call first asks the language model for the implied object noun, which becomes an additional retrieval query and later conditions grounding.

Three-way candidate retrieval. Three frozen retrievers independently rank all N frames: (1) SigLIP matches the original instruction to frame embeddings; (2) SigLIP matches the inferred object noun to frame embeddings; and (3) all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) matches the instruction to cached textual affordance inventories. We combine the rankings with reciprocal-rank fusion (Cormack et al., 2009) ($k = 10$, over each arm's top 30 frames) and retain the top eight frames. The first two arms share a visual encoder and hence correlated errors; the inventory arm is weak alone but can recover frames missed by both visual arms. The same configuration is used for every source.

Fixed-set visual reranking. The instruction and the slot tags [Frame 0]–[Frame 7] are followed by the eight candidate images in fused-rank order and given to Qwen3-VL-8B-Instruct, which selects the frame containing the object that best fulfills the goal. Candidate order is retained because the reranker uses it as an informative prior (Section 5.4). This stage sees only eight images and about 1.3k–2.2k prompt tokens (depending on frame aspect ratio) whether the memory contains 120 or 960 frames.

4.3 Grounding

The inferred noun conditions Grounding-DINO-tiny (Liu et al., 2024b) (box threshold 0.35) on the selected frame, and SAM 2.1-hiera-large (Ravi et al., 2025) refines the detected box into the final mask. The detector differs from the one used for automatically generated ground truth (Section 3.2). Every compared system except Affordance-R1 and A4-Agent, which produce their own masks, uses this common grounding stage, keeping most end-to-end differences focused on frame selection and referent inference. No component is trained or fine-tuned on MemAff-Bench.

5 Experiments

5.1 Setup

Compared systems. We evaluate three families of systems that can be pointed at a visual memory. **(1) One-shot VLM**: the entire memory is placed into a single request, every frame tagged with its index, and the model names the frame and the object (Qwen3-VL-8B (Bai et al., 2025), served with a 131k-token window, a deployment setting rather than an architectural limit; a larger model in Section 5.3). **(2) Off-the-shelf affordance systems**: Affordance-R1 (Wang et al., 2025), a single-image model run over the frame sequence, and A4-Agent (Zhang et al., 2025), an agentic pipeline with a Qwen3-VL-8B thinker and no context cap (Ego4D and PAP out of memory from 480 frames; Table 2). **(3) Retrieve-then-rerank**: MemAff (Section 4) and its strongest ablation, the same pipeline with a single SigLIP retriever in place of the three-way coarse stage. VLM inputs have a 512-pixel long side. Affordance-R1 and A4-Agent output their own masks (A4-Agent via Rex-Omni and SAM 2); every other system uses the same Grounding-DINO + SAM 2 stage and box threshold, so their E2E differences reflect frame selection and the object noun, not the segmenter. Missing outputs (overflow, out of memory) count as 0.

Protocol. All numbers are on the full benchmark (EPIC 2,589 / PAP 600 / Ego4D 439 tasks at 120 frames), pooled across sources with task weighting; per-source numbers, source-balanced pooling,

Table 2: Strict end-to-end mIoU pooled over the three sources (task-weighted) versus memory size. n = 3,628 / 3,607 / 3,589 / 3,543 tasks. "overflow": requests exceed the 131k-token window; "OOM": out of memory on an 80 GB GPU (Affordance-R1); both count as 0, as do A4-Agent's Ego4D and PAP requests at 480 and 960 frames, which run out of memory on a 48 GB GPU. †OOM verified on the 497-task EPIC subset. Last row: paired t vs. the SigLIP ablation (unrounded scores).

Memory size N	120	240	480	960
MEMAFF (three-way coarse + VLM rerank)	**0.144**	**0.123**	**0.089**	**0.070**
SigLIP coarse + same rerank (ablation)	0.122	0.096	0.069	0.057
One-shot Qwen3-VL-8B	0.122	0.087	0.052	overflow
A4-Agent (Qwen3-VL-8B thinker, no context cap)	0.044	0.017	0.005	0.003
Affordance-R1	0.003	0.001	OOM†	OOM†
Δ MEMAFF − SigLIP ablation (paired t)	+2.2 (5.2)	+2.6 (6.9)	+2.0 (5.5)	+1.3 (3.7)

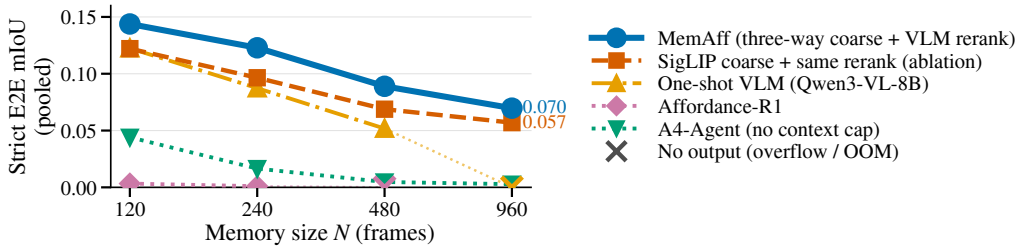

Figure 4: Strict E2E mIoU pooled over the three sources (task-weighted; EPIC 2,589 tasks) versus memory size (log scale). The one-shot VLM hits its context window at 960 frames (×); the off-the-shelf affordance systems are near zero at every size; only the retrieve-then-rerank systems degrade gracefully.

and a fixed 497-task EPIC subset give the same ordering (Appendix A). Significance is a paired t-test over tasks (E2E) or McNemar's test (frame accuracy); these are nominal statistics without correction for tasks that share source videos or recur across sizes. The 32B-reranker ablation uses the 497-task EPIC subset, on which main-table systems differ from their full-benchmark frame accuracy by at most 2 points.

