REFERCLAW: DISAMBIGUATING USER INTENT FOR EGOCENTRIC ASSISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Users of egocentric AI assistants often ask short, ambiguous questions whose intended referents cannot be uniquely resolved from video and dialogue history. For example, "Where did I leave it?" may refer to several previously handled objects. Existing VideoQA systems typically commit to a single interpretation, risking plausible but intention-misaligned answers. We introduce ReferClaw, an agentic framework that formulates egocentric streaming QA as an interactive process: the assistant asks clarifying questions to resolve user intent before answering. Three cooperating agents identify candidate interpretations, request clarification, and generate answers grounded in the resolved intent and visual evidence. A bounded, cost-aware policy balances expected information gain against user effort while limiting clarification turns and question complexity. ReferClaw learns from interaction through intent-specific skill updates and process-reward-guided policy optimization with agent-level credit assignment. Evaluations on augmented egocentric QA benchmarks demonstrate improved answer accuracy and joint answer-and-grounding performance over existing baselines, alongside zero-shot transfer across datasets. These findings support interactive clarification as a practical approach to resolving online user intent in egocentric streaming assistance.

1 INTRODUCTION

Egocentric cameras paired with multimodal AI assistants are becoming natural interfaces to the physical world, enabling continuous visual memory, navigation, and daily assistance. Recent benchmarks have advanced episodic memory (Bärmann & Waibel, 2022; Yang et al., 2025b), long-range understanding (Mangalam et al., 2023), and situated assistance (Xiao et al., 2025; Sun et al., 2025; Qin et al., 2026). Their standard QA protocols, however, evaluate answers against fixed questions and observation histories, rather than how an assistant acquires missing intent information. In practice, user requests are brief and ego-deictic, and the available context may support several interpretations. For "`Did she put it back?`" (Fig. 1), different persons may perform the same action at different times. Even when both encounters are visible, the intended one may remain unclear. An answer can therefore be compatible with the video yet irrelevant to what the user meant.

Despite this progress, the challenge remains largely underexplored. Existing approaches (Peng et al., 2025; Choi et al., 2026; Qin et al., 2026) exploit eye gaze and hand gesture as visual or textual prompts for disambiguating user intent. While effective, the gains remain limited for the complexity of real-world interactions. Recent VLMs instead explore proactive request for clarification (Jian et al., 2025; Cao et al., 2026; Yang et al., 2026), showing that models can ask rather than guess based on the utterance and a single observation. This may suffice for a static image, but may not for dynamic video stream, where different encounters can support different interpretations of the same question (Fig. 1). We therefore study how explicit spatio-temporal interpretations can guide what to ask, how to interpret the response, and which evidence to use for answering.

We formulate intent disambiguation as bounded information acquisition over grounded candidate interpretations. We propose ReferClaw which coordinates three agents around these candidates. The Intention Grounding Agent (IGA) constructs the candidate intents by grounding the question against the observed video, pairing each candidate with its supporting visual evidence, and scores the candidates by evidential support; the Cost-Aware Clarification Agent (CCA) generates and chooses the best clarification question by balancing expected information gain against respondent effort; and

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

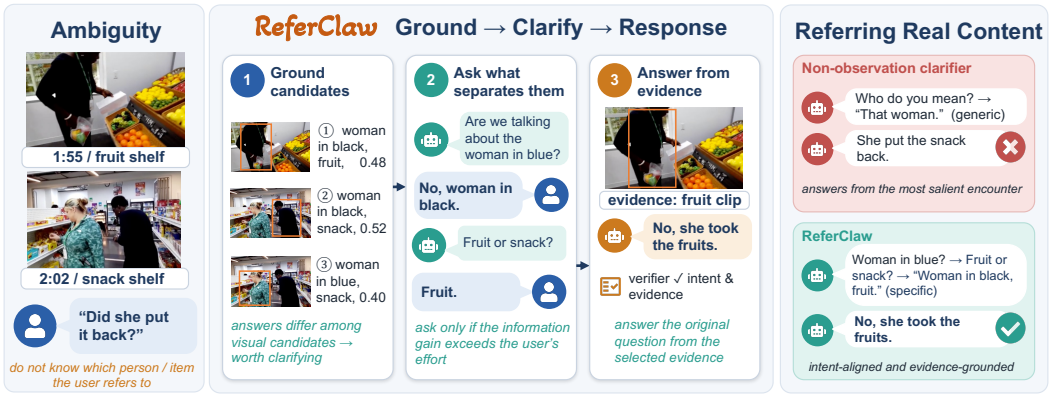

Figure 1: Different encounters can support different interpretations of the same question. ReferClaw asks for clarification to identify the intended encounter and select evidence for answering.

the Evidence-Fused Answering Agent (EFA) answers the original question from the resolved intent and its evidence. Their shared state allows clarification responses to refine the candidates or trigger re-grounding. A deterministic controller bounds clarification turns and question complexity, and decides when to ask or answer. ReferClaw also learns from past interactions to reduce recurring disambiguation errors. Following MetaClaw's two-timescale design (Xia et al., 2026), verified feedback produces intent-specific skills that take effect immediately and process rewards that guide updates to the responsible agent. The skills encode which contrasts distinguish competing interpretations, while agent-level updates target failures in grounding, clarification, or answering. In our evaluation protocol, the annotated target intent, supporting evidence, and answer correctness become available only after each prediction, so updates affect subsequent interactions.

To facilitate the study, we augment two recent egocentric streaming QA benchmarks BuddyVQA (Qin et al., 2026) and EgoBlind (Xiao et al., 2025) with ambiguity labels, target intents, and spatio-temporal visual evidence. Experiments separate the benefits of clarification from learning, test ambiguity across repeated encounters, and assess transfer to independent image-based clarification benchmarks ClearVQA (Jian et al., 2025) and ContextClarify (Cao et al., 2026). We additionally evaluate well-specified questions and human respondents to examine unnecessary interruptions and the applicability of simulated interactions. Our contributions are threefold:

- We formulate intent disambiguation in egocentric QA assistance as bounded information acquisition over spatio-temporally grounded interpretations.
- We develop ReferClaw, which coordinates grounding, clarification, and answering agents through shared candidates, and learns from feedback through intent-specific skills and agent-level updates.
- We augment two ego-streaming QA benchmarks with intent and grounding annotations, and evaluate clarification, continual adaptation, and transfer under controlled interaction protocols.

2 Related Work

Clarification and User Intent Disambiguation. Clarification establishes common ground by acquiring missing information and resolving alternative interpretations (Clark & Brennan, 1991; Saeidi et al., 2018; Xu et al., 2019; Rao & Daumé III, 2018; Min et al., 2020; Lee et al., 2023). Cooperative visual dialog agents (Das et al., 2017) and GuessWhat (de Vries et al., 2017) extend goal-directed question asking to images, and later methods select questions through uncertainty (Abbasnejad et al., 2019) or expected information gain (White et al., 2021). Recent work localizes plausible image regions for ambiguous questions (Chen et al., 2025a), teaches VLMs to request clarification on ClearVQA (Jian et al., 2025), and learns ask-or-answer policies on ContextClarify (Cao et al., 2026). Plug-and-Play Clarifier (Yang et al., 2026) addresses linguistic ambiguity, capture quality, and pointing gestures in egocentric interaction. Existing clarification benchmarks are largely image-based, motivating our augmentation of BuddyVQA and EgoBlind with intent annotations for video-based disambiguation (Sec. 4.1). Methodologically, prior works clarify from the utterance and a single observation; ReferClaw instead focuses on competing interpretations grounded

across video history, using shared candidates to connect clarification with evidence selection under bounded respondent effort.

Egocentric and Grounded Video QA. Ego4D (Grauman et al., 2022) evaluates episodic memory, while EgoSchema (Mangalam et al., 2023) tests long-range egocentric video understanding. EgoBlind (Xiao et al., 2025) emphasizes situated assistance, and BuddyVQA (Qin et al., 2026) evaluates ego-deictic and history-dependent questions over a causal stream. GroundVQA (Di & Xie, 2024) and NExT-GQA (Xiao et al., 2024) couple answering with temporal localization, GeLM (Chen et al., 2025b) develops grounded multi-hop reasoning, D^2VLM (Zeng et al., 2025) factorizes grounding and answering, and recent systems expose explicit temporal or spatio-temporal evidence for verifiable answers (Meng et al., 2026; Seo et al., 2026; Huang et al., 2026). Related systems support interactive video questioning (Wang et al., 2025a), pixel-level conversational grounding (Munasinghe et al., 2025), and streaming evidence retrieval (Di et al., 2025). These methods localize evidence under a single interpretation of the query, so stronger retrieval may confidently ground the wrong referent when several intentions remain plausible. ReferClaw represents the alternatives explicitly and acquires clarifications on which interpretation to guide evidence selection and answering.

Multi-Agent Collaboration. Multi-agent systems coordinate specialized roles through structured communication (Hong et al., 2024; Chen et al., 2024b; Liu et al., 2026; Dang et al., 2025). Interaction-based adaptation uses self-feedback, verbal reflection, and reusable skills or routines (Madaan et al., 2023; Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2024; 2025c; Tu et al., 2026). MetaClaw combines fast skill adaptation with slow policy optimization (Xia et al., 2026), while process supervision rewards intermediate steps (Lightman et al., 2024; Zhang et al., 2025b). ReferClaw adapts these mechanisms to intent disambiguation: skills encode ambiguity triggers and decisive contrasts, while process rewards guide updates to the grounding, clarification, and answering agents.

3 METHOD

Clarifying an ambiguous question requires three decisions that ordinary VideoQA does not: whether the question admits more than one visually supported interpretation, what to ask so that the user can resolve it at little effort, and how to answer from the interpretation the user selects. ReferClaw assigns each decision to a dedicated agent operating on shared grounded candidate interpretations, coordinated by a fixed controller that bounds the user's burden (Fig. 2), and learns from verified interaction feedback so that disambiguation mistakes are not repeated.

3.1 PROBLEM SETUP

For a question q asked at video time s_q, the assistant observes the causal prefix $V_{\le s_q} = \{v_s\}_{s=1}^{s_q}$ and episode dialogue D. The prefix stays fixed during clarification. At turn t, $h_t = \{(c_\ell, y_\ell)\}_{\ell=0}^{t-1}$ records clarification questions c and responses y. Intent is ambiguous when multiple visually supported interpretations remain plausible given $(V_{\le s_q}, q, D, h_t)$.

A candidate $z_i = (\widetilde{q}_i, e_i)$ pairs a question rewrite $\widetilde{q}_i$ with evidence $e_i = (\tau_i, r_i)$: an interval $\tau_i \subseteq [1, s_q]$ and referent box track r_i (empty for non-localizable information). For "Is it still on?" in Fig. 2, rewrites may specify the "stove" used at time 1:05 or the "faucet" at 2:45. The assistant maintains $\mathcal{Z}_t = \{z_i\}_{i=1}^{n_t}$, $n_t \le N_{\max}$, and belief p_t over these candidates plus an uncovered interpretation $z_\varnothing$. It asks c_t or terminates with an ANSWER $(a, \widetilde{q}, e, \gamma)$, where γ is the belief mass supporting the answer, a CONDITIONAL output with at most two interpretations and answers, or ABSTAIN.

For policy π and respondent type $x \in \{\mathrm{H}, \mathrm{L}\}$ (human or LLM), we seek

$$\max_{\pi} \ \mathbb{E}\big[U_{\text{ans}} + \lambda_g\, U_{\text{grd}} - \textstyle\sum_{t=0}^{N_{\text{ask}}-1} C_x(c_t)\big] \quad \text{s.t.} \quad N_{\text{ask}} \le T_{\max},\ |c_t| \le L_{\max},\ m(c_t) \le M_{\max}, \tag{1}$$

Here $U_{\text{ans}}(a, z^\star)$ is binary judged correctness for target intent $z^\star$, and $U_{\text{grd}}(e, e^\star) = \min(\text{tIoU}, \text{sIoU})$ measures overlap with target evidence $e^\star$, using temporal IoU alone when no target box exists. The weight λ_g balances grounding; $C_x(c_t) \in [0, 1]$ measures respondent cost. CONDITIONAL and ABSTAIN receive zero utility. The limits bound the number of questions N_{ask}, question length $|c_t|$ in tokens, and displayed options $m(c_t)$, preventing endless clarification and one-shot "super-questions" that overwhelm the user with many alternatives in a single turn.

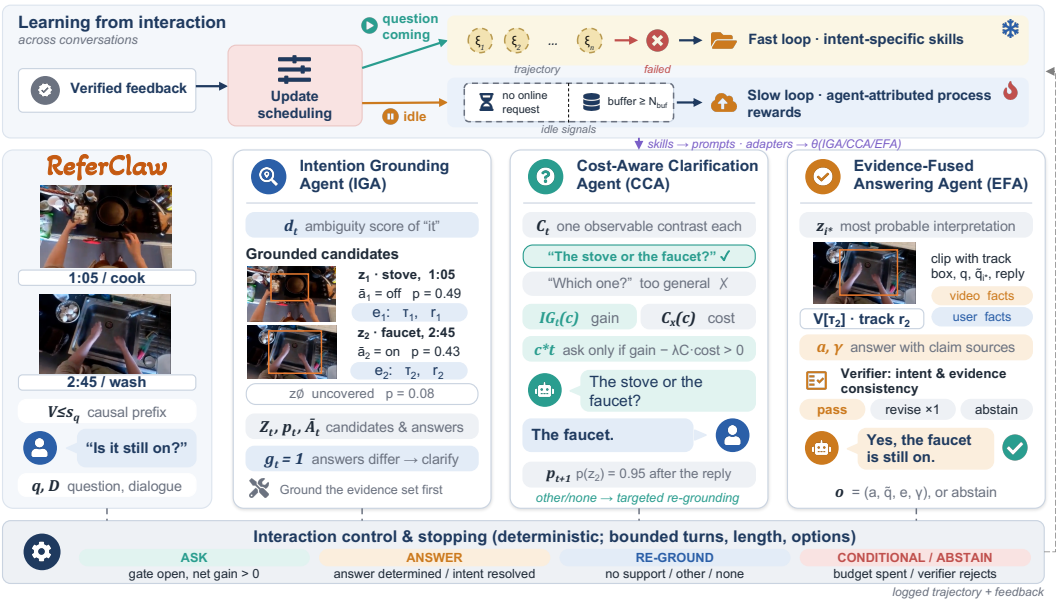

Figure 2: Overview of ReferClaw. The system coordinates three agents (IGA, CCA, EFA) to respectively ground candidate interpretations, requests clarification when useful, and verifies answers against the resolved intent and evidence. A fixed controller bounds interaction. *Top:* verified feedback updates prompt-level skills immediately and agent-specific LoRA adapters during idle periods.

3.2 Cooperating Agents and Interaction Controller

The three agents share $(\mathcal{Z}_t, p_t)$. Each uses a role-specific prompt and LoRA adapter on one frozen multimodal backbone; a fixed controller sequences their actions and enforces the interaction budget.

Intention Grounding Agent. IGA predicts an ambiguity score $d_t \in [0,1]$ and responsible span, then grounds before rewriting. Temporal retrieval, detection, and tracking produce up to eight evidence cards containing an interval–track pair, grounding score ρ_i, and visible-content description (Appendix B.1). Each card yields a rewrite $\widetilde{q}_i$, support score $s_i \in [0,1]$, and provisional answer $\bar{a}_i$. After merging duplicate entity–event cards and retaining at most $N_{\max}$ candidates, IGA initializes

$$p_t(z_i) = (1-\epsilon_t)\,\frac{\exp(s_i/\kappa)}{\sum_{j=1}^{n_t}\exp(s_j/\kappa)}, \qquad p_t(z_\varnothing) = \epsilon_t = \epsilon_0\left(1-\max_i \rho_i\right), \tag{2}$$

where κ is a validation-fitted temperature and ϵ_t reserves uncovered-intent mass when grounding support is weak. Clarification requires high ambiguity, no dominant candidate, and disagreement among provisional answers:

$$g_t = \mathbb{I}[d_t \geq \tau_d]\;\mathbb{I}[1-\max_i p_t(z_i) \geq \tau_u]\;\mathbb{I}[\mathrm{Dis}(\overline{\mathcal{A}}_t) \geq \tau_a], \tag{3}$$

Here Dis is the belief-weighted fraction of candidate pairs with different provisional answers $\overline{\mathcal{A}}_t = \{\bar{a}_i\}$. When all plausible candidates imply the same answer, clarification is unnecessary even if the intended referent remains uncertain.

