WHO MISSED THE JUMP? EVIDENCE-GROUNDED CREDIT ASSIGNMENT FOR MULTI-AGENT FIGURE SKATING REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable video reasoning requires both a correct answer and the evidence that supports it. In figure skating, a brief takeoff or landing interval can determine a judgment, yet answer accuracy and evidence localization need not improve together. Dividing the task among cooperating agents introduces a second ambiguity: an outcome reward does not identify where evidence delivery failed or which failures need correction. Our approach makes explicit evidence requirements a common reference for evaluation, diagnosis, and training. We introduce **SKATEEG**, with 7,200 video-grounded questions over 1,194 programs and 2,400 text-only controls. Its annotated records specify where to find the evidence and whether records are jointly required or sufficient alternatives. Our **EDGE** framework first diagnoses each record's earliest confirmed failure in request, acquisition, description, or use, then selects sufficient-path targets for negative task credit and localized hindsight. This selection avoids correcting unused alternatives when another path already supplies complete support. Positive task credit reinforces supported delivery, while unresolved targets receive zero task weight. On 600 independent failures, the first-break nominee matches repair-replay point rankings in 86.0% of cases, versus 62.5% for a same-information judge, without additional trajectory executions for diagnosis. Across three training seeds, the 7B system reaches 63.2% accuracy and 53.2% GroundedAcc, the proportion of test questions answered correctly with complete evidence support, exceeding MATPO by 2.9 and 8.3 percentage points, respectively. Matched negative-credit and alternative-path controls support the value of the target-selection rules. Together, these results connect explicit evidence requirements to learning from failed handoffs in a three-agent pipeline with annotated training evidence. Code and data will be released.

1 INTRODUCTION

Figure skating reasoning depends on brief, decisive events, such as blade edge at takeoff or free-foot contact at landing. A video model must locate these events, describe them faithfully, and connect them to a judgment. Yet answer accuracy and evidence localization can rank systems differently (Xiao et al., 2024). Among systems not trained on SKATEEG, Gemini 2.5 Pro reaches 57.1% accuracy with an evidence-window IoU of 0.35. MUPA reaches a higher IoU of 0.40 at 51.5% accuracy (Figure 1, upper right), motivating evaluation beyond answer correctness.

Multi-agent pipelines add a credit-assignment problem: an incorrect answer may originate before the answerer acts. Requesting landing frames for a takeoff-edge question can produce a faithful description that omits the needed evidence. An outcome reward does not distinguish the mistaken request from that faithful response. Existing approaches allocate credit across roles (Mo et al., 2025), score steps retrospectively (Tan et al., 2026; Li et al., 2026a), or infer information flow (Ren et al., 2026). Counterfactual methods resample actions at fixed histories (Chen et al., 2026b) or remove agents (Li et al., 2026b). We use explicit, question-specific evidence requirements to diagnose failed handoffs from logged requests, records, descriptions, and citations.

These requirements must distinguish joint support (AND) from sufficient alternatives (OR). Missing an AND record leaves support incomplete, whereas an unused OR alternative needs no correction when another path suffices. If every alternative fails, restoring one sufficient path can resolve the

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.

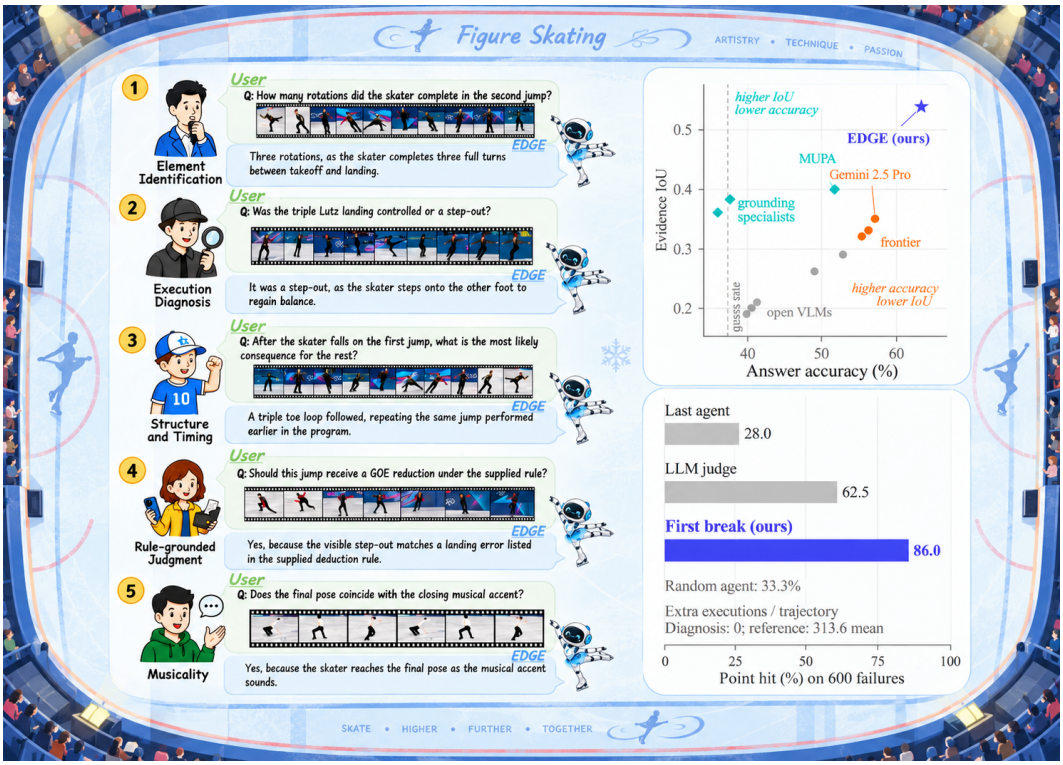

Figure 1: **Tasks and motivation in SKATEEG. Left:** five video-grounded tasks. **Upper right:** accuracy and evidence localization of systems not trained on SKATEEG, with trained EDGE for reference (Table 1). **Lower right:** attribution agreement and execution costs on 600 independent failures (Table 3; protocol: Appendix E.1).

deficit. We therefore separate record-level failure diagnosis from the selection of sufficient-path training targets.

We introduce **SKATEEG** to provide this common reference for evaluation and learning. It contains 7,200 video-grounded questions over 1,194 complete programs and 2,400 text-only rule controls. Existing figure skating benchmarks evaluate elements, execution, and artistry (Gao et al., 2025); SKATEEG adds source-grounded records specifying temporal windows, decisive-jump phases, modalities, and claims. Audited AND/OR annotations connect these records to complete answer support, enabling separate correctness, localization, and support evaluations while guiding training.

Our framework, **EDGE** (**E**volving agents with **D**elivery-**G**rounded **E**vidence attribution), traces each record through a planner–perception–reasoner pipeline. It identifies the first confirmed failure after confirmed passes in request, acquisition, description, or use, without separately penalizing downstream consequences. Sufficient-path selection retains confirmed AND breaks and, when all OR alternatives are confirmed broken, selects the record with the longest confirmed passing prefix. Break-Routed Policy Optimization (BRPO) routes negative task credit to selected actions and gates positive credit on complete evidence support; unresolved targets receive zero task weight. Link-Localized Hindsight Distillation (LHD) supplies the same selected actions with corrective information about their failed records.

Independent evaluation tests diagnosis and training in a single-pass, three-agent pipeline with annotated training evidence. On 600 failures, first-break agreement with repair-replay point rankings reaches 86.0%, versus 62.5% for a same-information judge and 28.0% for last-agent assignment (Figure 1, lower right). All three diagnosis methods require no extra trajectory executions; the adaptive replay reference averages 313.6 per case, while independent link annotations assess failure localization. Across three seeds, the 7B system exceeds MATPO by 2.9 accuracy points and 8.3 GroundedAcc points, the latter measuring correct, fully supported answers among all test questions. With positive support gating and LHD fixed, aligned negative credit gains 2.8 GroundedAcc points

over negative-only shuffle. Sufficient-path selection gains 7.4 GroundedAcc points over all-record targeting on 360 multi-path OR questions, jointly testing negative-credit and hindsight targets.

Our contributions are:

- **Evidence requirements for evaluation and learning.** SKATEEG provides record-level references and AND/OR support structures that connect complete answer support to the evidence needed during training.
- **Training targets from failed evidence handoffs.** EDGE combines first-break diagnosis with sufficient-path selection to determine negative-credit recipients and localized-hindsight content, while complete support governs positive credit.
- **Separate validation of diagnosis and learning.** Independent repair replay and link annotations evaluate the diagnostic targets; matched negative-credit and alternative-path controls test the training choices alongside gains in correct, fully supported answers.

2 SKATEEG: EVIDENCE REQUIREMENTS FOR FIGURE SKATING

SKATEEG provides question-specific references for both answer correctness and the evidence needed to support a judgment. It contains $7{,}200$ video questions over $1{,}194$ complete programs and $11{,}998$ technical element instances, plus $2{,}400$ text-only rule controls. The five dimensions are element identification (C1; $1{,}800$), execution diagnosis (C2; $1{,}800$), structure and timing (C3; $1{,}400$), rule-grounded judgment (C4; $1{,}400$), and musicality (C5; 800), illustrated on the left of Figure 1. Together, these tasks connect brief, phase-sensitive observations to element-level judgments, program structure, and movement–music alignment. Complete programs preserve the temporal context needed to compare elements and relate local events to program-level questions. The annotations make these evidence requirements explicit, supporting both system evaluation and diagnosis of failed evidence delivery. Fixed splits separate programs; corpus, licensing, inventory, and annotation details appear in Appendices B.1–B.4.

Records and support. An evidence record comprises a window, phase, single modality, and claim; the first three fields form its address. Resolved video questions have a nonempty reference set $E^*(q)$ with pure AND/OR support. For $\sigma(q) = \vee$, its records are alternative sufficient paths; for $\sigma(q) = \wedge$, the full set suffices and every record is necessary. Singletons use $\vee$; mixed or unconfirmed structures remain unresolved (Appendix B.4).

For movement–music alignment, visual and audio records may be jointly required; for element identification, RGB or pose may each suffice. The support labels distinguish missing necessary evidence from unused alternatives, which modality unions alone cannot resolve. These references support comparisons of evidence retrieval and complete answer support across systems, independently of the credit-assignment algorithm. Text-only controls have $E^*(q) = \emptyset$, are excluded from training, and permit empty evaluation requests without obtaining video support by vacuity.

Annotation and sufficiency. Annotators with skating competition or coaching experience review every question, with disagreements adjudicated by senior reviewers with judging or technical-panel experience. Pre-adjudication agreement reaches $\kappa = 0.87$ for answer validity and 0.82 for video-evidence support (Appendix B.4). Decisive jumps additionally receive source-grounded phase annotations. Mean takeoff and landing disagreement is 1.3 and 1.6 frames (about 43 and 53 ms at 30 fps), compared with 11.3 frames for conventional entry onset. These visually defined boundaries anchor the evidence windows, making temporal coverage checks sensitive to the brief events that determine technical judgments.

Evaluation controls. Audio commentary is removed and score overlays are masked in the benchmark. The template-only probe scores 27.3% against 25.0% chance. Empty-evidence answer rates span 30.5–43.7% across dimensions (Appendix B.6); these priors do not establish the theory's partial-evidence bounds. Distinct metrics assess accuracy, localization, citation, and complete support (Appendix B.7); rule controls are scored separately. Initialization, calibration, and confirmation trajectories are kept separate (Appendix E.1).

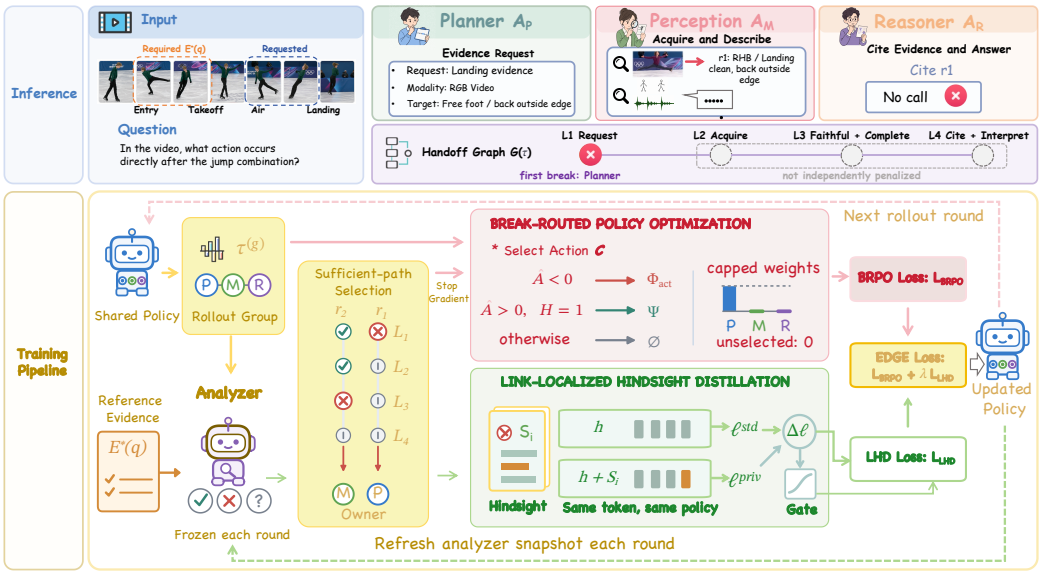

Figure 2: **EDGE diagnosis and training. Top:** A missed reference window yields a planner L_1 break without separate downstream penalties. **Bottom:** Sufficient-path selection filters frozen diagnoses for BRPO/LHD. BRPO routes negative credit to $\Phi_{\text{act}} \equiv \Phi_{\text{act}}^{\text{train}}$ and positive credit to Ψ when $H = 1$; otherwise, task weights are zero. LHD scores the same token with/without hindsight: $\Delta\ell$ is the privileged-minus-ordinary log-probability difference; λ weights LHD. Dashed arrows refresh rollouts and analyzer snapshots.

3 METHOD

EDGE turns evidence-delivery failures into training targets through first-break diagnosis and sufficient-path selection (Figure 2). BRPO assigns negative credit to selected actions and positive credit to fully supported delivery chains. LHD provides the same selected actions with corrective information about their failed links.

3.1 EVIDENCE ADDRESSES AND FIRST-BREAK ATTRIBUTION

Three agents $\mathcal{A} = \{A_P, A_M, A_R\}$ plan, acquire evidence, and answer a video question q. Agent i samples $a_{i,t} \sim \pi_{\theta_i}(\cdot \mid h_{i,t})$ from its local history, yielding joint trajectory τ and binary reward $R_{\text{task}}(\tau) = \mathbb{1}[\hat{y} = y^*]$, abbreviated $R(\tau)$. A record $r = \langle w, \phi, m, c \rangle$ contains a window, phase, single modality, and claim, with address $\text{addr}(r) = \langle w, \phi, m \rangle$. Each question uses $E^*(q)$ and $\sigma(q)$ from Section 2; cross-modal requirements use separate records. Perception returns record identifiers that the Reasoner cites when constructing its answer.

Requests, records, and consumer references define a directed handoff graph $G(\tau)$. For each reference record $r^* \in E^*(q)$, L_1 checks request coverage and L_2 checks acquisition of a matching address. Fixed temporal and modality tests operate on parsed fields, with phase encoded through annotated windows rather than a separate name test (Appendix C.2). Model-judged L_3 checks faithful description and inclusion of all decisive facts; L_4 checks citation and correct interpretation by A_R. Each predicate returns pass (1), fail (0), or unknown (?). We locate the first confirmed failure after confirmed passes as

$$k^*(r^*) = \min\{k : L_k(r^*) = 0,\ L_j(r^*) = 1 \text{ for all } j < k\} \tag{1}$$

when the set is nonempty; unknown predecessors block downstream attribution. Fully passed and unresolved records remain distinct in the diagnostic log. Break owners are $\text{own}(1) = A_P$, $\text{own}(2) = \text{own}(3) = A_M$, and $\text{own}(4) = A_R$; downstream links on the same record receive no separate penalty. One analyzer pass collects $(k^*, \text{own}(k^*), t, r^*)$ into the record-level diagnosis $\Phi^{\text{diag}}(\tau)$, including unused OR alternatives. Training then filters these diagnoses through the question's sufficient paths.

Repair optimality additionally depends on intervention scope and the post-repair evidence. For a fixed question, policy, and common replay prefix, $V_\tau^O(S)$ is expected success after oracle O repairs agents S, with $\Delta_i = V_\tau^O(\{i\}) - V_\tau^O(\emptyset)$. Minimal repair preserves valid outputs exactly; full replacement can change them, such as by narrowing a valid request window. Let p_0 denote the ordinary Reasoner's question-specific empty-evidence success rate, and p_{suff} a lower success bound under sufficient-evidence repair contexts.

Proposition 1 (First-break optimality under a single effective bottleneck)**.** *Fix $j = \mathrm{own}(k^*(r^*))$. Under the structural and distributional conditions* (A1)–(A4) *in Appendix C.1, the baseline lacks sufficient evidence, repair of j restores it with probability at least π_j, and other interventions either have gain at most u or remain evidence-insufficient. Let the baseline success lie within δ of p_0. At evidence-insufficient contexts, competing non-upstream repairs have success at most $p_0+\delta$, whereas repair of j has success at least $p_0 - \delta_j^-$, with $\delta_j^- \geq 0$. If*

$$\pi_j(p_{\mathrm{suff}} - p_0) - (1-\pi_j)\delta_j^- - \delta > \max\{u, 2\delta\}, \tag{2}$$

then $\arg\max_{i\in\mathcal{A}} \Delta_i = \{j\}$.

The guarantee concerns whole-answer support, rather than correction of one predicate. Exact preservation gives $u = 0$ for valid upstream outputs; passing a coverage predicate alone does not establish this. Appendix C.1 gives the proof and special cases, including residual refusal; Appendix E audits the conditions. Multiple broken records can implicate several agents; joint-repair and member-removal tests account for alternative OR paths. These candidates become training targets through the following sufficient-path selector.

3.2 Break-Routed Policy Optimization

BRPO combines group-relative advantages (Shao et al., 2024; Liu et al., 2026) with evidence-dependent action selection. At iteration k, frozen policies $\bar{\theta}_i^k$ generate G complete trajectories for each question. Only binary task correctness contributes to the normalized group-relative advantage:

$$\hat{A}^{(g)} = \frac{R_{\mathrm{task}}^{(g)} - \mu_q}{\sigma_q + \epsilon}, \quad \mu_q = \frac{1}{G}\sum_g R_{\mathrm{task}}^{(g)}, \quad \sigma_q^2 = \frac{1}{G}\sum_g (R_{\mathrm{task}}^{(g)} - \mu_q)^2. \tag{3}$$

Format and citation alignment use supervised training, without contributing reinforcement reward. All-correct and all-incorrect groups have zero task advantage, but remain eligible for evidence diagnosis and LHD (Appendix C.3).

The frozen analyzer collects reference records with all L_1–L_4 checks passed in $D(\tau)$ and their witnessed delivery actions in $\Psi(\tau)$. Training uses resolved, nonempty references with minimal sufficient sets $\mathcal{S}_q$ (Appendix B.7), giving the support gate $H(\tau) = \mathbb{1}[\exists S \in \mathcal{S}_q : S \subseteq D(\tau)]$. Thus OR requires $D(\tau) \neq \emptyset$, whereas AND requires $D(\tau) = E^*(q)$ for complete support.

A failed record need not imply missing support, because an alternative OR path may suffice. Set $S^\dagger = \emptyset$ whenever $H = 1$, including when an unused alternative is unresolved. When $H = 0$, AND retains its full reference set, while OR selects one record only if every alternative is confirmed broken. The OR choice maximizes the confirmed passing prefix, breaking ties by fixed record ID; unresolved alternatives cause abstention. Restricting Φ^{diag} to records in $S^\dagger$ yields Φ^{train}, with action projection $\Phi_{\mathrm{act}}^{\mathrm{train}}$. This record-level filtering governs negative credit and hindsight; unused alternatives remain in the diagnosis (Appendix C.3). Unknown links supply no targets, but confirmed AND breaks remain eligible when another record is unresolved. The task-advantage sign then determines which actions receive task credit:

$$\mathcal{C}(\tau^{(g)}) = \begin{cases} \Phi_{\mathrm{act}}^{\mathrm{train}}(\tau^{(g)}), & \hat{A}^{(g)} < 0, \\ \Psi(\tau^{(g)}), & \hat{A}^{(g)} > 0 \text{ and } H(\tau^{(g)}) = 1, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{4}$$

Confirmed insufficient or unresolved support blocks positive credit; correct uncited answers have $H = 0$ and receive no positive task credit.

Each complete role turn is one trainable action, giving $M_g = 3$ in our single-pass rollout; tool executions add no trainable actions. With $N_g = |\mathcal{C}(\tau^{(g)})|$, the selected actions receive the weights

$$\widetilde{w}_{i,t}^{(g)} = \begin{cases} \dfrac{M_g}{N_g}\mathbb{1}[(i,t) \in \mathcal{C}(\tau^{(g)})], & N_g > 0, \\ 0, & N_g = 0, \end{cases} \quad w_{i,t}^{(g)} = \min\{\widetilde{w}_{i,t}^{(g)}, w_{\max}\}, \quad A_{i,t}^{(g)} = \hat{A}^{(g)} w_{i,t}^{(g)}. \tag{5}$$

For nonempty selections, $\sum_{i,t} \widetilde{w}_{i,t}^{(g)} = M_g$ before clipping; empty sets give zero task weight without uniform fallback. Nonnegative weights preserve the advantage sign, $A_{i,t}^{(g)} \hat{A}^{(g)} \geq 0$, while allowing its magnitude to become zero (Ren et al., 2026). Selections, the support gate, and weights are detached from optimization. The clipped task objective broadcasts $A_{i,t}$ over generated tokens and masks environment tokens; Appendix C.3 specifies asymmetric clipping and KL regularization. Zero task weight leaves LHD, regularization, and updates through other trajectories or shared parameters possible.

3.3 Link-Localized Hindsight Distillation

LHD uses $\Phi^{\text{train}}(\tau)$ to provide selected actions with localized corrective information. For $(i,t) \in \Phi_{\text{act}}^{\text{train}}(\tau)$, hindsight S_i contains only that action's selected records and failed predicates. An L_1 break supplies the audited phase, window, and sufficient modalities; L_2 supplies the issued request and the evidence obtainable within it. For L_3, hindsight identifies false or omitted decisive facts; for L_4, it supplies delivered records and their correct reading (Table 8). Eligibility follows the sufficient-path filter regardless of task-advantage sign or variance, including for correct unsupported answers.