5.2 MAIN RESULTS

Table 2 and Figure 4 give the pooled strict E2E mIoU at the four memory sizes. Figure 5 shows per-task examples. Four observations carry the paper.

Single-pass systems hit a physical wall, and lifting the wall does not restore accuracy. At 960 frames every one-shot request exceeds the configured 131k-token window (1.5–2.5×10^5 tokens at 512-pixel frames) and the one-shot model produces no output on any source. A4-Agent, with no window cap, ingests all 960 EPIC frames yet selects the correct frame for only 5.2% of EPIC tasks at 120 frames and 0.9% at 960; Affordance-R1 runs out of memory from 240 frames on 48 GB and 480 on 80 GB. Failures thus combine hard resource limits with an inability to select among hundreds of frames.

Retrieval degrades gracefully and is the only family with a curve. MEMAFF falls from 0.144 to 0.070 over an 8× growth of memory, and it is the best main-table system at every size. Against the one-shot VLM at the sizes where the latter runs, it is +2.1, +3.6, and +3.7 points ahead (unrounded) at 120, 240, and 480 frames, at about 1.3k–2.2k reranking tokens per query instead of 18k to 127k (offline memory construction excluded).

The three-way coarse stage is a consistent improvement over instruction-only retrieval. Replacing instruction–image SigLIP with the three-way stage improves E2E at all four sizes (+1.3 to

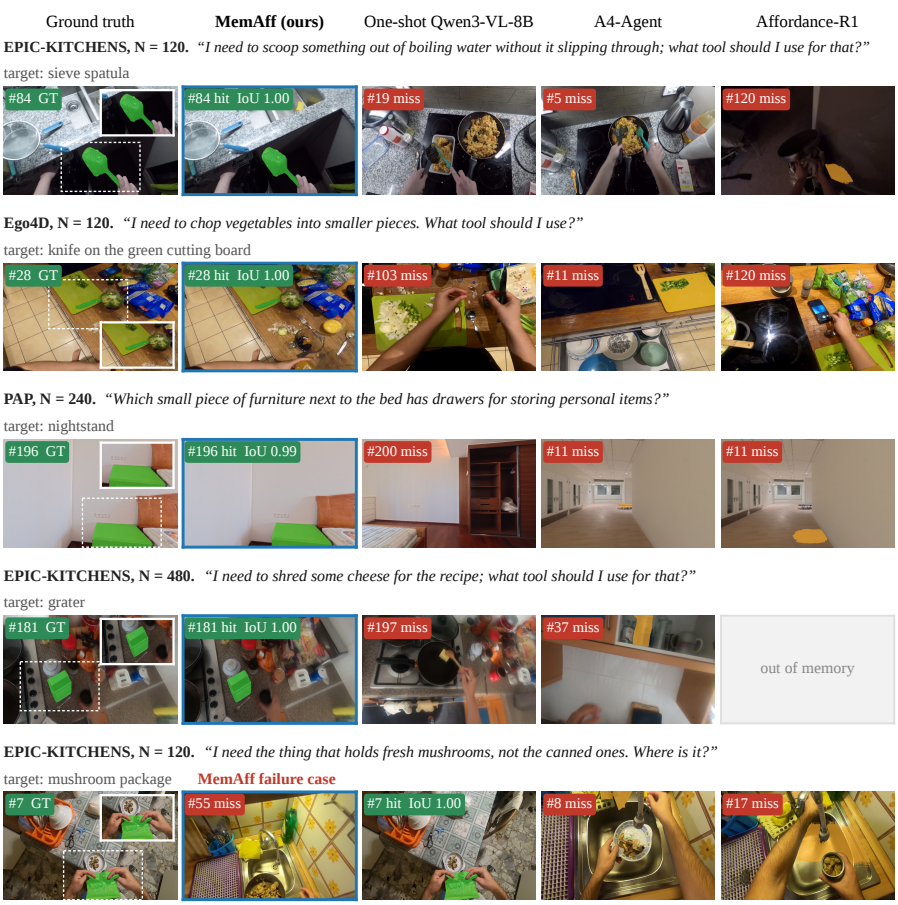

Figure 5: Five tasks (rows), labelled with source and memory size: the ground-truth frame and mask (inset magnifies the target), then each system's selected frame and mask. IoU is shown only when the selected frame is the target; grey cells mark no output; MEMAFF is outlined in blue. In rows 1–4 only MEMAFF retrieves the target. Row 5 is a MEMAFF failure: the small, partly clipped package of fresh mushrooms is ranked beyond position 30 by all three retrievers, so the reranker never sees it, while the one-shot model reading all 120 frames finds it.

+2.6 points, $t = 3.7$–6.9) and frame accuracy in all 12 source-size cells, nine significantly (Appendix Table 8). Section 5.4 shows that it is a robust cross-source choice rather than the optimum for every source.

Absolute scores are low because the metric multiplies two hard stages. Selecting the exact target frame out of N is the dominant loss; once the frame is right, masks are competent (0.45–0.78 IoU at 120 and 960 frames; Appendix Table 7). EPIC is the hardest source (small kitchen objects) and, being the largest, dominates the pooled score; source-balanced pooling gives $0.220 \rightarrow 0.121$ for MEMAFF (Appendix A).

5.3 LARGER BACKBONES

Is the one-shot failure specific to an 8B model with a 131k window? We ran the same one-shot protocol with Qwen3-VL-32B (262k window, so that 960 frames fit; this changes both model size and context) on all three sources in full, and compared it with MEMAFF on exactly the same tasks (Table 3).