Cost-Aware Clarification Agent. CCA proposes at most six questions $\mathcal{C}_t$, each contrasting candidates by identity, occurrence, location, or action. Each declares an option-based partition with an "other/none" option for $z_\varnothing$. The response model $p_x(y \mid c, z_i, h_t)$ predicts the respondent's answer class: a partition cell, other/none, uncertain, or invalid/missing. For LLM respondents, smoothed frequencies from eight backbone samples estimate this likelihood, conditioning on candidate z_i rather than on the annotated target intent, which only the respondent supplying actual responses (simulator or human) receives.

An actual response y_t updates the belief by Bayes' rule. Hypothetical responses define expected information gain (IG), measured as the reduction in Shannon entropy H (bits):

$$p_{t+1}(z_i) \propto p_t(z_i)\,p_x(y_t \mid c_t, z_i, h_t), \quad \mathrm{IG}_t(c) = H(p_t) - \textstyle\sum_y p_x(y \mid c, h_t)\,H\big(p_{t+1}(\cdot \mid c, y)\big), \tag{4}$$

where $p_x(y \mid c, h_t) = \sum_i p_t(z_i) p_x(y \mid c, z_i, h_t)$. IG is a one-step surrogate for answer utility, used only when the gate detects answer disagreement. CCA balances it against normalized effort $\overline{E}_x$, delay $\overline{T}_x$, and failure risk $\overline{F}_x$:

$$C_x(c) = w_E\,\overline{E}_x(c) + w_T\,\overline{T}_x(c) + w_F\,\overline{F}_x(c), \qquad c_t^\star = \arg\max_{c \in \mathcal{C}_t} u_t(c), \tag{5}$$

where $u_t(c) = \mathrm{IG}_t(c) - \lambda_C\, C_x(c)$ is the net utility of a question. CCA returns $c_t^\star$ together with $u_t(c_t^\star)$, which the controller uses to decide whether to ask, or *no-question* when the pool is empty or $u_t(c_t^\star) \le 0$. Human response likelihoods and costs are calibrated on the user study using question length, choice complexity (Hick, 1952; Hyman, 1953), response time, and failure rate. LLM costs use token counts, serving latency, and predicted invalid-format mass (Appendices B.2 and B.3).

Evidence-Fused Answering Agent. EFA selects $i^\star = \arg\max_i p_t(z_i)$ and generates $a = F_\omega(V[\tau_{i^\star}], r_{i^\star}, q, \widetilde{q}_{i^\star}, h_t)$ with confidence $\gamma = p_t(z_{i^\star})$. Its prompt separates observations from user responses and labels claims as `video` or `user`: clarification identifies the appliance, not whether it is on. A separate verifier prompt checks alignment with the rewrite and support for each claim in its corresponding source: `video` claims must be visible in the evidence card, and `user` claims must appear in h_t. It permits one revision before passing or abstaining, without changing the interpretation or retrieving evidence (Appendices F.6 and F.7).

Interaction Controller. In priority order, the controller abstains if support remains inadequate after one re-grounding; invokes EFA if $g_t = 0$ or $\max_i p_t(z_i) \ge \tau_p$; and asks if $g_t = 1$, $t < T_{\max}$, and net utility is positive. Otherwise, it returns CONDITIONAL when the top two candidates disagree and their belief margin is below τ_g, or invokes EFA. Responses update Eq. 4; other/none or an unsupported constraint triggers one constrained re-grounding. Uncertain or invalid responses leave belief unchanged and ban the asked partition (Appendix B.4).

3.3 LEARNING FROM INTERACTION

Following MetaClaw (Xia et al., 2026), our learning maintains a versioned skill library $\mathcal{S}_g$ and agent adapters θ_g indexed by skill generation g. Under our prequential protocol (Sec. 4.1), $z^\star$, $e^\star$, and the judge's verdict are revealed only *after* a prediction is logged. A trajectory whose answer is wrong, whose evidence misses $e^\star$, or that asked unnecessarily is a *verified failure*. Verified failures are attributed by type: coverage and grounding failures to IGA, non-separating or unnecessary questions to CCA, and inconsistent answers to EFA. A process-review prompt handles residual cases (91% agreement with manual labels; Appendix B.5). Verified trajectories feed two loops, scheduled by whether the assistant is currently serving requests (Fig. 2, top).

Fast Loop: Intent-specific Skills. Verified failures form support set $\mathcal{D}_g^{\mathrm{sup}}$, from which an evolver updates $\mathcal{S}_{g+1} = \mathcal{S}_g \cup E_{\mathrm{skill}}(\mathcal{D}_g^{\mathrm{sup}}, \mathcal{S}_g)$. Each agent-specific skill records an ambiguity trigger, candidate topology (*e.g.*, one object encountered twice), decisive contrast, corrected action, anti-pattern, and two regression tests against unnecessary asking. Up to three skills, retrieved by trigger–topology embedding similarity, immediately enter the responsible agent's prompt. For example, a CCA skill contrasts repeated encounters through the user's visible action rather than their order; Appendix B.6 provides the worked example.

Slow Loop: Agent-attributed Process Rewards. During idle periods, a query buffer $\mathcal{D}^{\mathrm{qry}}$ of 128 verified trajectories tagged with their skill generation, including successes and failures, triggers adapter updates. Each trajectory ξ receives

$$R_{\mathrm{proc}}(\xi) = r_{\mathrm{final}} + \alpha\, r_{\mathrm{ground}} + \beta \sum_t \Delta_t^\star - \lambda \sum_t C_x(c_t) - \mu\, r_{\mathrm{invalid}}, \tag{6}$$

where r_{final} combines correctness and verifier verdict, r_{ground} combines U_{grd} and target coverage, C_x is realized respondent cost, and r_{invalid} counts rule violations. Target progress $\Delta_t^\star = \log p_{t+1}(z^\star) - \log p_t(z^\star)$ rewards belief moving toward the intended interpretation. A fixed candidate scope prevents reward from candidate removal alone (Appendix B.7).

For each failed step, the attributed agent samples $G = 8$ alternatives from its logged prompt, including its original skills. The evaluation simulator completes them into trajectories; a reward model predicts R_{proc} (held-out Spearman 0.83). GRPO (Shao et al., 2024) updates only that agent's LoRA

adapter, with a KL penalty toward frozen Base. Promotion requires improvement on the target failure type without degrading overall accuracy, unnecessary asking, or mean cost on a fixed validation probe; otherwise the update is rolled back (Appendix B.7).

4 EXPERIMENTS

Our experiments answer four questions. **Q1:** Does targeted clarification improve intent-aligned and grounded answering over direct and interactive baselines (Sec. 4.2)? **Q2:** Does grounding the alternatives across time help most when the competing referents are separated in time (same object seen twice / similar objects seen at different moments) rather than visible at once (Sec. 4.4)? **Q3:** Are the gains obtained with limited user effort (Sec. 4.4)? **Q4:** Does learning from interaction reduce future errors beyond simpler adaptation (Secs. 4.2 and 4.4)?

4.1 EXPERIMENTAL SETUP

Benchmarks. To facilitate our study, we augment two closest-aim egocentric QA datasets with intent-ambiguity labels, minimal supporting temporal intervals and spatial boxes. BuddyVQA (Qin et al., 2026), our primary benchmark, with its causal streaming protocol, ego-deictic expressions, and history-dependent question chains; and EgoBlind-Intent (Xiao et al., 2025), with timestamped assistance questions from blind and visually impaired users, used for zero-shot transfer. We also test transfer on two independent image-based clarification benchmarks, ClearVQA (Jian et al., 2025) and ContextClarify (Cao et al., 2026), treating the image as a one-frame prefix with their official metrics and respondents. More benchmark augmentation details are in Appendix C.

Protocol and Metrics. A model may access only frames observed before the question timestamp and the preceding dialogue; gold labels, resolved questions, intervals, and boxes are never inputs at prediction time; only the latter three and the judge verdict are revealed to learning methods, after a prediction is logged. Answer accuracy uses each benchmark's multi-reference LLM-as-a-judge protocol over all test inputs, with CONDITIONAL and ABSTAIN counted as incorrect; temporal metrics are computed on grounding-eligible questions, spatial metrics and Acc@GQA (Xiao et al., 2024) (we further require correct answer with both T-IoU and S-IoU ≥ 0.5) on the box-annotated subset. Interactive evaluation additionally reports ambiguity-detection F1, ask precision/recall, clarification accuracy, unnecessary-ask and abstention rates, turns, question tokens, response time, latency, respondent cost, and the utility of Eq. 1. All numbers are means over three fixed stream orders, and all subset values of a row come from the same runs and checkpoint; the full protocol, the population of every metric, and confidence intervals are in Appendices C.4 and E.3.

Baselines and Implementation. We compare against direct MLLMs (Qwen Team, 2026), grounded and streaming VideoQA systems (Meng et al., 2026; Huang et al., 2026; Di et al., 2025; Wang et al., 2025a; Qin et al., 2026), and interactive clarification methods (Jian et al., 2025; Yang et al., 2026; Cao et al., 2026). Framework-level baselines and all ReferClaw variants use Qwen3.5-9B, under the same input protocol. To avoid conflating clarification, extra information, and online learning, we evaluate in two regimes. *Frozen inference* compares fixed systems, including **ReferClaw (Base)** and its asking-disabled control, which retains grounding, rewriting, and verification. *Learning from interaction* follows MetaClaw's online evaluation (Xia et al., 2026): feedback follows each logged prediction and updates affect only later inputs. We compare **ReferClaw (Skills)** (fast skill adaptation), **ReferClaw (Full)** (adding agent-attributed GRPO, Eq. 6), *Reflexion-style memory* (Shinn et al., 2023), and *MetaClaw-style* adaptation (generic skills, all-adapter updates, final-answer reward). All share the initial system, feedback, stream, and update budget. Details are in Appendix D.

4.2 MAIN RESULTS

Frozen Inference (Q1): Tab. 1 compares answer accuracy under the streaming protocol. Frozen ReferClaw reaches 66.2%, versus 58.5% for Clarify-or-Answer. Disabling asking while retaining grounding and verification lowers accuracy to 60.5%, giving a 5.7-point matched clarification gain. Improvements also appear on deictic and history-dependent questions: Base reaches 69.0% and 66.8%, respectively, versus 58.3% and 55.2% for Clarify-or-Answer. The category breakdown and published reference results appear in Tab. 1. **Learning From Interaction (Q4):** Starting from Base, intent-specific skills raise accuracy to 70.1%, and Full reaches 74.3%, versus 67.4% for Reflexion-style memory and 69.5% for MetaClaw-style adaptation. These comparisons measure adaptation

Table 1: Ego-streaming QA accuracy (%) on Augmented BuddyVQA. †: published results under the original protocol and split (not directly comparable, Appendix C.1). All other rows are comparable.

Method	Recall	Recog.	Advis.	Scrut.	Chain	Deictic	Overall
Human†	92.4	92.1	91.9	92.0	–	–	92.1
InternVL3.5-8B† (Wang et al., 2025b)	54.1	51.3	44.7	46.1	47.2	43.1	48.8
Flash-VStream† (Zhang et al., 2025a)	52.2	51.9	44.6	46.7	–	–	50.8
ReKV (LLaVA-OV-7B)† (Di et al., 2025)	54.6	54.1	44.9	44.4	50.9	48.6	52.7
MyBuddy (Qwen3.5-9B)† (Qin et al., 2026)	59.9	60.8	55.8	54.2	57.3	58.9	60.7
Frozen inference (Qwen3.5-9B)							
Qwen3.5-9B (Qwen Team, 2026)	52.8	55.9	46.6	49.1	46.8	47.9	51.4
MyBuddy (our reproduction) (Qin et al., 2026)	62.0	62.3	53.4	57.9	59.8	57.9	59.2
ClearVQA (Jian et al., 2025)	57.7	59.8	50.9	53.2	51.7	54.0	55.7
Plug-and-Play Clarifier (Yang et al., 2026)	59.1	61.3	52.9	55.2	53.9	57.1	57.4
Clarify-or-Answer (Cao et al., 2026)	60.4	62.1	53.8	56.5	55.2	58.3	58.5
ReferClaw (Base, asking disabled)	62.0$^{\uparrow 1.6}$	62.8$^{\uparrow 0.7}$	56.3$^{\uparrow 2.5}$	60.2$^{\uparrow 3.7}$	58.6$^{\uparrow 3.4}$	58.2$^{\downarrow 0.1}$	60.5$^{\uparrow 2.0}$
ReferClaw (Base)	**67.7**$^{\uparrow 5.7}$	**68.6**$^{\uparrow 5.8}$	**62.3**$^{\uparrow 6.0}$	**65.4**$^{\uparrow 5.2}$	**66.8**$^{\uparrow 8.2}$	**69.0**$^{\uparrow 10.8}$	**66.2**$^{\uparrow 5.7}$
Learning from interaction (same stream, feedback, and update budget)							
Base + Reflexion-style memory (Shinn et al., 2023)	68.9	69.7	63.5	66.7	68.0	70.1	67.4
Base + MetaClaw-style adaptation (Xia et al., 2026)	71.1	71.7	65.8	68.7	70.2	72.3	69.5
ReferClaw (Skills)	71.6$^{\uparrow 0.5}$	72.2$^{\uparrow 0.5}$	66.4$^{\uparrow 0.6}$	69.4$^{\uparrow 0.7}$	71.0$^{\uparrow 0.8}$	73.1$^{\uparrow 0.8}$	70.1$^{\uparrow 0.6}$
ReferClaw (Full)	**75.7**$^{\uparrow 4.1}$	**76.4**$^{\uparrow 4.2}$	**71.0**$^{\uparrow 4.6}$	**73.3**$^{\uparrow 3.9}$	**75.4**$^{\uparrow 4.4}$	**77.5**$^{\uparrow 4.4}$	**74.3**$^{\uparrow 4.2}$

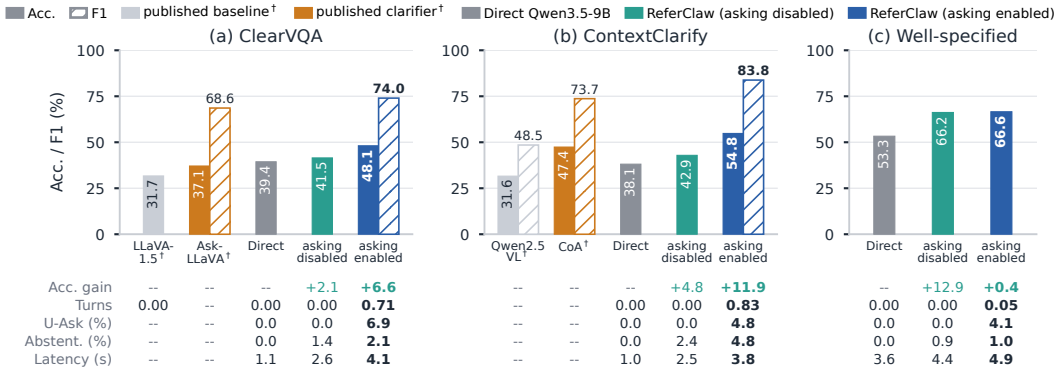

Figure 3: Clarification *vs.* No Clarification (ReferClaw uses frozen Base; †: published). Solid bars: accuracy; hatched bars: F1. Rows under each panel: accuracy gain over the preceding bar and interaction statistics.

under the same feedback and update budget. Sec. 4.4 further tests whether improvements also hold on an unchanged probe set.

Table 2: Results on EgoBlind-Intent (%).

Method	T-mIoU	S-mIoU	Ans.	Acc@GQA
Open-o3-Video	44.7	41.2	54.0	27.0
EG-Reasoner	46.5	39.8	54.8	27.6
ReferClaw (Skills)	48.9	41.6	58.8	30.1
ReferClaw (Full)	**50.1**	**42.6**	**60.7**	**31.6**

Generalization: Fig. 3 addresses two concerns: whether clarification helps beyond our own annotations, and whether it disrupts questions whose intent is already clear. F1 measures ambiguity detection on ClearVQA and the ask/answer decision on ContextClarify; Turns counts questions per input, U-Ask is the unnecessary-ask rate, and latency excludes respondent time. Without training skills or adapters on either image benchmark (protocols and populations in Appendix C.3), enabling clarification raises ClearVQA score from 41.5% to 48.1% and ContextClarify accuracy from 42.9% to 54.8%. These image-based tests assess clarification transfer, not temporal disambiguation. On well-specified BuddyVQA questions, accuracy changes by $+0.4$ points and the assistant asks on 4.1% of inputs. Confidence intervals are in Appendix E.3. **Joint Grounding and Cross-domain Transfer:** Correct answers may not be supported by the intended visual evidence. We therefore also evaluate Acc@GQA on our augmented benchmarks (Tab. 7, Appendix E.1). Frozen Base reaches 36.2% on BuddyVQA versus 35.0% for EG-Reasoner (Huang et al., 2026), while Full reaches 41.8%. Under zero-shot transfer to EgoBlind-Intent (Tab. 2, all thresholds, skills, and adapters frozen; full results for both benchmarks in Tab. 7), Full keeps a 4.0-point Acc@GQA and 5.9-point answer-accuracy advantage over EG-Reasoner, while both metrics drop by 10–12 points relative to BuddyVQA; the drop concentrates on blurred or off-center instances, whereas clarification quality transfers almost fully (Appendix E.2), so the gap lies in perception rather than in the clarification policy.