LHD combines on-policy distillation (Agarwal et al., 2024) with the confidence-gated scoring used in SEED (Wu et al., 2026). The same policy scores each sampled token $a_{i,t,\ell}$ from two contexts without resampling: ordinary $h_{i,t}$ and hindsight-augmented $\tilde{h}_{i,t} = \text{Augment}(h_{i,t}, S_i)$. Their log-probabilities ℓ^{std} and ℓ^{priv} define $\Delta = \text{sg}[\ell^{\text{priv}} - \ell^{\text{std}}]$ and $g = \sigma(\beta_{\text{lhd}}\Delta)$. This gate emphasizes tokens endorsed by hindsight in the on-policy loss:

$$\mathcal{L}_i^{\text{LHD}} = \mathbb{E}\Big[m_{i,t}\, m_\ell^{\text{tok}}\, g_{i,t,\ell}\big(\text{sg}[\ell_{i,t,\ell}^{\text{priv}}] - \ell_{i,t,\ell}^{\text{std}}\big)\Big], \tag{6}$$

where $m_{i,t} = \mathbb{1}[(i,t) \in \Phi_{\text{act}}^{\text{train}}(\tau)]$ selects filtered break actions and m^{tok} masks prompt and padding tokens. The privileged branch is detached, so gradients reach only the ordinary policy. Each agent minimizes $\mathcal{L}_i = \mathcal{L}_i^{\text{BRPO}} + \lambda_{\text{lhd}}\mathcal{L}_i^{\text{LHD}}$; decaying λ_{lhd} makes hindsight recede as training progresses.

3.4 Analyzer and Implementation

The analyzer is initialized on 150 annotated trajectories, frozen within each training round, and refreshed after each update. It extracts fields and judges semantic predicates against source-grounded facts; code executes fixed checks and assembles the first break. Generated labels cannot override fixed checks, and unknown judgments remain distinct from confirmed failures. Appendix C.4 details initialization and monitoring; Section 4.3 separates input errors from predicate execution, and Appendix F evaluates support and delivery localization.

Each question follows $A_P \rightarrow A_M \rightarrow A_R$ without revisiting preceding actions, making $G(\tau)$ acyclic. Agents share a backbone with role-specific low-rank adapters; Appendix D.4 tests independent backbones. All trained methods receive the same supervised record-format and citation alignment. Appendices C.2 and C.3 specify predicates, loss normalization, hyperparameters, and the training loop.

4 Experiments

4.1 Setup

Evaluation uses the fixed program-level test split of $1{,}440$ video questions, with $2{,}400$ rule controls scored separately. Trained baselines include single-agent GRPO, Video-R1 (Feng et al., 2025),

Table 1: **Accuracy and evidence localization.** Test results on 1,440 video questions (C1–C5: Section 2). Trained rows show three-seed means; ± denotes sample SD of overall accuracy. IoU is undefined without evidence localization; matching follows Appendix B.7.

Method	Backbone	Task accuracy (%) ↑					Overall (%)	Evidence
		C1	C2	C3	C4	C5	Acc. ↑	IoU ↑
Proprietary, zero-shot								
GPT-5	—	55.6	55.8	60.0	60.0	45.0	56.2	0.33
Gemini 2.5 Pro	—	55.0	57.2	59.3	59.3	53.8	57.1	0.35
Not trained on SKATEEG								
VideoMultiAgents	Gemini	58.6	58.1	58.2	59.6	55.0	58.2	—
MUPA	7B	50.3	51.7	55.4	55.0	41.3	51.5	0.40
EDGE, untrained	Gemini	58.9	60.0	60.7	60.4	58.1	59.7	0.45
Trained on SKATEEG								
Video-R1	7B	62.2	57.0	51.9	54.8	47.5	55.8 ± 0.9	—
Single agent	7B	64.8	58.8	54.8	56.8	51.3	58.3 ± 0.8	0.41
SEED	7B	64.9	59.3	55.0	57.6	51.3	58.6 ± 0.9	0.46
MAGRPO	7B	64.6	58.1	53.3	55.8	49.8	57.4 ± 0.7	0.44
MASIRL	7B	65.2	59.6	55.8	56.4	53.1	58.9 ± 0.8	0.49
MATPO	7B	65.8	60.8	57.9	58.3	54.8	60.3 ± 0.6	0.47
C3 (adapted)	7B	66.1	61.4	58.6	58.9	55.6	60.9 ± 0.6	0.50
IAPO (adapted)	7B	65.8	60.8	57.9	58.9	54.4	60.4 ± 0.5	0.50
EDGE	7B	**67.7**	**62.7**	**61.1**	**62.3**	**59.6**	$\mathbf{63.2} \pm 0.5$	**0.54**

SEED (Wu et al., 2026), MAGRPO (Liu et al., 2026), MASIRL (Zhang et al., 2026), MATPO (Mo et al., 2025), and adapted C3 (Chen et al., 2026b) and IAPO (Ren et al., 2026). They share Qwen2.5-Omni-7B, tools, split, and format alignment. External credit mechanisms and adaptations are specified in Appendix D.1; local ablations hold the EDGE loop fixed. Evaluation uses 1,200 updates and three paired seeds, with a separate generation ceiling because equal updates do not imply equal compute (Appendix D.5).

Accuracy scores answers; IoU scores record localization. Among correct answers, Any-citation requires one reference-record citation; CompleteSupport (CS) requires full AND/OR support. GroundedAcc (GA) is the fraction of all test questions answered correctly with full support (Appendix B.7). Every method–seed output receives double annotation and adjudication; the 300-question subset evaluates the analyzer (Appendix F).

4.2 Overall Performance and Evidence Quality

EDGE's accuracy gains were accompanied by larger gains in fully supported correct answers. It reached 63.2% accuracy and 0.54 evidence IoU, leading the trained baselines across all five task dimensions (Table 1). Relative to MATPO, accuracy increased by 2.9 percentage points (95% paired program-bootstrap interval: [1.7,4.1]). GroundedAcc rose from 44.9% to 53.2%, an 8.3-point gain (95% interval: [6.6,10.0]), while CompleteSupport increased from 74.5% to 84.1% (Table 2). Both intervals condition on the three trained seeds; Appendix D.2 reports seed dispersion and dimension-level results. Comparisons with adapted C3 and IAPO showed the same pattern: EDGE's accuracy margins were 2.3/2.8 points, compared with 5.8/5.9 points for GroundedAcc. These comparisons favor more complete evidence support alongside improved accuracy under the disclosed baseline adaptations and information access.

4.3 Independent Attribution Validation

We evaluated 600 failures outside initialization and the 250-trace calibration set, freezing the analyzer, overlap threshold $\eta = 0.5$, prompts, and replay protocol. Common-prefix replay and blind link annotation supplied independent references (Appendix E.1). First-break point hit reached 86.0%, versus 62.5% for the same-information judge (23.5-point gain; paired 95% interval: 19.0–28.0; Table 3). Unique-winner agreement was 90.0% on 80.0% of traces; 12.0% had ties and 8.0% remained unresolved. Set hit reached 91.0%, whereas full break-set agreement was 78.6%, distinguishing

Table 2: **Evidence quality of structured-record systems.** Any-cit. and CS condition on correct answers; GA uses all test questions. Mod.-hit and Mod.-cov. range over $[0, 1]$; other metrics are percentages.

Method	Answer support (%)			Modality coverage		Partial requests
	Any-cit. ↑	CS ↑	GA ↑	Mod.-hit ↑	Mod.-cov. ↑	C5-part. ↓
EDGE, untrained	79.8	71.4	42.6	0.60	0.57	28.1
Single agent	70.4	62.9	36.7	0.57	0.53	34.0
SEED	76.3	68.5	40.1	0.60	0.57	30.2
MAGRPO	71.2	63.1	36.2	0.58	0.54	34.8
MASIRL	77.5	70.2	41.3	0.61	0.57	31.0
MATPO	80.6	74.5	44.9	0.63	0.60	26.3
C3 (adapted)	83.2	77.8	47.4	0.66	0.63	24.6
IAPO (adapted)	84.3	78.3	47.3	0.65	0.62	23.3
EDGE	**88.4**	**84.1**	**53.2**	**0.69**	**0.67**	**17.3**

Table 3: **Independent attribution confirmation** (%). Diagnoses add no executions; random rates are analytic. Headers give sample sizes; definitions and references appear in Appendix E.1.

Method	Point hit ↑ $n = 600$	Unique hit ↑ $n = 480$	Set hit ↑ $n = 600$	Link exact ↑ $n = 588$
Random agent	33.3	33.3	42.7	–
Last agent (A_R)	28.0	27.5	38.0	20.4
LLM judge (GPT-5)	62.5	65.0	69.0	51.4
First break (ours)	**86.0**	**90.0**	**91.0**	**78.6**
Oracle diagnostic ceiling				
First break, oracle L_3/L_4	94.8	95.8	96.0	92.2

Table 4: **Credit assignment and target selection** (%). Accuracy: three-seed mean ± SD. External adaptations and matched local controls are specified in Appendix D.3.

Configuration	Acc. ↑	Any-cit. ↑	CS ↑	GA ↑	C5-part. ↓
Credit algorithms without hindsight					
Uniform task credit	57.6 ± 0.6	71.5	63.4	36.5	34.6
C3 (adapted)	60.9 ± 0.6	83.2	77.8	47.4	24.6
IAPO (adapted)	60.4 ± 0.5	84.3	78.3	47.3	23.3
BRPO only	60.6 ± 0.5	86.9	81.7	49.5	20.8
Hindsight and matched target-selection controls					
Uniform credit + LHD	60.1 ± 0.6	80.2	73.6	44.2	19.4
Negative-only shuffle	61.6 ± 0.5	86.2	81.8	50.4	19.0
Dependency-ranked targets	62.2 ± 0.5	86.7	82.6	51.3	18.3
All-record OR targets	61.9 ± 0.5	86.1	81.0	50.1	18.1
Full EDGE	$\mathbf{63.2 \pm 0.5}$	**88.4**	**84.1**	**53.2**	**17.3**

agent selection from exact link diagnosis. On adjudicated structural inputs, fixed L_1/L_2 checks agreed on 720/720 and 540/540 evaluable decisions; runtime extraction errors remain in the headline results (Appendix E.2).

Diagnosis required no additional executions, while the replay reference averaged 313.6 per trajectory; independent resampling reproduced 77 of 80 unique winners. For failures spanning agents, Planner–Perception joint repair raised success from 36.0% to 70.0%, compared with single-agent gains of 2.0 and 3.0 points (Appendix E.3). This joint-repair effect supports retaining the implicated set when several evidence records fail across agents.

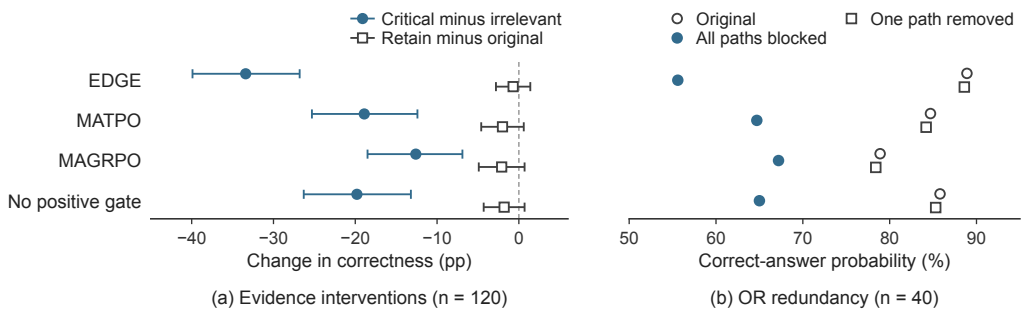

Figure 3: **Selective evidence dependence.** **(a)** Evidence-intervention contrasts on 120 questions; bars: pointwise 95% program-bootstrap intervals. **(b)** OR redundancy on 40 common questions; means without intervals. Both panels use 16 draws per question/condition from seed-13 policies (Appendix F.3).

4.4 Negative-Credit Alignment and Sufficient-Path Selection

Without hindsight, BRPO exceeded C3 and IAPO by 2.1–2.2 GroundedAcc points at similar accuracy (Table 4). Because these credit rules also differ in positive gating and annotation access, the matched controls isolate negative-credit localization more directly. Shuffling negative weights reduced accuracy by 1.6 points and GroundedAcc by 2.8 points, with positive gating, abstention, and LHD retained. Dependency ranking narrowed these deficits to 1.0 and 1.9 points with the model and reference access fixed, preserving each trajectory's target count and weight multiset. EDGE and negative-only shuffle both had 44.8% task-credit coverage, and the gains persisted at the common token ceiling (Appendices D.3 and D.5). These controls support the training benefit of evidence-aligned negative targets, complementing Proposition 1's conditional guarantee.

Sufficient-path selection improved accuracy and GroundedAcc by 4.9 and 7.4 points on 360 multi-path OR questions, compared with 0.0 and 0.3 points on AND questions (Table 23). The concentration of gains on alternative paths supports filtering unused records from corrective targets. This comparison changes negative credit and hindsight jointly, assessing sufficient-path filtering as their shared selection rule. Removing the positive gate lowered CompleteSupport by 16.3 points (Table 25), identifying supported-answer reinforcement as a separate contributor to the full method.

4.5 Does the Reasoner Depend on the Evidence?

We tested evidence dependence by intervening on frozen Reasoners using common questions with faithful sufficient evidence, selected before observing replayed answers. Critical removal reduced EDGE's correct-answer probability by 34.3 points, compared with 0.9 for irrelevant removal. The adjusted decline was 33.4 points, versus 19.8 without the positive gate (Figure 3a). Retaining a sufficient path changed EDGE by -0.7 points (95% interval: $[-2.8, 1.4]$), consistent with a small average change without establishing equivalence. OR redundancy provided a complementary test: removing one path changed accuracy from 88.9% to 88.6%, whereas blocking all paths lowered it to 55.6% (Figure 3b). Together, these interventions support selective dependence on sufficient evidence within the eligible questions; the matched gate ablation tests the corresponding training effect. Independent delivery and credit audits provide further validation in Appendix F.

5 Conclusion

SkateEG specifies answer support through evidence records and AND/OR requirements. EDGE routes task credit and localized hindsight through first-break diagnosis and sufficient-path selection. By separating record-level failures from task-level repair needs, this selection avoids correcting unused alternatives when a sufficient evidence path is already complete. Independent replay and link annotations separately validate the diagnostic targets. Matched ablations support improved grounding in a single-pass, three-agent pipeline with annotated training evidence. The negative-credit and OR controls support the value of selecting both the responsible actions and the evidence paths that warrant correction. Together, these results make explicit evidence requirements a common basis for evaluating answer support and directing agent learning.

AI USE STATEMENT

Generative AI tools were used solely to improve the grammar, wording, and readability of the manuscript. They were not used for research methodology design, experiment design, theoretical development, result interpretation, or other research tasks requiring disclosure under the ICLR 2027 AI Policy. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

SKATEEG supports research on evidence-grounded video reasoning; it is not intended for automated competitive judging. Its release protocol restricts redistribution of broadcast material and includes age-based media exclusions and a removal-request process. Appendix B.2 details data provenance, licensing, access restrictions, privacy safeguards, and intended use.

REPRODUCIBILITY STATEMENT

Code, trained checkpoints, and the SKATEEG release (questions, evidence records, AND/OR labels, phase boundaries, extracted features, source URLs, and the human evaluation labels) will be made publicly available. Dataset construction is described in Appendices B.1–B.5, training configurations and hyperparameters in Appendices C.3 and C.5, baseline adaptations in Appendix D.1, evaluation and annotation protocols in Appendices B.7, E.1, and F.1, and the proof of Proposition 1 in Appendix C.1.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Seoyeon An, Hyeonseo Jang, Minsu Kim, Chanho Lee, Younghan Park, and Kangwook Lee. AgentVidBench: A multi-hop video question answering benchmark for evaluating MLLM agents. *arXiv preprint arXiv:2609.21386*, 2026.

Boyu Chen, Zikang Wang, Zhengrong Yue, Kainan Yan, Chenyun Yu, Yi Huang, Zijun Liu, Yafei Wen, Xiaoxin Chen, Yang Liu, Peng Li, and Yali Wang. VideoChat-M1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33772–33783, 2026a.

Haodong Chen, Haojian Huang, Xinxiang Yin, and Dian Shao. FineQuest: Adaptive knowledge-assisted sports video understanding via agent-of-thoughts reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2909–2918, 2025. doi: 10.1145/3746027.3754885.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859v2*, 2026b. URL https://arxiv.org/abs/2603.06859v2.

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.

Chengqi Dong, Chuhuai Yue, Hang He, Yandong Liu, Fenghe Tang, S. Kevin Zhou, Xiaohan Wang, Jiajun Chai, and Guojun Yin. TAPO: Tool-aware policy optimization via credit transfer for multimodal search agents. *arXiv preprint arXiv:2606.05784*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-72670-5_5.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Rong Gao, Xin Liu, Zhuozhao Hu, Bohao Xing, Baiqiang Xia, Zitong Yu, and Heikki Kälviäinen. FSBench: A figure skating benchmark for advancing artistic sports understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13595–13605, 2025. doi: 10.1109/CVPR52734.2025.01269.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8590–8597, 2023. doi: 10.1145/3581783.3613795.

Xiaolong Jin, Dingmin Wang, Vijay Lingam, and Varun Kumar. AHEAD: Adaptive hindsight with environment-augmented distillation for agentic RL. *arXiv preprint arXiv:2608.24114*, 2026.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, Kazuki Kozuka, and Ehsan Adeli. VideoMultiAgents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*, 2025.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents. *arXiv preprint arXiv:2605.19447*, 2026a.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563v5*, 2026b. URL https://arxiv.org/abs/2603.21563v5.

Shenglan Liu, Aibin Zhang, Yunheng Li, Jian Zhou, Li Xu, Zhuben Dong, and Renhao Zhang. Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2163–2171, 2021. doi: 10.1609/aaai.v35i3.16314.

Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. LLM collaboration with multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32150–32158, 2026. doi: 10.1609/aaai.v40i38.40487.

Zhanfeng Mo, Xingxuan Li, Yuntao Chen, and Lidong Bing. Multi-agent tool-integrated policy optimization. *arXiv preprint arXiv:2510.04678*, 2025.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC20)*, 2020.

Bo Ren, Yirong Mao, Yi Yang, and Wenhui Que. IAPO: Influence-aware policy optimization for credit assignment in multi-turn service agents. *arXiv preprint arXiv:2608.24588*, 2026.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096956.

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.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents. *arXiv preprint arXiv:2603.08754*, 2026.

Ryota Tanaka, Tomohiro Suzuki, and Keisuke Fujii. 3D pose-based temporal action segmentation for figure skating: A fine-grained and jump procedure-aware annotation approach. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports (MMSports)*, pp. 17–26, 2024. doi: 10.1145/3689061.3689077.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, 2025a. doi: 10.18653/v1/2025.findings-emnlp.50.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SEED: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, and Hanjie Chen. SPORTU: A comprehensive sports understanding benchmark for multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Haotian Xia, Haonan Ge, Junbo Zou, Hyun Woo Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, Xiyang Lin, Christopher Lai, Shengjie Zhang, Junwen Miao, Shichao Chen, Rhys Tracy, Vicente Ordonez, Weining Shen, and Hanjie Chen. SportR: A benchmark for multimodal large language model reasoning in sports. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with MLPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2901–2909, 2023. doi: 10.1609/aaai.v37i3.25392.

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 (CVPR)*, pp. 13204–13214, 2024. doi: 10.1109/CVPR52733.2024.01254.

Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590, 2020. doi: 10.1109/TCSVT.2019.2927118.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. OPID: On-policy skill distillation for agentic reinforcement learning. *arXiv preprint arXiv:2606.26790*, 2026a.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. SVAgent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24062–24072, 2026b.

Calvin Yeung, Tomohiro Suzuki, Ryota Tanaka, Zhuoer Yin, and Keisuke Fujii. AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5945–5956, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Mingwen Zhang, Jisheng Dang, Minqiang Yang, Bimei Wang, Bin Hu, and Tat-Seng Chua. Multi-agent self-improving reinforcement learning for video reasoning. *arXiv preprint arXiv:2608.28675*, 2026.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15085–15099, 2023. doi: 10.1109/ICCV51070.2023.01385.

Junbo Zou, Haotian Xia, Zhen Ye, Shengjie Zhang, Christopher Lai, Vicente Ordonez, Weining Shen, and Hanjie Chen. DeepSport: A multimodal large language model for comprehensive sports video reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2511.12908*, 2025.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. VideoBrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*, 2026.

A Related Work

Skating and evidence-grounded video understanding. Skating datasets cover program scoring (Xu et al., 2020; Xia et al., 2023; Ji et al., 2023), temporal action segmentation (Liu et al., 2021), and 3D jumps (Tanaka et al., 2024; Yeung et al., 2025). FSBench (Gao et al., 2025) extends to multimodal questions on elements, execution, and artistry, alongside broader sports benchmarks and agents (Xia et al., 2025; 2026; Chen et al., 2025; Zou et al., 2025). NExT-GQA (Xiao et al., 2024) documents the gap between answer accuracy and temporal grounding. Grounding-specialized models (Yu et al., 2023; Wang et al., 2025a), temporal-grounding reinforcement learning (Wang et al., 2025b), and agentic verification (Dang et al., 2025) address this gap. VideoBrain's failure analysis identifies imprecise evidence sampling (Zou et al., 2026); AgentVidBench scores evidence acquisition against trajectory milestones (An et al., 2026). SkateEG specifies record addresses, modality requirements, and AND/OR support, combining question-level probes with source-grounded review. These requirements support both evaluation and training-credit assignment.

Multi-agent video reasoning. Training-free systems coordinate modality specialists (Kugo et al., 2025), memory-augmented tools (Fan et al., 2024), cross-modal evidence checks (Yang et al., 2026b), or parallel reasoning paths (Dang et al., 2025). VideoChat-M1 (Chen et al., 2026a) jointly trains agents on answer and collaboration rewards; MASIRL (Zhang et al., 2026) trains a grounder with a frozen verifier. Single-agent reinforcement learning is another route (Feng et al., 2025; Li et al., 2025; Zou et al., 2025). EDGE diagnoses reference records' request, acquisition, description, and use, then filters sufficient-path targets for task credit and hindsight.

Credit assignment for cooperating language agents. MAGRPO (Liu et al., 2026) shares a group-relative advantage (Shao et al., 2024); MATPO (Mo et al., 2025) allocates reward across planner and worker roles. Hindsight rescoring (Tan et al., 2026), bounded action reweighting (Li et al., 2026a), and sign-preserving tool credit (Dong et al., 2026) provide finer assignments. C3 (Chen et al., 2026b) samples alternative actions under a fixed history with a leave-one-out baseline; CCPO (Li et al., 2026b) compares joint outcomes with agent-removal counterfactuals. IAPO (Ren et al., 2026) routes advantages along typed dependencies extracted from completed rollouts with mass conservation, sign preservation, and uniform fallback. BRPO instead grounds candidates in explicit evidence records, normalizes nonempty action sets before clipping, preserves advantage signs, and assigns zero task weight without a confirmed target. Tables 1 and 2 include task-adapted C3 and IAPO; Table 4 separates those full credit rules from a matched dependency-ranked negative-target control. The free-form inferred-graph configuration remains an interface ablation (Table 25), not an IAPO reproduction. C3 already evaluates credit quality through interventions. Our diagnostic target is the first failed handoff per record, read without extra executions; agent repair replay and independent link annotations evaluate its repair value and localization separately.