With a large enough window, 960 frames no longer overflow, so the hard wall is deployment-dependent; reliable selection, however, remains unresolved, with 32B accuracy of only 0.061 (EPIC)

to 0.205 (Ego4D) at 960 frames. On identical tasks, EPIC and Ego4D differ by at most 2.1 points at any size, all within noise; PAP is tied at 120 frames and MEMAFF leads by 3.6 to 8.0 points from 240 frames on (significant at 240 and 960). Pooled over all 14,367 paired instances, MEMAFF improves frame accuracy from 0.165 to 0.176 ($\sigma = 3.0$), while using 14–116$\times$ fewer prompt tokens per query and an 8B model.

Conversely, replacing only the 8B reranker with Qwen3-VL-32B gives *no* improvement: frame accuracy is unchanged within noise on EPIC, 3.4–3.7 points *lower* on Ego4D ($\sigma = 2.0$–2.3 at every size), and 5.0 points lower on PAP at 120 frames ($\sigma = 2.8$; Appendix Table 10). It follows the candidate order less often (32% vs. 40% top-ranked picks) without gaining accuracy. Any reranker is capped by its candidates: targets never ranked in the top eight account for 65–84% of MEMAFF's lost E2E score at 960 frames and 35–59% at 120, where reranking errors also matter (Appendix Table 11). Together with the order-ablation below, this locates the main bottleneck at large memories in candidate recall.

Table 3: Frame Acc@1, one-shot Qwen3-VL-32B (262k window) vs. MEMAFF on the same tasks (counts as in Table 1). 32B prompts: 18k/23k/32k tokens at 120 frames to 147k/185k/254k at 960 (EPIC/Ego4D/PAP). Pooled over 14,367 paired task–size instances: 0.165 vs. 0.176 ($\sigma = 3.0$).

Source	System	120	240	480	960
EPIC	32B one-shot	0.177	0.130	0.094	0.061
	MEMAFF	0.184	0.143	0.095	0.062
Ego4D	32B one-shot	0.403	0.367	0.294	0.205
	MEMAFF	0.387	0.346	0.275	0.224
PAP	32B one-shot	0.452	0.275	0.224	0.137
	MEMAFF	0.452	0.349	0.260	0.217

5.4 ABLATIONS

Coarse retrieval arms. Appendix Table 8 compares five coarse configurations with the same reranker. Adding the text arm to the two SigLIP arms improves nine cells, is unchanged after rounding in one, and slightly hurts two EPIC cells; its clearest gains are on PAP and Ego4D (up to 5.2 points). On EPIC, noun-only retrieval is stronger than the three-way fusion at 120, 480, and 960 frames, significantly so at 480.

The reranker uses candidate order as a prior. When the ground-truth frame is the top coarse candidate the reranker selects it 76% of the time ($n = 162$); when it is fifth, 9% ($n = 34$; PAP, 120 frames). Shuffling the eight candidates costs 8.7 points on PAP (0.4 on EPIC; text-only coarse stage), so raising recall@8 without moving the target up gains little. Finally, reducing the reranker's 512-pixel input to 384 or 256 never helps significantly and hurts on Ego4D and EPIC (Appendix Table 9).

6 DISCUSSION AND CONCLUSION

We formulated Memory-Grounded Affordance Prediction and built MEMAFF-BENCH, 3,628 tasks from three real-image sources with memories from 120 to 960 frames. Single-pass systems hit context or accelerator limits, and removing the cap does not restore accuracy: a 32B one-shot model reaches only 0.061–0.205 frame accuracy at 960 frames. A retrieve-then-rerank pipeline degrades more smoothly at far lower token cost, yet the best strict E2E still falls from 0.144 to 0.070.

Our analyses locate the main bottleneck before the visual reranker: a 32B Qwen reranker yields no significant improvement, and at large memories most of the lost score comes from targets the coarse stage never ranks. A trained frame selector (Appendix B) raises PAP recall@8 by 7.7–14.0 points but, at 120 frames, lowers it by 19.2 points on Ego4D and 15.9 on EPIC. Progress therefore requires retrieval representations designed for functional similarity, or selector training data broad enough to cover diverse real visual memories.

Limitations. Targets are always observed (no abstention), masks are object-level rather than part-level, and ground-truth provenance differs across sources. Uniqueness is not re-checked at larger sizes; a lenient automatic check flags a possible second referent in 15% (PAP) to 28% (EPIC) of sampled 960-frame tasks. Ego4D deduplication uses MEMAFF's own frame inventories, 31 EPIC tasks (1.2%) have empty instructions, and token savings exclude memory construction (Appendix D).

ETHICS STATEMENT

MEMAFF-BENCH is derived from existing research datasets and inherits their licenses, privacy considerations, and geographic and activity biases. We will distribute only material permitted by the source licenses and document provenance and processing for each source. The benchmark is intended for evaluating embodied memory and affordance systems; deployment in safety-critical settings requires additional testing for absent targets, uncertainty, and failure recovery.

AI USE STATEMENT

Large language models were used in two distinct roles. As *components of this work*, they support benchmark construction (instruction generation, rewriting, and uniqueness checking) and are components of evaluated systems; these uses are described in the paper. As *writing and coding assistance*, the authors used generative AI to help implement experiment scripts, aggregate results, audit consistency, and edit prose. All AI-assisted outputs, experimental results, and claims were reviewed and verified by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The benchmark construction and quality-control procedures are described in Section 3; model components and retrieval fusion are described in Section 4, with implementation details and hyperparameters in Appendix D; and full per-source results and ablations are provided in the appendix. We plan to release the benchmark annotations, construction and evaluation code, and configuration files in anonymized supplementary materials.