Table 3: Agent and communication ablation on Augmented BuddyVQA; each row removes one agent output or cross-agent message with all else fixed. Clar.: clarification attempts after which the EFA selects the target intent; Ask: fraction of candidate-bearing inputs with a question; Cost: $\sum_t C_x(c_t)$ per asked input. The last four columns average the per-input terms of Eq. 1 over candidate-bearing inputs, with Utility $= \overline{U}_{\text{ans}} + 0.5\,\overline{U}_{\text{grd}} - \overline{\Sigma C}$ (Appendix C.4). Superscripts: change relative to the full model.

Variant	Ans.	Acc@GQA	Clar.	Ask	Cost	$\overline{U}_{\text{ans}}$	$\overline{U}_{\text{grd}}$	$\overline{\Sigma C}$	Utility
ReferClaw (Full)	**74.3**	**41.8**	**81.4**	**38.7**	**0.42**	**0.732**	**0.328**	**0.163**	**0.733**
Intention Grounding Agent									
− IGA candidates (global video to CCA/EFA)	$68.0^{\downarrow 6.3}$	$30.5^{\downarrow 11.3}$	$68.2^{\downarrow 13.2}$	$42.0^{\uparrow 3.3}$	$0.48^{\uparrow 0.06}$	0.666	0.264	0.202	$0.596^{\downarrow 0.136}$
IGA temporal only (− spatial tracks)	$72.1^{\downarrow 2.2}$	$37.1^{\downarrow 4.7}$	$75.6^{\downarrow 5.8}$	$40.0^{\uparrow 1.3}$	$0.44^{\uparrow 0.02}$	0.714	0.302	0.176	$0.689^{\downarrow 0.044}$
Evidence-Fused Answering Agent									
EFA on global video (− selected evidence)	$71.0^{\downarrow 3.3}$	$35.2^{\downarrow 6.6}$	$81.1^{\downarrow 0.3}$	38.7	0.42	0.701	0.289	0.163	$0.683^{\downarrow 0.050}$
− resolved-intent rewrite and history in EFA	$71.4^{\downarrow 2.9}$	$38.8^{\downarrow 3.0}$	$73.5^{\downarrow 7.9}$	38.7	0.42	0.704	0.315	0.163	$0.699^{\downarrow 0.034}$
− answer–intent–evidence verifier	$72.7^{\downarrow 1.6}$	$39.7^{\downarrow 2.1}$	81.4	38.7	0.42	0.714	0.319	0.163	$0.711^{\downarrow 0.022}$
Cross-agent communication and control									
CCA without IGA candidates	$70.1^{\downarrow 4.2}$	$36.3^{\downarrow 5.5}$	$65.2^{\downarrow 16.2}$	$45.0^{\uparrow 6.3}$	$0.51^{\uparrow 0.09}$	0.690	0.293	0.230	$0.607^{\downarrow 0.126}$
EFA without CCA-updated posterior	$70.8^{\downarrow 3.5}$	$37.3^{\downarrow 4.5}$	$69.8^{\downarrow 11.6}$	38.7	0.42	0.702	0.304	0.163	$0.691^{\downarrow 0.041}$
− controller and posterior stopping	$69.9^{\downarrow 4.4}$	$36.9^{\downarrow 4.9}$	$71.2^{\downarrow 10.2}$	$52.1^{\uparrow 13.4}$	$0.66^{\uparrow 0.24}$	0.682	0.297	0.344	$0.486^{\downarrow 0.247}$

4.3 Ablation Study

Agents and Shared Candidates: Tab. 3 evaluates ReferClaw (Full) under matched stream order, backbone, interaction limit, and training budget. Grounded candidates support both intent resolution and clarification: replacing IGA's candidate evidence with global video lowers Acc@GQA by 11.3 points, while withholding its candidates from CCA reduces clarification success by 16.2 points. The EFA ablations show that selected evidence, the resolved-intent rewrite and history, and verification each contribute to answer quality. Posterior-based stopping also improves accuracy while reducing respondent cost. Together, these results support sharing grounded alternatives across question selection and answering. Appendix E.4 (Fig. 8) maps the four core designs to case types: grounded candidates and contrastive clarification account for the gains on referents separated in time and for staying silent when history resolves the reference, evidence-fused answering for questions that require reasoning after the referent is resolved, and learning from interaction for failure types that recur along the stream, so the designs are complementary rather than redundant. **Isolating Clarification:** Disabling asking while holding the entering model state and history fixed lowers Full accuracy from 74.3% to 66.9%, measuring the immediate benefit of clarification. A separate online run that never asks reaches 64.8%, reflecting the combined effects on answering and subsequent learning. A top-two clarification heuristic reaches 71.0% at higher respondent cost, supporting selective question choice. Removing the ambiguity gate or cost penalty increases asking frequency, while generic questions reduce clarification success from 81.4% to 64.9%. Full controls and further ablations are in Appendices E.5 and E.6.

4.4 Further Analysis

1. Direct Test of Temporal Ambiguity (Q2). Fig. 7 tests whether the benefit is concentrated in temporal ambiguity by partitioning the referentially underspecified questions into five groups (sizes and assignment rules in Appendix C.2). Frozen Base exceeds Clarify-or-Answer by 19.1 points for repeated encounters with the same object (G2) and 16.1 for similar objects at different times (G3), versus 7.8 for objects visible simultaneously (G1); the Full-minus-Clarify-or-Answer margins annotated in the figure follow the same pattern. When history already resolves the ambiguity (G4), the baselines still ask on about a quarter of the inputs (27.9% for Plug-and-Play Clarifier, 24.4% for Clarify-or-Answer), whereas Base asks on 7.0% and Full on 6.5%; all methods degrade when the target is absent from the proposals (G5). The result visualization (Fig. 9, Appendix E.8) shows both effects on one episode: ReferClaw names two temporally separated encounters in one question and answers the history-resolved follow-up without asking. A separate paired-intent test presents identical video and questions with two mutually exclusive target answers, changing only the respondent's clarification. Full answers both presentations correctly on 71.3% of pairs (Base: 64.2%), versus 38.8% for Clarify-or-Answer and 33.3% for Plug-and-Play Clarifier; deterministic direct answering scores zero because it receives identical inputs. This tests whether answers follow the clarified intent rather than a fixed interpretation; clarification-quality metrics on the ambiguous set and example traces are in Appendices E.7 and E.8.

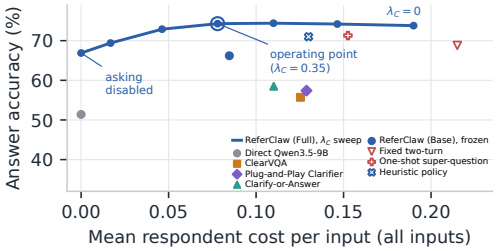

Figure 4: Accuracy–respondent-cost trade-off on Augmented BuddyVQA; the curve sweeps λ_C of ReferClaw (Full).

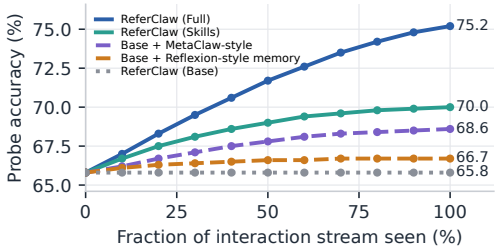

Figure 5: Fixed-probe accuracy after each 10% of the stream under identical feedback and update budgets.

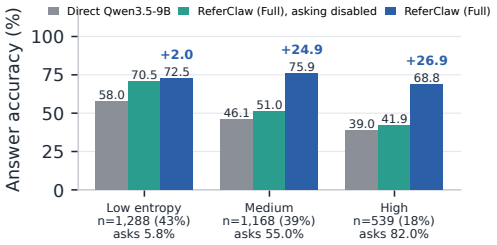

Figure 6: Accuracy by initial candidate entropy on candidate-bearing questions (labels: ask rate and share per bin).

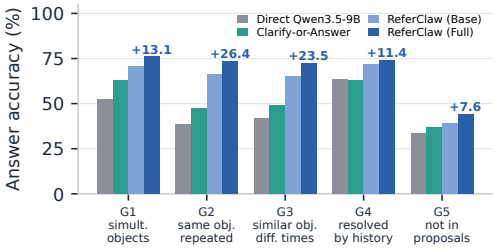

Figure 7: Accuracy by ambiguity group G1–G5 (Sec. 4.4); labels: ReferClaw (Full) minus Clarify-or-Answer.

2. Efficiency and Respondent Effort (Q3). Fig. 4 averages respondent cost over all inputs; the Full curve sweeps λ_C from 0 (right) to asking disabled (left). Filled markers denote frozen systems and hollow markers Full-based policy controls. Base outperforms frozen clarification baselines at lower cost. Full asks on 18.6% of inputs, averaging 0.22 questions per input. Its questions average 20.6 tokens and take humans 5.1 s to answer, against 24.0 tokens and 6.9 s for Clarify-or-Answer. Removing the cost penalty more than doubles cost without improving accuracy; a fixed two-turn policy also costs more and performs worse. Gains over the matched asking-disabled replay are larger for high-entropy candidates (Fig. 6; Appendix E.9). Timing and cost components are in Appendix E.10. Human-respondent results are reported in Appendix E.11.

3. Learning Dynamics (Q4). On an unchanged held-out probe, Skills improves over Base by 4.2 points and Full by 9.4, compared with 0.9 for Reflexion-style memory and 2.8 for MetaClaw-style adaptation (Fig. 5). Full also cuts failure recurrence from 34.2% to 10.7% with positive backward transfer (+4.1), versus 21.3% and +1.1 for MetaClaw-style adaptation. Targeted ablations further show that a generic skill schema reduces answer accuracy by 1.8 points and updating all adapters rather than the attributed agent reduces it by 2.0 (Appendix E.6). Since the MetaClaw-style variant alters the skill schema, the adapter-update scope, and the reward jointly, its gap to Full confounds the three factors; the single-factor ablations above disentangle them and attribute a gain to each. Learning curves, recurrence rates, and backward-transfer results are reported in Appendix E.12, and a manual audit of residual errors is provided in Appendix E.9.

5 Conclusion

We studied intent ambiguity in egocentric assistance, where video and dialogue evidence supports multiple possible referents. ReferClaw explicitly represents competing interpretations with spatio-temporal evidence, allowing the same grounded candidates to determine whether to clarify, what to ask, and which evidence to use. Three cooperating agents make these decisions under a fixed interaction budget, while verified feedback is distilled into intent-specific skills and agent-attributed process rewards to reduce recurring errors. On augmented ego-streaming assistance benchmarks, ReferClaw asks more useful clarification questions than strong interactive baselines, improves intent-aligned and jointly grounded answers with lower user effort, and provides the largest gains when competing referents are temporally separated. The learned skills and parameters further improve fixed probes and reduce repeated failures, while the frozen system transfers to two independent image-based clarification benchmarks without degrading well-specified questions.

Ethics Statement

This work adheres to the ICLR Code of Ethics. All video and image data come from publicly released benchmarks (BuddyVQA, EgoBlind, ClearVQA, and ContextClarify) and are used under their licenses. The user study (Appendix B.3) involved adult volunteers who gave informed consent, answered clarification questions about public clips, and were compensated; only anonymized responses and response times were recorded. Because ReferClaw targets assistive use, including questions from blind and visually impaired users, an answer to the wrong interpretation could mislead a user who cannot verify the evidence, and clarification imposes effort. The system therefore grounds answers in verifiable evidence, abstains or gives conditional answers when the intent cannot be resolved, and bounds the number, length, and cost of questions; it remains a research prototype and should not be deployed for safety-critical assistance without validation.

We have taken care to avoid any possible biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code, prompts, trained LoRA adapters, the skill library, and our augmented annotations (intent-ambiguity labels, target intents, supporting intervals, evidence-frame boxes, diagnostic groups, and the paired-intent set) will be provided upon acceptance. The method is specified in Sec. 3 and Appendix B (controller transition table, response likelihood and cost, process reward, promotion rule); all hyperparameters are listed in Tab. 6 and the hardware and optimization setup in Appendix D. The evaluation protocol, including what the assistant and the simulated user receive, when feedback is revealed, and the population of every metric, is documented in Appendix C.4, and all prompts are given verbatim in Appendix F. All benchmarks and backbone or tool models are publicly available. Human-respondent results depend on the recruited participants and are reported separately. We believe these measures will enable other researchers to reproduce our work and further advance the field.

References

Ehsan Abbasnejad, Qi Wu, Qinfeng Shi, and Anton van den Hengel. What's to know? uncertainty as a guide to asking goal-oriented questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4155–4164, 2019. doi: 10.1109/CVPR.2019.00428.

Leonard Bärmann and Alex Waibel. Where did i leave my keys?—episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1559–1567. IEEE, 2022.

Zongwan Cao, Bingbing Wen, and Lucy Lu Wang. Clarify or answer: Reinforcement learning for agentic VQA with context under-specification. *arXiv preprint arXiv:2601.16400*, 2026.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. Acknowledging focus ambiguity in visual questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1228–1238, 2025a.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024a.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025b. doi: 10.1609/aaai.v39i2.32214.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, 2024b.

Yura Choi, Roy Miles, Rolandos Alexandros Potamias, Ismail Elezi, Jiankang Deng, and Stefanos Zafeiriou. Do you see what i am pointing at? gesture-based egocentric video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18534–18544, 2026.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006.

Jisheng Dang, Huilin Song, Junbin Xiao, Bimei Wang, Han Peng, Haoxuan Li, Xun Yang, Meng Wang, and Tat-Seng Chua. MUPA: Towards multi-path agentic reasoning for grounded video question answering. *arXiv preprint arXiv:2506.18071*, 2025.

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2951–2960, 2017.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503–5512, 2017. doi: 10.1109/CVPR.2017.475.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12943, 2024. doi: 10.1109/CVPR52733.2024.01229.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video KV-cache retrieval. In *International Conference on Learning Representations*, 2025.

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.

William E. Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952. doi: 10.1080/17470215208416600.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking verifiable video question answering with grounded temporal evidence. *arXiv preprint arXiv:2606.24797*, 2026.

Ray Hyman. Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45(3):188–196, 1953. doi: 10.1037/h0056940.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 3619–3638, 2025. doi: 10.18653/v1/2025.acl-long.182.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. Asking clarification questions to handle ambiguity in open-domain QA. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11526–11544, 2023. doi: 10.18653/v1/2023.findings-emnlp.772.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023.

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.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3 Video: Grounded video reasoning with explicit spatio-temporal evidence. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5783–5797, 2020. doi: 10.18653/v1/2020.emnlp-main.466.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. VideoGLaMM: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19036–19046, 2025.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of MLLM: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2025.

Hangyu Qin, Junbin Xiao, Shenglang Zhang, and Angela Yao. Companion-style QA assistance in ego-vision. *arXiv preprint arXiv:2609.06721*, 2026.

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

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2737–2746, 2018. doi: 10.18653/v1/P18-1255.

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.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding DINO 1.5: Advance the "edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2087–2097, 2018. doi: 10.18653/v1/D18-1233.

Jinhwan Seo, Kyubeom Han, Jumin Lee, Junhyug Noh, and Sung-eui Yoon. What you ask is what you ground: Bridging question intent to temporal evidence for grounded videoqa. *arXiv preprint arXiv:2608.15708*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2512–2522. IEEE, 2025.

Haoqin Tu, Jianwen Chen, Zijun Wang, Siwei Han, Juncheng Wu, Hardy Chen, Haonian Ji, Kaiwen Xiong, Jiaqi Liu, Peng Xia, et al. Visualclaw: A real-time, personalized agent for the physical world. *arXiv preprint arXiv:2606.16295*, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22274–22284, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, 2025c.

Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. Open-domain clarification question generation without question examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 563–570, 2021. doi: 10.18653/v1/2021.emnlp-main.44.