Hindsight distillation. Privileged-context distillation scores on-policy samples using hindsight (Wu et al., 2026; Yang et al., 2026a). Extensions provide stepwise corrections (Jin et al., 2026), select distillation targets (Li et al., 2026a), or refresh self-analysis with the policy (Wu et al., 2026). LHD distributes hindsight across agents by failed links retained after sufficient-path selection. We compare refreshed and frozen analyzers using low-budget replay during training (Appendix D.6) and separate final agent/link confirmation (Section 4.3).

B SkateEG Details

B.1 SkateEG: Corpus Details

The 1,194 complete competition programs preserve element context and broadcast replays (Figure 4): 561 men and 633 women, with 728 free-skating and 466 short programs. The 11,998 timed technical elements comprise 6,494 jumps, 3,582 spins, and 1,922 step sequences. Official grade-of-execution marks are negative for 4,319 elements and non-negative for 7,679; negative instances supply execution-diagnosis items.

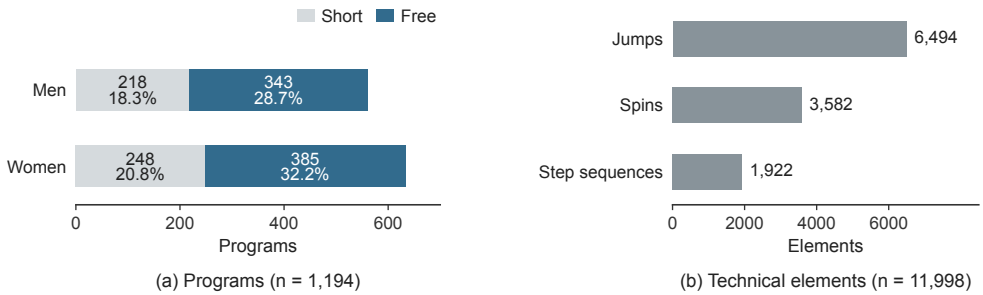

Figure 4: **Source corpus composition. (a)** Short and free programs by competition division; labels give counts and percentages of all 1,194 programs. **(b)** The 11,998 timed technical elements by type.

Videos use 30 fps and a 720-pixel short side. All systems receive the same RGB, pose, and commentary-removed audio streams, with no text stream. Pose uses ViTPose-H (Xu et al., 2022)

followed by MotionBERT (Zhu et al., 2023), producing 17-joint Human3.6M coordinates (Ionescu et al., 2014) rather than overlays. The lifting model is fine-tuned on AthletePose3D (Yeung et al., 2025) to accommodate competitive-sport motion. Appendix B.6 specifies audio processing.

B.2 Datasheet, Licensing, and Ethics

Release and provenance. The public release contains our question–answer pairs, annotations, score labels, and rule-grounding materials under CC BY-NC 4.0, extracted visual and audio features, and source URLs with precise timestamps. Raw broadcasts and frames are not publicly redistributed, following prior skating benchmarks (Xia et al., 2023; Xu et al., 2020; Ji et al., 2023). Trimmed clips require a signed data use agreement limiting use to non-commercial research and prohibiting redistribution and identity inference. The datasheet records each asset class's source, copyright holder, terms, and release status. Broadcast rights remain with their holders; ISU rule documents are cited, not redistributed.

Privacy and intended use. Footage comes from publicly broadcast senior ISU competitions. Public ISU biographies are checked against event dates: skaters under 18 at recording are excluded from redistributed media and referenced only by URL and timestamp. No identity annotations beyond the broadcast are added, and no identity-inference task is defined. Athletes or representatives may request removal, propagated through versioned releases. SkateEG supports research on evidence-grounded video reasoning, not automated competitive judging. Official grade-of-execution labels record human officials' judgments rather than objective technique.

B.3 Dimensions and Question Inventory

Table 5: Full composition. Structure entries describe the dominant labels within each dimension. "Both" denotes OR and AND on different questions, not a mixed logical structure within one question.

	Dimension	#Q	Dominant phase	$\sigma(q)$	p_0 (%)
C1	Element Identification	1,800	takeoff–air	$\vee$	30.5
C2	Execution Diagnosis	1,800	landing / takeoff	both	37.6
C3	Structure and Timing	1,400	across elements	$\wedge$	40.1
C4	Rule-Grounded Judgment	1,400	element window	$\vee$ + rules	43.7
C5	Musicality	800	non-jump / program	$\wedge$ cross-modal	35.9
Video-grounded		**7,200**			37.3
RG	Rule Grounding (text)	2,400	—	—	—
Total		**9,600**			

C1 identifies elements, rotations, and edges; C2 diagnoses execution quality; C3 tests element order and program structure; C4 applies rules to observed execution; and C5 tests movement–music alignment and presentation. RG tests rule recall, distinctions, and judgment without video, separately from C4. Questions were balanced across evidence requirements, adding underrepresented landing/entry phases, cross-modal requirements, and temporal conjunctions. Figure 5 summarizes these distributions. Its modality union is $U_q = \{m(r) : r \in E^*(q)\}$, where $m(r)$ is the record's modality: RGB–pose may denote OR alternatives or a joint AND requirement, distinguished by $\sigma(q)$. The support audit below defines the corresponding question-level labels.

C3 and C5 together contain 1,320 of the 2,020 AND questions (65.3%), while accounting for 2,200 of the 7,200 video questions (30.6%). Conjunctive support is therefore concentrated in structure-and-timing and musicality questions.

Example questions. Each example is shown with its address set; windows are given relative to program start.

- **C1, disjunctive.** *How many rotations did the skater complete on the second jump of the opening combination?*
  E^*: $\langle [31.4, 32.1],\ \text{takeoff–air},\ \text{Pose} \rangle$.

Table 6: Dimension and logical structure counts for the 7,200 video questions; the test panel has 1,440 questions.

Dimension	Corpus OR	Corpus AND	Test OR	Test AND
C1	1650	150	330	30
C2	1430	370	286	74
C3	680	720	136	144
C4	1220	180	244	36
C5	200	600	40	120
Total	5180	2020	1036	404

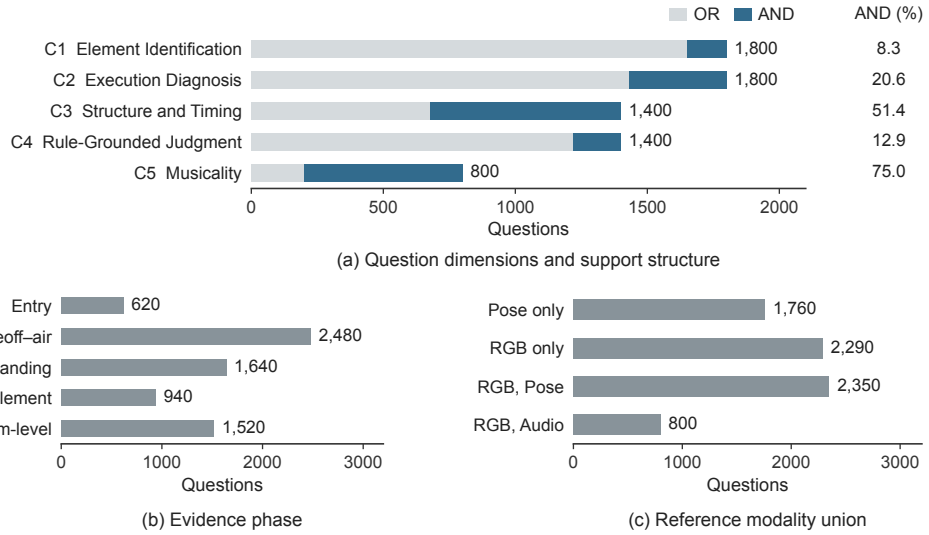

Figure 5: **Question and evidence composition. (a)** Corpus OR/AND counts by dimension; bar-end labels give totals and the right column gives within-dimension AND percentages. **(b)** Evidence-phase and **(c)** reference-modality-union marginals over the same 7,200 video questions. Modality union does not imply joint necessity. The 2,400 text-only rule controls are separate. Exact structure counts and the test split: Table 6.

- **C2, disjunctive.** *Was the landing of the triple Lutz controlled, or was there a step-out?*
  E^*: $\langle[78.9, 79.6],\ \text{landing},\ \text{RGB}\rangle$.
- **C3, temporally conjunctive.** *Which element immediately followed the step sequence, and did it repeat a jump already performed?*
  E^*: $\{\langle[96.2, 104.7],\ \text{program},\ \text{RGB}\rangle, \langle[33.0, 34.2],\ \text{takeoff–air},\ \text{RGB}\rangle\}$, $\sigma = \wedge$.
- **C4, disjunctive with rules.** *Should the grade of execution for this jump be reduced, and on what ground?*
  E^*: $\langle[52.7, 53.5],\ \text{takeoff–air},\ \text{Pose}\rangle$, with the applicable rule passage supplied as text.
- **C5, cross-modal conjunctive.** *Does the final pose of the choreographic sequence coincide with the closing musical accent?*
  E^*: $\{\langle[142.0, 146.5],\ \text{program},\ \text{RGB}\rangle, \langle[142.0, 146.5],\ \text{program},\ \text{Audio}\rangle\}$, $\sigma = \wedge$. The two records share a window and differ only in modality, which is how a cross-modal requirement is expressed.

B.4 Annotation Procedure and Quality Control

A proposer identifies relevant rules, elements, and segments; a generator writes questions, answers, and distractors; and a verifier checks correctness and support. Annotators with skating competition or coaching experience review every item for answer correctness, evidence support, category, distractors, and technique. Senior reviewers with judging or technical-panel experience adjudicate

disagreements against source rules and video. Compensation exceeds the local minimum hourly wage.

Layer 1: evidence addresses. Annotators construct a candidate reference set for each video question and audit which records supply its answer-decisive facts, keeping the question, options, and applicable rule context fixed. The accepted $E^*(q)$ is nonempty and has one of two pure forms:

- **OR:** review every listed record in isolation. Each must support the complete answer without the other records. Singleton reference sets use this label by convention.
- **AND:** verify that the full set supports the complete answer and that removing each record loses an indispensable fact. Record the missing fact for every deletion. An AND set contains at least two records.

Answerability after one deletion establishes redundancy, not singleton sufficiency. These source-grounded judgments concern available facts, independently of sampled model correctness. Timed elements automatically anchor C1, C2, and scoring-related C4 addresses, with sampled human verification. Annotators select decisive timed elements for C3 and other C4 items; C5 addresses are recorded during authoring. Modality probes (Appendix B.5) provide separate statistical labels: positive tests do not certify complete factual support, and negative tests do not establish logical necessity.

Mixed forms such as $(a \wedge b) \vee c$ or alternative multi-record conjunctions require reauthoring or adjudication before inclusion. References remain unresolved when no subset has confirmed probe sufficiency or support judgments are incomplete or disputed. They are never converted to OR, AND, or text-only labels. Probe-positive subsets with probe-negative supersets remain flagged as non-monotone measurements without forced relabeling; dimension averages cannot supply missing per-question decisions. Re-auditing preserves frozen questions and answers. Unresolved references supply no evidence-derived training targets and retain evaluation denominators and coverage accounting (Appendix B.7).

Layer 2: jump phase boundaries. Only jumps decisive for a question receive phase annotations. Takeoff is the last frame before the blade leaves the ice; landing is the first frame after return. Annotators mark entry onset independently, using a fixed $1.5\,\mathrm{s}$ pre-takeoff interval only when judgment is unavailable. Table 7 reports agreement at $30\,\mathrm{fps}$.

Table 7: Inter-annotator agreement on jump phase boundaries, over a doubly-annotated tenth of the annotated jumps, at $30\,\mathrm{fps}$. The two visually defined boundaries agree an order of magnitude more tightly than the conventional one.

Boundary	#Double-annotated	Frame error	ms	Segment IoU
Takeoff	320	1.3	≈43	0.92
Landing	320	1.6	≈53	0.93
Entry	320	11.3	≈377	0.67

Takeoff and landing anchor required windows and thereby affect L_1/L_2 temporal checks (Section 3.1); phase-name equality is not an additional condition. Entry remains descriptive: its conventional onset does not set coverage boundaries, avoiding roughly $0.4\,\mathrm{s}$ of annotation noise.

Before adjudication, answer and evidence-support agreement reach $\kappa = 0.87$ and 0.82, on 9,600 answer-validity and 7,200 video-support judgments, respectively. Mean evidence-window IoU is 0.76 across video items, including broader program-level windows. A blind stratified 20% audit of 1,440 questions finds 11 residual errors (0.76%); independent official-score extractors agree on the complete score tuple for 1,180/1,194 programs (98.8%). The support audit validates listed paths, not all possible alternative evidence. Modality sufficiency remains operationally tied to the probes and sampling budget; an unconfirmed subset may still contain the information needed to answer.

B.5 Modality Sufficiency Protocol

Fix each question's evidence input w_q, including disjoint windows where needed. Both Qwen2.5-Omni-7B (Xu et al., 2025) and Gemini 2.5 Pro receive the question, options, and each of the seven

nonempty subsets of $\{\mathrm{RGB}, \mathrm{Pose}, \mathrm{Audio}\}$ restricted to w_q. For every probe–subset pair, draw $N = 8$ independent responses at fixed stochastic decoding settings; count correct option letters, treating unparseable outputs as incorrect. Retain versions, prompts, settings, and individual draws. Main answer-accuracy evaluation instead uses greedy decoding.

Fixed reference and joint test. Let b_d be the fixed reference for dimension d, using the empty-evidence Reasoner rates in Table 5: 0.305, 0.376, 0.401, 0.437, and 0.359 for C1–C5. These are benchmark thresholds, distinct from either probe's unknown empty-evidence success probability on an individual question and from the question-specific p_0 in Proposition 1. For probe $b \in \{1, 2\}$, let $x_{b,qMw}$ be its number of correct responses. The one-sided binomial p-value and the joint p-value are

$$p_{b,qMw} = \sum_{j=x_{b,qMw}}^{8} \binom{8}{j} b_{d(q)}^{j} (1 - b_{d(q)})^{8-j}, \qquad p_{qMw} = \max_{b \in \{1,2\}} p_{b,qMw}.$$

The maximum p-value tests whether at least one probe's success probability is at most the fixed reference, without assuming probe independence. It measures threshold exceedance; a same-question evidence-versus-empty gain requires a matched control.

Question-level correction and labels. The testing family $\mathcal{T}_q$ contains the seven pairs (M, w_q) for one question. If alternative evidence windows are evaluated, all such conditions must be included in the family before observing probe outputs. Write $m_q = |\mathcal{T}_q|$, with $m_q = 7$ for the fixed-input protocol, and sort the joint p-values as $p_{q,(1)} \le \cdots \le p_{q,(m_q)}$. Apply Holm's correction (Holm, 1979):

$$\widetilde{p}_{q,(j)} = \min\left\{1, \max_{1 \le \ell \le j} (m_q - \ell + 1) p_{q,(\ell)}\right\}, \qquad \mathrm{suff}(q, M, w) = \mathbb{1}[\widetilde{p}_{qMw} \le 0.05].$$

Holm controls question-level family-wise error despite dependent subsets. A zero label includes probe disagreement and means unconfirmed sufficiency at this budget, not absence of information.

For the questions $\mathcal{Q}_d$ in dimension d, the reported fraction is

$$S_{d,M} = \frac{1}{|\mathcal{Q}_d|} \sum_{q \in \mathcal{Q}_d} \mathrm{suff}(q, M, w_q).$$

The 5×7 cells aggregate these decisions, not pooled accuracy or per-dimension significance. For agreement, apply Holm separately to each probe over the same $\mathcal{T}_q$ and compute agreement and Cohen's κ across paired (q, M, w_q) decisions. This diagnostic differs from the joint final label. Probe agreement is 91.4% ($\kappa = 0.81$).

	single stream			*combined streams*				
	RGB	Pose	Audio	RUP	RUA	PUA	All	Δ
C1 Element identification 1,800 questions · p_0 30.5%	0.42	**0.78**	0.02	0.83	0.43	0.79	0.84	+0.05
C2 Execution diagnosis 1,800 questions · p_0 37.6%	0.51	**0.64**	0.03	0.74	0.52	0.65	0.75	+0.10
C3 Structure and timing 1,400 questions · p_0 40.1%	**0.71**	0.29	0.04	0.76	0.72	0.31	0.77	+0.05
C4 Rule-grounded judgment 1,400 questions · p_0 43.7%	**0.58**	0.55	0.03	0.67	0.59	0.56	0.68	+0.09
C5 Musicality 800 questions · p_0 35.9%	**0.24**	0.06	0.19	0.26	0.81	0.23	0.82	**+0.57**

Figure 6: Modality-sufficiency fractions over the seven subsets. Each cell reports the question-level confirmed fraction $S_{d,M}$, not probe answer accuracy. Rows give question counts and the dimension-specific empty-evidence reference (shown as p_0); bold marks the best single stream, and Δ compares the best pair with it. Musicality is outlined.

Figure 6 summarizes the confirmed fractions. Pose leads on element identity (0.78 versus RGB's 0.42), while RGB leads on program-level questions (0.71). C5's best pair exceeds its best singleton

by 57 percentage points, versus 5–10 elsewhere; the full triple adds at most one point over the best pair. These marginal fractions describe confirmed questions, not information absence or individual joint necessity. Question-level labels and source support determine the latter. The low C1–C4 audio fractions complement the speech-removal audit below, rather than measuring audio-only accuracy or certifying complete commentary removal.

B.6 Shortcut Controls

Media and item controls. HTDemucs v4 (Rouard et al., 2023) removes the vocal stem while retaining musical accompaniment. Automatic speech recognition finds transcribable speech in fewer than 1% of processed windows, reducing answer leakage through commentary. The C1–C4 audio-sufficiency fractions (0.02–0.04) are a separate diagnostic. Score displays and protocol overlays are masked; mark-revealing metadata and text/OCR streams are withheld. Answer positions are balanced ($25.1/24.8/25.0/25.1$ for A/B/C/D). A question-and-options template probe scores 27.3% versus 25.0% chance.

The empty-evidence Reasoner is sampled 8 times per item. Its rates (Table 5) range from 30.5% on C1 to 43.7% on C4, with weighted mean 37.3%. Domain priors can therefore support answer selection beyond the template-only probe, even without observed evidence.

These empty-evidence rates describe domain priors. Branch-specific bounds for true partial evidence and repaired contexts are evaluated separately in Appendix E.3.

B.7 Evaluation Protocol

Splits and matched conditions. The $1{,}194$ programs are split $836/119/239$ into training, validation, and test, yielding roughly $5{,}040/720/1{,}440$ video questions. Splitting by program keeps correlated items together. All methods, including untrained references, share these frozen splits; all $2{,}400$ text-only rule questions form a separate control set outside the video-test metrics. Inference budgets, frame/pose/audio tools, and trained baselines' format alignment (Appendix C.5) are matched. C1–C5 contain $360/360/280/280/160$ test questions and the OR/AND strata contain $1{,}036/404$ questions, proportional to the full-data marginals.

Metrics. We use the names Any-citation (Any-cit.), CompleteSupport (CS), GroundedAcc (GA), Modality hit (Mod.-hit), and Complete modality coverage (Mod.-cov.) throughout the tables. Accuracy, Any-cit., CS, GA, and C5-part. are reported in percent; evidence IoU, Mod.-hit, and Mod.-cov. use the $[0, 1]$ scale. Accuracy is the fraction of correct answers. Evidence IoU scores localized windows against a sufficient reference path, as defined below. Accuracy, IoU, Any-citation, and modality hit have dimension breakdowns; CompleteSupport is also reported separately for AND and OR questions. These breakdowns distinguish single-record support from conjunctions across records. Record matching uses resolved, nonempty $E^*(q)$ on video-grounded questions. Unresolved references retain their questions through the bounds below; text-only rule questions are evaluated separately.

Citation and complete support. Let $Y_q = \mathbb{1}[\hat{y}_q = y_q^*]$ and let $\mathcal{J}_q \subseteq E^*(q)$ contain the reference records matched by valid citations in the answer. Matching uses record identifiers and the fixed address checks, before judging claim faithfulness or interpretation. *Any-citation* is

$$\text{Any-citation} = 100\,\frac{\sum_q Y_q \mathbb{1}[\mathcal{J}_q \neq \emptyset]}{\sum_q Y_q}. \tag{7}$$

Its denominator is the number of correct answers, and one matched citation suffices. CompleteSupport uses the same audited pure structures as the training gate. Define the family of minimal sufficient record sets by

$$\mathcal{S}_q = \begin{cases} \{E^*(q)\}, & \sigma(q) = \wedge, \\ \{\{r\} : r \in E^*(q)\}, & \sigma(q) = \vee. \end{cases} \tag{8}$$

OR members are verified singleton alternatives; $E^*(q)$ lists the available reference paths, without requiring every alternative to be used. AND requires every record jointly. Unresolved references leave $\mathcal{S}_q$ unassigned, never an empty sufficient set that could pass by vacuous inclusion.

For one method and evaluation seed, let D_q^* contain the reference records whose four handoffs pass the human reference checks: the right evidence was requested and acquired, its description is correct and complete, and the answer cites and correctly interprets it. The full-test annotation protocol is specified in Appendix F. For AND questions, the use check includes the required joint interpretation; individually correct citations cannot pass it when their combination is misinterpreted.

The reference status H_q^* is 1 if some $S \in \mathcal{S}_q$ is contained in D_q^*, 0 if every sufficient path has a confirmed failed handoff, and ? otherwise. A confirmed path suffices for OR even if another path is unresolved. An unresolved reference structure yields ?. Set $h_q^- = \mathbb{1}[H_q^* = 1]$ and $u_q = \mathbb{1}[H_q^* =?]$. The reported confirmed-support scores are

$$\text{CompleteSupport} = 100\,\frac{\sum_q Y_q h_q^-}{\sum_q Y_q}, \qquad \text{GroundedAcc} = \frac{100}{N}\sum_q Y_q h_q^-, \tag{9}$$

where N counts evaluated video questions. CompleteSupport is conditional on correctness; GroundedAcc counts correct, fully supported answers over all questions. Both allow OR alternatives and require annotations beyond Any-citation. Appendix F evaluates the predicted training gate H and delivery set Ψ against these references.

Multiple-record evidence IoU. Let $\widehat{\mathcal{W}}_q$ contain the localized window–modality pairs returned by a method. Structured systems use the addresses of records delivered to the Reasoner; other localization interfaces use their explicit returned spans and declared streams, extracted by a fixed parser. A whole input video's duration is not substituted for a missing localization. Each interval has positive duration and uses the same video-relative clock. Explicit multi-interval outputs are split at their own interval boundaries, before reference matching. They are not split at gold boundaries. Exact duplicate window–modality pairs count once.