REFERENCES

Tommaso Apicella, Alessio Xompero, and Andrea Cavallaro. Visual affordance prediction: Survey and reproducibility. *arXiv preprint arXiv:2505.05074*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. In *International Journal of Computer Vision (IJCV)*, 2022.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: VIdeo segmentations and object relations. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3D AffordanceNet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2021.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 538–547, 2019.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

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.

Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys*, 54(3):1–35, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. LOCATE: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10922–10931, 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024b. arXiv:2303.05499.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2252–2261, 2022.

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 (NeurIPS)*, 2023.

Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1374–1381. IEEE, 2015.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22, 2023.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. AffordanceLLM: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7587–7597, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.00714.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Killian Steunou, Anas Filali Razzouki, Khalil Guetari, Mounîm A. El-Yacoubi, and Yannis Tevissen. PEEK: Picking essential frames via efficient knowledge distillation. In *British Machine Vision Conference (BMVC)*, 2026.

Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-R1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*, 2025.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv preprint arXiv:2605.27365*, 2026.

Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Jianbing Shen. RAGNet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11980–11990, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. SeqAfford: Sequential 3D affordance reasoning via multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1691–1701, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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. *Transactions on Machine Learning Research*, 2024a. arXiv:2406.16852.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Zixin Zhang, Kanghao Chen, Hanqing Wang, Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Litao Guo, and Ying-Cong Chen. A4-Agent: An agentic framework for zero-shot affordance reasoning. *arXiv preprint arXiv:2512.14442*, 2025.

Zixin Zhang, Chenfei Liao, Hongfei Zhang, Harold Haodong Chen, Kanghao Chen, Zichen Wen, Litao Guo, Bin Ren, Xu Zheng, Yinchuan Li, Xuming Hu, Nicu Sebe, and Ying-Cong Chen. Panoramic affordance prediction. *arXiv preprint arXiv:2603.15558*, 2026.

A Additional Results

A.1 Per-source end-to-end scores

Table 4 reports strict E2E mIoU per source for MemAff and its SigLIP ablation. The three sources differ by a factor of four to five in absolute difficulty, which is why the pooled score of Table 2 is dominated by EPIC (2,589 of 3,628 tasks).

Table 4: Strict E2E mIoU per source, MemAff / SigLIP ablation, at the four memory sizes, with the paired t statistics of the difference.

Source	N=120	N=240	N=480	N=960	paired t (120/240/480/960)
EPIC	0.083 / 0.071	0.070 / 0.052	0.048 / 0.034	0.034 / 0.024	2.9 / 4.7 / 4.3 / 3.5
Ego4D	0.252 / 0.227	0.228 / 0.188	0.195 / 0.146	0.159 / 0.133	1.6 / 3.3 / 3.3 / 1.7
PAP	0.325 / 0.264	0.279 / 0.225	0.196 / 0.169	0.170 / 0.155	4.5 / 3.9 / 2.0 / 1.1

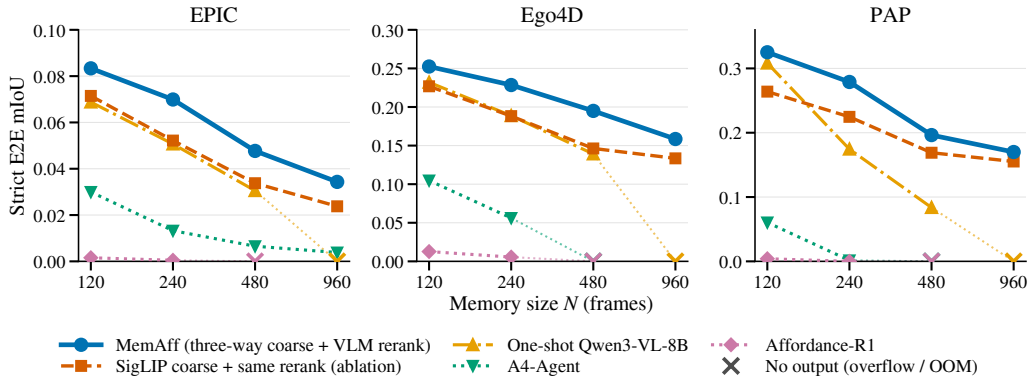

Figure 6: Strict E2E mIoU per source versus memory size for the five main-table systems (same runs as Tables 2 and 4). × marks settings with no output: one-shot overflow at 960 frames, Affordance-R1 out of memory from 480 frames, and A4-Agent on Ego4D and PAP from 480 frames, where it runs out of memory on a 48 GB GPU. MemAff leads on every source at every size; the pooled curve in Figure 4 is dominated by EPIC, the largest and hardest source.

Grounding threshold. With a Grounding-DINO box threshold of 0.25 instead of 0.35 for every system grounded by Grounding-DINO, pooled strict E2E rises by 0.1–1.0 points; MemAff still leads the one-shot model by 1.6, 3.5, and 3.7 points at 120, 240, and 480 frames and the SigLIP ablation at every size ($t = 3.8$–7.3). This check does not involve the native grounders of Affordance-R1 and A4-Agent.

A.2 Alternative pooling schemes

The main table weights every task equally, so the largest source dominates. Table 5 weights the three sources equally instead, and Table 6 restricts EPIC to a fixed random subset of 497 evaluation tasks on which all five systems were run, so that the baselines' EPIC cost stays affordable. Both keep MemAff first and preserve the ordering of the other systems; at 120 frames, the one-shot model and the SigLIP ablation are within 0.5 points under all three summaries.

A.3 Decomposition of the end-to-end score

Mask IoU on hit frames is averaged over the tasks each system retrieves correctly, a subset that changes with source and scale; it is not an oracle-frame segmentation comparison on a common image set.