Peng Xia, Jianwen Chen, Xinyu Yang, Haoqin Tu, Jiaqi Liu, Kaiwen Xiong, Siwei Han, Shi Qiu, Haonian Ji, Yuyin Zhou, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. Metaclaw: Just talk—an agent that meta-learns and evolves in the wild. *arXiv preprint arXiv:2603.17187*, 2026.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024. doi: 10.1109/CVPR52733.2024.01254.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0886. Datasets and Benchmarks Track.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1618–1629, 2019. doi: 10.18653/v1/D19-1172.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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, 2025b.

Sicheng Yang, Yukai Huang, Weitong Cai, Shitong Sun, You He, Jiankang Deng, Hang Zhang, Jifei Song, and Zhensong Zhang. Plug-and-play clarifier: A zero-shot multimodal framework for egocentric intent disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17921–17929, 2026. doi: 10.1609/aaai.v40i21.38851.

Wenzheng Zeng, Difei Gao, Mike Zheng Shou, and Hwee Tou Ng. Factorized learning for temporally grounded video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20683–20693, 2025.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21059–21069, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025b. doi: 10.18653/v1/2025.findings-acl.547.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

A LLM USAGE

LLMs were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to refine the language, improve readability, and ensure clarity across various sections of the paper. The model helped with tasks such as rephrasing sentences, checking grammar, and improving the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

B METHOD DETAILS

B.1 MODULE INSTANTIATION AND CANDIDATE GROUNDING

IGA, CCA, and EFA share a Qwen3.5-9B backbone but use separate LoRA adapters, which keeps the visual representation fixed while allowing agent-specific updates. IGA uses the released Ground-VQA checkpoint for coarse temporal proposals, Grounding DINO 1.5 Pro (Ren et al., 2024) for open-vocabulary boxes, and SAM 2 Hiera-Large (Ravi et al., 2025) to propagate the selected box through the retrieved interval. CCA and EFA operate on the resulting evidence cards and Qwen3.5-9B video tokens. BGE-M3 (Chen et al., 2024a) encodes skills and trajectories for cosine retrieval. The process reward model initializes from Qwen3.5-9B, adds a scalar value head, and is trained to regress the computed reward of Eq. 6 on query trajectories with feedback (held-out Spearman

0.83, MAE 0.09 on a $[0, 1]$-scaled reward). The *predictive respondent model* used by CCA to estimate response likelihoods is the Qwen3.5-9B backbone under the predictive-respondent prompt of Appendix F; it never receives the annotated target intent. The *evaluation simulator* that generates the actual responses during testing and learning rollouts is Qwen3-32B (Yang et al., 2025a); it receives the hidden target intent and its evidence but never the system posterior, candidate scores, or reference answer. Eight samples per (c, z_i) pair estimate each response-class distribution.

B.2 Response Likelihood and Respondent Cost

Response Classes and Likelihood. Free-text responses are assigned to the nearest response class (a partition cell, "other/none", *uncertain*, or invalid/missing) by BGE-M3 embedding similarity. For the LLM respondent, the likelihood $p_{\mathrm{L}}(y \mid c, z_i, h_t)$ is the class frequency over eight samples of the predictive respondent model per (c, z_i) pair, with additive smoothing $\delta = 0.05$ so that no class receives zero probability. For the human respondent, $p_{\mathrm{H}} = (1 - \kappa_{\mathrm{H}} - f_{\mathrm{H}})\, p_{\mathrm{L}} + \kappa_{\mathrm{H}}\, \mathrm{Unif} + f_{\mathrm{H}}\, q_{\mathrm{F}}$, where q_{F} is a normalized distribution over the failure classes, with κ_{H}, f_{H}, and q_{F} fitted as described in Appendix B.3. For both respondent types the failure-class likelihood depends on the question only, not on z_i, so a failure response leaves the belief unchanged under Eq. 4, which is also what the controller executes. The uncovered pseudo-candidate $z_\varnothing$ has the likelihood $p_x(y \mid c, z_\varnothing, h_t) = (1 - f(c))\, \mathbb{I}[y = \text{other/none}] + f(c)\, q_{\mathrm{F}}(y)$, where $f(c)$ is the candidate-independent failure mass of the question: a user whose intent is not among the displayed candidates selects "other/none" unless the response fails. Entropy, expected information gain, and the posterior update are all computed over the $n_t + 1$ elements $\{z_1, \ldots, z_{n_t}, z_\varnothing\}$.

Cost Components. For a human respondent, $\overline{E}_{\mathrm{H}}$ averages the normalized question length, the expected response length under the predictive respondent, and the choice complexity $\log_2(m(c) + 1)/\log_2(M_{\max} + 1)$; $\overline{T}_{\mathrm{H}}$ is the response time predicted by a linear regression on question length and option count fitted on the calibration half of the user study ($\overline{T}_{\mathrm{H}} = 1.0 + 0.12\,|c| + 0.55\,m(c)$ seconds); and $\overline{F}_{\mathrm{H}}$ is the predicted mass of the invalid, missing, and uncertain response classes. For an LLM respondent, $\overline{E}_{\mathrm{L}}$ counts prompt and completion tokens, $\overline{T}_{\mathrm{L}}$ is measured serving latency, and $\overline{F}_{\mathrm{L}}$ is the predicted invalid-format or self-inconsistent mass. A failed response already yields little information gain; $\overline{F}_x$ additionally charges the wasted turn and the user's time. All components are normalized to $[0, 1]$ with the 5th and 95th percentiles of the validation split, and the weights (w_E, w_T, w_F) are fitted by grid search on validation utility separately for the two respondent types, with a shared λ_C (Tab. 6).

B.3 Human Response Model Calibration

The 600-interaction user study (10 participants, 300 stratified ambiguous examples, each example answered under ReferClaw and under Clarify-or-Answer by different participants) serves two purposes that we keep separate by example. The 300 examples are split once into a *calibration* half (150 examples) and a *held-out evaluation* half (150 examples). On the calibration half we use the ReferClaw interactions to fit the human response model

$$p_{\mathrm{H}}(y \mid c, z_i, h_t) = (1 - \kappa_{\mathrm{H}} - f_{\mathrm{H}})\, \hat{p}_{\mathrm{L}}(y \mid c, z_i, h_t) + \kappa_{\mathrm{H}}\, \mathrm{Unif}(y) + f_{\mathrm{H}}\, q_{\mathrm{F}}(y),$$

where $\hat{p}_{\mathrm{L}}$ is the predictive respondent's class distribution, Unif is uniform over all response classes, and q_{F} is a distribution supported on the three failure classes (*uncertain*, *invalid*, *no response*) with proportions $(0.55, 0.25, 0.20)$ estimated from the study; each component is a normalized distribution, so p_{H} sums to one. "Other/none" is an ordinary answer class, not a failure class, because it is the correct response when the intended referent is missing from the options. The fitted weights are $\kappa_{\mathrm{H}} = 0.18$ and $f_{\mathrm{H}} = 0.06$. Response time is fitted on the same half as $\overline{T}_{\mathrm{H}} = 1.0 + 0.12\,|c| + 0.55\,m(c)$ seconds ($R^2 = 0.61$), with $|c|$ the question length in tokens and $m(c)$ the number of options. On the held-out half, the calibrated model lowers the Brier score of response prediction from 0.31 (uncalibrated simulator frequencies) to 0.22, and its predicted response times correlate with measured times at $r = 0.74$. These parameters define the likelihoods and costs used when the respondent is human and the human-calibrated profile of the evaluation simulator; simulator-based likelihoods and costs use the simulator's own statistics. The human comparison in Appendix E.11 uses only the held-out half for both methods; the Clarify-or-Answer interactions of the calibration half enter only the descriptive response-time statistics of Tab. 11.

B.4 INTERACTION CONTROL ALGORITHM

The interaction controller of Sec. 3.2 is a deterministic state machine over the state $(t, \mathcal{Z}_t, p_t, h_t, \mathcal{B}_t, n_{\mathrm{rg}})$, where $\mathcal{B}_t$ is the set of banned partitions and $n_{\mathrm{rg}} \in \{0, 1\}$ counts re-grounding passes. Tab. 4 lists its transitions; at every turn the first row whose condition holds is executed, and the implementation, the pseudocode, and the main text share this table. All limits ($T_{\max}$, $L_{\max}$, $M_{\max}$, no repeated partitions) are enforced before generation, so every trajectory terminates within $T_{\max}$ turns. Two procedures are shared by all rows. ANSWERPATH($i^\star$) runs the EFA on candidate $i^\star$ and then the verifier: *pass* returns the answer; *revise* regenerates once and returns the answer if the second verdict is *pass*, otherwise ABSTAIN; *abstain* returns ABSTAIN. REGROUND($\cdot$) runs the IGA once more, with relaxed detector thresholds (initial state) or restricted to the constraint stated in the response (after "other/none"); it may be invoked at most once per question ($n_{\mathrm{rg}} \leq 1$), new candidates absorb the mass of $z_\varnothing$ in proportion to their grounding scores, and an empty result triggers row T1 on the next evaluation.

Table 4: Priority-ordered controller transitions for one user question. Rows are tested top to bottom at every turn t; the first satisfied row is executed. $i^\star = \arg\max_i p_t(z_i)$; $(1), (2)$ index the two most likely candidates.

	Condition	Action	Output / next state
T1	$\mathcal{Z}_t = \emptyset$	if $n_{\mathrm{rg}} = 0$: REGROUND, $n_{\mathrm{rg}} \leftarrow 1$, re-evaluate; else terminate	ABSTAIN
T2	$g_t = 0$ **or** $\max_i p_t(z_i) \geq \tau_p$	ANSWERPATH($i^\star$); if $g_t = 0$ because all plausible candidates share an answer, EFA receives their pooled evidence and γ is their summed mass	ANSWER or ABSTAIN
T3	$g_t = 1$, $t < T_{\max}$, CCA returns $c_t^\star$ with $\mathrm{IG}_t - \lambda_C C_x > 0$ and partition $\notin \mathcal{B}_t$	ask $c_t^\star$; receive y_t; $h_{t+1} \leftarrow h_t \cup \{(c_t^\star, y_t)\}$; then (a) y_t uncertain, invalid, or missing: $p_{t+1} \leftarrow p_t$, $\mathcal{B}_{t+1} \leftarrow \mathcal{B}_t \cup \{\mathrm{partition}(c_t^\star)\}$; (b) y_t = other/none or unmatched constraint: REGROUND(y_t) if $n_{\mathrm{rg}} = 0$, else ABSTAIN; (c) otherwise p_{t+1} by Eq. 4, drop candidates with $p_{t+1} < 0.02$; $t \leftarrow t + 1$	continue at T1
T4	otherwise ($t = T_{\max}$, or no admissible question with positive utility)	if $\bar{a}_{(1)} \neq \bar{a}_{(2)}$ and $p_t(z_{(1)}) - p_t(z_{(2)}) < \tau_g$: ANSWERPATH($(1)$) and ANSWERPATH($(2)$), each labeled by its interpretation; else ANSWERPATH($i^\star$)	CONDITIONAL (if both branches pass; ABSTAIN if either fails) or ANSWER/ABSTAIN

Three consequences of the table are used in the analysis. First, every formal answer, including each branch of a CONDITIONAL output, is produced by the same EFA-then-verifier path, so verifier statistics are comparable across output types. Second, a question is asked only through T3, so the number of questions per input is $N_{\mathrm{ask}} \leq T_{\max}$ and a question can be asked only if $n_0 \geq 2$ (with $n_0 = 1$ the gate is closed). Third, uncertain and invalid responses never change the belief, which matches the candidate-independent failure-class likelihood used when scoring questions (Appendix B.2), so the expected information gain computed before asking and the update executed after asking are consistent.

B.5 FAILURE ATTRIBUTION RULES AND THEIR VALIDATION

A verified failure is attributed to one agent by deterministic rules on the logged trajectory wherever possible: $z^\star$ absent from $\mathcal{Z}$, or present but grounded with $\mathrm{IoU} < 0.5$ against $e^\star$, attributes to IGA (coverage or grounding failure); $z^\star$ present but placed in the same partition cell as the top competitor of the asked question, or a question asked when all plausible provisional answers already agreed, attributes to CCA (non-separating or unnecessary question); $z^\star$ selected with correct evidence but a wrong or unsupported answer attributes to EFA (inconsistent answer). Only trajectories that match none of these rules are attributed by the process-review prompt of Appendix F. Two authors independently attributed 200 randomly sampled verified failures using the trajectory and feedback (inter-annotator agreement $\kappa = 0.86$; disagreements adjudicated). The deterministic rules cover 84% of the failures and agree with the adjudicated label in 94% of covered cases; the process-review prompt handles the remaining 16% with 78% agreement, for a combined agreement of 91%. Most resid-

ual disagreements concern trajectories in which the target intent was covered with borderline IoU (0.4–0.5) and the question also failed to separate it, *i.e.*, cases with two plausible responsible agents.

B.6 Skill Schema and a Worked Example

Each skill stores an ambiguity trigger, a candidate topology, the decisive visual contrast, the corrected action, an anti-pattern, a retrieval key, and two regression tests (Appendix F gives the evolver prompt and schema). Retrieval uses BGE-M3 cosine similarity between the current trigger–topology description and stored keys with threshold 0.45; at most three skills are injected into the responsible agent's prompt. The following trace illustrates what is learned beyond the initial prompt, which already instructs CCA to contrast occurrences rather than repeat a noun. *Failure:* for a mug picked up twice in the same kitchen within 40 seconds, CCA asked "the first or the second time?"; the user answered *uncertain* and the final answer was wrong. *Feedback:* $z^\star$ was the second encounter, during which the user was pouring. *Stored skill (CCA):* trigger "same object, repeated encounters, same location, gap < 60 s"; corrected action "contrast by the user's action visible in each track (holding *vs.* pouring), not by order or location"; anti-pattern "ordinal or location contrast for repeated encounters at one place". *Later behavior:* on a subsequent kettle question with the same topology, CCA asked "when you were filling it or when you were pouring?", the response resolved the intent, and the answer was correct. Such action-contrast skills account for 23% of the CCA skills learned on Augmented BuddyVQA.

B.7 Process Reward, Adapter Update, and Promotion

Process Reward. The target-progress term $\Delta_t^\star = \log p_{t+1}(z^\star) - \log p_t(z^\star)$ of Eq. 6 is evaluated on the candidate scope that existed *before* turn t: candidates dropped or added after the response are re-inserted with their pre-update mass for the purpose of this term, so a change of scope cannot earn reward by itself, and $\Delta_t^\star$ is 0 when $z^\star$ was not covered at turn t. Directly computed terms are r_{ground}, $\sum_t C_x(c_t)$, $\sum_t \Delta_t^\star$, and r_{invalid}; r_{final} combines the judge verdict with the verifier's verdict. The process reward model predicts R_{proc} for sampled alternative trajectories whose feedback is not yet available.

Adapter Update. When the query buffer holds 128 trajectories and no online request is being served, each failed step is re-executed: the attributed agent samples $G = 8$ alternative structured outputs from its current adapter given the exact logged input, the process reward model scores each completed trajectory, and the adapter is updated with GRPO, *i.e.*, group-normalized advantages under a clipped policy objective (clip ratio 0.2) with a KL penalty (coefficient 0.02) toward the frozen Base adapter. Only the attributed adapter changes. A candidate adapter is promoted only if, on a fixed validation probe of 500 ambiguous and 500 well-specified examples, it improves accuracy on its target failure type by at least one point, lowers overall probe accuracy by at most 0.5 points, and does not raise the unnecessary-ask rate or mean cost by more than 0.01; otherwise it is rolled back and the buffer is retained for the next consolidation window.

C Benchmarks, Annotation Protocol, and Evaluation Details

C.1 Augmented Benchmarks

Augmented BuddyVQA. BuddyVQA (Qin et al., 2026) is our primary benchmark because its causal streaming protocol, ego-deictic expressions, and history-dependent question chains jointly exercise visual memory, referent tracking, and dialogue dependence. The original release contains 1,012 videos and 21,574 questions with a 704/308 video split for training and test; we annotate every question with an intent-ambiguity label, a minimal supporting temporal interval, and a box on the clearest evidence frame when the referent is spatially localizable (otherwise the instance is marked non-localizable). The original test split has 6,571 questions on the 308 test videos. We exclude 331 of them (5.0%): 214 whose timestamp lies within the first 30 seconds of the video, so that no meaningful causal prefix exists, and 117 for which the two annotators could not agree on a reference answer under the annotated interpretation after adjudication. The resulting test split contains 6,240 questions (Recall 1,790, Recognition 1,650, Advisory 1,420, Scrutinization 1,380). The published BuddyVQA results quoted in Tab. 1 (†) were obtained by their authors under the

original protocol on all 6,571 test questions; their type scores therefore do not satisfy the 6,240-question weighting, and we do not compare against them at the type level or compute differences to them. Of these, 1,398 are *referentially underspecified* in wording; 1,197 of them (19.2%; Recall 344, Recognition 318, Advisory 272, Scrutinization 263) remain intent-ambiguous given the prefix and the dialogue history and form the ambiguous population of every metric, whereas the other 201 are resolved by the history (group G4 below) and are labeled well-specified, giving 5,043 verified-unambiguous questions. 4,655 questions (74.6%) are grounding-eligible, *i.e.*, have an annotated supporting interval, and 3,914 of these additionally have a spatial box (741 are temporal-only).