For a sufficient set S, let $\mathcal{W}(S) = \{(w(r), m(r)) : r \in S\}$ be its distinct address pairs. Coincident reference addresses are counted once for localization, while their different factual requirements remain separate for complete support. Let $\mathfrak{M}(S, q)$ contain all partial one-to-one matchings between $\mathcal{W}(S)$ and $\widehat{\mathcal{W}}_q$ with equal modalities. Then

$$I_q = \max_{S \in \mathcal{S}_q} \frac{1}{|\mathcal{W}(S)|} \max_{\mu \in \mathfrak{M}(S,q)} \sum_{(a,b)\in\mu} \frac{|w(a) \cap w(b)|}{|w(a) \cup w(b)|}, \tag{10}$$

where $|w|$ is duration. Unmatched reference windows contribute zero; empty predictions score zero. AND averages all required localization units, while OR selects the best singleton path. One predicted interval cannot match two same-modality reference windows: one exact hit on a two-window AND set scores 0.5, while a complete OR alternative can score 1.

IoU uses neither η nor the containment override and does not change L_1/L_2 or the support gate. Extra predictions incur no penalty, so record counts and tool budgets accompany the score. IoU measures localization, not content faithfulness or evidence use, and averages equally over questions, including incorrect answers, within each seed.

Modality hit and complete modality coverage. Let M_q be the union of requested modalities and let $\mathcal{M}_q = \{\{m(r) : r \in S\} : S \in \mathcal{S}_q\}$ be the question's sufficient modality combinations. Modality hit uses any required modality, while Complete modality coverage requires a sufficient combination:

$$\begin{aligned} h_q^{\text{any}} &= \mathbb{1}[M_q \cap (\bigcup_{U \in \mathcal{M}_q} U) \neq \emptyset], \\ h_q^{\text{complete}} &= \mathbb{1}[\exists U \in \mathcal{M}_q : U \subseteq M_q]. \end{aligned} \tag{11}$$

Mod.-hit and Mod.-cov. average h_q^{any} and h_q^{complete}, respectively, over all evaluated video questions, including incorrect answers. Combinations derive from audited record-support families, not dimension-level probe fractions. Figure 5c's union $U_q = \bigcup_{U \in \mathcal{M}_q} U$ need not itself be minimal: RGB/pose OR alternatives allow either stream, whereas an RGB–pose AND set requires both. An audio-only request for RGB–audio AND gives $h_q^{\text{any}} = 1$ but $h_q^{\text{complete}} = 0$. Neither score checks temporal coverage, including two required RGB windows. C5-part. is the percentage of all C5 questions requesting one but not both jointly necessary modalities. Request counts and tool budgets accompany Mod.-cov. to distinguish sufficient selection from requesting every stream.

Scoring and aggregation. Correct answers without valid citations contribute zero to Any-citation and CompleteSupport; incorrect answers are excluded from their conditional denominators and contribute zero to GroundedAcc. For a method with a localization interface, a valid empty response or an output with no parseable valid intervals scores zero IoU. A missing evaluation log or unresolved reference instead leaves the score unknown. A method with no corresponding output interface is marked undefined for that metric; annotators do not generate evidence outputs on its behalf. Invalid or abstaining answers in available outputs count as incorrect. A missing answer log leaves Y_q unknown: report the missing count and withhold the complete accuracy and conditional-support summaries until the log is recovered. It is not an observed model failure. The support bounds below assume the answer outcomes Y_q are available.

Unresolved support judgments remain in both applicable denominators. The scores in Eq. (9) are lower endpoints, with upper endpoints obtained by adding $100 \sum_q Y_q u_q / \sum_q Y_q$ to CompleteSupport and $100 \sum_q Y_q u_q / N$ to GroundedAcc. These are bounds for unresolved labels, not confidence intervals. Report resolved-reference coverage over all N outputs and separately among correct answers. Unknown IoU or modality scores likewise retain their questions in the denominator and contribute $[0, 1]$ to the aggregate bounds. Any-citation uses the same bound construction for unknown citation matches among correct answers. No gold answer or modality label is fabricated to close a missing reference.

Conditional support is undefined when a seed has no correct answers. For each method, each evaluation seed first produces its own numerator, denominator, coverage, and bounds on the same frozen question set. Report the mean and sample standard deviation across training seeds, rather than pooling their correct-answer denominators; an undefined seed is reported explicitly. An untrained checkpoint contributes its single evaluation output set. GroundedAcc is computed from joint question-level outcomes, never by multiplying rounded cross-seed means of accuracy and CompleteSupport. Program-clustered intervals resample the same source programs across methods, retaining their questions and seed outputs together. They describe sampling variation conditional on the human references, separately from seed variation and unresolved-label bounds.

C Methodology Details

C.1 Proof of Proposition 1

Intervention and evidence event. Fix a question q, a task-policy checkpoint, and a common prefix h_τ before the earliest candidate intervention. All branches use the same sampling law and budgets. Let O be a specified repair operator and define

$$V_\tau^O(S) = \mathbb{E}[R \mid h_\tau, \mathrm{do}(A_i \leftarrow O_i,\ i \in S)], \qquad \Delta_i = V_\tau^O(\{i\}) - V_\tau^O(\emptyset). \tag{12}$$

Here S contains agents; $\Phi^{\mathrm{diag}}(\tau)$ retains the finer agent-action and record indices used to select training targets. Write $C_q(E) = 1$ when the faithful evidence delivered and used satisfies the question: $E \supseteq E^*(q)$ for conjunction, and $E \cap E^*(q) \neq \emptyset$ for disjunction. The success probability is allowed to depend on the full reachable context and the reasoner used by the branch. The reference p_0 is the ordinary Reasoner's empty-evidence success rate for this question and checkpoint. It is not replaced by an oracle's refusal rate or by a dimension-level average.

$O_{\min}$ changes failed outputs while copying already valid outputs exactly. O_{full} replaces the output with a canonical annotated version, which can change a valid request's width or its sampled frames. Both respect information access: perception uses content available through the actual request, and reasoning uses the records actually delivered. Joint repairs are executed in planner–perception–reasoner order on the inputs produced by that branch. Reference answer labels are never supplied to a repair branch as privileged inputs. An oracle Reasoner's answer is derived from its permitted records and rules. The reference annotations determining validity are independent of the analyzer being evaluated.

Conditions for a single effective bottleneck. Let $j = \mathrm{own}(k^*(r^*))$. The following conditions concern the replay distribution, not just one observed failed execution.

(A1) Structural scope. The baseline has $C_q(E) = 0$ almost surely. Partition the other agents into upstream agents U and the remaining agents D. For $i \in D$, repairing i leaves $C_q(E) = 0$

almost surely: the unresolved bottleneck cannot be bypassed through another sufficient path or an information channel exposed only to the oracle. Membership is specified structurally, without selecting the observed replay winner.

(A2) Bounds by intervention branch. Fix $\delta \geq 0$ and $\delta_j^- \geq 0$. At each reachable context with $C_q(E) = 0$, the baseline's conditional success probability lies in $[p_0 - \delta, p_0 + \delta]$. For repair of j, only the lower bound $p_0 - \delta_j^-$ is required on its remaining insufficient contexts. For each competing repair $i \in D$, only the upper bound $p_0 + \delta$ is required. These bounds concern the reasoner actually used by each branch and include true partial evidence on conjunctive questions. A bound on a zero-probability conditioning event is unneeded.

(A3) Upstream bound. For $i \in U$, $\Delta_i \leq u$, with $u \geq 0$. Under $O_{\min}$, $u = 0$ follows if every reachable upstream output is already valid and hence copied unchanged, with all other mechanisms fixed. Passing a coverage predicate alone does not establish this property for O_{full}.

(A4) Restoration. Repairing j gives $C_q(E) = 1$ with probability at least π_j, where $0 \leq \pi_j \leq 1$. At each such context, success is at least $p_{\text{suff}} > p_0$. This lower bound ranges over the sufficient sets actually reachable after repair, including alternative disjunctive paths.

Proof. By (A1)–(A2), $p_0 - \delta \leq V_\tau^O(\emptyset) \leq p_0 + \delta$. For $i \in D$, its upper bound and the baseline's lower bound give $\Delta_i \leq 2\delta$. For $i \in U$, (A3) gives $\Delta_i \leq u$. If $\alpha_j = \Pr(C_q(E) = 1 \mid h_\tau, O_j) \geq \pi_j$, then (A2) and (A4) give

$$V_\tau^O(\{j\}) \geq \alpha_j p_{\text{suff}} + (1 - \alpha_j)(p_0 - \delta_j^-) \tag{13}$$

$$\geq \pi_j p_{\text{suff}} + (1 - \pi_j)(p_0 - \delta_j^-), \tag{14}$$

$$\Delta_j \geq \pi_j(p_{\text{suff}} - p_0) - (1 - \pi_j)\delta_j^- - \delta. \tag{15}$$

The second inequality uses $p_{\text{suff}} > p_0$ and $\delta_j^- \geq 0$. Under Eq. (2), this lower bound strictly exceeds $\max\{u, 2\delta\}$, an upper bound for every $i \neq j$. Thus j is the unique maximizing agent. For $u = 0$, the condition is $\pi_j(p_{\text{suff}} - p_0) > (1 - \pi_j)\delta_j^- + 3\delta$. With $\delta_j^- = \delta$, this reduces to the special case $\pi_j(p_{\text{suff}} - p_0) > (4 - \pi_j)\delta$. □

An insufficient-evidence response receives $R = 0$. A competing oracle in D that returns this response whenever required facts are missing satisfies its upper bound on those contexts, regardless of its distance below p_0. No competitor lower bound is used in the proof. This does not certify its upper bound at other insufficient contexts or establish (A1).

If the first-break repair itself can leave insufficient evidence, its residual lower bound must be checked separately. The choice $\delta_j^- = p_0$ always supplies the nonnegative-success lower bound and allows strict refusal. It gives $\Delta_j \geq \pi_j p_{\text{suff}} - p_0 - \delta$. The sharper choice $\delta_j^- = \delta < p_0$ requires positive residual success and cannot describe a branch that refuses on all of those contexts. If restoration is certain, the residual term vanishes. These choices concern the declared branch distribution, not a parameter fitted to make an observed replay winner satisfy the inequality.

For conjunctive questions with breaks on distinct agents, repairing one may leave another required record unavailable. If all single-agent branches remain insufficient and each has success at most $p_0 + \delta$, the baseline lower bound gives gains at most 2δ without selecting a unique winner. Stochastic re-execution can also recover an unmodified failure, so multiple observed breaks do not imply zero single-agent gain in every replay protocol. The same issue applies to (A1): an observed failure does not imply that the baseline replay distribution is always insufficient. When these conditions fail, the proposition supplies no ranking guarantee.

The record-level construction supplies the candidate owner pool $B(\tau) = \{i : (k, i, t, r) \in \Phi^{\text{diag}}(\tau)\}$. Its answer-level role is evaluated through $V_\tau^O(S)$ for joint interventions and removal of individual members (Appendix E.3). OR recovery can use a subset of this pool; training uses the sufficient-path targets of Appendix C.3. The reward remains answer correctness. Equal BRPO weights on the selected actions are evaluated by the separate training ablation in Section 4.4.

C.2 PREDICATE DEFINITIONS

Coverage and matching. For a planner request ρ carrying a window $w(\rho)$ and a requested modality set $M(\rho)$,

$$\mathrm{cover}(\rho, r^*) \iff \Big[\mathrm{IoU}\big(w(\rho), w(r^*)\big) \geq \eta \ \vee \ w(r^*) \subseteq w(\rho)\Big] \ \wedge \ m(r^*) \in M(\rho), \tag{16}$$

with $\eta = 0.5$. Address matching $\mathrm{addr}(r) \simeq \mathrm{addr}(r^*)$ uses the same temporal criterion together with modality equality. These are the complete coverage and matching conditions; the phase field adds no independent equality or compatibility test. The disjunction with containment matters: a planner requesting a generously wide window that covers the decisive instant should not be penalized for low IoU, since the perception agent can still acquire the right evidence within it.

Inputs and deterministic execution. The runtime first extracts requests, tool returns, delivered record IDs, and their parent-action IDs from the log. Explicit tool metadata takes precedence over generated descriptions. Structured fields are read directly; the analyzer extracts fields from remaining text and retains their source spans. This extraction can disagree with an independent reading of the same log. The reference addresses are frozen annotations in both branches. Timestamps are converted to source-video seconds using tool timestamps and clip offsets. Intervals are half-open with positive duration; intersection and containment use unrounded endpoints, including equality at a boundary. Modality names use a fixed vocabulary and alias map. No phase-name or content judgment enters address matching.

For each reference ID, L_1 takes the disjunction of coverage over issued requests. Given $L_1 = 1$, L_2 checks for a delivered record with a matching address and a valid parent tool return under a covering request. All matching record IDs are retained in fixed ID order; content quality cannot remove an address match. A valid empty collection fails the corresponding check. If a necessary field or parent mapping is unreadable, the result is unknown unless another candidate establishes a pass. Later links require confirmed passes. Both checks are executed by code, and their states are immutable when the model supplies the semantic judgments. Code then constructs Eq. (1) and the owning action from the retained IDs.

When several records match one reference, each candidate retains its own description and citation chain. Any complete chain completes that reference; otherwise an unresolved candidate blocks a confirmed first break. If every candidate terminates at a confirmed failure, the longest passing prefix determines the reference's break, with fixed-ID tie-breaking. This prevents combining a faithful description from one record with a citation to another. Execution logs retain the raw spans, extracted fields, candidate IDs, predicate states, and parser/checker versions. Appendix E.2 separates execution agreement on these fields from end-to-end diagnosis.

For jump evidence tied to takeoff or landing, the annotated visual events anchor the required window $w(r^*)$. Phase therefore affects L_1 and L_2 through that window and the temporal criterion above. The field ϕ remains part of the record address for description and hindsight, but its name alone cannot pass or fail a link. A broad request can cover both events, and matching phase names cannot compensate for an incorrect window or modality. Holding windows and modalities fixed, changing only a phase name leaves both checks unchanged.

Takeoff and landing are annotated to within roughly $50\,\mathrm{ms}$, whereas entry onset is a convention with an order of magnitude looser agreement (Appendix B.4). The entry timestamp is not used as an automatic boundary for the coverage checks. Entry-related questions retain their annotated evidence windows, and entry remains available as descriptive context for hindsight. No additional phase-specific tolerance is introduced.

The overlap threshold is fixed at $\eta = 0.5$ after calibration on 250 trajectories; Appendix E.4 reports its sensitivity.

Model-judged predicates. L_3 checks both factual correctness and completeness relative to the question; "faithful" in the link labels includes both requirements. The source-grounded reference specifies the decisive facts assigned to each reference record before inspecting the evaluated description. A matched record passes only if its relevant claims are supported by the source and its description conveys all of those facts, allowing equivalent wording. It need not describe every visible detail. For an AND question, each record must convey its assigned contribution; it need not

answer the whole question alone. For an OR question, each alternative has the single-record support requirement of Appendix B.4.

A confirmed false statement or omission of an assigned decisive fact fails L_3. For example, "free foot low" does not convey a visible touch of the ice. When L_1 passes but no acquired record matches the required address, the first break is at L_2. An unavailable or ambiguous source comparison instead leaves L_3 unknown. L_4 checks citation and interpretation of the facts actually delivered. An omitted fact cannot be supplied by the reference answer or treated as a Reasoner-only failure after incorrectly passing L_3.

Each check returns pass, fail, or unknown. Unreadable fields, an unresolved content comparison, or an unknown predecessor cannot supply a confirmed first-break target. A parseable answer with no valid citation is a definite use failure; it enters Φ^{diag} at L_4 only when the preceding links pass. Irrespective of its first-break location, an absent citation cannot establish a complete delivery chain.

The decision instructions are fixed during development. Unknown judgments are logged separately from confirmed failures and confirmed insufficient support. The independent link evaluation and its missing-prediction column appear in Appendix E.2; a missing prediction is retained in the evaluation denominator. Predicate confidence controls target eligibility, while the token gate in Eq. (6) controls the strength of hindsight on already eligible actions.

C.3 Training Algorithm and Objectives

Sufficient-path selection. The diagnostic set Φ^{diag} retains every confirmed record-level first break. Let $\mathcal{F}(\tau) = \{r \in E^*(q) : k^*(r) \text{ is defined}\}$. A record in $\mathcal{F}$ is confirmed broken even when links downstream of its first failure are unevaluated. A record in neither D nor $\mathcal{F}$ is unresolved. For the resolved, nonempty pure structures used in training, select the target record set

$$S^\dagger(\tau) = \begin{cases} \emptyset, & H = 1, \\ E^*(q), & H = 0,\ \sigma(q) = \wedge, \\ \{r^\dagger\}, & H = 0,\ \sigma(q) = \vee,\ \mathcal{F}(\tau) = E^*(q), \\ \emptyset, & \text{otherwise.} \end{cases} \tag{17}$$

For the OR branch, define $p_\tau(r) = k^*(r) - 1$ and choose $r^\dagger = \arg\max_{r \in E^*(q)} p_\tau(r)$, resolving ties by the smallest fixed reference-record ID. IDs are assigned with the annotations and retained under list reordering. This selects the longest confirmed passing prefix; it orders observed progress rather than oracle repair value. The $H = 1$ branch takes precedence even if an unused OR alternative is unknown. With no complete OR path, any unresolved alternative instead makes selection abstain. AND retains its confirmed breaks even when other records are unknown.

Filtering precedes action projection and hindsight construction:

$$\begin{aligned} \Phi^{\mathrm{train}}(\tau) &= \{(k, i, t, r) \in \Phi^{\mathrm{diag}}(\tau) : r \in S^\dagger(\tau)\}, \\ \Phi^{\mathrm{train}}_{\mathrm{act}}(\tau) &= \{(i, t) : (k, i, t, r) \in \Phi^{\mathrm{train}}(\tau)\}. \end{aligned} \tag{18}$$

Each S_i uses only the tuples for action (i, t) retained in Φ^{train}, even if excluded records have the same owner. The selector leaves D, H, Ψ unchanged for positive credit. Its outputs are detached and fixed during the round. Empty filtered targets suppress negative task credit and LHD; positive credit still follows H, Ψ. KL and full-batch loss denominators are retained.

Hindsight context by break position. The privileged context supplied to an agent uses only records retained in Φ^{train}, with content fixed by the selected failed link (Table 8).

Each role receives only the selected records and link-specific context in Table 8.

Reward groups and extra sampling. Binary answer correctness is the sole task reward. Schema and citation-format indicators are logged separately. Both constant-outcome group types have $\hat{A} = 0$, but confirmed path-filtered breaks remain eligible for LHD, including unsupported correct answers. All groups are diagnosed before selection and retained in token counts and loss denominators.

Table 8: Link-specific privileged context for localized hindsight.

Break	Recipient	Privileged context
L_1	A_P	phase and window of r^*; the modality subset that suffices
L_2	A_M	the issued request; the evidence obtainable within it
L_3	A_M	the misstated or omitted decisive facts in $c(r)$, grounded in the acquired source
L_4	A_R	the delivered records and their correct reading

If fewer than half of the base question groups contain a nonzero routed task advantage or an eligible LHD target, one fresh batch adds 12 question groups (96 trajectories) to the eight base groups (64 trajectories). This is a single addition at 1.5 times the base question count, without replacement or recursive refilling. Local controls apply the trigger to their own enabled targets. Coverage, unresolved attribution, and effective LHD mass are reported in Tables 22 and 33.

Token losses and normalization. Each role's turn-level advantage is broadcast over its generated tokens; prompt, padding, and environment tokens are excluded. Task and LHD sums use the full generated-token denominator for that role, including zero-credit and LHD-ineligible actions. Matched controls retain this convention so abstention does not amplify the remaining terms. The current-round snapshot $\bar{\theta}_i^k$ supplies the KL reference. Asymmetric ratio clipping uses $\epsilon_{\text{high}} > \epsilon_{\text{low}}$, widening the positive trust region following Yu et al. (2025).

Clipped task objective and KL estimator. Let $m_{i,t,\ell}$ exclude prompt, padding, and environment tokens and $Z_i = \sum_{g,t,\ell} m_{i,t,\ell}^{(g)}$. On tokens sampled from the frozen round policy, define

$$\rho = \frac{\pi_{\theta_i}(a \mid h)}{\pi_{\bar{\theta}_i}(a \mid h)}, \qquad d = -\log \rho, \qquad K = \rho(\exp d - d - 1).$$

The task loss is

$$\mathcal{L}_i^{\text{BRPO}} = \frac{1}{Z_i} \sum_{g,t,\ell} m_{i,t,\ell}^{(g)} \left[-\min\{\rho A_{i,t}^{(g)}, \text{clip}(\rho, 0.80, 1.28) A_{i,t}^{(g)}\} + 0.005K \right]. \tag{19}$$

For the full-support rollout sampler below, the expectation of K under the reference token distribution equals $D_{\text{KL}}(\pi_{\theta_i} \| \pi_{\bar{\theta}_i})$ at the stored history. Gradients pass through ρ and d, not through the reference or advantages. KL applies to every generated token, including zero-credit actions. The same denominator applies to the masked LHD sum. A role with $Z_i = 0$ has no loss that round. Two optimizer passes allow the policy to move away from its collection snapshot before the second surrogate evaluation.

Rollouts, including video decoding and pose estimation, account for roughly 80% of wall-clock time. Retaining usable LHD targets in zero-advantage groups recovers a training signal from these generated trajectories.

Weight cap. With $M_g = 3$, selecting one, two, or three actions gives weights 3, 1.5, or 1 before clipping. The default $w_{\max} = 2$ changes only singleton weights; every cap at least three is equivalent to no clipping. Empty sets retain zero weight, and clipping can reduce total weight below M_g. Appendix D.4 reports cap sensitivity, and Appendix G.2 gives numerical examples.

Algorithm 1 keeps diagnoses, routing masks, and hindsight fixed through both optimizer passes. Implementation and decoding settings are collected in Appendix C.5; worked examples appear in Appendix G.2.