Table 5: Source-balanced (macro) pooling: the mean of the three per-source scores, so that each source counts equally regardless of its size. Affordance-R1 at 480 and 960 frames is out of memory and counted as 0. Compare with Table 2.

Memory size N	120	240	480	960
MEMAFF	0.220	0.192	0.146	0.121
SigLIP ablation	0.187	0.155	0.116	0.104
One-shot Qwen3-VL-8B	0.192	0.138	0.084	0.000
A4-Agent	0.065	0.024	0.002	0.001
Affordance-R1	0.006	0.002	0.000	0.000

Table 6: Task-weighted pooling with EPIC restricted to a fixed random subset of 497 evaluation tasks, on which all five systems were run. Affordance-R1 at 480 and 960 frames is out of memory and counted as 0. Compare with Table 2.

Memory size N	120	240	480	960
MEMAFF	0.225	0.194	0.150	0.124
SigLIP ablation	0.190	0.159	0.116	0.105
One-shot Qwen3-VL-8B	0.194	0.138	0.081	0.000
A4-Agent	0.064	0.022	0.001	0.003
Affordance-R1	0.006	0.002	0.000	0.000

Table 7: Decomposition of MEMAFF's strict E2E into frame accuracy (Acc@1) and mask IoU on hit frames, per source, at the smallest and largest memory. Random frame selection gives Acc@1 $= 1/N$.

	$N = 120$			$N = 960$		
Source	Acc@1	mIoU|hit	E2E	Acc@1	mIoU|hit	E2E
PAP	0.452	0.720	0.325	0.217	0.783	0.170
Ego4D	0.387	0.652	0.252	0.224	0.707	0.159
EPIC	0.184	0.453	0.083	0.062	0.552	0.034

Table 8: Frame Acc@1 after the same VLM rerank for the five coarse configurations, all three sources, all four memory sizes. In parentheses: McNemar σ of that arm against the three-way fusion (positive = fusion better). Ego4D is the full set and PAP the 600/585-task cohort; EPIC is the full evaluation set, 2,589 tasks (2,588 at 960 frames).

Source	N	n	① instr	② noun	③ text	①⊕②	①⊕②⊕③ (ours)
PAP	N120	600	0.377 ($+3.9\sigma$)	0.352 ($+5.0\sigma$)	0.352 ($+4.4\sigma$)	0.400 ($+2.8\sigma$)	0.452
PAP	N240	585	0.309 ($+2.2\sigma$)	0.285 ($+3.4\sigma$)	0.244 ($+4.5\sigma$)	0.321 ($+1.7\sigma$)	0.349
PAP	N480	585	0.236 ($+1.3\sigma$)	0.197 ($+3.6\sigma$)	0.166 ($+4.8\sigma$)	0.253 ($+0.4\sigma$)	0.260
PAP	N960	585	0.207 ($+0.6\sigma$)	0.152 ($+4.1\sigma$)	0.137 ($+4.4\sigma$)	0.193 ($+1.8\sigma$)	0.217
Ego4D	N120	439	0.346 ($+2.0\sigma$)	0.362 ($+1.3\sigma$)	0.269 ($+5.3\sigma$)	0.367 ($+1.2\sigma$)	0.387
Ego4D	N240	433	0.279 ($+4.1\sigma$)	0.277 ($+3.8\sigma$)	0.215 ($+5.7\sigma$)	0.296 ($+3.6\sigma$)	0.346
Ego4D	N480	415	0.210 ($+3.6\sigma$)	0.219 ($+2.9\sigma$)	0.171 ($+4.8\sigma$)	0.234 ($+2.7\sigma$)	0.275
Ego4D	N960	370	0.192 ($+1.8\sigma$)	0.173 ($+3.0\sigma$)	0.127 ($+4.8\sigma$)	0.205 ($+1.3\sigma$)	0.224
EPIC	N120	2589	0.155 ($+4.2\sigma$)	0.192 (-1.1σ)	0.102 ($+10.6\sigma$)	0.187 (-0.5σ)	0.184
EPIC	N240	2589	0.105 ($+5.7\sigma$)	0.138 ($+0.9\sigma$)	0.057 ($+12.5\sigma$)	0.133 ($+1.7\sigma$)	0.143
EPIC	N480	2589	0.069 ($+5.0\sigma$)	0.108 (-2.3σ)	0.032 ($+10.6\sigma$)	0.095 ($+0.1\sigma$)	0.095
EPIC	N960	2588	0.046 ($+3.7\sigma$)	0.070 (-1.7σ)	0.018 ($+9.3\sigma$)	0.064 (-0.4σ)	0.062

A.4 COARSE-STAGE ABLATION, FULL TABLE

B TRAINED FRAME SELECTOR

Following distilled keyframe selectors (Steunou et al., 2026), we trained a small instruction-conditioned scorer on 3,000 composed 120-frame PAP episodes. In-domain it is effective: PAP recall@8 rises from 0.652 to 0.780 at 120 frames and from 0.385 to 0.462 at 960. Across sources it fails: at 120 frames recall@8 drops by 16–19 points on Ego4D and EPIC, and neither a larger scorer

Table 9: Frame Acc@1 when the reranker sees candidates at 256, 384, or 512 pixels (long side); coarse ranking and candidate set are identical. In parentheses: McNemar σ against 512 px (positive = 512 px better).