Augmented EgoBlind-Intent. EgoBlind (Xiao et al., 2025) complements BuddyVQA with timestamped assistance questions from blind and visually impaired users under imperfect viewpoints, motion blur, and partial observation. We select 2,036 QA instances from the official test set and add the same ambiguity, interval, and box annotations. 612 instances are *intent-relevant* because their ego-deictic, elliptical, or context-dependent wording requires explicit referent resolution; among these, 418 satisfy the stricter answer-changing criterion used for BuddyVQA and form the ambiguous subset on which clarification metrics are computed, and 1,447 instances are grounding-eligible. Because the two datasets differ in how much of the intent-relevant population is answer-changing, we never pool them. The benchmark is reserved for zero-shot transfer: no instance is used for prompt selection, threshold tuning, skill synthesis, or policy optimization.

C.2 Annotation Protocol

Guidelines. Annotators watched only the video prefix up to the question timestamp together with the preceding dialogue. For each question they (i) listed every entity, moment, or place that the question could refer to and that was visible in the prefix; (ii) for each interpretation wrote a resolved question and a short provisional answer from the video; (iii) marked the question intent-ambiguous if at least two interpretations were visually supported given the prefix *and* the dialogue history and their provisional answers differed, and marked it history-resolved if the wording alone admitted several interpretations but a prior dialogue turn selected one of them; and (iv) annotated the minimal supporting interval and, when applicable, a single box on the clearest evidence frame for each interpretation (no box tracks are annotated; spatial IoU is therefore evaluated on that frame, Appendix C.4). The reference answer of the original dataset was shown only afterwards to select the annotated target intent among the enumerated interpretations; 3.1% of BuddyVQA and 4.6% of EgoBlind candidates for which the reference answer matched none of the enumerated interpretations were excluded from the ambiguous set and kept as ordinary questions. Two annotators labeled every instance independently (Cohen's $\kappa = 0.78$ on BuddyVQA and 0.74 on EgoBlind for the ambiguity label); disagreements were adjudicated by a third annotator, and 86% of adjudicated resolved questions were judged equivalent to at least one annotator's proposal. Because intent is assigned retrospectively, all results should be read as accuracy with respect to the annotated target intent rather than the original user's intention.

Diagnostic Groups. The groups of Sec. 4.4 partition the 1,398 referentially underspecified questions; they are assigned in the following priority order so that every question belongs to exactly one group. G4 (201) if the dialogue history contains a prior mention that uniquely selects one interpretation; these questions are labeled well-specified and are used to measure unnecessary asking. Among the remaining 1,197 intent-ambiguous questions, G5 (131) if the annotated target interval has IoU < 0.3 with all eight GroundVQA proposals, a tool-level condition shared by all methods; otherwise G1 (412) if the competing referents are distinct entities whose supporting intervals overlap, G2 (356) if the competing interpretations share the same entity track at disjoint intervals, and G3 (298) if the competing referents are distinct entities of the same category at disjoint intervals. The paired-intent set contains 240 videos for which two enumerated interpretations both have a valid reference answer in the original dataset (from different questions on the same video) and the two reference answers are judged mutually exclusive by the benchmark judge; we present the same ambiguous question twice with the respondent instructed with each target in turn, with the model, history, and random seed reset between the two presentations.

C.3 CLEARVQA AND CONTEXTCLARIFY

ClearVQA (Jian et al., 2025) contains 4,025 test questions over single images in three ambiguity categories (referential 1,938, intent under-specification 1,095, spelling 992), an unambiguous contrast set for ambiguity detection, and a VQA-score protocol with and without clarification; we report VQA score, ambiguity-detection F1 on its contrast set, and its published LLaVA-1.5-13B and Ask-LLaVA-13B (SFT+DPO) numbers. ContextClarify (Cao et al., 2026) contains 275 ambiguous image–question pairs with human-verified clarifications and, for each, two contrast variants: a *context-completed* variant in which the appended context resolves the ambiguity (gold decision ANSWER) and an *irrelevant-context* variant in which plausible but non-resolving context is appended, so that the item remains ambiguous (gold decision CLARIFY). Its official test split has 42 ambiguous items and 84 variants (42 of each kind). We report end-to-end accuracy on the 42 ambiguous items, the CLARIFY-*vs.* -ANSWER F1 of the decision over all 126 test items (84 CLARIFY positives: the originals and the irrelevant-context variants), the unnecessary-ask rate on the 42 context-completed variants, and the published prompting and GRPO-CR numbers. For both benchmarks we treat the image as a one-frame prefix, disable temporal proposals, keep Grounding DINO for spatial candidates, and use each benchmark's official metrics and respondent protocol (ClearVQA's clarification-response simulator and ContextClarify's oracle context provider). No skill or adapter is trained on either benchmark.

C.4 EVALUATION PROTOCOL AND METRIC DEFINITIONS

This subsection documents the protocol behind every number in the paper: what the assistant receives, what the simulated user knows, what is revealed after a prediction and to whom, when the assistant asks, answers, gives alternatives, or abstains, and which examples enter each metric.

What Does the Assistant Receive? Before producing any output for a test question, a method receives exactly four things: the causal video prefix $V_{\leq s_q}$ decoded at 1 fps up to the question timestamp, the question q, the preceding dialogue of the same episode (earlier questions, the method's own earlier answers, and earlier clarification exchanges), and the interaction budget ($T_{\max} = 2$, $L_{\max} = 24$, $M_{\max} = 4$). Future frames, the gold ambiguity label, the annotated target intent $z^\star$, its resolved question, the supporting interval and box $e^\star$, the diagnostic-group label, and the reference answer are never inputs at prediction time, for any method and in any regime. On ClearVQA and ContextClarify the prefix is the single image.

What Does the Simulated User Know? Two respondent roles are kept strictly separate (prompts in Appendix F). The *evaluation simulator* (Qwen3-32B; Llama-3.3-70B in Appendix E.11) answers the clarification questions that are actually asked during testing and learning rollouts. It receives the original question, the clarification history, the selected question with its options, the hidden target intent $z^\star$ as a resolved question together with a textual description of its evidence (object, place, and time of the intended encounter), and a respondent profile calibrated on the user study. It never receives the reference answer, the system posterior, candidate scores, the system's preferred option, or the video itself, and its rules forbid revealing more than the clarification asks. The *predictive respondent* used internally by CCA to estimate response likelihoods is the assistant's own Qwen3.5-9B backbone prompted once per candidate z_i "as a user whose intent is z_i"; it never sees $z^\star$. Human participants in the user study see only the video prefix, the original question, the target intent description they are asked to hold, and the clarification question. On ClearVQA and ContextClarify we use the benchmarks' official respondents (ClearVQA's clarification-response simulator and ContextClarify's oracle context provider).

What Is Revealed after a Prediction, and Who Learns from It? After a method has logged its final output for interaction k, the evaluator computes the judge verdict and grounding IoUs. Under the prequential protocol, the following *verified feedback* is then made available: $z^\star$ (resolved question and evidence $e^\star$) and the binary judge verdict U_{ans}. The reference answer string itself is not revealed, so no method can memorize answers. Only adapting conditions consume the feedback: ReferClaw (Skills) and ReferClaw (Full) with their learning-loop ablations and the online no-ask run, the Reflexion-style memory, and the MetaClaw-style variant. Feedback for interaction k may affect interactions $k{+}1$ onward only; frozen methods (all baselines in the frozen block, ReferClaw (Base), and both asking-disabled controls) receive the feedback stream but ignore it.

Table 5: Population and success criterion of every reported metric on Augmented BuddyVQA (EgoBlind-Intent in parentheses). Cand.: candidate-bearing inputs; asked: inputs with at least one question.

Metric	Averaged over	Success / definition
Ans. Acc. (Overall; type; Chain, Deictic)	all 6,240 (2,036) inputs; that type or subset	judge verdict w.r.t. target intent; CONDITIONAL/ABSTAIN wrong
Ambiguous / well-specified acc.	1,197 / 5,043 (418 / –)	Overall $= (1{,}197\,\mathrm{Amb} + 5{,}043\,\mathrm{Well})/6{,}240$
T-mIoU, T-R@.5, Ans. (Tab. 7)	4,655 (1,447) grounding-eligible	mean T-IoU; T-IoU ≥ 0.5; judge verdict
S-mIoU, S-R@.5, Acc@GQA	3,914 (1,203) box-annotated	S-IoU on annotated frame; S-IoU ≥ 0.5; correct and T-, S-IoU ≥ 0.5
Amb-F1; Ask-P / Ask-R	all inputs (1,197 ambiguous positives)	F1 of the ambiguity flag; precision / recall of asking
Unnecessary-ask rate	5,043 well-specified (Fig. 3); 201 G4 (Sec. 4.4)	at least one question asked
Clar.-Acc. (= Clar. Succ.); Intent-Acc.	asked ambiguous, per method or variant (612 asked by all five); all 1,197	posterior mode, selected by EFA, is $z^\star$; answer addresses $z^\star$
Ask rate; Turns (all / cand. / asked)	all / cand. / asked inputs	asked at least once; N_{ask} (0 if none, $\leq T_{\max}$)
Q tokens; human time; GPU time, latency	questions; study responses (else regression); all inputs	tokens / question; s / response; s / input, respondent excluded
Cost (asked / all); Utility	asked / all inputs; cand. inputs	$\sum_t C_x(c_t)$ ("–" if never asked); $U_{\mathrm{ans}} + 0.5\,U_{\mathrm{grd}} - \sum_t C_x(c_t)$
Abstention; paired-intent; probe	all inputs; 240 pairs; 1,000 held-out	CONDITIONAL/ABSTAIN; both correct; verdict at checkpoint

When Does the Assistant Ask, Answer, Give Alternatives, or Abstain? The decision follows the controller transitions of Tab. 4 (Appendix B.4): a method answers directly when the gate is closed or one candidate dominates (T2), asks only through T3, returns a CONDITIONAL output (one separately verified answer per top-two interpretation) when no further question is admissible but the two interpretations still disagree (T4), and ABSTAINs when no supported candidate or verified answer remains. CONDITIONAL and ABSTAIN outputs are scored as incorrect in every accuracy column and reported separately as the abstention rate (Appendix E.13).

Answer Correctness and Grounding. Answer accuracy uses each benchmark's multi-reference LLM-as-a-judge protocol; a stratified sample of 500 judgments was audited manually (97.2% agreement). On ambiguous questions the judge compares the answer with the reference answer of the *annotated target intent*, so an answer that is correct for a different interpretation counts as wrong. Temporal IoU is computed between the predicted interval and the annotated minimal interval. Spatial ground truth is a single box on the annotated evidence frame (Appendix C.2); spatial IoU is the IoU between that box and the predicted track's box on the same frame, and it is 0 when the predicted interval does not contain that frame. A question is *grounding-eligible* when it has an annotated interval, and *box-annotated* when it additionally has an annotated box ($r^\star \neq \varnothing$). Temporal-only questions (no box, *e.g.*, "Did I lock the door?") are excluded from all spatial metrics and from Acc@GQA rather than scored as S-IoU $= 0$; in Eq. 1 they contribute $U_{\mathrm{grd}} = \mathrm{tIoU}$.

Populations and Definitions. Tab. 5 lists, for every reported metric, the fixed set of examples it is averaged over and what counts as a success. The 1,197 *intent-ambiguous* and 5,043 *well-specified* questions (the latter including the 201 history-resolved G4 questions) partition the 6,240 test inputs. *Candidate-bearing inputs* are the 2,995 questions (48.0%; 1,160 intent-ambiguous, 1,835 well-specified) for which the frozen Base IGA returns at least two candidates. This stratum is used only for evaluation, so that turns, Utility, and the entropy-bin analysis compare methods on the same inputs; every method constructs its own candidates at test time, and questions asked outside the stratum count toward the ask rate over all inputs. The 1,197 intent-ambiguous questions are the positives of Ask-P and Ask-R; the 201 G4 questions count toward the unnecessary-ask rate instead.

Runs and Aggregation. Every number in a table comes from the same three runs (three fixed stream orders, identical seeds across methods), and the same checkpoint is used for all subsets of a row, so subset and overall values always refer to the same predictions; reported means are means of per-run values. GPU time includes candidate generation, response sampling, information-gain scoring, verification, and skill retrieval. Standard deviations and confidence intervals are in Appendix E.3.

D Additional Experimental Details

D.1 Baselines and Fair Comparison

We compare against complementary methods that isolate direct video answering, evidence retrieval, and active disambiguation. The *direct-answer MLLM* is Qwen3.5-9B (Qwen Team, 2026). *Grounded and streaming VideoQA* baselines include ReKV (Di et al., 2025), RGA3 (Wang et al., 2025a), and MyBuddy (Qin et al., 2026); these respectively test streaming memory retrieval, object-centric spatio-temporal referring, and ego-deictic reasoning with dialogue history. For joint answering and grounding (Tabs. 2 and 7), we further compare GeLM (Chen et al., 2025b), Open-o3-Video (Meng et al., 2026), GroundFormer (Seo et al., 2026), and EG-Reasoner (Huang et al., 2026), which return temporal or spatio-temporal evidence with each answer. *Interactive clarification* baselines include the ask-capable ClearVQA model (Jian et al., 2025), Plug-and-Play Clarifier (Yang et al., 2026), and Clarify-or-Answer with GRPO-CR (Cao et al., 2026). Framework-level baselines that permit backbone substitution use Qwen3.5-9B, whereas architecture-specific methods use their official checkpoints. Every method receives the identical decoded frame stream, causal video prefix, and maximum visual-token budget, and the same access to the dialogue history; fixed-frame architectures uniformly subsample from that same prefix. The history a method sees at question k of an episode consists of the earlier questions of that episode together with the method's *own* earlier answers and clarification exchanges, so histories are identical across methods in access rights and in the user-side content but not in the assistant-side content; controls and ablations that must be compared at the same state are instead run as *paired replays* on the logged entering state of the reference run (below). Gold ambiguity labels, resolved questions, spatial boxes, and future frames are hidden at prediction time. All interactive methods share the same respondent, maximum number of turns, and token and option budgets. We use released code and default prompts where available, changing only the input adapter required to accept a video clip instead of a single image. The *asking-disabled* controls run IGA and EFA with the verifier but never invoke CCA; they are paired replays of the reference run (same adapter and skill version at each stream position, same entering history and seed), differing only in that the current question is never asked (Appendix E.5).

D.2 Learning-from-Interaction Conditions

Following MetaClaw's dynamic online evaluation (Xia et al., 2026), we evaluate all adapting systems under identical inference and interaction budgets. For every ordered stream, interaction k is first answered and scored by the current system, verified feedback (the annotated target intent, its evidence, and the judge's correctness verdict) is revealed only afterward, and any resulting skill or parameter update can influence interactions $k+1$ onward but never the current or preceding examples. The model, skill library, and optimization state are initialized identically at the start of each stream order and then carried forward without resets inside that stream. **ReferClaw (Base)** runs the frozen pipeline. **ReferClaw (Skills)** additionally enables fast adaptation: after each verified failure, the skill evolver synthesizes a short skill, and the three most relevant skills are retrieved by cosine similarity and inserted into the responsible agent's prompt. **ReferClaw (Full)** combines skill adaptation with process-reward-guided policy optimization; every verified trajectory (successes and failures) enters the query buffer together with the skill version under which it was produced; skills take effect as soon as they are synthesized, so a buffer of 128 trajectories spans several skill generations. When the buffer is full and the server is not processing an online request, each failed step is re-executed with its own logged prompt, including the skills that were injected at the time, and only the LoRA adapter attributed to that step is updated with GRPO (Shao et al., 2024); the adapter version is constant within one consolidation window. The *Reflexion-style memory* baseline stores a free-text reflection for each verified failure and retrieves the three most similar reflections into every agent prompt. The *MetaClaw-style* baseline uses MetaClaw's generic skill schema (task, insight, action) without candidate topology or decisive contrast, and updates all three adapters with GRPO on the final-answer reward plus the same cost penalty; it receives the same update budget as ReferClaw (Full).