Algorithm 1 One training round of BRPO + LHD

Require: policies $\theta_P, \theta_M, \theta_R$; evidence annotations and fixed predicate checks
$\bar{\theta}_i \leftarrow \theta_i$ for each agent i
for each question q in the batch **do**
  sample G complete trajectories from the frozen policies
  compute $R^{(g)}_{\text{task}} = \mathbb{1}[\hat{y}^{(g)} = y^*]$; log format indicators separately
  compute $\hat{A}^{(g)}$ by (3) ▷ zero in both constant-outcome groups
  for each trajectory $\tau^{(g)}$ **do**
    extract log fields and IDs; execute and lock L_1/L_2 by fixed code
    judge L_3/L_4 on matched chains; assemble the graph and predicate states
    construct record diagnosis Φ^{diag}, completed records D, delivery actions Ψ, and gate H
    select $S^\dagger$ by (17); construct Φ^{train} and $\Phi^{\text{train}}_{\text{act}}$ by (18)
    select $\mathcal{C}$ by (4); compute w and A by (5)
    if some $A^{(g)}_{i,t} \neq 0$ **then**
      accumulate the clipped BRPO task term on selected actions
    end if
    for each $(i,t) \in \Phi^{\text{train}}_{\text{act}}(\tau^{(g)})$ **do**
      construct S_i from its tuples in Φ^{train}; score the same tokens with and without hindsight
      accumulate the masked LHD term of (6)
    end for
  end for
end for
count groups with task or LHD targets; if fewer than half, resample once at multiplier 1.5
process added groups by the same rules; retain the configured KL regularizer
take two optimizer passes per agent with the round snapshot fixed

C.4 Analyzer Initialization and Monitoring

Initialization. Writing $X^{\text{priv}}_i(\tau)$ for the privileged context of role i and $\Phi^{\text{diag},*}$ for the reference record diagnosis, the initialization objective is the likelihood of the structured target,

$$\mathcal{L}^{\text{init}}_i = -\sum_{\ell} \log \pi_{\theta_i}\big(\Phi^{\text{diag},*}_\ell \mid X^{\text{priv}}_i(\tau), \Phi^{\text{diag},*}_{<\ell}\big). \tag{20}$$

This initialization teaches the structured diagnosis format. At execution, the learned stage supplies field extraction and semantic judgments, while fixed checks constrain the final record, link, and turn assignment. The generated diagnosis cannot override these checks. It provides no question-answering supervision. It is run once, before reinforcement learning begins (Wu et al., 2026), using one epoch over 150 trajectories collected after format alignment and before RL, with the optimizer in Appendix C.5. Optimizer moments are reset before the reinforcement-learning schedule begins.

Refresh and monitoring. After each update, $\bar{\theta}^{k+1}_i \leftarrow \theta^{k+1}_i$ copies the policy snapshot without another optimization stage. Every 200 updates, each seed generates 100 fresh development trajectories. The refreshed and initialization-frozen analyzers score the same trajectories against $K = 3$ per-agent replay. Each checkpoint therefore costs $100 \times 3 \times 3 = 900$ executions; six subsequent checkpoints cost 5,400 per seed (about 7% of rollout volume), plus 900 at initialization. This stream excludes confirmation programs; partition rules appear in Appendix E.1.

Figure 7 compares agreement at each checkpoint. Changes across checkpoints also reflect the current policy's sampled trajectories and low-budget replay references. Link-frequency curves summarize analyzer predictions; independent final link labels are supplied by Appendix E.2.

Audit outputs. Training logs retain task reward, format flags, advantage, predicate states, Φ^{diag}, D, H, Ψ, $S^\dagger$, Φ^{train}, its action projection, $\mathcal{C}$, weights, and both loss masks. Stored structural fields and semantic judgments permit re-execution of routing decisions. Confirmed insufficient support, unresolved judgments, and missing negative targets remain distinct. Appendix E.2 evaluates break localization and execution consistency; Appendix F evaluates support gating and witnessed delivery.

A link error can preserve the owner while changing hindsight, so the two evaluations retain record-level labels.

C.5 Implementation and Sampling

Configuration. All experiments run on $4\times$ A100 (80 GB). The shared backbone is Qwen2.5-Omni-7B (Xu et al., 2025), chosen because the perception agent must form claims over audio as well as vision; the three roles differ by LoRA adapters (Hu et al., 2022) of rank 32 applied to attention and MLP projections, trained in bf16 with gradient checkpointing and ZeRO-2 sharding (Rajbhandari et al., 2020). Rollout and update alternate on the same devices. Table 9 lists the training settings.

Table 9: EDGE and local-control training specification. The 64 trajectories are the base batch before conditional resampling. External methods retain their stated objectives; sampler and run bindings are specified in Appendix C.5.

Group	Setting	Value
Rollout	group size G	8
	questions per batch	8
	trajectories per update	64
	max frames per acquisition	64
	program-level sampling	2 fps
	max context	32k tokens
BRPO	$\epsilon_{\text{low}}/\epsilon_{\text{high}}$	0.20 / 0.28
	β_{KL}	0.005
	$w_{\max}$	2.0
LHD	β_{lhd} (gate temperature)	1.0
	λ_{lhd} initial	0.5
	λ_{lhd} schedule	linear to 0 over second half
Optim.	learning rate (all roles)	1×10^{-5}
	warmup / schedule	20 outer updates / cosine
	LoRA rank / α	32 / 64
	outer updates / passes per batch	1,200 / 2
	wall-clock	$\approx$ 72 h

Sampling and parameter scope. The base backbone is frozen; only the three role adapters are trained, with LoRA dropout zero. AdamW uses moments $(0.9, 0.999)$, numerical stabilizer 10^{-8}, zero weight decay, and global gradient-norm cap 1.0. The advantage stabilizer is 10^{-6}. Each outer update makes two optimizer passes over the collected batch while holding its reference snapshot, sampled tokens, advantages, attribution masks, and hindsight targets fixed. The reference refreshes after the second pass, giving 2,400 optimizer steps per role at the 1,200-update endpoint. Warmup, cosine decay, and the LHD coefficient use outer-update time. A microbatch contains one trajectory per GPU; gradient accumulation spans 16 microbatches for the base batch and 40 when all 12 extra question groups are added. Partial microbatches retain their actual token counts. Each role is normalized by its full-batch generated-token count, then the three role losses are averaged. Sampling and length caps appear in Appendix C.5; external objectives and their task-specific adaptations are specified in Appendix D.1.

Supervised format alignment. Before reinforcement learning, all methods with the structured interface receive one supervised epoch on 400 reformatted trajectories. Targets teach record schemas and identifier citations without correcting answer content or adding question-answering supervision. The optimizer is specified above. A disjoint 720-question validation panel supplies eight draws per checkpoint (5,760 perception turns and answers). Before/after alignment uses one fixed initialization; final RL rates average three seeds. Table 29 reports schema/citation validity before alignment (62.3/48.7%), after alignment (97.5/95.1%), and after training. Format indicators are monitoring metrics rather than reinforcement rewards or sufficient-support labels.

Sampling and decoding. Table 10 specifies units and counts. Paired seeds align adapter initialization and data order where compatible, not generated trajectories. Probe settings are fixed across each question's modality conditions; ordinary-policy replay and Reasoner interventions retain their respective samplers across conditions. Deterministic oracle templates and minimal-repair copies are logged separately.

Table 10: Sampling units and counts. Decoding parameters are fixed by use below.

Use	Specified sampling unit
Task training	Eight base questions per update. EDGE: eight trajectories each, with at most 12 extra groups. IAPO: eight independent trajectories each, no extra groups. C3: a $2 \times 2 \times 2$ branching tree, eight terminal leaves per question, no extra groups.
Main benchmark	One greedy output per question and checkpoint; API temperature zero.
Sufficiency and empty evidence	Eight draws per question and modality subset for each probe; empty-evidence Reasoner draws recorded separately.
Calibration and monitoring	Five replays per agent on the diagnostic set; three per agent in training monitoring.
Independent repair validation	Auxiliary repairs: 64 per condition. Independent confirmation: 16 to 256 per repair branch, with 64 baseline repetitions; independent audit: 256.
Positive evidence interventions	16 Reasoner draws per eligible prefix and condition.

Training seeds are 13, 37, and 73, with stream offsets 1,000 for question order, 2,000 for rollouts, 3,000 for probes, 4,000 for replay/interventions, 5,000 for analyzer generation, and 6,000 for hindsight. Each stochastic call also keys its stream by question, update, condition, and repetition. Training rollouts use temperature 1.0, top-$p = 1$, and no top-k truncation. Probes, ordinary-policy replays, and evidence interventions use temperature 0.7, top-$p = 1$, and no top-k truncation. Benchmark outputs, analyzer labels, and hindsight text use greedy decoding; greedy calls consume no sampling RNG. All use repetition penalty 1.0 and stop at EOS or the output cap. Oracle templates and exact copies do not use a language-model sampler.

For the shared trained backbone, Planner/Perception/Reasoner output caps are 512/2,048/1,536 tokens; single-role baselines use 4,096. Probes have a 1,536 token cap, and analyzer/hindsight text a 1,024-token cap. Untrained API multi-agent runs use 1,024/4,096/4,096 role caps; single-call API baselines use 8,192, including billed reasoning tokens where returned by the provider. These are generation caps, separate from the 32k context limit. Interventions and ordinary replay preserve the evaluated policy's role cap. A truncated or invalid answer scores incorrect. Provider versions and their actual decoding support are part of the execution export.

Execution records and shared budgets. Each method–seed result binds the code version, backbone and adapter initialization, task-policy and analyzer checkpoints, reward, optimizer, sampler, and evaluation export. Exports identify program splits, question IDs, frozen references, window parsers, and human annotation versions. C3 additionally binds its replay provider and branching allocation; IAPO binds its extractor, prompt, parser, and fallback. The dependency-ranked control binds its shared analyzer and separate graph prompt.

Repeated EDGE rows use the same three final-policy output sets. C3 and IAPO rows in Table 4 likewise reuse the checkpoints and outputs of Tables 1–2, because both already exclude hindsight. Local and interface ablations use their own configurations and inputs; published-backbone, reduced-scale, cap, alignment, and budget-checkpoint studies retain separate runs. Partition and annotation accounting are centralized in Appendices E.1 and F.1.

Each trained structured system has a cap of four requested window–modality pairs and four localized output pairs per question, with common tools and per-acquisition frame limits. Invalid or empty outputs contribute zero counts. Appendix D.5 defines update counting and the secondary generation ceiling. Seed variation, program sampling, and unresolved-label bounds are reported separately under Appendix B.7.

D SUPPLEMENTARY COMPARISONS AND ABLATIONS

D.1 BASELINE IMPLEMENTATIONS AND ADAPTATIONS

Table 11 aligns resources while preserving each external method's stated mechanism. Greedy benchmark evaluation permits 6 tool calls, 32k context tokens, and at most 64 frames per acquisition, with program views sampled at 2 fps. These are ceilings where an interface supports the operation; actual consumption appears in Table 32. EDGE's tool budget is separate from its three role actions. Option letters are scored by exact match; unparseable outputs are incorrect. Public proprietary APIs use temperature zero. Stochastic experiments follow Appendix C.5.

Table 11: External-method adaptations. The defining mechanism is preserved; backbones, training data, and supported interfaces are aligned as listed.

Method	Kept	Changed
VideoMultiAgents	Independent modality agents aggregated by an organizer; no training	Backbone → Gemini 2.5 Pro; pose and audio agents augment the published agent set
MUPA	Multi-path grounding and answering with a reflective aggregator	None; released 7B weights, frames sampled as in our pipeline
Video-R1	Single-pass video RL with temporal contrast (T-GRPO)	Backbone → Qwen2.5-Omni-7B; pose and audio as pre-extracted streams; trained on SKATEEG
SEED	One agent; episode-level hindsight skills; self-evolving on-policy distillation	Backbone → Qwen2.5-Omni-7B; our tool interface; analyzer given the same annotations as ours
MAGRPO	Centered joint-return policy gradient; no ratio clipping or KL	Our three roles, terminal task reward, and structured interface
MASIRL	One trainable component guided by a frozen verifier	Grounder mapped to planner and perception; reasoner frozen; 7B
MATPO	Planner and worker roles in one model, with reward distributed between them	Planner ↔ our planner and reasoner; worker ↔ perception; Qwen2.5-Omni-7B
C3 (adapted)	Same-prefix action alternatives, ordinary-policy continuation, reward-only leave-one-out credit	Three multimodal roles; $2 \times 2 \times 2$ branching; shared 7B adapters and token-level optimizer
IAPO (adapted)	Typed dependency features, sign-conditioned length-normalized weights, uniform fallback	Our equation-based implementation; three role nodes, 7B adapters, eight samples and shared optimizer

Video-R1 receives full-program pose and audio streams as pre-extracted inputs; it makes no selective tool calls or evidence records, so record-based metrics are undefined. SEED receives the same training annotations as EDGE, including $E^*(q)$, phase boundaries, and sufficiency labels. MASIRL's trainable grounder covers planner and perception functions while its verifier/reasoner remains frozen, retaining its span-calibration loss. MATPO's planner requests and answers, and its worker perceives, preserving its two-role reward allocation. The single-agent baseline uses MAGRPO's codebase with one role and the same tools and records, but a GRPO update rather than published MAGRPO. Their comparison varies both role count and optimization. DeepSport (Zou et al., 2025), VideoChat-M1 (Chen et al., 2026a), and SVAgent (Yang et al., 2026b) are omitted because public code was unavailable at the evaluation snapshot.

External MAGRPO versus the local uniform control. MAGRPO follows Liu et al. (2026), Eqs. 1–2, with the objective and collection rules in Table 12. Its sequence log probabilities exclude prompt and environment tokens and are not length-normalized. The local control retains EDGE's optimizer schedule. Their 0.2-point accuracy difference combines the listed objective and sampling changes.

C3 and IAPO implementation boundaries. The C3 configuration follows Chen et al. (2026b) and the authors' public implementation, with `reward_only` and leave-one-out estimation; a group-mean fallback without checkpoint continuation is not this comparison. IAPO is our implementation

Table 12: MAGRPO and the local uniform control. Both use terminal binary correctness and no LHD. Their 57.4 and 57.6 accuracy values correspond to separate configurations.

Setting	MAGRPO	Uniform task credit
Advantage	$R^{(g)} - \mu_q$	$(R^{(g)} - \mu_q)/(\sigma_q + \epsilon)$
Action weights	One for every role	One for every role
Task objective	Advantage times sequence log probability	Clipped token-ratio surrogate
Task reduction	Mean over responses; sum log probabilities within a response	Divide by all generated tokens of the role
Ratio clipping	None	$[1 - 0.20, 1 + 0.28]$
KL coefficient	0	0.005, current-round reference
Group collection	Fixed base batch of fresh groups	EDGE's conditional extra batch

of the equations in Ren et al. (2026); we could not verify an official code release at the comparison snapshot. Both are explicitly adapted configurations, not exact reproductions of the published systems. They share the 7B backbone, three role adapters, structured interface, format alignment, tools, binary task reward, and test split. They use the learning rate, two optimizer passes, token masks, $[0.80, 1.28]$ ratio clip, and 0.005 current-round KL in Appendix C.3. Neither uses LHD, EDGE's support gate, target abstention, or its action-weight cap. Each role loss divides by that role's actual generated-token count, and the three losses are averaged. Sampling and credit estimation differ as below.

C3 freezes the behavior policy for collection and optimization. At each question, two Planner actions branch into two Perception actions each and then two Reasoner actions each: eight terminal rewards and up to $2 + 4 + 8 = 14$ distinct generated actions. For an action a at a restored history h, let $\mathcal{L}(h, a)$ be its descendant leaves and define

$$\widehat{Q}(h, a) = \frac{\sum_{\ell \in \mathcal{L}(h,a)} R_\ell}{|\mathcal{L}(h, a)|}, \qquad \widehat{A}(h, a) = \widehat{Q}(h, a) - \widehat{Q}(h, a'),$$

where a' is its sibling sampled from the identical history. Subtree returns average four, two, and one leaves at the three roles. This balanced allocation is our three-role adaptation of fixed-history leave-one-out credit; shared descendants are correlated and are not independent replay replicates. Each distinct action enters its owner's loss once per optimizer pass, with its own advantage sign and no group-SD rescaling. Reused prefixes contribute neither duplicate losses nor duplicate generated-token counts. A snapshot contains resource IDs, selected windows/frames, tool cache, messages, record IDs, role adapter, and decoder state. Branch-specific RNG streams produce alternatives; deterministic media preprocessing is cached by input. Changing a request reacquires the matching observations. Later agents follow the ordinary behavior policy, not oracle replacements. Missing or schema-invalid role outputs terminate the branch with reward zero; missing descendants remain zero-reward leaves and incur no generated-token loss. This preserves the fixed allocation without dropping failures.

IAPO represents complete Planner, Perception, and Reasoner outputs as three ordered action nodes. Tool observations supply evidence and inherit the action that elicited them; they receive no policy loss. Following the paper's Eqs. 3–7, support-use and failed-use edges define

$$\phi_i^+ = \log(1 + d_i^+ + d_i^{\text{in}}), \qquad \phi_i^- = \log\big(1 + e_i(1 + d_i^- + d_i^{\text{in}})\big).$$

Here d_i^+ and d_i^- count outgoing edges of each type, d_i^{in} counts incoming support and direct user input, and e_i marks an observable error. The published length-weighted standardization, exponential reweighting, and renormalization use $c = 0.25$, $\beta_+ = 0.05$, and $\varepsilon = 10^{-6}$. Negative credit uses m_i^-; positive credit uses error-discounted m_i^+. Thus $\sum_i L_i w_i = \sum_i L_i$ over generated action lengths before the shared role-wise loss reduction; uninformative features give uniform weights. A zero trajectory advantage still gives zero task credit. Our rollout has three role turns rather than a service dialogue, eight rather than 16 samples per question, and one final checkpoint rather than a last-three-checkpoint average. These and the shared optimizer are changes to the original setup.

The IAPO extractor is frozen Qwen3-32B, with greedy decoding, thinking disabled, a 32k context limit and a 1,024-token output cap. Its input is the question, ordered action text and structured

tool log, without answer keys, reference records, support labels, or EDGE diagnoses. The graph prompt instructs it to identify concrete information used by each later action, attach source and consumer spans, distinguish support from reuse of an observable failure, and reject thematic-only links. For overlapping sources it retains the most informative source, breaking ties by earliest action; distinct information units may have different sources. The JSON fields are `edges` (source, target, type, source span, consumer span), `errors` (action, error span), and `direct_input` (action, question span). The parser requires valid forward IDs, exact input spans, allowed types, deduplicated edges, and an error marker for each failed-use source. One invalid extraction triggers a single repair call with its validation errors; another failure yields an empty graph and uniform weights. Graphs are cached by the full rollout, extractor checkpoint, and prompt version and fixed through both optimizer passes. The local dependency control uses the same schema with reference-aware errors (Appendix D.3).

Table 13: Credit-estimation inputs. All configurations receive terminal task reward and visible trajectories. Reference annotations are training-only; test support labels never enter optimization.

Configuration	Extra inputs	Credit estimator
C3 (adapted)	Restorable tool/media state	Ordinary-policy continuations; no diagnostic model
IAPO (adapted)	None	Frozen Qwen3-32B dependency extractor
BRPO / EDGE	Training reference records, phase and sufficiency labels	Refreshed EDGE analyzer
Dependency-ranked targets	Same reference inputs as EDGE	Same analyzer checkpoint; separate graph extraction without first-break outputs

Table 13 makes the information difference explicit. The external comparison evaluates task-adapted algorithms; the reference-aware local control tests negative-target choice with model and information access held fixed. Neither ordinary-policy marginal advantage nor graph influence is identical to the oracle-repair gain in Proposition 1. The free-form inferred-graph row remains a separate interface ablation.

External-method reference configurations. Table 14 compares the listed external-method reference configurations with their adaptations; C3 and IAPO use only the task-adapted configurations specified above. Adapted scores increase by 1.2–6.1 points, with the largest gain for MASIRL's scale change. MATPO and Video-R1 gain 7.1 and 6.2 C5 points, respectively; each adaptation changes several factors, so these are package-level comparisons. SEED gains 1.2 points with matched annotation access and remains 4.6 below EDGE. Table 1 uses the stronger configuration for these paired comparisons; local ablations, rather than these external margins, isolate EDGE components.

Proprietary backbones. Table 15 substitutes three backbones in the two untrained agentic systems. Gemini accepts audio natively; GPT-5 and Claude receive structured beat/accent times and tempo changes. Average accuracy varies by at most 1.4 points, and EDGE's untrained pipeline leads VideoMultiAgents by at least 1.5 throughout. C5 differs by 5.0–5.6 points, with backbone and audio representation changing together. The trained system's advantage over its untrained pipeline is 3.5–4.8 accuracy points and 0.09–0.10 IoU.

Reference models. Table 16 evaluates open, proprietary, and grounding-specialized models without SKATEEG training under the common metrics.

Grounding-specialized models have higher IoU but lower accuracy than strong general models, whose IoU remains near 0.35. This separates answering from temporal localization, as observed in general-domain benchmarks (Xiao et al., 2024; Dang et al., 2025); it does not measure evidence-use frequency.

D.2 Evidence Quality and Statistical Variation

Table 17 decomposes the main evidence comparison. EDGE's IoU margin over MATPO is 0.06 on element identification, 0.08 on temporal structure, and 0.09 on musicality; its musicality Modality-

Table 14: Reference versus adapted external configurations. Δ is adapted minus reference. All trained configurations use the SKATEEG training split.

Method	Configuration	Avg.	C5	IoU	Δ Avg.
VideoMultiAgents	Reference: GPT-4o, three agents	55.7	44.4	—	
	Adapted (main table)	58.2	55.0	—	+2.5
MUPA	Released 2B	48.9	39.4	0.37	
	Released 7B (main table)	51.5	41.3	0.40	+2.6
Video-R1	Released checkpoint, zero-shot	49.3	38.8	—	
	Published backbone, trained	54.6	41.3	—	
	Adapted (main table)	55.8	47.5	—	+1.2
SEED	Reference: no privileged annotations	57.4	49.4	0.43	
	Adapted (main table)	58.6	51.3	0.46	+1.2
MASIRL	Reference scale: 2B	52.8	46.0	0.45	
	Adapted (main table)	58.9	53.1	0.49	+6.1
MATPO	Reference: Qwen3-8B, text tool outputs	55.9	47.7	0.42	
	Adapted (main table)	60.3	54.8	0.47	+4.4

Table 15: Untrained agentic systems under three proprietary backbones. The last column is the margin of our trained 7B system (63.2%) over the untrained pipeline with that backbone.

	VideoMultiAgents		Ours, untrained			
Backbone	Avg.	C5	Avg.	C5	IoU	Trained margin
GPT-5	57.3	50.0	58.9	53.1	0.44	+4.3
Gemini 2.5 Pro	58.2	55.0	59.7	58.1	0.45	+3.5
Claude Opus 4.5	56.8	49.4	58.4	52.5	0.44	+4.8

Table 16: Reference models on SKATEEG, no training. The weighted p_0 is 37.3%. Grounding-specialized systems invert the usual pattern: higher evidence IoU at lower accuracy.

	Model	Acc. (%)	Evidence IoU	Mod.-hit
Open	Qwen3-VL-8B	41.2	0.21	0.34
	Qwen3-VL-235B-A22B	52.6	0.29	0.41
	InternVL3.5-8B	40.5	0.20	0.33
	InternVL3.5-38B	48.9	0.26	0.38
	Qwen2.5-Omni-7B	39.8	0.19	0.37
Prop.	GPT-5	56.2	0.33	0.46
	Gemini 2.5 Pro	57.1	0.35	0.48
	Claude Opus 4.5	55.3	0.32	0.45
Gr.	Grounded-VideoLLM	37.6	0.38	0.31
	SeViLA	35.9	0.36	0.29

hit margin is 0.13. The larger gains occur where records must be combined across time or modalities. The untrained pipeline is close on C3 (0.4 accuracy points and 0.01 IoU), but its C4 Any-citation is only 64.5% despite strong accuracy. MAGRPO exceeds the single agent in all three dimension-level metrics in Table 17 while trailing in accuracy. These comparisons locate the answer–evidence gap, without identifying individual answers' dependence on priors or isolating role decomposition.