Source	N	n	256px	384px	512px(main)
PAP	N120	600	0.430 ($+1.4\sigma$)	0.443 ($+0.6\sigma$)	0.452
PAP	N240	585	0.344 ($+0.4\sigma$)	0.338 ($+1.0\sigma$)	0.349
PAP	N480	585	0.251 ($+0.9\sigma$)	0.265 (-0.6σ)	0.260
PAP	N960	585	0.210 ($+0.7\sigma$)	0.229 (-1.3σ)	0.217
Ego4D	N120	439	0.360 ($+1.6\sigma$)	0.362 ($+1.8\sigma$)	0.387
Ego4D	N240	433	0.296 ($+3.2\sigma$)	0.323 ($+1.9\sigma$)	0.346
Ego4D	N480	415	0.246 ($+1.9\sigma$)	0.255 ($+1.6\sigma$)	0.275
Ego4D	N960	370	0.216 ($+0.6\sigma$)	0.205 ($+1.5\sigma$)	0.224
EPIC	N120	2589	0.166 ($+3.1\sigma$)	0.181 ($+0.6\sigma$)	0.184
EPIC	N240	2589	0.136 ($+1.4\sigma$)	0.137 ($+1.3\sigma$)	0.143
EPIC	N480	2589	0.088 ($+1.8\sigma$)	0.087 ($+2.2\sigma$)	0.095
EPIC	N960	2588	0.063 (-0.2σ)	0.063 (-0.1σ)	0.062

Table 10: Frame Acc@1 when only the reranking model is changed from Qwen3-VL-8B to Qwen3-VL-32B; the coarse ranking and the eight candidates are identical. In parentheses: McNemar σ (negative = the 32B reranker is worse). EPIC is the 497-task subset.

Source	N=120	N=240	N=480	N=960
EPIC	0.197→0.175 (-1.3σ)	0.149→0.149 (0.0σ)	0.101→0.085 (-1.5σ)	0.078→0.066 (-1.3σ)
Ego4D	0.387→0.351 (-2.0σ)	0.346→0.309 (-2.1σ)	0.275→0.241 (-2.2σ)	0.224→0.189 (-2.3σ)
PAP	0.452→0.402 (-2.8σ)	0.349→0.337 (-0.8σ)	0.260→0.267 ($+0.5\sigma$)	0.217→0.219 ($+0.1\sigma$)

nor adding Ego4D videos to training rescues EPIC. Whether broader real-video training would transfer remains open, so the training-free pipeline remains our reference system.

The selector is a low-rank residual bilinear scorer over frozen SigLIP embeddings (2.6×10^5 parameters) whose output equals the frozen cosine similarity at initialization, so training starts from the frozen ranking. The episodes are rendered from the *training* split of PAP-12K (9,020 panoramas, disjoint from the 101 panoramas used by the benchmark); the 2,700 training episodes give 16,091 tasks; the loss is a listwise softmax with the target frame carrying 0.8 of the mass and the other views of the same room sharing 0.2; early stopping uses the 300 held-out episodes, which share panoramas with the training episodes (only the benchmark panoramas are disjoint). Table 12 reports coarse recall@8 of the single instruction–image retriever before and after training; for comparison, the three-way fusion used by MEMAFF reaches $0.763/0.638/0.525/0.456$ on PAP, $0.629/0.531/0.434/0.362$ on Ego4D, and $0.455/0.348/0.251/0.184$ on EPIC.

B.1 RERANK RESOLUTION

B.2 REPLACING THE RERANKER WITH A LARGER MODEL

B.3 WHERE THE END-TO-END SCORE IS LOST

Plugged into the full pipeline on PAP, the trained selector raises frame accuracy from $.452/.349/.260/.217$ to $.475/.395/.320/.248$ (paired McNemar $\sigma = 1.3/2.5/3.6/2.0$). A larger variant with a two-layer transformer over the episode behaves identically in and out of domain. Adding the Ego4D benchmark videos to the training set (split by video, early stopping on held-out videos, EPIC untouched) does not repair the transfer: EPIC recall@8 at 120 frames goes from $.382$ frozen to $.254$, better than the PAP-only selector's $.223$ but still far below frozen; oversampling the real-video episodes to match the composed ones makes it worse ($.168$), with the training videos memorized (recall@1 $.81$). Because this variant trains on benchmark-source video, it is a diagnostic rather than a training-free result, and it does not isolate data coverage from optimization or architecture as the cause of poor transfer.

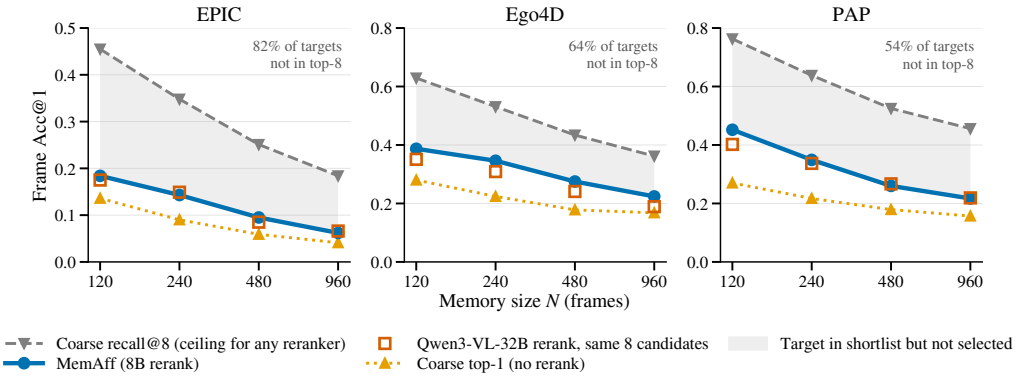

Figure 7: Where frame accuracy is lost, per source: three-way coarse recall@8 (the ceiling for any reranker over eight candidates), MEMAFF Acc@1 with the 8B reranker, the Qwen3-VL-32B reranker on the same eight candidates (EPIC on the 497-task subset; Table 10), and coarse top-1 without reranking. The shaded band is the accuracy lost inside the shortlist. At 960 frames, 54–82% of targets are outside the top eight, so candidate recall bounds every reranker; the 32B reranker never exceeds the 8B one by more than noise.