D.3 Optimization and Hyperparameters

Videos are processed causally at 1 fps; the top retrieved intervals are resampled at 4 fps and 448×448 resolution for fine grounding. The three agent adapters and the PRM are trained in bfloat16 on

Table 6: ReferClaw implementation and hyperparameters.

Grounding	1 fps causal stream; retrieved clips at 4 fps, 448×448; 8 temporal proposals; detection/text thresholds $0.25/0.20$ ($0.15/0.12$ for the single re-grounding pass)
Candidates	$N_{\max}=4$ after merging; $\kappa=0.5$; $\epsilon_0=0.2$; drop threshold 0.02; question pool $\lvert\mathcal{C}_t\rvert\le 6$ (3 multiple-choice, 3 open-ended)
Thresholds	$\tau_d=0.60$, $\tau_u=0.25$, $\tau_a=0.30$, $\tau_p=0.80$, $\tau_g=0.15$
Budget	$T_{\max}=2$, $L_{\max}=24$ displayed tokens, $M_{\max}=4$ options incl. *other/none* (one-shot stress test: $L_{\max}=64$, $M_{\max}=8$); generation cap 32 tokens per candidate question (longer candidates discarded) and 128 answer tokens
Respondent	8 samples per (c,z_i), smoothing $\delta=0.05$, free text mapped to classes by BGE-M3; $\lambda_g=0.50$, $\lambda_C=0.35$; $(w_E,w_T,w_F)=(0.45,0.25,0.30)$ human, $(0.55,0.15,0.30)$ simulator; components normalized to $[0,1]$ by validation 5th/95th percentiles
Learning	PRM weights $(\alpha,\beta,\lambda,\mu)=(0.50,0.50,0.35,1.00)$; one failure triggers skill synthesis (≤ 6 failures per call), BGE-M3 top-3 retrieval at cosine ≥ 0.45, ≤ 3 injected skills; LoRA rank 16, $\alpha_{\mathrm{LoRA}}=32$, dropout 0.05; query buffer 128; $G=8$; KL 0.02 to Base; clip 0.2; lr 5×10^{-6}, batch 64, one epoch per window; promotion probe 500+500 validation examples (target type $+1.0$, probe ≥ -0.5, unnecessary-ask and cost $\le +0.01$)
Decoding	IGA/CCA temperature 0.2, top-$p=0.9$; EFA/verifier temperature 0

eight NVIDIA H100 80GB GPUs and served with vLLM (Kwon et al., 2023). AdamW uses $\beta_1=0.9$, $\beta_2=0.95$, weight decay 0.01, gradient clipping at 1.0, a 3% linear warm-up, and cosine decay. Hyperparameters are selected once on the Augmented BuddyVQA validation split and then fixed for all test benchmarks. Cost weights (w_E,w_T,w_F) and λ_C are chosen by grid search over $\{0.1,\dots,0.6\}^3$ (normalized to sum to one) and $\{0,0.1,0.2,0.35,0.5,0.8\}$ to maximize validation utility (Eq. 1); the human and simulator cost weights are fitted separately and differ mainly in w_T (0.25 for humans, 0.15 for the simulator). The complete configuration is in Tab. 6.

E Additional Results

E.1 Grounded Answering Results

Table 7: Grounded VQA (%) on grounding-eligible questions of Augmented BuddyVQA and zero-shot Augmented EgoBlind-Intent. T/S: temporal/spatial. T-mIoU, T-R@.5, and Ans. are computed on all grounding-eligible questions, S-mIoU, S-R@.5, and Acc@GQA on the box-annotated subset (populations in Appendix C.4). Temporal-only baselines share our frozen Grounding DINO 1.5/SAM 2 spatial head. Superscripts on Ans. and Acc@GQA: change over the preceding row.

	Augmented BuddyVQA						Augmented EgoBlind-Intent (zero-shot)					
Method	T-mIoU	T-R@.5	S-mIoU	S-R@.5	Ans.	GQA	T-mIoU	T-R@.5	S-mIoU	S-R@.5	Ans.	GQA
GeLM (Chen et al., 2025b)	46.1	44.4	33.7	32.0	56.2	25.3	41.3	39.1	30.5	28.2	50.2	20.1
RGA3 (Wang et al., 2025a)	44.8	42.7	41.9	40.7	58.0	28.9	40.1	37.9	38.4	36.5	52.1	23.7
Open-o3-Video (Meng et al., 2026)	49.3	47.2	45.5	43.9	60.1	33.4	44.7	42.4	41.2	39.5	54.0	27.0
GroundFormer (Seo et al., 2026)	52.7	51.1	38.6	37.0	59.4	31.7	47.2	45.0	34.8	32.4	53.6	25.4
EG-Reasoner (Huang et al., 2026)	51.6	49.9	43.2	41.7	61.0	35.0	46.5	44.1	39.8	37.7	54.8	27.6
ReferClaw (Base, frozen)	54.9	53.3	46.8	45.1	64.9$^{\uparrow 3.9}$	36.2$^{\uparrow 1.2}$	47.6	45.9	40.4	38.6	56.9$^{\uparrow 2.1}$	28.4$^{\uparrow 0.8}$
ReferClaw (Skills)	56.6	55.1	48.5	47.0	68.6$^{\uparrow 3.7}$	38.9$^{\uparrow 2.7}$	48.9	47.2	41.6	39.8	58.8$^{\uparrow 1.9}$	30.1$^{\uparrow 1.7}$
ReferClaw (Full)	**58.3**	**57.0**	**50.3**	**48.7**	**72.9**$^{\uparrow 4.3}$	**41.8**$^{\uparrow 2.9}$	**50.1**	**48.4**	**42.6**	**40.9**	**60.7**$^{\uparrow 1.9}$	**31.6**$^{\uparrow 1.5}$

E.2 Per-Category Transfer Breakdown

On Augmented EgoBlind-Intent, ReferClaw (Full) answer accuracy over all 2,036 instances (Tab. 2 uses the 1,447 grounding-eligible ones) is 66.8% on instances without a blur or off-center flag ($n=1{,}291$) and 54.8% on flagged instances ($n=745$), against 61.3% and 43.6% for EG-Reasoner; Acc@GQA on the box-annotated subset is 35.9% *vs.* 24.7% for the two groups. Clarification metrics transfer more completely than grounding: Ask-P and Ask-R are 82.6% and 79.4% on the 418 ambiguous instances (85.8% and 83.1% in-domain), and clarification success is 78.0% (81.4% in-domain).

E.3 Confidence Intervals

Standard deviations over the three stream orders are at most 0.6 points for every accuracy in Tabs. 1 and 7. 95% paired-bootstrap confidence intervals resampled at the video level (10,000 resamples)

are $[73.2, 75.3]$ for ReferClaw (Full), $[69.0, 71.1]$ for ReferClaw (Skills), $[65.1, 67.4]$ for ReferClaw (Base), $[57.3, 59.7]$ for Clarify-or-Answer, $[58.0, 60.4]$ for the MyBuddy reproduction, and $[50.2, 52.6]$ for Direct Qwen3.5-9B (answer accuracy), and $[40.3, 43.4]$, $[37.4, 40.5]$, $[34.7, 37.7]$, and $[33.5, 36.5]$ for the Acc@GQA of ReferClaw (Full), (Skills), (Base), and EG-Reasoner. Differences between ReferClaw (Full) and every baseline, and between ReferClaw (Base) and every frozen baseline, exclude zero. For ContextClarify, the 95% interval of the enabled-minus-disabled difference on 42 items is $[+2.4, +21.4]$ points.

E.4 Which Design Helps Which Case



Figure 8: Which core design helps which case. Rows are the four core designs of ReferClaw; columns are the case types defined by the annotations (diagnostic groups G1–G5, well-specified questions, questions that require state or action reasoning after the referent is resolved, and failure types that recur along the interaction stream). Cell shading gives the qualitative level of help, and cell notes quote the result that supports it.

Fig. 8 summarizes, per core design, the cases in which it makes a difference; the pattern is complementary rather than uniform. Grounded candidate construction is decisive where the competing readings are separated in time: because each candidate is anchored in its own interval and track, the two encounters of one object (G2) or two similar objects seen at different moments (G3) become distinct hypotheses that a question can name, which is where the advantage over the closest clarification baseline is largest (26.4 and 23.5 points against 13.1 when the referents are visible simultaneously, Fig. 7); removing the candidate set costs 11.3 Acc@GQA points and dropping the spatial tracks another 4.7 (Tab. 3). The same design also carries most of the improvement on well-specified questions, where a single grounded candidate delivers selected evidence without any question (12.9 points over direct answering with asking disabled, Fig. 3), whereas it cannot help when the referent lies outside the proposals (G5), which remains the weakest group and whose headroom the coverage oracle bounds at 2.6 points (Tab. 8).

Cost-aware contrastive clarification determines *whether* and *what* to ask. Its gate is what keeps the system silent when the history already resolves the reference (G4, asked on 6.5% of inputs against 24.4% for Clarify-or-Answer, with 11.4 points higher accuracy) and on well-specified questions (unnecessary-ask rate around 4%), and its information-gain scoring concentrates questions where the candidate belief is uncertain (asked on 82.0% of high-entropy inputs but 5.8% of low-entropy ones, Fig. 6). What is asked matters as much as when: generic questions reduce clarification success from 81.4% to 64.9% (Tab. 9), and a heuristic that always asks about the top-two contrast recovers

most of the accuracy only at higher cost (0.65 against 0.46 turns per candidate-bearing input, Tab. 8). Clarification cannot repair a coverage failure, so on G5 it asks without benefit.

Evidence-fused answering acts after the intent is resolved and is therefore most visible on questions whose answer requires reasoning about state or action once the referent is known: Advisory and Scrutinization questions gain 8.5 and 8.9 points over Clarify-or-Answer (Tab. 1), replacing the selected clip by the global video costs 6.6 Acc@GQA points, and the paired-intent test, on which the same question must be answered differently for two intents, rises from 38.8% to 71.3% because the answer is produced from the selected evidence rather than from the most salient encounter. The verifier contributes a smaller but consistent 1.6 points by rejecting answers unsupported by the card.

Learning from interaction is the only design whose effect grows along the stream: recurrence of verified failure types falls from 34.2% to 10.7% and backward transfer is positive (4.1 points, Tab. 13), so its help concentrates on failure types that repeat. Its per-group effect is largest again on repeated encounters, driven by the action-contrast skills that account for 23% of the learned CCA skills (Appendix B.6), and it improves the Advisory and Scrutinization types by 8.7 and 7.9 points from Base to Full, whereas G5 improves only modestly and G4 mostly through fewer unnecessary questions (7.0% to 6.5%). Taken together, the four designs address different case types, which is why removing any one of them degrades a different column of Fig. 8 rather than uniformly lowering accuracy.

E.5 Controls that Isolate the Value of Clarification

Table 8: Controls that isolate the value of clarification (Augmented BuddyVQA, all test inputs). Turns: per candidate-bearing input; Cost: per asked input ("–" when never asked). Top block: policies sharing the learned skills and adapters of ReferClaw (Full); middle: the frozen Base pair and the online no-ask run; bottom: oracles using gold annotations at prediction time.

Variant	Ans. Acc.	Acc@GQA	Turns	Cost ↓	What it measures
ReferClaw (Full)	74.3	41.8	0.46	0.42	–
ReferClaw (Full), asking disabled (state-matched replay)	66.9	39.9	0.00	–	immediate effect of asking, learned states
Heuristic policy (always ask top-2 contrast when g_t=1)	71.0	39.6	0.65	0.52	value of IG−cost scoring
Single agent, same structured steps, one adapter	71.6	39.9	0.50	0.44	value of role separation
ReferClaw (Base)	66.2	36.2	0.47	0.45	–
ReferClaw (Base), asking disabled (state-matched replay)	60.5	34.8	0.00	–	immediate effect of asking, frozen system
Online no-ask run (Base + both loops, CCA disabled throughout)	64.8	38.3	0.00	–	learning without ever asking
Oracle: gold intent inserted into candidate set	76.9	44.6	0.41	0.40	candidate coverage ceiling
Oracle: gold resolved question as input	79.8	47.3	0.00	–	cost of intent resolution

Both state-matched replays keep IGA, EFA, and the verifier but never invoke CCA: each question is re-run with the checkpoint, skill version, entering history, and seed that the reference run had at that stream position, so reference and replay differ only on inputs with $n_0 \geq 2$. Because the replayed states of ReferClaw (Full) were themselves learned from earlier clarifications, the 7.4-point difference (74.3 *vs.* 66.9) is the *immediate* effect of asking at matched states; the frozen pair gives 5.7 points. The *online no-ask run* instead measures learning without any clarification: an independent prequential run from Base with both loops but CCA disabled reaches 64.8%, 4.3 points above the frozen asking-disabled pipeline and 9.5 below ReferClaw (Full), the latter being the total contribution of clarification including its effect on what is learned. The heuristic policy asks about the top-two candidates whenever the gate of Eq. 3 opens and therefore asks on every gate-open input (52.1% of candidate-bearing inputs against 38.7%); the single-agent control executes the same structured steps as one prompt sequence on one adapter. The oracle rows provide diagnostic headroom: inserting the gold intent adds 2.6 accuracy points and supplying the gold resolved question adds 5.5. These interventions do not provide an additive causal decomposition of the remaining errors.

E.6 Fine-Grained Ablation

Cost-Aware Clarification. Removing the ambiguity gate makes the system ask on 70.0% of candidate-bearing inputs instead of 38.7%, raises the cost per asked input to 0.69, and reduces accuracy by 1.8 points because well-specified inputs are needlessly perturbed (lowest Utility, 0.387). Removing the cost penalty preserves most of the accuracy (73.8) but asks on every gate-open input (52.1%, the same fraction as the one-shot policy and the no-stopping variant of Tab. 3) at nearly twice the cost per asked input (0.76). The one-shot super-question, run with a relaxed length budget ($L_{\max} = 64$, $M_{\max} = 8$), loses 3.0 points despite a Cost of 0.61, and generic questions reduce clarification success from 81.4% to 64.9%.

Table 9: Fine-grained ablation on Augmented BuddyVQA; columns as in Tab. 3. Within each block only the named component is changed.

Variant	Ans. Acc.	Acc@GQA	Clar. Succ.	Ask (%)	Cost (asked) ↓	$\overline{U}_{\text{ans}}$	$\overline{U}_{\text{grd}}$	$\overline{\Sigma C}$ ↓	Utility ↑
ReferClaw (Full)	**74.3**	**41.8**	**81.4**	**38.7**	**0.42**	**0.732**	**0.328**	**0.163**	**0.733**
Cost-aware clarification									
− ambiguity gate	72.5	40.4	73.6	70.0	0.69	0.710	0.320	0.483	0.387
− expected information gain (confidence-only trigger)	71.9	40.0	70.1	42.0	0.47	0.708	0.316	0.197	0.669
− cost penalty ($\lambda_C = 0$)	73.8	41.4	82.0	52.1	0.76	0.732	0.326	0.396	0.499
One-shot super-question (relaxed length budget)	71.3	39.4	68.5	52.1	0.61	0.708	0.313	0.318	0.547
Generic rather than contrastive clarification	71.1	38.8	64.9	40.0	0.45	0.702	0.312	0.180	0.678
Learning from interaction: fast loop									
− fast loop	71.2	38.6	77.2	40.0	0.43	0.704	0.307	0.172	0.685
Random skill retrieval	72.0	39.4	78.0	40.0	0.44	0.715	0.315	0.176	0.696
Top-1 rather than top-3 retrieval	73.3	40.7	80.2	39.0	0.43	0.727	0.321	0.168	0.719
Generic rather than intent-specific skill schema	72.5	39.9	78.8	39.0	0.43	0.715	0.318	0.168	0.706
− support–query skill-generation separation	72.2	39.5	78.5	39.0	0.43	0.709	0.318	0.168	0.700
Learning from interaction: slow loop									
− slow loop (Skills only)	70.1	38.9	79.4	40.0	0.43	0.691	0.312	0.172	0.675
Final-answer-only reward	72.6	39.4	79.1	43.0	0.47	0.716	0.310	0.202	0.669
− grounding reward	73.1	37.6	80.3	39.0	0.43	0.722	0.303	0.168	0.706
− target-progress reward $\Delta_t^\star$	72.9	40.7	76.8	45.0	0.49	0.716	0.313	0.221	0.651
Undirected entropy reduction ΔH_t instead of $\Delta_t^\star$	73.4	40.9	78.9	41.0	0.44	0.721	0.324	0.180	0.703
Update all agent adapters	72.3	39.4	78.3	42.0	0.46	0.713	0.312	0.193	0.676

Learning Loops. Disabling the fast loop costs 3.1 accuracy points; random retrieval and a generic skill schema trail the full model by 2.3 and 1.8 points, so the gain comes from targeted intent skills rather than a longer prompt, and removing the support–query separation costs 2.1 points. The slow loop adds 4.2 points over Skills only: removing the grounding reward causes the largest Acc@GQA decrease (41.8 to 37.6), removing the target-progress term $\Delta_t^\star$ mainly reduces clarification success (81.4% to 76.8%), replacing it by an undirected entropy reduction costs 0.9 points and 2.5 points of clarification success, and updating every adapter is 2.0 points below agent-attributed updates.