Table 17: Evidence metrics by dimension (C1–C5 as in Table 1). IoU and Modality hit aggregate video questions; Any-citation is conditional on correct answers, with each seed using its own denominator. All columns follow Appendix B.7.

	C1	C2	C3	C4	C5	All
Evidence window IoU						
Ours, untrained	0.46	0.48	0.44	0.47	0.38	0.45
Single agent	0.49	0.44	0.30	0.43	0.33	0.41
SEED	0.53	0.49	0.36	0.48	0.37	0.46
MAGRPO	0.52	0.47	0.34	0.46	0.35	0.44
MASIRL	0.57	0.52	0.39	0.51	0.41	0.49
MATPO	0.55	0.50	0.37	0.49	0.39	0.47
C3 (adapted)	0.58	0.52	0.40	0.53	0.43	0.50
IAPO (adapted)	0.57	0.52	0.41	0.52	0.44	0.50
EDGE	**0.61**	**0.57**	**0.45**	**0.56**	**0.48**	**0.54**
Modality hit (Mod.-hit)						
Ours, untrained	0.55	0.60	0.66	0.63	0.54	0.60
Single agent	0.64	0.57	0.56	0.53	0.49	0.57
SEED	0.67	0.60	0.59	0.56	0.52	0.60
MAGRPO	0.65	0.58	0.57	0.54	0.50	0.58
MASIRL	0.68	0.61	0.60	0.57	0.53	0.61
MATPO	0.70	0.63	0.62	0.58	0.58	0.63
C3 (adapted)	0.72	0.65	0.63	0.61	0.70	0.66
IAPO (adapted)	0.71	0.64	0.62	0.60	0.69	0.65
EDGE	**0.74**	**0.68**	**0.67**	**0.63**	**0.71**	**0.69**
Any-citation (% of correct answers citing a record of $E^*(q)$)						
Ours, untrained	85.4	84.3	82.4	64.5	79.6	79.8
Single agent	76.0	72.6	65.2	66.0	67.1	70.4
SEED	81.2	78.6	71.4	71.9	74.4	76.3
MAGRPO	77.1	72.7	65.8	67.2	68.2	71.2
MASIRL	82.2	79.3	72.9	73.6	75.3	77.5
MATPO	85.6	82.6	75.9	76.5	77.9	80.6
C3 (adapted)	87.0	85.5	84.8	78.2	74.2	83.2
IAPO (adapted)	86.5	86.8	**87.0**	79.4	75.9	84.3
EDGE	**91.1**	**89.4**	85.6	**86.2**	**88.1**	**88.4**

Table 18 evaluates the whole-answer condition by structure. EDGE leads MATPO by 14.3 CompleteSupport points on AND questions and 7.7 on OR questions. Removing the positive gate lowers AND support by 24.1 points. AND requires all constituent records; OR retains alternative sufficient paths. C3 and IAPO reach 67.4/68.9 on AND and 81.5/81.6 on OR, below EDGE's 80.6 and 85.5. The larger AND deficit identifies conjunction completion as a remaining weakness even when an alternative credit estimator improves answer accuracy. This comparison alone does not assign that difference to negative routing, since support gating and reference access also differ.

The paired EDGE–MATPO differences in accuracy, Any-citation, CompleteSupport, and GroundedAcc are 2.9, 7.8, 9.6, and 8.3 points. Their respective 95% program-clustered intervals are [1.7,4.1], [5.9,9.7], [7.7,11.5], and [6.6,10.0]. These use 10,000 paired resamples of source programs, retaining all questions and seeds within a sampled program; they address program sampling conditional on these seeds and resolved labels. Seed dispersion in Table 19 is a separate quantity. The interval targets concern these four primary comparisons, without a claim of simultaneous coverage across all ablations.

EDGE's higher localization and coverage scores accompany 2.11 requested and 2.06 delivered pairs, compared with MATPO's 2.13 and 2.08. Output-volume inflation therefore does not explain this

Table 18: CompleteSupport (%) within OR and AND strata. Overall scores weight per-seed correct-answer counts.

Method	OR	AND
MAGRPO	67.8	51.2
MATPO	77.8	66.3
C3 (adapted)	81.5	67.4
IAPO (adapted)	81.6	68.9
Dependency-ranked targets	84.9	76.3
No positive gate	72.3	56.5
EDGE	85.5	80.6

Table 19: Sample standard deviations across the three evaluation seeds for the strict-support tables. Units are percentage points; means appear in Tables 2, 4, and 25.

Configuration	Any-cit.	CS	GA	C5-part.
Single agent	0.5	0.4	0.8	0.4
SEED	0.4	0.5	0.3	0.4
MAGRPO	0.5	0.5	0.2	0.4
MASIRL	0.3	0.5	0.3	0.4
MATPO	0.5	0.5	0.1	0.0
C3 (adapted)	0.4	0.1	0.5	0.4
IAPO (adapted)	0.5	0.1	0.4	0.4
Dependency-ranked targets	0.4	0.2	0.6	0.4
EDGE	0.1	0.5	0.7	0.4
Uniform task credit	0.1	0.5	0.1	0.4
Uniform credit + LHD	0.1	0.4	0.2	0.0
BRPO only	0.1	0.5	0.1	0.4
Negative-only shuffle	0.1	0.4	0.6	0.4
No positive gate	0.1	0.4	0.6	0.0
No zero-variance LHD	0.1	0.5	0.8	0.4
Negative unknown fallback	0.1	0.3	0.3	0.4
Shuffled task weights	0.2	0.4	0.1	0.4
Frozen analyzer	0.1	0.4	0.2	0.4
One cross-agent break	0.1	0.5	0.7	0.4
All cross-agent actions	0.1	0.4	0.2	0.4
All-record OR targets	0.1	0.3	0.6	0.0

Table 20: Localization output volume and sample standard deviations for the trained structured systems. Requests count distinct window–modality pairs; delivered addresses count deduplicated output pairs. Volumes are per-question means over three seeds. SDs use the native $[0, 1]$ metric scales.

Method	Requests	Delivered	IoU SD	Hit SD	Cov. SD
Single agent	2.17	2.10	0.01	0.006	0.006
SEED	2.14	2.07	0.01	0.006	0.007
MAGRPO	2.21	2.16	0.01	0.006	0.006
MASIRL	2.15	2.09	0.01	0.006	0.006
MATPO	2.13	2.08	0.01	0.006	0.006
C3 (adapted)	2.14	2.09	0.01	0.006	0.006
IAPO (adapted)	2.16	2.10	0.01	0.006	0.006
EDGE	2.11	2.06	0.01	0.006	0.007
All-record OR targets	2.35	2.26	0.01	0.006	0.006

comparison; C3 and IAPO similarly request 2.14 and 2.16 pairs. These averages do not isolate the quality of every request.

D.3 Credit Assignment and Sufficient-Path Controls

Local controls share EDGE's format alignment, backbone, role adapters, tools, data, loss denominators, weight cap, and optimizer schedule, changing only the rules in Table 21. Three paired seeds are used throughout; implementation settings are in Appendix C.5.

Table 21: Local ablation configurations. Full EDGE uses the routing and LHD rules of Section 3.2; all unlisted settings are shared.

Configuration	Change from full EDGE
Uniform task credit	Set task weights to one on all actions, bypassing routing and the positive gate; disable LHD.
Uniform credit + LHD	Use the same uniform task weights; retain path-filtered LHD.
BRPO only	Set $\lambda_{\mathrm{LHD}} = 0$; preserve both routing branches.
Negative-only shuffle	Permute full negative-weight vectors within equal-advantage, equal-target-count strata; preserve positive credit, empty sets, and path-filtered LHD targets.
Dependency-ranked targets	For eligible negative trajectories, replace selected actions by reference-aware dependency ranking; preserve target count, capped weight multiset, positive credit, abstention, and the original LHD targets.
All-record OR targets	On OR, replace Φ^{train} by Φ^{diag}, except retain empty targets when $H = 0$ and an alternative is unresolved. AND and positive credit are unchanged.
No positive gate	For $\hat{A} > 0$, set task weights to one, bypassing both H and Ψ.
No zero-variance LHD	Suppress LHD only when $\sigma_q = 0$, including all-correct groups.
Negative unknown fallback	For $\hat{A} < 0$, unresolved attribution or OR support with empty $\Phi^{\mathrm{train}}_{\mathrm{act}}$ receives uniform task credit.
Shuffled task weights	Permute task weights within the fixed strata below; preserve LHD.
Frozen analyzer	Keep the initialization analyzer for all training rounds.
One cross-agent break	On negative cross-agent multi-break trajectories, keep only the earliest action in $\Phi^{\mathrm{train}}_{\mathrm{act}}$.
All cross-agent actions	On those same eligible trajectories, distribute task credit over every trainable action.

All controls use 1,200 updates and paired seeds 13/37/73. All-record OR selection changes record targets before action projection and hindsight, including unused alternatives when $H = 1$, while preserving unresolved-OR abstention. Negative-only shuffle uses an additional RNG stream (offset 7,000), keyed by seed and update and fixed across both optimizer passes.

Negative-credit localization. Negative-only shuffle first computes the full EDGE diagnosis, sufficient-path targets, and capped weights on its own batch. It changes only task weights for trajectories with $\hat{A}^{(g)} < 0$ and $N_g > 0$. Within each stratum of equal advantage and selected-action count, draw one uniform permutation p of the trajectory indices and assign the entire three-role vector:

$$\mathbf{w}_g^{-,\mathrm{shuffle}} = \mathbf{w}_{p(g)}^{-}, \qquad \mathbf{w}_g^{-} = (w_{P,g}, w_{M,g}, w_{R,g}).$$

With binary reward and fixed $G = 8$, the number of correct trajectories in the question group identifies the exact negative-advantage stratum. The permutation is applied after any extra batch is collected; strata pool all eligible trajectories in that update. Singletons and identical vectors may remain unchanged, and permutations are not redrawn to force a change.

Equal N_g preserves each trajectory's target count and capped weight sum. Moving whole vectors preserves each role's weight multiset, nonzero count, and summed negative advantage within the stratum. Positive weights, empty negative sets, and zero advantages remain untouched. The unpermuted Φ^{train} supplies the LHD mask and hindsight; only a separate task-weight array is permuted. Same-batch sampling eligibility, LHD targets, and KL configuration remain unchanged. The two optimizer passes reuse the same permutation. Across training runs, policy evolution can change the generated trajectories and resulting coverage.

Negative-only shuffle reaches 61.6% accuracy, 81.8% CompleteSupport, and 50.4% GroundedAcc (Table 4). Paired accuracy gains for EDGE are 1.7/1.5/1.7 points across seeds 13/37/73; GroundedAcc gains are 2.6/2.8/2.8. Their means and sample SDs from unrounded differences are 1.6 ± 0.1 and 2.8 ± 0.1 points, with 95% program-bootstrap intervals [0.8,2.4] and [1.8,3.8]. The same 10,000-resample protocol conditions on these trained seeds. This isolates negative-credit alignment with positive gating and LHD retained.

The dependency-ranked control computes EDGE's full diagnosis and weights on its own collected batch. Only when $\hat{A}^{(g)} < 0$ and $N_g > 0$ does it replace the selected actions. A separate graph call uses the same analyzer checkpoint, context limit, greedy decoder, and reference inputs as EDGE; it receives neither the analyzer's link decisions nor Φ^{train}. The graph schema above is retained, but error spans may additionally cite a contradiction or omission against the training reference facts. This reference-aware extension is a local control, not native IAPO. Rank the three actions by descending ϕ_i^-, then fixed order P, M, R. If π_g is that ordering and $v_{g,1} \geq \cdots \geq v_{g,N_g} > 0$ are the original nonzero capped weights, set

$$w^{-,\text{dep}}_{\pi_g(j),g} = \begin{cases} v_{g,j}, & j \leq N_g, \\ 0, & j > N_g. \end{cases}$$

This preserves the target count and weight multiset within each trajectory. Ties, including a graph with all-zero features, use the same fixed order; parse failure follows the one-retry empty-graph rule. Empty EDGE target sets remain empty. Positive and zero advantages, original LHD records and masks, sampling eligibility, loss denominators, and optimizer settings remain fixed on the same inputs. The graph and selected weights are detached and reused for both passes. Unlike the batch permutation, ranking can change each role's weight multiset and token-weighted mass, which are recorded separately.

Dependency-ranked targets reach 62.2% accuracy, 51.3% GroundedAcc, and 82.6% CompleteSupport. EDGE retains margins of 1.0, 1.9, and 1.5 points, respectively, under matched eligibility and reference access. Ordered first-break checks distinguish initiating failures from downstream reuse of faulty information; the tested three-node graph ranks observable dependencies. The comparison concerns these target locators, not every dependency method or the oracle-repair optimum.

To measure whether the graph distinguishes candidate targets, a fixed-input audit samples 100 eligible negative EDGE trajectories per seed: 34/33/33 from the 50-update windows ending at 200/600/1,200, uniformly before graph extraction. The 300 trajectories contain 200 singleton, 75 two-action, and 25 three-action target sets. Graph ranking changes 126/300 sets (42.0%). At the selection boundary, 45/300 cases (15.0%) tie, including 30 all-zero error-feature graphs; four of those follow parse failure after one retry. All remain in the denominator and use fixed-ID tie breaking. These checks measure target distinguishability, not correctness against a human or counterfactual reference; they add no human support labels.

Credit identity and coverage. The shuffled control computes full-method weights as a shadow mask on its own batch, then permutes them within fixed advantage-sign, role, and advantage-magnitude strata. Each stratum retains its weight multiset and nonzero count; LHD targets are unchanged. This preserves local credit amounts while breaking their evidence alignment. Different policies generate different trajectories, so global cross-run coverage can differ. For action population $\mathcal{T}$, task-credit coverage is $|\{(i,t) \in \mathcal{T} : A_{i,t} \neq 0\}|/|\mathcal{T}|$, measuring coefficients rather than nonzero PPO gradients. LHD eligibility follows path filtering and precedes token gating and the coefficient schedule. Unknown attribution has a trajectory denominator.

On identical collected batches, negative-only shuffle changes the action set for 22,500/38,000, 23,100/38,500, and 23,400/39,000 eligible trajectories (59.2/60.0/60.0%), retaining singleton and unchanged strata in the denominator. All four preservation checks hold across 3,600 updates (Table 22b). Thus target identity changes while same-input task eligibility, positive weights, and LHD targets are preserved.

Since actions differ in length, equal action weights need not give equal token-weighted mass. For each role, sum $\sum_{g:\hat{A}^{(g)}<0} |\hat{A}^{(g)}| w_{i,g} \ell_{i,g}$ over training, where $\ell_{i,g}$ is its generated-token count. The shuffled-to-shadow ratios, averaged across seeds, are 0.998/1.004/1.001 for $P/M/R$. All role–seed ratios lie within 0.5% of one. These are coefficient-mass checks, not gradient norms or counts of active clipped-PPO terms.

Table 22: Training coverage and permutation audit. (a) Coverage is pooled over training batches within each seed and then averaged, in percentages. Task+ and Task− use all generated actions as denominator and sum to Task. LHD denotes eligibility; unknown is an unresolved analyzer judgment, not an oracle-replay tie. (b) Counts of updates satisfying each constraint across three training runs.

(a) Coverage across each configuration's training runs

Configuration	Task+	Task−	Task	LHD	Unknown
Uniform task credit	34.9	34.3	69.2	0.0	6.7
Uniform credit + LHD	34.7	33.8	68.5	45.8	6.5
BRPO only	26.9	18.3	45.2	0.0	6.8
Full EDGE	26.7	18.1	44.8	46.1	6.2
Negative-only shuffle	26.6	18.2	44.8	46.0	6.3
No positive gate	43.0	18.4	61.4	45.9	6.5
No zero-variance LHD	26.5	18.0	44.5	30.3	6.8
Negative unknown fallback	26.6	23.1	49.7	46.0	6.2
Shuffled task weights	26.8	18.2	45.0	45.9	6.3
Frozen analyzer	26.2	17.9	44.1	42.0	9.1
All-record OR targets	26.3	23.4	49.7	55.0	6.4

(b) Negative-only permutation on identical input batches

Preservation check	Matching updates / tested
Positive weights and LHD targets	3,600/3,600
Empty negative sets and zero advantages	3,600/3,600
Per-trajectory target count and weight sum	3,600/3,600
Per-role weight multiset	3,600/3,600

Full EDGE and shuffled weights have similar task coverage (44.8 versus 45.0) and LHD eligibility (46.1 versus 45.9), but differ by 2.3 accuracy and 7.0 GroundedAcc points. Removing the positive gate instead raises task coverage to 61.4 while lowering GroundedAcc. Unknown and abstained trajectories remain in coverage and test denominators. Unlike negative-only shuffle, the shuffled-task-weight control permutes both advantage branches, so its performance gap measures their joint evidence alignment.

OR target selection. The OR comparison partitions the fixed test questions by their annotated support structure, before inspecting policy outputs. The split in Table 23 uses 360 multi-path OR questions, 676 single-path OR questions, and 404 AND questions. The first two strata partition the 1,036 OR questions by their reference-record IDs. All policies use the same questions. Acc. and GA use every question in a stratum; CS uses that seed's correct answers, including unsupported ones. Means average the three seeds.

Table 23: Support-structure comparison. EDGE aggregates reproduce its overall totals. All scores are percentages.

Stratum	Questions	Policy	Acc.	CS	GA
Multi-path OR	360	EDGE	63.3	87.0	55.1
		All-record OR	58.4	81.6	47.7
Single-path OR	676	EDGE	63.3	84.7	53.6
		All-record OR	63.2	81.2	51.3
AND	404	EDGE	63.0	80.6	50.7
		All-record OR	63.0	80.1	50.4

The gains concentrate on multi-path OR: 4.9 accuracy and 7.4 GroundedAcc points, compared with 0.1/2.3 on single-path OR and 0.0/0.3 on AND. Shared parameters allow the OR training change to affect other strata. Because the ablation changes both negative credit and hindsight, it tests sufficient-path filtering jointly rather than either component in isolation.

On multi-path OR questions, EDGE requests 1.29 pairs and delivers 1.24 addresses per question, versus 2.24 and 2.13 for all-record training. The request reduction is 42.4% using displayed means. Together with higher GroundedAcc, this pattern favors completing sufficient evidence over acquiring unused alternatives; counts alone do not identify necessary requests.

OR target audit. The audit applies both selectors to the same stored EDGE batches, holding trajectories, analyzer outputs, and advantages fixed. It compares record tuples before action projection and the records supplied to hindsight. Multi-path OR accounts for 26.7% of collected trajectories. On the identical batches, all-record shadow routing raises negative task coverage from 18.1% to 21.9% and LHD eligibility from 46.1% to 53.4%, with positive coverage fixed at 26.7%. These contrasts isolate target selection; Table 22a instead compares each trained policy's own trajectories.

For independent references, sample 100 multi-path OR trajectories per seed: 34/33/33 from 50-update windows ending at updates 200/600/1,200, uniformly within each window's eligible population before inspecting support, correctness, or advantage. Two annotators label all records and adjudicate disagreements, blinded to the selector and analyzer predictions. Human predicates define the reference target set by the same fixed-ID selector. A trajectory has an off-target record if either corrective mask or hindsight includes a record outside that set. Without a complete path, an unknown reference alternative requires abstention. The counts in Table 24 describe the sampled population, not sampling quotas.

Table 24: Independent target audit on 300 training trajectories. Entries count trajectories with at least one off-target record under the human reference selector. Both columns use identical EDGE trajectories.

Reference state	n	EDGE	All-record shadow
Complete OR path	120	2	104
Use failure; alternative unrequested	60	1	59
Other resolved OR failures	90	1	83
Unresolved OR, no complete path	30	1	1
Total	300	5	247

Off-target rates are 5/300 (1.7%) for EDGE and 247/300 (82.3%) for all-record shadow routing; among supported trajectories they are 2/120 (1.7%) and 104/120 (86.7%). EDGE retains 146/150 human-selected failed records (97.3%), versus 147/150 (98.0%). Record-level scoring detects irrelevant hindsight even when alternatives share an owner. The remaining EDGE errors reflect disagreement with human predicates, separate from execution consistency (Appendix E.2). These checkpoint-balanced rates describe the audit, not every training step.

All-record training triggers extra sampling at 225 rather than EDGE's 255 updates, yet uses more task and LHD-eligible tokens because requests and hindsight are longer. Appendix D.5 reports the full costs.

D.4 Interface, Hindsight, and Stability Ablations

Tables 25 and 26 extend the main matched ablations with interface, annotation, and dimension breakdowns. The core-ablation IoU/Modality-hit pairs are 0.50/0.64 for BRPO only, 0.51/0.66 for uniform credit + LHD, 0.52/0.67 without the positive gate, and 0.51/0.66 for the frozen analyzer.

BRPO, LHD with uniform task credit, and their combination improve accuracy over uniform task credit by 3.0, 2.5, and 5.6 points, respectively (Table 4). Freezing the analyzer costs 1.4 points; removing the positive gate lowers CompleteSupport by 16.3 points. Removing phase metadata lowers C1 and C2 by 2.6 and 2.3 points, with smaller changes elsewhere. The intended contrast removes explicit phase labels and event boundaries from analyzer and hindsight inputs while retaining annotated windows, modality tests, and the L_1/L_2 overlap threshold. Its interpretation depends on that input configuration; there is no separate phase predicate. Removing sufficiency labels costs 7.1 C5 points and at most 2.5 elsewhere. Without the positive gate, C4 accuracy rises by 0.4 while overall Any-citation falls by 13.8 points. This is consistent with correct but unsupported answers; individual credit decisions require the positive-rollout audit in Appendix F. Excluding only all-incorrect LHD

Table 25: Supplementary interface and mechanism ablations. Scores are percentages except IoU and Mod.-hit in $[0, 1]$. Accuracy in panel (b) is mean $\pm$ sample SD; the core credit comparison is Table 4.