Table 11: Each task of MEMAFF (8B reranker) is assigned to the first stage at which it fails: the target frame is not among the eight coarse candidates; it is a candidate but not selected; the frame is right but no box is detected; or the box is found but the mask has IoU < 0.5. Entries are each stage's share of the lost strict E2E score, $1 -$ E2E, summing per-task shortfalls; rows need not sum to 100%, and the remainder (up to 4.7%) is imperfect masks with IoU ≥ 0.5. These are loss shares, not failure frequencies.

Source	N	Strict E2E	Not in top 8	Wrong candidate	No box	Poor mask
EPIC	120	0.083	59%	30%	2%	9%
	960	0.034	84%	13%	0%	2%
Ego4D	120	0.252	50%	32%	4%	13%
	960	0.159	76%	16%	1%	6%
PAP	120	0.325	35%	46%	4%	10%
	960	0.170	65%	29%	0%	3%

Table 12: Coarse recall@8 of the frozen retriever and of the trained selector. PAP is in-domain (different panoramas, same renderer and instruction style); Ego4D and EPIC are out-of-domain.

Source	N=120	N=240	N=480	N=960
PAP (in-domain)	0.652→**0.780**	0.561→**0.638**	0.451→**0.591**	0.385→**0.462**
Ego4D (transfer)	0.513→0.321	0.423→0.240	0.330→0.171	0.297→0.149
EPIC (transfer)	0.382→0.223	0.271→0.144	0.188→0.093	0.126→0.059

C BENCHMARK SAMPLES AND ADDITIONAL QUALITATIVE RESULTS

Figure 8 shows tasks from the human-audited subset of each source, and Figure 9 extends the qualitative comparison of Figure 5 with two further MEMAFF failures and three large-memory settings.

D PROTOCOL DETAILS

Known annotation issues. The evaluation set contains 31 EPIC tasks with empty instructions (1.2% of its 120-frame set); they remain in the reported denominators and are effectively unanswerable. An automatic, deliberately lenient VLM check of 120 sampled tasks per source at 960 frames flags a possible second referent in 18/120 (15%) PAP and 33/120 (28%) EPIC tasks; these

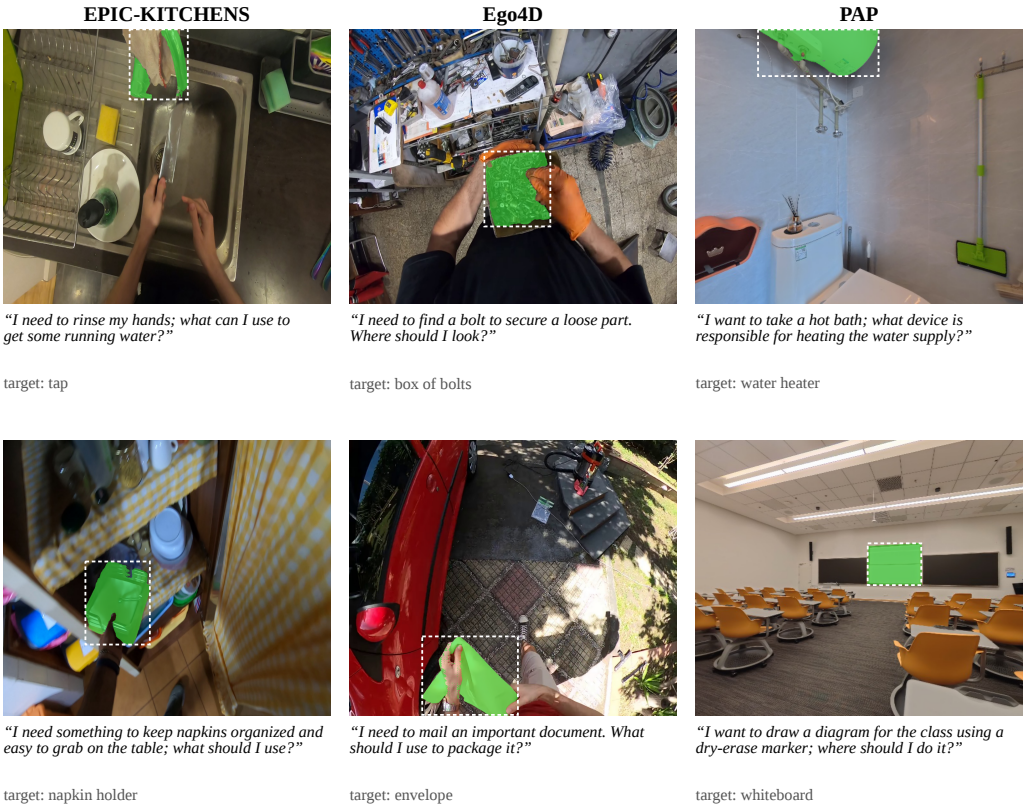

Figure 8: Sample tasks from the human-audited subset, two per source: the annotated target frame with its object mask (green; dashed box marks the mask extent), the intent-style instruction, and the target object. Instructions name a goal rather than the object. EPIC and Ego4D frames are egocentric video frames; PAP frames are perspective renders of indoor panoramas.

rates are upper bounds rather than human judgments. Ego4D deduplication (Section 3.2) removes frames using the same Qwen3-VL-8B inventories that MEMAFF's text arm retrieves over, which may favor that arm on Ego4D. Shared SAM-family refinement of EPIC/Ego4D annotations and of all compared systems' predictions may also favor similar mask boundaries.