E.7 Clarification Quality

Table 10: Clarification quality on Augmented BuddyVQA (%). Amb-F1: F1 of the ambiguity flag; Ask-P/R: precision and recall of asking, with the 1,197 intent-ambiguous questions as positives; Clar.-Acc.: target intent selected after clarification, over the questions each method asked (asked) or over the 612 asked by all five methods (common); Intent-Acc.: final answer addresses the target intent, over all intent-ambiguous questions.

Method	Amb-F1	Ask-P	Ask-R	Clar.-Acc. (asked)	Clar.-Acc. (common)	Intent-Acc.
ClearVQA (Jian et al., 2025)	72.6	68.8	76.2	66.1	64.1	58.3
Plug-and-Play Clarifier (Yang et al., 2026)	75.4	72.0	76.8	69.5	67.8	61.0
Clarify-or-Answer (Cao et al., 2026)	78.1	76.3	79.6	73.7	71.9	64.9
ReferClaw (Base)	81.0	82.4	80.9	77.5	76.3	72.8
ReferClaw (Full)	**84.3**	**85.8**	**83.1**	**81.4**	**80.2**	**77.6**

In Tab. 10, ReferClaw improves Amb-F1 by 6.2 points and ask precision by 9.5 over Clarify-or-Answer, so its gain is not obtained by asking more often. Because methods ask on different subsets, we also report clarification accuracy on the 612 ambiguous examples asked by all five methods; the ordering is unchanged and the gap widens slightly (80.2 *vs.* 71.9). Intent accuracy over all ambiguous examples, which penalizes both wrong resolutions and failures to ask, rises from 64.9% to 77.6%. Internally, the realized target progress $\mathrm{TP} = \sum_t \Delta_t^\star$ (the target-progress term of Eq. 6, in nats, computed on the fixed candidate scope of Appendix B.7) averages 0.62 per asked question for ReferClaw (Full), and 92% of asked questions with $\mathrm{TP} > 0.5$ end with the target intent selected, confirming that belief moves toward the intended interpretation rather than toward a confidently wrong one.

E.8 Example Traces

Qualitative Comparison. Fig. 9 follows an episode in which the user takes two capsules from a bottle on the kitchen counter and, later, one tablet from a pill box at the coffee corner, and then asks "How many did I take?" and "Where did I leave it?". Direct answering merges the two encounters, and grounded answering localizes its evidence in the capsule encounter, an error that carries over to the follow-up. Clarify-or-Answer obtains the intended referent with a generic question, but because the reply is not tied to an encounter in the video it still answers from the more salient one, and it asks again although the dialogue already resolves "it". ReferClaw grounds both encounters before

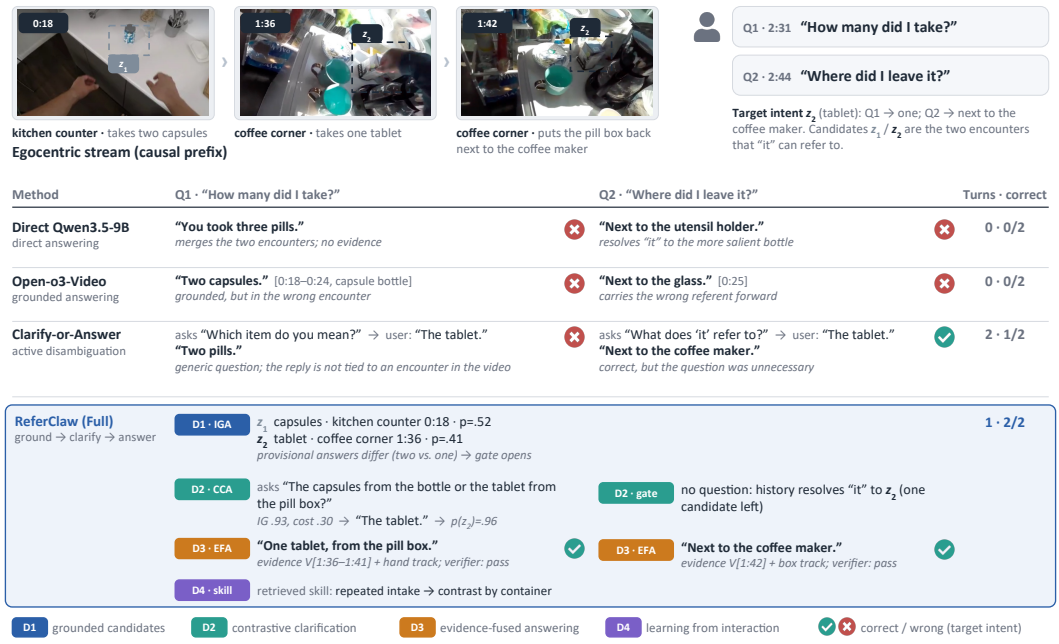

Figure 9: Qualitative comparison on a two-question episode. z_1, z_2: grounded candidates (dashed boxes: their tracks); tags mark the core design behind each ReferClaw step (D1–D4 as in Fig. 8); ✓/×: correct or wrong for the target intent z_2.

asking (D1), so a single contrastive question separates them (D2), guided by a skill learned earlier in the stream that contrasts repeated intakes by container (D4); it then answers from the selected clip under verification (D3) and answers the follow-up without a question because the resolved referent leaves a single candidate.

Successful Trace. Video prefix 1:50–2:30 (kitchen, then study). Question at 2:30: "Is it still on?" IGA: $d_0 = 0.81$; candidates z_1 (stove, 1:50–2:07, kitchen) with $p = 0.41$ and z_2 (desk lamp, 2:20–2:30, study) with $p = 0.52$, $p(z_\varnothing) = 0.07$; provisional answers differ ($\bar{a}_1$ = no, $\bar{a}_2$ = yes; Dis $= 0.92$), $g_0 = 1$. CCA: pool of five questions; selected c_0 = "The stove or the desk lamp?" (options: stove, desk lamp, other/none), $\mathrm{IG}_0 = 0.88$ bits, $C_\mathrm{L} = 0.28$, net utility 0.78. Response: "The lamp." Posterior: $p_1(z_2) = 0.94$. Controller: $\max p_1 \geq \tau_p$, ANSWER. EFA on $V[2{:}20, 2{:}30]$ with the lamp track: "Yes, the desk lamp is still on" (`video`); verifier *pass*. Correct.

Failure Trace. Same episode; question at 2:32: "Did I unplug it?" after a dialogue turn about the toaster. IGA: candidates z_1 (toaster, 2:03) $p = 0.60$ and z_2 (phone charger, 2:26) $p = 0.32$; $d_0 = 0.72$, Dis $= 0.85$, $g_0 = 1$. CCA selected "The toaster or the phone charger?" Response: "The toaster." EFA answered "Yes, you unplugged it" from a frame at 2:03 in which the hand occludes the socket; the verifier passed because the claim was labeled `video` and the card described "hand at the toaster plug". Annotated answer: no. Attribution: EFA (target covered and grounded; answer unsupported). This trace belongs to the occlusion category of the failure audit and illustrates the verifier's limitation as a prompted consistency check.

E.9 Model Dynamics and Failure Modes

The Value of Interaction Is Concentrated in Uncertain Cases. Fig. 6 bins the 2,995 candidate-bearing test questions by the initial entropy $H(p_0)$ of the frozen Base candidate set; the low-, medium-, and high-entropy bins contain 1,288, 1,168, and 539 questions (43.0%, 39.0%, 18.0%). To attribute the gain to asking rather than to learned grounding, each bin compares ReferClaw (Full) with its paired asking-disabled replay (Tab. 8) and with direct answering. ReferClaw asks on only 5.8% of the low-entropy bin, where asking adds 2.0 points (72.5 *vs.* 70.5), whereas it asks on 55.0% and 82.0% of the medium- and high-entropy bins, where asking adds 24.9 and 26.9 points (75.9 *vs.*

51.0 and 68.8 *vs.* 41.9); the gains over direct answering are 14.5, 29.9, and 29.9 points. Because the replay fixes the entering history, the two branches differ only on these 2,995 inputs, over which asking adds 15.4 points, corresponding to the 7.4-point Overall difference of Tab. 8. This monotonic pattern explains why adaptive stopping dominates a uniform interaction depth: most inputs need little assistance, but the difficult tail benefits strongly from a well-chosen question.

Residual Failures Point to Perception Rather than to the Clarification Loop. A manual audit of 500 incorrect ReferClaw (Full) predictions attributes 31% to severe occlusion or motion blur, 27% to multiple visually near-identical referents, 23% to information absent from the observed prefix, 12% to long-range identity drift, and 7% to invalid or noisy user responses. The first and fourth categories are grounding failures, and the third cannot be repaired by a better question. Future work should combine stronger persistent tracking with requests for *new visual evidence* (*e.g.*, asking the user to look toward the shelf), which our formulation can express as an action with cost but which we have not evaluated.

E.10 Efficiency and Respondent Cost

Table 11: Accuracy–cost trade-off on Augmented BuddyVQA. Ask rate: all inputs / candidate-bearing inputs; Outside: inputs outside the candidate-bearing stratum on which the method asked; Turns: per input over all / candidate-bearing / asked inputs; Tok./q.: tokens per question; Human time: seconds per response, measured in the user study or predicted by the calibrated regression (‡); GPU time: seconds per input; Cost: per asked input / per input. The one-shot super-question uses a relaxed length budget.

Method	Acc.	Ask rate (%)	Outside	Turns (all / cand. / asked)	Tok./q.	Human time (s)	GPU time (s)	Cost (asked) ↓	Cost (all) ↓
Direct Qwen3.5-9B	51.4	0.0 / 0.0	0	0.00 / 0.00 / –	–	–	3.6	–	0.000
ClearVQA (Jian et al., 2025)	55.7	21.2 / 43.1	35	0.28 / 0.57 / 1.32	23.3	6.0‡	5.0	0.59	0.125
Plug-and-Play Clarifier (Yang et al., 2026)	57.4	20.4 / 41.7	28	0.28 / 0.58 / 1.38	22.2	5.9‡	5.4	0.63	0.129
Clarify-or-Answer (Cao et al., 2026)	58.5	20.0 / 41.0	20	0.26 / 0.53 / 1.28	24.0	6.9	5.1	0.55	0.110
Fixed two-turn clarification	68.8	25.0 / 52.1	0	0.50 / 1.04 / 2.00	18.4	4.9‡	7.8	0.86	0.215
One-shot super-question (relaxed budget)	71.3	25.0 / 52.1	0	0.25 / 0.52 / 1.00	47.5	11.1‡	5.8	0.61	0.153
ReferClaw (Base)	66.2	18.8 / 39.2	0	0.23 / 0.47 / 1.21	21.7	5.3‡	4.9	0.45	0.085
ReferClaw (Full)	**74.3**	**18.6 / 38.7**	0	**0.22 / 0.46 / 1.19**	**20.6**	5.1	5.0	**0.42**	**0.078**

Tab. 11 gives the operating points of Fig. 4, whose horizontal axis is the cost per input over all inputs. The fixed two-turn and one-shot policies ask on every gate-open input (25.0% of all inputs), so their turn counts bound what a gate-triggered policy can spend. ReferClaw (Full) reaches the highest accuracy while asking on the fewest inputs among the interactive methods, with the fewest questions per asked input and the shortest questions, at a GPU time (5.0 s) within the range of the other interactive systems. Because frozen baselines and the adapting ReferClaw (Full) are different regimes, the figure also marks the frozen ReferClaw (Base), which dominates every frozen baseline on both axes (66.2% at 0.085 against 58.5% at 0.110 for Clarify-or-Answer).

E.11 Human Respondents

Table 12: Human-respondent evaluation on the 150 held-out examples of the user study (10 participants). Confusing: fraction of questions rated unanswerable or confusing. Sim. agreement: fraction of interactions in which the simulator's response falls in the same class as the human's. The last two columns give simulator accuracy on the same examples.

Method	Intent resolved (%)	Ans. Acc. (%)	Resp. time (s)	Confusing (%)	Sim. agreement (%)	Acc. Qwen3-32B	Acc. Llama-3.3-70B
Clarify-or-Answer	70.0	57.3	6.9	12.7	85.9	58.1	57.6
ReferClaw (Full)	**80.0**	**72.0**	**5.1**	**4.3**	88.6	73.7	73.4

Tab. 12 makes the human study visible on the held-out half, which was not used to fit the response model or the response-time regression. With human respondents, ReferClaw resolves the target intent on 80.0% of examples and answers 72.0% correctly, against 70.0% and 57.3% for Clarify-or-Answer, while requiring 1.8 fewer seconds per response and producing a third as many questions rated unanswerable or confusing. Simulator accuracy is 1.7 points higher than human accuracy for ReferClaw and 0.8 for the baseline, so the simulator is slightly optimistic for both, and response-class agreement between simulator and humans is 86–89%. Replacing the evaluation simulator with Llama-3.3-70B, which was not used for any design decision, changes accuracy by at most 0.5 points for either method, suggesting that the question-selection policy is not tuned to the particular simulator used during development.

E.12 Learning Dynamics

Table 13: Learning dynamics on Augmented BuddyVQA. Q1–Q4: prequential accuracy on stream quartiles; Probe: accuracy on a fixed held-out probe of 1,000 examples after the stream; Recurrence: verified errors followed by the same failure type within 200 questions; BWT: accuracy change on first-quartile examples re-evaluated after the stream; Update: GPU-hours per 1,000 interactions.

Condition	Q1	Q2	Q3	Q4	Probe	Recurrence ↓	BWT ↑	Update GPU-h ↓
ReferClaw (Base)	66.0	66.1	66.3	66.4	65.8	34.2	0.0	0.0
Base + Reflexion-style memory	66.0	67.2	67.9	68.5	66.7	27.8	+0.4	0.2
Base + MetaClaw-style adaptation	66.5	68.9	70.6	72.1	68.6	21.3	+1.1	3.1
ReferClaw (Skills)	66.2	69.3	71.5	73.3	70.0	18.5	+1.9	0.3
ReferClaw (Full)	**66.5**	**72.6**	**77.0**	**81.2**	**75.2**	**10.7**	**+4.1**	**2.8**

The quartile columns show prequential accuracy along the stream, while the probe column evaluates the final checkpoint on an unchanged held-out set so that the gain cannot reflect changing example difficulty. Skills improve the probe at 0.3 GPU-hours per 1,000 interactions because they take effect without weight updates; the slow loop's 2.8 GPU-hours are incurred asynchronously between online requests.

Fig. 5 plots the probe accuracy of every checkpoint (after each 10% of the stream, mean over the three orders; ReferClaw (Base) stays at 65.8%). ReferClaw (Full) improves at every checkpoint, with the largest increments in the first half of the stream, where most first-of-type failures occur and the skill library grows fastest; the Skills-only curve flattens after 60% of the stream, when new failures mostly repeat already-stored topologies, whereas the slow loop keeps consolidating gains until the end.

E.13 Coverage

ReferClaw (Full) returns an ANSWER on 96.1% of Augmented BuddyVQA test inputs, a CONDITIONAL output on 2.3%, and ABSTAIN on 1.6%; accuracy on the answered population is 77.3%, and 61% of CONDITIONAL outputs contain the correct answer paired with the target intent (counted as incorrect in all tables). Clarify-or-Answer answers 98.9% of inputs with 59.1% accuracy on the answered population; Direct Qwen3.5-9B answers 100%. Reporting accuracy over all inputs, as in the main tables, therefore does not favor ReferClaw through selective abstention.