(a) Interfaces, annotations, and hindsight

	Acc. (%)	IoU	Mod.-hit	Any-cit. (%)	C5-part. ↓
Free-form, uniform credit	55.9	—	—	—	—
Free-form, inferred graph	59.3	0.47	0.61	78.9	29.6
Undifferentiated hindsight, uniform credit	58.8	0.48	0.62	75.9	28.8
− role slicing	61.0	0.49	0.64	85.0	22.1
− phase privilege	61.7	0.49	0.68	86.1	17.9
− sufficiency table	61.2	0.53	0.61	85.7	26.9
− all-incorrect groups in LHD	62.0	0.52	0.67	87.1	19.0

(b) Remaining routing and analyzer controls

Configuration	Acc.	Any-cit.	CS	GA	C5-part.
No positive gate	62.3 ± 0.5	74.6	67.8	42.2	18.1
No zero-variance LHD	61.8 ± 0.6	86.5	82.0	50.7	21.7
Negative unknown fallback	61.9 ± 0.6	85.6	80.6	49.9	19.6
Shuffled task weights	60.9 ± 0.5	81.8	75.9	46.2	20.4
Frozen analyzer	61.8 ± 0.6	86.3	79.8	49.3	19.6
One cross-agent break	62.2 ± 0.5	87.1	82.8	51.5	19.0
All cross-agent actions	62.4 ± 0.5	86.8	82.2	51.3	18.5

Table 26: Accuracy by dimension for the component and training ablations. Parenthesized: change from the full method. C1–C5 as in Table 1.

	C1	C2	C3	C4	C5	Avg.
Full method	67.7	62.7	61.1	62.3	59.6	63.2
− LHD	65.7	60.4	57.6	60.2	55.2	60.6
No positive gate	67.0	61.7	59.2	62.7 (+0.4)	57.5	62.3
− role slicing	66.2	60.8	58.3	60.2	56.0	61.0
− phase	65.1 (−2.6)	60.4 (−2.3)	60.5	61.5	59.2	61.7
− sufficiency	65.2	61.2	60.5	61.9	52.5 (−7.1)	61.2
Dependency-ranked targets	66.7	61.9	60.4	60.7	58.1	62.2
− all-incorrect	67.2	61.9	58.9	61.0	57.5	62.0
Frozen analyzer	66.9	61.8	59.8	59.3 (−3.0)	57.9	61.8

groups differs from suppressing both constant-outcome group types, because all-correct groups can contain confirmed evidence failures.

Table 27: Sensitivity to the routing weight cap with $M_g = 3$. Gradient variance is reported relative to $w_{\max} = 1.5$. All caps at least three define one uncapped condition.

$w_{\max}$	End-to-end acc. (%)	Rel. grad. variance
1.5	61.6	1.00
2.0	**63.2**	1.14
No clipping ($w_{\max} \geq 3$)	62.9	1.43

Weight-cap sensitivity. The default cap gains 0.3 accuracy points over no clipping while reducing relative gradient variance by 20.3% (Table 27). Variance is the trace of the task-only adapter-gradient sample covariance over 100 fixed development batches per seed, before global gradient clipping. The final EDGE checkpoint rescores the same batches under each cap; variance ratios are normalized within seed and averaged. This fixed-policy diagnostic isolates weighting, whereas accuracy comes from separately trained policies. The cap acts after diagnosis and cannot alter labels on fixed inputs.

Parameter sharing and role-specific hindsight. Table 28 compares localized and full-trajectory hindsight within the shared 7B, rank-32 adapter configuration and within three independent 3B backbones that fit the same hardware. This tests the benefit of role-specific hindsight under each implementation.

Table 28: Role-specific hindsight against undifferentiated hindsight (the full trajectory given to every agent), within shared 7B and independent 3B configurations. Δ is the accuracy difference within each configuration; backbone size and sharing both change between configurations.

Backbone	Hindsight	End-to-end acc. (%)	Evidence IoU	Δ
Shared 7B + 3 LoRA	undifferentiated	61.0	0.49	—
Shared 7B + 3 LoRA	role-specific	63.2	0.54	+2.2
3× independent 3B	undifferentiated	51.3	0.41	—
3× independent 3B	role-specific	54.7	0.48	+3.4

Localized hindsight improves accuracy by 2.2 and 3.4 points, respectively, with higher IoU in both settings. Backbone size and sharing differ between settings, so the cross-setting contrast cannot isolate parameter sharing.

Table 29: Format validity on a fixed validation panel. Values are percentages; each evaluation set contains 5,760 outputs.

Checkpoint	Schema	Citation format
Before alignment	62.3	48.7
After alignment	97.5	95.1
After RL: EDGE	98.4	96.4
After RL: no positive gate	98.2	94.8
After RL: uniform task credit	97.8	94.5

Table 29 separates interface validity from complete evidence support. The common supervised alignment protocol is specified in Appendix C.5.

D.5 Training and Inference Cost

Table 30 consolidates full-run costs. Negative-only shuffle and EDGE use similar task tokens, LHD eligibility, and GPU time; extra sampling occurs at 253.3 versus 255 updates on average.

C3's prefix reuse reduces policy generation; EDGE does not have a universal compute advantage over checkpoint-based credit. IAPO avoids counterfactual continuations but its 32B graph extractor

Table 30: Full-training cost, averaged over three seeds at 1,200 updates. Task tokens include all policy-generated branches once; LHD counts eligible generated tokens before two-context rescoring. GPU-hours include supervised stages, media/tool computation, graph extraction, hindsight, and optimization on the same A100 hardware; they are summed device-hours, not wall time.

Configuration	Task tokens (M)	LHD tokens (M)	GPU-hours
C3 (adapted)	214.0	0.0	286
IAPO (adapted)	222.0	0.0	318
Negative-only shuffle	238.3	35.0	291
Dependency-ranked targets	237.8	35.0	303
All-record OR targets	252.0	45.0	316
EDGE	237.3	35.0	290

has a separate cost, counted even when parsing fails. EDGE's advantage here is stronger complete evidence support, with the fixed-token comparison below separating that pattern from generation volume. The local dependency control retains EDGE's diagnosis for LHD and eligibility and adds graph extraction, explaining its extra cost.

Matched generation budget. The primary endpoint is 1,200 outer updates, each comprising all role updates from one collected batch. Within-batch optimizer passes and reuse are counted separately, and test outcomes select neither checkpoints nor seeds. Cumulative generation $N_{m,s}(k)$ counts all sampled task-policy tokens, including invalid outputs, unused groups, the optional extra batch, and every distinct C3 alternative and continuation. Cached prefixes are counted only when generated, not once per descendant leaf. Prompt/environment tokens and analyzer/hindsight work are accounted separately. For the development-fixed grid $\mathcal{B}$, the secondary ceiling is $B^* = \max\{b \in \mathcal{B} : b \le \min_{m,s} N_{m,s}(1200)\}$. Each run uses its latest saved checkpoint $k_{m,s} = \max\{k : N_{m,s}(k) \le B^*\}$, without interpolation or exceeding the ceiling. The comparison is unavailable if either set is empty for a run. The secondary comparison covers EDGE, MATPO, no positive gate, shuffled weights, all-record OR targets, negative-only shuffle, C3, IAPO, and dependency-ranked targets. Its grid is $\mathcal{B} = \{25, 50, \ldots, 300\}$ million tokens, with checkpoints every 25 outer updates. The minimum final count is 205 million, so $B^* = 200$ million. Table 31 gives the selected checkpoints and unused allowance; the next checkpoint exceeds the ceiling in every run. Equal ceilings do not imply equal realized tokens or wall time. Final task-token counts for seeds 13/37/73 are 250/252/254 million for all-record OR, 237/240/238 for negative-only shuffle, 212/214/216 for C3, 220/222/224 for IAPO, and 237/238/238.4 for dependency ranking. All exceed the common ceiling. Auxiliary checkpoints receive output-specific support labels.

Table 31: Generation-ceiling checkpoints. Triples follow seeds 13/37/73; task tokens are in millions, and Acc./GA are three-seed means (%). The shortfall is 200 minus the listed count; the next checkpoint uses 200.1–205.0 million tokens.

Configuration	Outer update	Tokens	Acc.	GA
C3 (adapted)	1100/1100/1125	197.2/198.1/199.0	60.6	47.0
IAPO (adapted)	1050/1075/1075	197.1/198.6/199.3	60.1	46.8
Dependency-ranked targets	1000/1000/1000	197.4/199.0/198.2	61.9	50.7
EDGE	1000/1000/1000	197.0/199.2/197.9	62.9	52.8
MATPO	1150/1150/1175	197.4/199.3/198.6	60.2	44.8
No positive gate	1025/1025/1025	195.3/198.2/196.5	62.0	42.0
Shuffled task weights	1000/975/975	199.5/198.3/196.6	60.6	45.8
All-record OR targets	950/950/950	197.5/198.4/199.1	61.6	49.7
Negative-only shuffle	1000/1000/1000	197.5/199.4/198.1	61.3	49.8

At the common generation ceiling, EDGE retains accuracy/GroundedAcc margins of 2.7/8.0 over MATPO, 1.6/3.0 over negative-only shuffle, and 1.3/3.1 over all-record OR targets. Its GroundedAcc margins are 5.8–6.0 over adapted C3 and IAPO and 2.1 over dependency ranking. These comparisons separate the support gains from sampled task-token volume. Analyzer, hindsight, tool, and optimizer costs remain separate from this ceiling and are included in Table 30.

Inference cost. Table 32 counts tokens across all model calls, including visual, audio, and billed reasoning tokens. Latency is unbatched wall time per question. Cost uses the evaluation-time API rates of $1.25/$10 per million input/output tokens for GPT-5 and Gemini, or measured batched 7B throughput on one A100 at an assumed $2/GPU-hour. Shared pose/music preprocessing is cached once per program and excluded.

Table 32: Per-question inference cost on the video-grounded test set. Tokens are summed over all model calls and include visual, audio, and reasoning tokens. Accuracy is from Table 1.

System	Backbone	Calls	Tokens in / out (k)	Latency (s)	$ per 1k q.	Acc. (%)
GPT-5	—	1	29.0 / 4.2	24	78	56.2
Gemini 2.5 Pro	—	1	30.0 / 2.9	19	67	57.1
VideoMultiAgents	Gemini	5	62.0 / 7.4	47	152	58.2
Ours, untrained	Gemini	3	35.0 / 5.1	31	95	59.7
MATPO	7B	2	31.0 / 1.9	9.6	0.90	60.3
EDGE	7B	3	35.0 / 2.3	11.8	**1.01**	**63.2**

EDGE costs $1.01 per thousand questions versus $95 for its Gemini-driven pipeline, about 94 times less, with 3.5 points higher accuracy and 2.6 times lower latency. Both read 35k tokens; the difference is largely model price, not less input. A training run including supervised stages costs about 290 GPU-hours ($580), recovered after roughly 6,200 questions at a $0.094 per-question saving. This excludes annotation, engineering, and GPU idle time; API prices include provider margin. Token and latency comparisons are less sensitive to these accounting assumptions.

D.6 Training Dynamics and Retention

All-incorrect/all-correct groups account for 31/6% early in training and 9/22% at its end. Figure 7 tracks these outcomes alongside attribution agreement; Table 33 reports usable targets and effective hindsight mass.

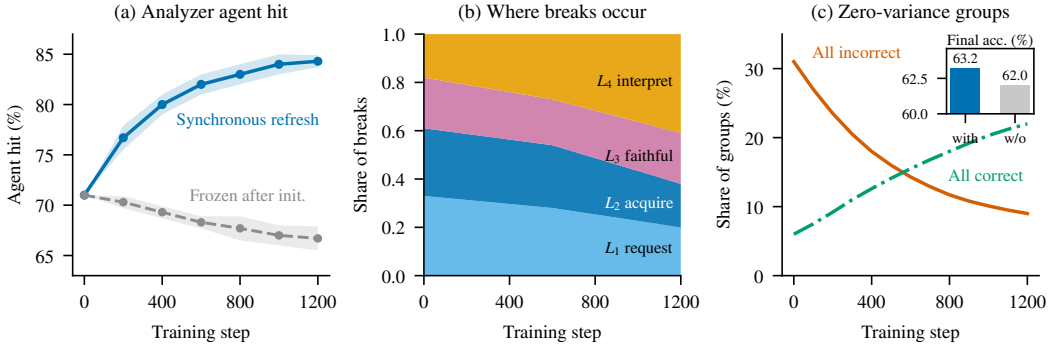

Figure 7: Training dynamics. **(a)** Analyzer agent hit on trajectories freshly sampled from the current policy at each checkpoint and scored against the development-stream $K = 3$ oracle reference, for the synchronously refreshed analyzer and one frozen after initialization; bands show the sample standard deviation of three sets of 100 trajectories per checkpoint. **(b)** Share of breaks at each link, from the analyzer's own attributions. **(c)** Share of groups in which every trajectory fails or every trajectory succeeds; inset: final accuracy when all-incorrect groups do and do not contribute to LHD.

Figure 7a compares refreshed and frozen analyzers on the same fresh development trajectories at each checkpoint, using $K = 3$ replay references. Both start at 71.0%; the refreshed analyzer reaches 84.3% while the frozen one falls to 66.7%. Across checkpoints the sampled policy outputs and low-budget references change, so the decline does not isolate a source of error. Panel (b) shows analyzer-predicted composition: L_1/L_2 falls from 61% to 38%, while L_4 rises from 18% to 41%. Independent final-sample link labels (Appendix E.2) do not validate that training-time shift. The frozen-analyzer ablation's 1.4-point accuracy deficit, largest on C4 (-3.0), supports refreshing

without establishing composition change as its cause. Excluding only all-incorrect LHD groups costs 1.2 accuracy points; suppressing both constant-outcome group types costs 1.4.

Table 33 distinguishes path-filtered targets from loss strength. Effective mass averages $\lambda_{\mathrm{LHD}} m_{i,t} m_{\ell}^{\mathrm{tok}} g_{i,t,\ell}$ over generated agent tokens. The coefficient reaches zero at the final step, so eligible hindsight then has zero effective mass. Constant-outcome groups can retain confirmed targets despite zero task advantage.

Table 33: Diagnostics at fixed training checkpoints. Group shares and LHD target fractions are percentages. LHD targets within constant-outcome groups count groups with at least one confirmed target. Mass is dimensionless.

Diagnostic	Step 200	Step 600	Step 1,200
All-incorrect groups	23.5	14.3	9.0
All-correct groups	9.2	15.5	22.0
LHD target within all-incorrect groups	88.0	91.0	94.0
LHD target within all-correct groups	18.5	12.1	6.8
Task-credit action coverage	34.7	47.3	50.1
LHD eligible-action coverage	51.8	46.3	39.2
Unresolved trajectories	10.8	6.2	3.9
λ_{LHD}	0.5	0.5	0.0
Effective LHD token mass	0.182	0.153	0.000

Text-only rule retention. All 2,400 rule questions are held out from training and have $E^*(q) = \emptyset$, so they supply no record-level first-break targets. The planner remains active under the common inference budget and may issue an empty request. The shared post-format-alignment, pre-RL backbone is the starting checkpoint. Table 34 shows lower rule accuracy for every trained system; EDGE's 0.6-point decline is the smallest reported, with a 0.4-point seed standard deviation. Its empty-request rate is 92.3% on rule questions and 1.1% on video questions. On the 7.7% of rule questions receiving a video request, accuracy is 66.1%, versus 71.1% for empty requests. The groups contain different questions, so this contrast is associative.

Table 34: Accuracy on the three rule-grounding tasks before and after training, and the rate at which the planner issues an empty request on rule questions and (false empty) on video-grounded questions. Trained systems: mean over three seeds.

	Recall 950	Distinction 760	Judgment 690	All 2,400	Empty (%)	False empty (%)
Before training	76.2	70.8	65.4	71.4	94.1	0.8
Single agent	75.5	69.6	63.8	70.3	89.6	1.7
MAGRPO	75.7	69.9	64.0	70.5	88.1	2.4
MATPO	75.8	69.9	64.2	70.6	89.0	1.9
EDGE	75.9	70.1	64.4	$70.8_{\pm 0.4}$	92.3	1.1

E Attribution Validation

This appendix evaluates agent repair value, independent link localization, and conditions underlying Proposition 1. Training effects are isolated in Section 4.4; correct-answer support has a separate reference population in Appendix F.

E.1 Samples, Repairs, and Statistical Resolution

Partitions. The 400 diagnostic trajectories are stratified by dimension and roughly 7 : 3 by evidence structure. Of these, 150 initialize the analyzer using independent record/turn/link annotations; 250 question-disjoint trajectories calibrate η and supply the diagnostics below. Calibration uses $K = 5$ replays per agent (15 additional executions), with 84.4% first-break agreement. A separate 100 development trajectories fix oracle prompts, tolerances, and budgets. Independent

confirmation samples 600 held-out failures uniformly from a frozen policy, one trajectory per question, before enrichment by link, structure, or repair success. Their questions and programs exclude initialization, calibration, development, and training monitoring. Policy and analyzer checkpoints, $\eta = 0.5$, oracle versions, baseline prompts, and scoring rules are frozen before inspecting confirmation labels or outcomes; confirmation never selects checkpoints or supplies further tuning. Separate auxiliary samples contain 120 upstream cases, 120 targeted cross-agent multi-break trajectories, and 200 AND partial-evidence questions spanning all dimensions, including conjunctive C3/C5 and high-prior C4. OR alternatives are audited in the content-intervention subset of Appendix F. They exclude confirmation programs and do not determine its protocol.

Permitted repairs. The operators in Appendix C.1 differ on already valid outputs: $O_{\min}$ copies them exactly, whereas O_{full} generates canonical replacements. Planner repairs use annotated addresses. Perception describes only evidence accessible through the request reached in its branch. Reasoner repairs apply an expert-authored rule procedure to delivered facts, without missing media or the reference answer. Its released rule program, required fields, and answer template specify each question type. Generation refuses when every sufficient path lacks a required fact, yielding $R = 0$; an intact OR alternative prevents refusal. Under $O_{\min}$, copying a valid output takes precedence over generation. The ordinary Reasoner may instead answer from priors, so its p_0 differs from strict-oracle success on missing-field inputs. Experts author repairs before seeing method predictions, using source content and addresses without answer-option labels, and strip final-answer hints. Joint repairs follow the actual $P \to M \to R$ inputs.

Replay and stopping. Every branch starts from a common prefix before the earliest candidate intervention with the same checkpoint, tools, budgets, and sampling law. The baseline receives no repair; logged failures may recover on replay. Operators act on their own branch's reached context. Repetition seeds are independent and recorded; equal seed identifiers across altered inputs do not ensure identical continuations. A local restart with a matched sham would estimate a different contrast; all repair results here use the common prefix defined above. Decoder settings, generation limits, and oracle/copy behavior are specified in Appendix C.5.

The full-replacement confirmation reference uses cumulative $K \in \{16, 32, 64, 128, 256\}$ per repair, advancing all three unresolved branches together, and 64 baseline repetitions. At each look, each probability has a two-sided exact binomial interval with miscoverage $0.05/(3 \times 5)$, giving 95% simultaneous coverage across agents and looks per trajectory. Stopping uses reference outcomes alone. The common baseline cancels in agent rankings. For intervals $[\ell_i, u_i]$, define $C = \{i : u_i \geq \max_j \ell_j\}$. A unique winner requires $\ell_i > \max_{j \neq i} u_j$. Practical equivalence requires every pairwise difference interval $[\ell_i - u_j, u_i - \ell_j]$ within C to lie in the development-fixed $[-0.05, 0.05]$ and separation from excluded agents. Other cases continue, retaining C and an unresolved flag at the cap. A broad set alone does not establish equivalence. Point rankings use final success fractions, breaking exact ties in the fixed order P, M, R.

Auxiliary repair comparisons use $K = 64$ per condition for aggregate effects. At probability 0.5, independent differences have standard error $\sqrt{0.5/64} = 0.088$ and approximate 95% half-width 17.3 points; a 4-point half-width would require about 1,201 repetitions before multiplicity correction. These studies cannot certify close individual winners or uniform 4-point bounds.

Metrics and audit. A method's single-agent nominee owns its chronologically earliest predicted break, with simultaneous ties ordered P, M, R. The full attribution set is retained for routing, link exact match, and joint repair. Abstention is a miss. Point and candidate-set hit use all 600 trajectories; unique hit also reports identifiable coverage. Set hit tests nominee membership in C, with set size and unresolved mass; random set hit is $\mathbb{E}|C|/3$. Aggregate 95% intervals use 10,000 paired source-program bootstrap resamples, conditional on reference decisions. Training-seed variability, simultaneous replay uncertainty, and replication are evaluated separately.

A prespecified random 100-trajectory audit receives new independent $K = 256$ batches and single-look 95% simultaneous exact intervals across the three agents, plus a separate $K = 5$ comparison. On all 100, repair-value shortfall compares the frozen primary replay nominee with the method nominee on the new batch; it is loss to the optimum only when that reference choice is optimal. Intervals additionally resample repetitions within each sampled program, using the same resampled branch outcomes for every method. All branch probability intervals are retained.

Logs retain program/question/trajectory IDs, partition, checkpoints, oracle version, prefix, repair set, seeds, inputs, addresses, predicates, evidence arrival, correctness, calls, and runtime. Branches distinguish ordinary Reasoner responses, oracle generation, minimal copies, refusal, and the (A2) bound being checked. Structural eligibility is fixed independently of replay rankings, with eligible, ineligible, and undetermined cases retained.

E.2 Independent Agent and Link Confirmation

EDGE and the GPT-5 judge receive identical completed trajectories and evidence annotations, with their configurations fixed on development data. The last-agent heuristic predicts R/L_4; random-agent expectations are analytic. Oracle L_3/L_4 predicates provide an annotation-assisted ceiling. No method receives confirmation replay outcomes. Table 3 gives the primary agent and whole-trajectory link comparisons.

Table 35: Reference resolution on the 600-trajectory natural failure sample. Practical equivalence requires the prespecified 5-point interval criterion. Unresolved cases remain in all-sample scores.

Reference status	n	Share (%)	Candidate-set size
Unique winner identified	480	80.0	1
Two practically equivalent candidates	72	12.0	2
Unresolved at the budget cap	48	8.0	3
All	600	100.0	1.28 (mean)

EDGE's point, unique-winner, and set-hit counts are respectively 516/600, 432/480, and 546/600; the judge's are 375/600, 312/480, and 414/600. Their 95% intervals are [82.7,88.9] versus [58.0,66.8], [86.7,92.6] versus [60.4,69.6], and [87.8,93.7] versus [64.7,73.2]%. Paired advantages are 23.5 points [19.0,28.0] overall and 25.0 [20.0,30.0] on unique winners. Among 72 equivalent-pair cases, set membership is correct on 66 for EDGE and 54 for the judge. The 48 unresolved three-agent sets contribute uninformative hits to both methods and remain in the denominator.

The independently screened, identifiable subset contains 400/600 trajectories (66.7%), all within the 480 unique winners. EDGE matches 368/400 (92.0%, [88.5,94.6]) and oracle predicates 388/400 (97.0%, [94.7,98.6]). On the other 80 unique cases they match 64 and 72. Across all 600, oracle predicates match 569 point rankings (94.8%) and 576 candidate sets (96.0%). Eligibility audits do not certify uniform theoretical bounds.

Table 36: Reference replication and independent repair-value audit. Replication uses the 80 primary unique-winner cases within the prespecified random 100-trajectory audit. Repair-value shortfall uses all 100 and is measured in percentage points against the frozen primary replay nominee.