Counting and pairing. The scale totals $3{,}628 + 3{,}607 + 3{,}589 + 3{,}543 = 14{,}367$ are task–size instances, not distinct tasks. Systems are paired within each source–size cell; across sizes, queries recur and some disappear. The released per-task predictions support a matched analysis on the task intersection across sizes.

Implementation details. All VLM stages use Qwen3-VL-8B-Instruct served with vLLM (131,072-token context, temperature 0, at most 200 output tokens); retrieval uses google/siglip-large-patch16-384 and all-MiniLM-L6-v2; fusion uses $k = 10$ over each arm's top 30 and keeps $K = 8$; reranking images have a 512-pixel long side, and an unparsable answer defaults to slot 0. Grounding keeps the highest-scoring Grounding-DINO-tiny box above 0.35 and refines it with SAM 2.1-hiera-large. One-shot and MEMAFF are served with vLLM (paged attention); Affordance-R1 and A4-Agent run through their authors' transformers-based code. A4-Agent's thinker holds the entire memory in one non-paged prefill (93k–124k tokens at 480 frames with 512-pixel frames), sharing the GPU with its Rex-Omni and SAM 2 spotter, which is why it runs out of memory on the 48 GB GPUs used for Ego4D and PAP; on an 80 GB GPU it ingests 480 frames but still runs out of memory at 960. Affordance-R1's Qwen2.5-VL vision encoder (transformers 4.51) materializes a dense attention mask over all patches of a request, which grows quadratically with the number of frames (38 GB at 240 frames and 153 GB at 480 with 336-pixel frames), so it runs out of memory at

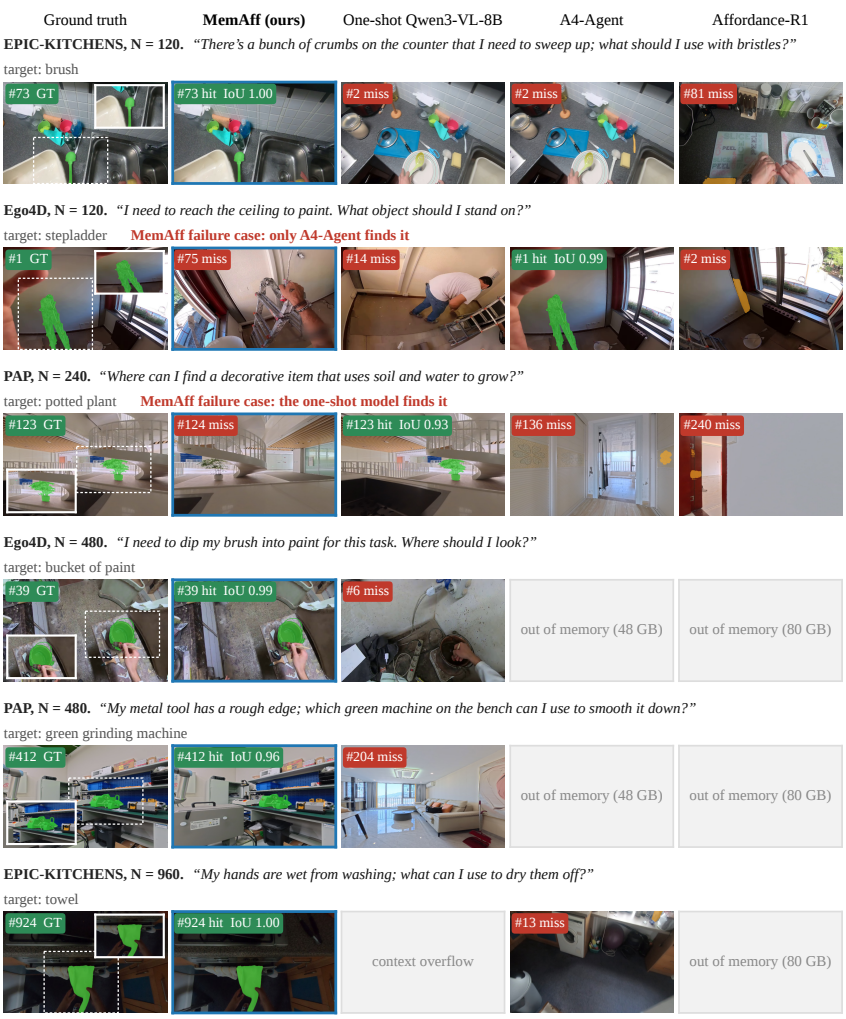

Figure 9: Additional qualitative examples (layout as in Figure 5), including memories of 480 and 960 frames. Rows 2 and 3 are MEMAFF failures: only A4-Agent finds the stepladder, and the one-shot model finds the potted plant that MEMAFF's reranker passes over. At 480 and 960 frames the one-shot model selects a wrong frame or overflows its window, A4-Agent runs out of memory on Ego4D and PAP, and Affordance-R1 is out of memory, while MEMAFF still recovers the target in the three examples shown. Failure labels as in Table 2: overflow = the 131k-token window; out of memory = GPU memory, on 80 GB for Affordance-R1 and 48 GB for A4-Agent (Appendix D).

240 frames on 48 GB and at 480 frames on any single GPU. Runs were distributed over 48 GB (A40) and 80 GB (A800) GPUs; this does not affect MEMAFF (eight images per reranking request) or the vLLM-served one-shot model (bounded by its 131k-token window), and every out-of-memory cell is reported with the card size on which it occurred. The selector is trained with AdamW (learning rate 3×10^{-4}, weight decay 0.01) for 20 epochs, selecting the checkpoint by validation recall@8, with seed 0. The fixed EPIC subset is drawn with a seeded random sample. We will release task manifests, per-scale frame lists, verbatim prompts, per-task predictions, and selector splits with the code.