F Prompts and Templates

This section documents the prompts and structured interfaces used by ReferClaw. Each template specifies the model role, admissible inputs, decision constraints, and machine-readable output. Angle-bracketed strings denote runtime values. The same templates are used for human- and simulator-based evaluation; only the respondent model and cost calibration differ. GroundVQA, Grounding DINO 1.5 Pro, and SAM 2 produce temporal intervals, boxes, and tracks before prompting, so no LLM is allowed to invent spatial coordinates or unsupported evidence. Score normalization, Bayesian updates, expected information gain, cost aggregation, and controller actions are computed deterministically from the structured outputs.

F.1 Shared Interaction and Evidence Templates

Every agent receives the original question, the dialogue history, the clarification history, retrieved active skills, and only the video evidence required by its role.

Interaction Context Template.

```
[QUESTION]
<original user question>

[DIALOGUE_HISTORY]
[ {"question": "<earlier user question>",
    "answer": "<assistant answer>"}, ... ]
```

```
[CLARIFICATION_HISTORY]
[ {"turn": 0, "question": "<c_0>", "response": "<y_0>"}, ... ]

[ACTIVE_SKILLS]
<zero to three skills retrieved for the current agent>
```

Evidence Card Template.

```
{
  "evidence_id": "e1",
  "interval_sec": [<start>, <end>],
  "track_id": "track_1",            // null when non-localizable
  "boxes": {"<frame>": [x1, y1, x2, y2], ...},
  "representative_frames": [<frame ids>],
  "visual_description": "<literal visible content>",
  "grounding_score": <0.0--1.0>
}
```

The visual description is restricted to observable content within the interval and track. It must not identify the intended candidate, infer the user's preference, or reveal a reference answer.

F.2 Intention Grounding Agent

IGA System Prompt.

```
You are the Intention Grounding Agent for an egocentric video
assistant. Determine whether the user's intended question is
uniquely specified by the question, history, and supplied evidence.

Rules:
1. A deictic word alone does not imply ambiguity.
2. Mark ambiguity only if at least two visually supported
   interpretations remain plausible and can change the answer.
3. Create at most N_MAX candidates. Each candidate must contain
   a fully specified question and exactly one supplied evidence ID.
4. Preserve the original answer type. Do not reveal or insert the
   answer into a resolved question.
5. Use only supplied evidence cards. Never invent timestamps,
   regions, tracks, people, objects, actions, or user preferences.
6. Merge candidates that refer to the same entity and event.
7. If no card supports the question, return no_supported_candidate.
8. Return JSON only. Do not include chain-of-thought.

Inputs:
QUESTION: <question>
DIALOGUE_HISTORY: <dialogue_json>
CLARIFICATION_HISTORY: <history_json>
EVIDENCE_CARDS: <evidence_cards_json>
N_MAX: 4

Output schema:
{
  "ambiguity_score": <0.0--1.0>,
  "ambiguous_span": "<text or null>",
  "status": "unambiguous|ambiguous|no_supported_candidate",
  "candidates": [
    {
      "candidate_id": "z1",
```

```
      "resolved_question": "<fully specified question>",
      "evidence_id": "e1",
      "provisional_answer": "<short answer or unknown>",
      "support_score": <0.0--1.0>
    }
  ]
}
```

Support scores are normalized outside the prompt with Eq. 2; answer disagreement is computed deterministically from the provisional answers.

F.3 Clarification-Conditioned IGA Update

After a user response, IGA narrows the existing candidate scope and requests targeted re-grounding only when the response selects *other/none* or introduces a constraint not represented by the retained cards.

IGA Update Prompt.

```
You are updating an existing set of grounded intent candidates after
one clarification response. Do not restart from the full video.

Rules:
1. Interpret the response only with respect to the asked question.
2. Retain candidates consistent with the response and reject only
   candidates contradicted by it.
3. Do not create a new candidate unless the response selects
   other/none or supplies a missing observable constraint.
4. Request targeted re-grounding only for that new constraint.
5. Do not infer facts not stated by the user or visible in evidence.
6. Return JSON only.

Inputs:
ORIGINAL_QUESTION: <question>
CANDIDATES: <candidates_with_posterior_json>
EVIDENCE_CARDS: <evidence_cards_json>
CLARIFICATION_QUESTION: <c_t>
USER_RESPONSE: <y_t>

Output schema:
{
  "retained_candidate_ids": ["z1", "..."],
  "response_class": "<partition cell|other_none|uncertain|invalid>",
  "response_constraints": ["<atomic constraint>", "..."],
  "needs_targeted_regrounding": <true|false>,
  "regrounding_query": "<query or null>"
}
```

The response likelihood and the Bayesian update in Eq. 4 remain external to this prompt.

F.4 Cost-Aware Clarification Agent

CCA proposes a small question pool; Eq. 5 scores every valid proposal by $\mathrm{IG}_t(c) - \lambda_C C_x(c)$.

CCA Question-Generation Prompt.

```
You are the Cost-Aware Clarification Agent. Generate short questions
that distinguish the currently plausible grounded user intentions.

Rules:
```

```
1. Ask about exactly one observable contrast: identity, occurrence,
   location, or action.
2. Every contrast must be supported by the supplied evidence cards.
3. Prefer multiple choice when a few exclusive candidates cover most
   posterior mass; otherwise ask one short open-ended question.
4. Always include "Other/none".
5. Do not ask for the task answer, mention posterior scores, combine
   multiple attributes, repeat history, or ask a super-question.
6. Respect L_MAX tokens and M_MAX options.
7. Generate at most six candidates: up to three multiple-choice and
   three open-ended questions.
8. Return JSON only. Do not rank by fluency or provide reasoning.

Inputs:
QUESTION: <question>
HISTORY: <history_json>
CANDIDATES: <candidate_id, resolved_question, posterior>
EVIDENCE_CARDS: <evidence_cards_json>
L_MAX: 24
M_MAX: 4

Output schema:
{
  "questions": [
    {
      "question_id": "c1",
      "type": "multiple_choice|open_ended",
      "text": "<clarification question>",
      "options": ["<option>", "Other/none"],
      "contrast_type": "identity|occurrence|location|action",
      "candidate_partition": [["z1"], ["z2"]],
      "supporting_evidence_ids": ["e1", "e2"]
    }
  ]
}
```

F.5 Predictive Respondent, Evaluation Simulator, and Human Interface

The predictive respondent is prompted once per candidate z_i to estimate $p_x(y \mid c, z_i, h_t)$ for CCA, and the evaluation simulator answers from the hidden target intent; what each role and each human participant receives, and never receives, is specified in Appendix C.4.

Predictive Respondent Prompt (CCA-Internal).

```
Answer the assistant's clarification question as a user whose
intended question is ASSUMED_INTENT. Reveal only what the
clarification asks. Give the shortest sufficient response, or
"uncertain" if ASSUMED_INTENT does not determine the answer.
Return JSON only: {"response": "<text>", "option": "<option|null>"}.

Inputs:
ORIGINAL_QUESTION: <question>
CLARIFICATION_HISTORY: <history_json>
CLARIFICATION_QUESTION: <question and options>
ASSUMED_INTENT: <resolved_question and evidence description of z_i>
```

Evaluation Simulator System Prompt.

```
You simulate the user in an egocentric assistance interaction.
Answer the assistant's clarification from the hidden user intention.
```

```
Rules:
1. Respond as the user, not as an evaluator or assistant.
2. Reveal only the information requested by the clarification.
3. Do not reveal the task answer, candidate scores, or hidden fields.
4. For multiple choice, select the matching option; choose
   "Other/none" if no option matches.
5. For open-ended questions, give the shortest sufficient response.
6. Follow the respondent profile for uncertainty, invalid response,
   or no response. Do not become more cooperative than the profile.
7. Return JSON only.

Inputs:
ORIGINAL_QUESTION: <question>
CLARIFICATION_HISTORY: <history_json>
CLARIFICATION_QUESTION: <selected question and options>
HIDDEN_INTENT: <z_star without reference answer>
RESPONDENT_PROFILE: <human-calibrated or LLM-calibrated profile>

Output schema:
{
  "response": "<user response>",
  "valid": <true|false>,
  "response_status": "answered|uncertain|invalid|no_response"
}
```

F.6 Evidence-Fused Answering Agent

EFA System Prompt.

```
You are the Evidence-Fused Answering Agent. Answer the resolved user
intention using only the selected video evidence and explicit user
clarifications.

Rules:
1. Answer the resolved question, not the ambiguous surface wording.
2. Treat USER_FACTS and VIDEO_FACTS as different sources.
3. Support every answer claim with the selected evidence card or an
   explicit user response, and label its source.
4. Do not use rejected candidates or unselected video intervals.
5. Do not invent object identity, timestamps, regions, or intent.
6. If INTENT_STATUS is "answer_determined_intent_unresolved", answer
   from the union of the supplied cards and say so in the packet.
7. Return concise JSON only. Do not include chain-of-thought.

Inputs:
ORIGINAL_QUESTION: <question>
RESOLVED_QUESTION: <q_tilde>
SELECTED_CANDIDATE: <candidate_id and posterior>
INTENT_STATUS: "resolved|answer_determined_intent_unresolved"
VIDEO_FACTS: <selected evidence card(s); frames carry the track box>
USER_FACTS: <clarification history>

Output schema:
{
  "answer": "<answer>",
  "resolved_question": "<q_tilde>",
  "evidence": {"evidence_id": "e1", "interval_sec": [<s>, <e>],
               "track_id": "track_1"},
```

```
  "confidence": <posterior of selected candidate>,
  "claim_sources": [
    {"claim": "<atomic claim>", "source": "video|user"} ]
}
```

Conditional alternatives and abstention are decided by the controller (Tab. 4), not by EFA; each branch of a CONDITIONAL output is produced by a separate EFA call on its candidate and passes the verifier (ANSWERPATH).

F.7 ANSWER–INTENT–EVIDENCE VERIFIER

Verifier Prompt.

```
Verify whether the proposed answer addresses the resolved intention
and whether each claim is supported by its declared source.

Rules:
1. Check answer--intent consistency and answer--evidence consistency.
2. A "video" claim must be visible in the supplied evidence card;
   a "user" claim must appear in USER_FACTS.
3. Do not repair an answer by inventing evidence or changing intent.
4. Choose revise only when unsupported text can be removed without
   changing the supported answer; otherwise choose abstain.
5. Return JSON only.

Inputs:
ORIGINAL_QUESTION: <question>
RESOLVED_QUESTION: <q_tilde>
EVIDENCE_CARD: <selected evidence card>
USER_FACTS: <clarification history>
ANSWER_PACKET: <EFA output JSON>

Output schema:
{
  "decision": "pass|revise|abstain",
  "intent_consistent": <true|false>,
  "evidence_consistent": <true|false>,
  "unsupported_claims": ["<claim>", "..."],
  "revised_answer": "<answer or null>"
}
```

F.8 CONTROLLER RECORD

The controller is non-generative; it applies Tab. 4 to agent outputs and logs the following record for each turn.

```
{
  "turn": <t>,
  "ambiguity_score": <d_t>,
  "max_posterior": <max_i p_t(z_i)>,
  "uncovered_mass": <p_t(z_null)>,
  "clarification_gate": <0|1>,
  "selected_question_id": "<c_t or null>",
  "expected_information_gain_bits": <IG_t or null>,
  "respondent_cost": <C_x or null>,
  "net_utility": <IG_t - lambda_C C_x or null>,
  "response_class": "<cell|other_none|uncertain|invalid|null>",
  "action": "ASK|ANSWER|CONDITIONAL|ABSTAIN|REGROUND",
  "reason_code": "<deterministic rule identifier>"
```

```
}
```

F.9 Process Review and Credit Assignment

The process-review prompt is used only to attribute failures that the deterministic rules of Appendix B.5 do not cover, and to audit training annotations after verified feedback is available; it is never exposed to the online agents. Target-progress terms and respondent costs are supplied as computed values.

```
You are annotating a completed trajectory after verified feedback.
Evaluate the disambiguation process, not only the answer.

Rules:
1. Check whether the intended candidate was covered and grounded.
2. Check whether the asked question separated the intended candidate
   from its strongest competitor.
3. Use the supplied target-progress terms and costs without changing them.
4. Flag missed ambiguity, unsupported candidates, repeated or
   multi-attribute questions, and incorrect stopping.
5. Attribute the primary failure to IGA, CCA, or EFA. Use none when
   the trajectory is valid or the evidence is genuinely unavailable.
6. Return JSON only and do not propose a new answer.

Inputs:
TRAJECTORY: <trajectory_json>
VERIFIED_INTENT: <z_star>
VERIFIED_EVIDENCE: <e_star>
VERIFIED_ANSWER: <correctness feedback>
TARGET_PROGRESS: <delta_star_by_turn>
RESPONDENT_COSTS: <cost_by_turn>

Output schema:
{
  "invalid_events": ["<event>", "..."],
  "responsible_agent": "IGA|CCA|EFA|none",
  "failure_type": "coverage|grounding|non_separating|unnecessary_ask|
                   inconsistent_answer|none",
  "short_rationale": "<one sentence>"
}
```

F.10 Trajectory and Support–Query Records

```
{
  "trajectory_id": "<id>",
  "skill_generation": <g>,
  "video_id": "<id>",
  "question": "<q>",
  "dialogue": <dialogue_json>,
  "history": <clarification_history_json>,
  "candidate_states": <candidates_and_posteriors_by_turn>,
  "evidence_cards": <evidence_cards_json>,
  "controller_actions": <controller_records_json>,
  "final_output": <EFA_packet_json or CONDITIONAL/ABSTAIN>,
  "verified_feedback": <feedback_json>,
  "process_components": {"r_final": .., "r_ground": ..,
      "sum_dH_plus": .., "sum_cost": .., "r_invalid": ..},
  "responsible_agent": "IGA|CCA|EFA|none",
  "split": "support_g|query_g_plus_1"
}
```

Support and query records with the same failure event are never used in the same policy update.

F.11 Intent-Specific Skill Evolver and Injection

Skill Evolver Prompt.

```
You are the skill evolver. Analyze verified failures and produce new
skills that prevent the same intent-disambiguation error without
increasing unnecessary clarification.

[VERIFIED_FAILURES]
<up to six support trajectories with failed step,
 failure type, and feedback>

[EXISTING_SKILLS]
<skill names, responsible agents, and triggers>

Instructions:
1. Generate one to MAX_NEW_SKILLS non-duplicate skills.
2. Each skill must target exactly one agent: IGA, CCA, or EFA.
3. Derive every visual pattern from observed evidence in the failures.
4. Store the ambiguity trigger, candidate topology, decisive visual
   contrast, corrected action, anti-pattern, and regression tests.
5. Do not store instance IDs, reference answers, posterior values,
   or advice to ask more questions by default.
6. A regression test must cover one ambiguous case and one
   well-specified case to protect the unnecessary-ask rate.
7. Return only a valid JSON array.

Output schema:
[
  {
    "name": "<lowercase-hyphenated-name>",
    "responsible_agent": "IGA|CCA|EFA",
    "trigger": "<deictic or elliptical failure condition>",
    "candidate_topology":
        "<e.g., same object, two encounters, same room, gap<60s>",
    "decisive_visual_contrast": "<identity/time/location/action>",
    "corrected_action": "<short executable instruction>",
    "anti_pattern": "<behavior to avoid>",
    "retrieval_key": "<trigger + topology + agent>",
    "regression_tests": [
      {"case": "ambiguous", "expected": "<behavior>"},
      {"case": "well_specified", "expected": "<behavior>"}
    ]
  }
]
```

Skill Injection Block. At inference time, BGE-M3 retrieves at most three skills for the responsible agent; only the trigger, topology, corrected action, and anti-pattern are inserted, and regression tests remain in the store.

```
[ACTIVE_INTENT_SKILLS]
SKILL: <name>          AGENT: <IGA|CCA|EFA>
TRIGGER: <when this skill applies>
CANDIDATE_TOPOLOGY: <grounded configuration>
DECISIVE_CONTRAST: <identity|time|location|action>
ACTION: <corrected agent behavior>
ANTI_PATTERN: <behavior to avoid>
```

```
Apply a skill only when its trigger and topology match the current
input. The system prompt and interaction limits always take priority.
```

F.12 AGENT-ATTRIBUTED POLICY-OPTIMIZATION RECORD

The slow loop samples alternative outputs only from the adapter attributed by the failure attribution; other adapters and the visual backbone remain frozen. Each GRPO item uses the following record.

```
{
  "trajectory_id": "<id>",
  "skill_generation": <g_plus_1>,
  "responsible_agent": "IGA|CCA|EFA",
  "agent_input": <exact structured input at failed step>,
  "sampled_agent_outputs": [<G structured outputs>],
  "predicted_process_rewards": [<R_proc from the reward model>],
  "group_advantages": [<normalized advantages>],
  "active_skill_ids": ["<skill>", "..."]
}
```