Audit measure	Result
New $K = 256$ batch identifies the same unique winner	77/80 (96.3%)
Independent $K = 5$ point ranking selects that winner	65/80 (81.3%)
EDGE repair-value shortfall [95% interval]	2.1 [1.3, 3.0] pp
LLM judge repair-value shortfall [95% interval]	8.6 [6.9, 10.4] pp

The three unreplicated unique winners remain unresolved on the independent batch. EDGE's smaller repair-value shortfall connects reference agreement to success on a randomly selected audit, without conditioning on EDGE's correct diagnoses.

Confirmation uses 188,160 executions, mean 313.6, nearest-rank 90th percentile 448, and maximum 832 per trajectory. A uniform 256-per-agent budget with the same baseline costs 499,200, so adaptation saves 62.3%. The independent audit adds 76,800 executions and its five-replay comparison 1,500. EDGE adds no trajectory execution but incurs log-analysis computation. The lower-right panel of Figure 1 uses this independent confirmation protocol and its mean reference cost; the $K = 5$ calibration diagnostics remain separate.

Independent link annotation. Two annotators independently inspect source video/audio, annotated addresses and phases, requests, tool returns, delivered descriptions, and the factual answer,

Table 37: Adaptive reference allocation. All three repaired agents use the listed budget; each trajectory also receives 64 baseline executions. Audit repetitions are accounted for separately.

Repetitions per repair	Trajectories	Share (%)	Executions per trajectory
16	120	20.0	112
32	120	20.0	160
64	150	25.0	256
128	150	25.0	448
256	60	10.0	832
Mean	600	100.0	313.6

blinded to method predictions and replay rankings. They independently resolve time units, clip offsets, modalities, and record/action mappings from the raw logs. An independent checker applies the same frozen interval and membership rules to these adjudicated fields. The end-to-end comparison therefore includes input extraction and alignment; the shared-input control below isolates predicate execution. Neither comparison selects the threshold. For L_3, annotators record source observations and any false or omitted decisive fact. For L_4, they assess the answer's use of delivered records and applicable rules, without filling factual gaps from the reference. Gold answers may inform adjudication but never enter repair branches.

Each record receives its first confirmed failed handoff after confirmed passes, with its turn and owning action. An unknown predecessor leaves downstream attribution unresolved. Inaccessible downstream links are not applicable; multiple record breaks, no-break states, and ambiguity are retained. A third expert adjudicates without method predictions, preserving unresolved cases. Labels retain pre-adjudication agreement, class counts, coverage, and complete trajectory sets. Independent confirmation labels never update training. Of 600 trajectories, 588 have complete adjudicated sets with 720 broken records; 12 remain unresolved (98.0% coverage). Pre-adjudication agreement on these records is 90.8% ($\kappa = 0.88$). The natural sample contains 420 OR and 180 AND trajectories, of which 410 and 178 have resolved sets. Multiple record breaks occur in 40/410 OR cases (9.8%) and 92/178 AND cases (51.7%); all 132 have two broken records, giving 720 records in 588 trajectories. Distinct owners occur in 16 OR and 64 AND cases. These counts describe record-level failures, including redundant OR paths; multiple breaks alone do not establish that a joint repair is necessary.

Table 38: Link confusion on the same 720 adjudicated broken records. Rows are reference labels and columns are EDGE predictions; entries are counts. (a) End-to-end diagnosis uses runtime-extracted fields. (b) The input control substitutes adjudicated structural fields, with the semantic model frozen. ∅ denotes abstention or a missing predicted break. Full-set exact match additionally penalizes extra predicted breaks.

(a) End-to-end diagnosis

Reference	n	L_1	L_2	L_3	L_4	∅
L_1 requested	180	**171**	5	1	1	2
L_2 acquired	180	6	**163**	7	2	2
L_3 faithful/complete	210	2	20	**169**	16	3
L_4 used	150	1	2	19	**126**	2

(b) Adjudicated structural inputs

Reference	n	L_1	L_2	L_3	L_4	∅
L_1 requested	180	**180**	0	0	0	0
L_2 acquired	180	0	**180**	0	0	0
L_3 faithful/complete	210	0	0	**191**	16	3
L_4 used	150	0	0	19	**129**	2

End-to-end EDGE correctly labels 629/720 records (87.4%), with macro-F1 87.9% counting missing predictions as false negatives. Recalls for L_1–L_4 are 95.0, 90.6, 80.5, and 84.0%. These rates include extraction and record alignment. The 20 $L_3 \rightarrow L_2$ errors preserve the agent but change

its hindsight and can change the selected OR path; the 35 L_3/L_4 confusions change both link and owner. Exact record–turn–link set match also penalizes extra entries: EDGE matches 462/588, the judge 302/588, and oracle predicates 542/588 (78.6, 51.4, and 92.2%). Counting unresolved trajectories as mismatches gives EDGE 462/600 (77.0%). These failure-conditioned labels are separate from the positive-delivery references in Appendix F.

For all 720 records, replay the fixed checks from the runtime's serialized fields, retaining unknown and not-applicable states. The independent checker reconstructs candidates from those same fields and reproduces all 720 state vectors. This establishes execution agreement conditional on the inputs. Table 38b instead supplies both implementations with the adjudicated structural fields. The reference addresses, model, prompts, and decoding seeds remain fixed. Candidate chains and semantic judgments are recomputed without reference semantic labels. Agreement with the reference is 720/720 for L_1 and 540/540 for L_2 after L_1 passes. The 180 records failing L_1 have no downstream L_2 decision.

The input control resolves 51 of the 91 end-to-end record errors: 47 cross-link errors involving L_1/L_2 and four missing early-link predictions. Accuracy becomes 680/720 (94.4%) and macro-F1 94.6%. The remaining 40 errors comprise 35 L_3/L_4 confusions and five semantic abstentions or missed breaks. Table 39 groups the 51 resolved errors by the earliest differing input operation. Each entry is paired with the raw source span, extracted and adjudicated fields, candidate IDs, and the two check outputs. The input control changes these records from incorrect to correct and introduces no new record errors.

Table 39: Sources of the 51 errors resolved by structural-input substitution. Columns give the reference first break; each record has one primary cause.

Input operation	L_1	L_2	L_3	L_4	Total
Time unit or clip-offset extraction	4	5	8	1	18
Modality-field extraction	2	4	5	1	12
Request/return/record alignment	1	6	9	1	17
Unreadable runtime field	2	2	0	0	4
Total	9	17	22	3	51

Shared-field agreement isolates execution; adjudicated-input substitution isolates structural extraction and alignment. Headline scores and training retain runtime inputs. The oracle-L_3/L_4 ceiling also retains them, explaining why its trajectory-level full-set score can remain below complete agreement.

On the same 300 OR training trajectories as Table 24, re-executing stored fields and semantic judgments reproduces all diagnoses, support and delivery sets, selected targets, weights, LHD masks, and hindsight records (300/300). Hindsight text is reused rather than regenerated. The five disagreements with human target references remain: correct execution does not remove extraction or semantic errors. Substituting adjudicated fields in training would be a separate intervention.

E.3 Repair Assumptions and Joint Effects

Valid upstream replacement. The 120 upstream cases satisfy L_1 and first fail at L_2 or L_3, including broad requests with distracting frames. Conditions are baseline, canonical Planner replacement, valid windows approximately twice and four times the reference width within video bounds, and minimal/full first-break repair, at fixed frame count and inference budget. Exact message comparison checks that minimal repair preserves valid upstream outputs.

Full first-break repair gains 22.8 points over canonical upstream replacement (Table 40). The upstream 1.2-point benefit shows that valid-window changes can still alter perception. This average is distinct from the conditional upstream bound u; minimal repair instead preserves valid messages by definition.

All 7,680 minimal valid-Planner replacements copy their messages exactly. Full first-break repair delivers a faithful sufficient input path in 6,144/7,680 replays (80.0%); 4,723 of those succeed (76.9%), with 0/1,536 successes in the residual insufficient stratum. Minimal first-break repair gives

Table 40: Upstream and fixed-input comparisons, each with 64 draws per question. Upstream gains use the 120-case no-repair baseline; AND gains use ordinary Reasoner empty input. All values are percentages or percentage points.

Condition	Accuracy	Gain
No repair	37.5	0.0
Minimal repair of valid Planner	37.5	0.0
Canonical repair of valid Planner	38.7	1.2
Valid Planner window: $2\times$ width	37.9	0.4
Valid Planner window: $4\times$ width	37.0	-0.5
Minimal repair of first-break owner	60.8	23.3
Full repair of first-break owner	61.5	24.0
Fixed-input AND audit (200 questions)		
Empty evidence	37.3	0.0
Unrelated evidence	38.1	0.8
Correct but insufficient subset	39.3	2.0
Complete sufficient evidence	68.0	30.7

6,048/7,680 sufficient arrivals (78.8%), with 4,669/6,048 successes (77.2%) and 0/1,632 residual successes. These input-arrival counts do not additionally certify the theorem's answer-use condition. For a branch whose residual success is zero, the safe lower-bound choice is $\delta_j^- = p_0$, not the small-deviation special case. Neither aggregate stratum identifies a uniform conditional success bound.

The diagnostic summaries use a pooled empty-evidence rate of 0.373, full-evidence success of 0.68, and a lowest downstream pass rate of 0.74. Empty and unrelated-input controls have dimension-level deviations at most 0.04. These aggregates do not establish the context-wise bounds of (A2)–(A4), which also cover true partial evidence, branch-specific Reasoners, and every reachable sufficient path. The following audit separates those conditions.

Partial evidence and refusal. On the 200-question AND audit, fixed-input conditions include empty and unrelated evidence, each true record, proper subsets, leave-one-out sets, and the complete set. Table 40 uses one nonempty proper subset per question, selected by a fixed record-ID hash before replay. The panel has 100 two-record and 100 three-record questions; all subsets are tested, reusing the same draws when singleton and leave-one-out inputs coincide. Subset summaries first average within each question, then across questions. The selected-subset row reuses its corresponding subset draws. A confirmed OR singleton is already sufficient and never enters this insufficient-subset condition. Model, options, format, and output budget are fixed. Ordinary Reasoner responses, strict-oracle generation, and minimal copies have separate logs and scores. The ordinary Reasoner's $d_q(E) = p_q(E) - p_{0,q}$ retains both positive and negative deviations. Oracle refusal and success are measured against the same ordinary $p_{0,q}$, not a recentered zero-success baseline.

On each empty, unrelated, and selected partial-input condition, strict-oracle generation refuses in 12,800/12,800 calls with zero successes. Ordinary Reasoner single-record and leave-one-out accuracy is 38.8% and 40.1%, respectively, using question-weighted subset means. The selected proper subset yields 39.3%, a 2.0-point gain over empty input, compared with 30.7 points for the complete set.

Branch audits separately check the baseline interval, competitors' upper bounds, the repaired owner's residual lower bound, restoration, and success on reachable sufficient paths. An empty residual stratum needs no bound. These aggregate fixed-input measurements do not certify uniform bounds over replay contexts or substitute ordinary-policy rates for strict-oracle behavior.

Conditional repair margin. For $\delta_j^- = \delta$, $\pi_j \geq 0.70$, $p_{\text{suff}} - p_0 \geq 0.25$, $\delta \leq 0.04$, and $u \leq 0.02$, Eq. (15) gives $\Delta_j \geq 0.70(0.25) - 1.30(0.04) = 0.123$, exceeding the competitor bound 0.08 by 4.3 points. Competing refusal branches need only an upper bound; the repaired owner's residual contexts must satisfy their lower bound.

Table 41: Sensitivity in the special case $\delta_j^- = \delta$, with $\pi = 0.74$, $m = 0.68 - 0.373 = 0.307$, and $u \leq 2\delta$. $L = \pi m - (2-\pi)\delta$. This is an algebraic sensitivity calculation, conditional on the repaired owner's residual lower bound.

Assumed δ	Lower bound L	Competitor bound 2δ	Margin
0.04	0.1768	0.0800	$+0.0968$
0.06	0.1516	0.1200	$+0.0316$
0.08	0.1264	0.1600	-0.0336
0.10	0.1012	0.2000	-0.0988

For Table 41's inputs, $\delta < \pi m/(4-\pi) \approx 0.0697$ suffices; larger deviations invalidate this certificate without deciding the empirical ranking. If the repaired owner instead refuses on every residual insufficient context, set $\delta_j^- = p_0$. With $\pi_j = 0.74$, $p_{\text{suff}} = 0.68$, $p_0 = 0.373$, and $\delta = 0.04$, the lower bound becomes $0.74(0.68) - 0.373 - 0.04 = 0.0902$, exceeding the competitor bound 0.08 by 0.0102 when $u \leq 0.08$. Both calculations give conditional bounds under their respective branch assumptions. The pooled 0.373 does not establish a question-specific p_0.

Joint repair. The 120 targeted multi-break trajectories evaluate all eight subsets of $\{P, M, R\}$ under both operators, retaining success, sufficient-evidence arrival, and residual failures. AND and OR cases are separated; the 20 cross-agent calibration cases are not assumed all conjunctive. Reasoner replacements refuse according to evidence reached in their branch. Upstream re-execution can restore a path, so common-prefix Reasoner-repair success need not be zero.

Table 42: Joint repair on 120 targeted Planner–Perception cases (80 AND, 40 OR), with 64 draws per condition. Full and minimal are the two repair operators. Arrival is faithful sufficient input under full repair. Gains use the common baseline; all entries are percentages or points.

Set	Full	Gain	AND	OR	Minimal	Arrival
$\emptyset$	36.0	0.0	35.0	38.0	36.0	44.0
P	38.0	2.0	36.0	42.0	37.8	46.0
M	39.0	3.0	37.5	42.0	38.8	47.0
R	36.0	0.0	35.0	38.0	36.0	44.0
PM	70.0	34.0	72.0	66.0	69.4	84.0
PR	39.0	3.0	37.0	43.0	38.8	47.0
MR	40.0	4.0	38.5	43.0	39.6	48.0
PMR	71.0	35.0	73.0	67.0	70.5	85.0

For $B = \{P, M\}$, joint repair raises success from 36.0% to 70.0%, while isolated gains are 0.0–3.0 points. Removing P or M loses 31.0 or 32.0 points; adding R gains 1.0. The interaction $V(PM) - V(P) - V(M) + V(\emptyset) = 0.29$ identifies a joint effect beyond summed isolated gains. Practical equivalence still requires its prespecified interval. Effective candidate sets use development-fixed point criteria: at least 20 points of success gain and evidence-restoration gain, at least 10 points lost by deleting each member, and no more than 5 points gained by adding the remaining agent. These operational criteria do not certify statistical equivalence. All 120 targeted references select PM; this enriched sample does not describe the natural failure distribution. EDGE exactly matches 102/120 sets (85.0%), with six supersets and 12 singleton omissions. Its micro precision/recall is 228/234 and 228/240 (97.4/95.0%), and mean per-trajectory Jaccard is 93.3%. Oracle predicates match 114/120 (95.0%), with three supersets and three omissions, giving 237/240 precision and recall (98.8%) and mean Jaccard 97.9%. Reference membership is resolved for all 120; other samples retain unknowns and allow alternative candidate sets. Full PM repair leaves 2,304/7,680 failures: 1,229 without sufficient input and 1,075 despite sufficient input. Minimal PM reaches 69.4%, below full repair's 70.0% but above every singleton. The interaction holds for both structures; its pooled effect is not a universal single-agent optimum. Joint repair value does not determine training weights; Section 4.4 isolates routing while fixing the other training rules.

E.4 Calibration and Threshold Sensitivity

Table 43 calibrates the overlap threshold on 250 trajectories excluded from analyzer initialization. Lower thresholds move breaks from request coverage (L_1) toward acquisition (L_2); higher thresholds reverse that allocation.

Table 43: Calibration sensitivity to the overlap threshold η on 250 trajectories. Attribution accuracy is the agent-level hit rate against oracle replay; the last two columns give the share of breaks assigned to L_1 and L_2, the remainder falling on L_3–L_4.

η	Attribution acc. (%)	L_1 share	L_2 share	L_3–L_4 share
0.3	78.4	0.19	0.34	0.47
0.4	82.8	0.24	0.30	0.46
0.5	**84.4**	0.28	0.26	0.46
0.6	82.8	0.35	0.20	0.45
0.7	79.6	0.44	0.13	0.43

Agreement with replay peaks at $\eta = 0.5$, which is fixed before independent confirmation. The L_3/L_4 share changes little across the sweep because these predicates do not use the threshold. Confirmation selects neither η, oracle prompts, nor analysis thresholds.

The same 250 calibration trajectories comprise 174 OR and 76 AND cases. Table 44 reports $K = 5$ agreement and rank correlation; replay's 100% self-agreement is a normalization convention.

Table 44: Calibration agreement with $K = 5$ per-agent replay. Overall agent hit and rank correlation are diagnostic summaries on the 250 trajectories used for the threshold sweep. Cost counts additional executions.

Method	$\vee$ hit	$\wedge$ hit	Overall	ρ	Cost
Random	33.3	33.3	33.3	0.00	0
Last agent	31.0	27.6	30.0	0.09	0
LLM judge (GPT-5)	64.4	51.3	60.4	0.46	0
First break	89.1	73.7	84.4	0.78	0
First break, oracle L_3/L_4	95.4	81.6	91.2	0.89	0
Oracle replay (reference)	100	100	100	1.00	15

Of these trajectories, 49 contain several breaks: 29 have one shared owner, with agent agreement 86.2%, and 20 span agents. In the latter, implicated agents' mean replay gain is 0.06 and mean difference 0.04, too small to establish a unique winner with five replays. This motivates set-valued repair without establishing optimal uniform training weights. When rule-decided links hold, calibration agreement is 81.4% on L_3 and 78.9% on L_4. The calibration abstention category covers 4.8% overall and 9.1% in C4; it includes correct uncited answers alongside unresolved judgments. This aggregate diagnostic category does not define training eligibility. Training logs distinguish confirmed insufficient support from unknown judgment. Both block positive credit, but path-filtered confirmed breaks remain eligible for LHD; neither invokes uniform fallback. C4's $p_0 = 43.7\%$ and agent agreement 79.3% also expose an L_4 boundary: correctly reading evidence but misapplying a rule is not a separate fifth link.

F Evidence Support and Dependence

Citation overlap, complete factual support, and behavioral evidence dependence have distinct tests. Metric definitions are in Appendix B.7.

F.1 Full-Test Human References

The reference protocol assigns two independent human ratings to every output on all 1,440 test questions, including incorrect answers, for each method and evaluation seed receiving strict scores. A third expert adjudicates disagreements; unresolved labels remain unknown. Frozen windows,

decisive facts, and pure AND/OR requirements are shared across methods; delivery and answer-use labels are specific to each question–method–seed output and cannot be copied between outputs. These test questions and programs exclude training, analyzer initialization, calibration, and protocol development. Labels never select checkpoints or update policies.

Annotators inspect questions/options, source media, frozen requirements, rules, requests, tool returns, delivered descriptions, citations, and answers. Method/seed identities, analyzer predictions, replay outcomes, and other annotators' decisions are hidden, although output style may reveal identity. Neither the training analyzer nor an automatic judge supplies these labels. Annotators resolve structural fields against the source logs independently of the analyzer's extraction. Code evaluates L_1/L_2 on those fields with frozen η and containment; phase enters through annotated windows. L_3 requires source-faithful descriptions conveying every assigned decisive fact. L_4 requires citation and correct use of delivered facts, jointly for AND questions. Gold answers and facts visible only to annotators cannot fill missing descriptions.

The record ledger retains pass, fail, unknown, and not-applicable links, first identifiable breaks, source observations, and disagreement reasons. An upstream confirmed break blocks a record despite unevaluable later links; ambiguity alone is not failure. Adjudication produces D_q^* and three-valued H_q^*, retaining every confirmed sufficient alternative. The projection Ψ_q^* retains witness record IDs and owning actions; incomplete action references stay uncertain. Labels retain versions, pre-adjudication agreement, and resolved coverage over all outputs and correct answers.

All methods share address matching and sufficient-set rules. Mod.-hit measures any-modality hit, separately from complete modality coverage. Inference budgets and request counts remain method-specific. Scores are stratified by dimension and AND/OR structure, with training-seed variation, unresolved-label bounds, and program-clustered uncertainty reported separately. For s_m evaluation sets per method, annotation covers $1{,}440 \sum_m s_m$ outputs, each rated twice with adjudication as needed.

The final-checkpoint panel contains 64 evaluation sets, 92,160 outputs, and 184,320 independent ratings before adjudication. It comprises:

- The core panel: one untrained set, six trained methods at three seeds, and ten local variants at three seeds (49 sets, 70,560 outputs, 141,120 ratings).
- All-record OR and negative-only shuffle: three sets, 4,320 outputs, and 8,640 ratings each. These bring the subtotal to 55 sets, 79,200 outputs, and 158,400 ratings.
- C3, IAPO, and dependency-ranked targets: nine sets, 12,960 outputs, and 25,920 ratings.

Before adjudication, 4,234 core support decisions disagree (6.0%; $\kappa = 0.88$). All 70,560 core references, including correct answers, are resolved with no missing answer exports, so their support bounds coincide. All 4,320 negative-only-shuffle references are also resolved. Other configurations retain their own agreement, adjudication, and coverage records.

The 300-question delivery diagnostic reuses these full-test labels. The 300-trajectory OR target audit needs 600 initial ratings; log-based permutation and graph diagnostics need none. Auxiliary checkpoints and interface studies have separate exports. In particular, the three-seed budget checkpoints of C3, IAPO, and dependency-ranked targets add 12,960 outputs and 25,920 ratings outside the final-checkpoint totals.

F.2 Positive Delivery and Credit

The diagnostic samples 300 questions uniformly before inspecting outputs, sharing frozen IDs across EDGE, MATPO, MAGRPO, and the no-gate ablation for all seeds. It reuses full-test human labels. Per-seed correct answers define positive-delivery strata; incorrect answers remain in joint-outcome scores. Rare-boundary enrichment is reported separately.

Human references D^*, H^*, Ψ^* evaluate record delivery, support gating, and witnessed actions. Gate precision/recall retain resolved confusion counts, unknown references split by admission, and assignment bounds. Unknown predictions are not admitted. Record scoring precedes action projection, since a correctly identified owner can supply the wrong witness. Unresolved coverage accompanies record/action metrics. Evaluation outputs measure gate eligibility; a separate $\hat{A} > 0$ training

sample measures actual positive credit. Neither low admission alone nor incomplete references certify accuracy.

The diagnostic uses 190 correct EDGE outputs from one prespecified seed on the 300 questions, with 160 supported and 30 insufficient references. The actual checkpoint/seed accompanies the export. All three analyzers inspect these same outputs, isolating diagnosis from the policy's evidence distribution.

Table 45: Support-gate audit. TP/FP/FN/TN compare predicted H with independent H^* on 190 correct outputs from one EDGE seed. Precision and recall are percentages. All rows share 160 supported and 30 insufficient reference cases.

Analyzer	TP	FP	FN	TN	Precision	Recall
Refreshed EDGE	156	2	4	28	98.7	97.5
Frozen analyzer	142	8	18	22	94.7	88.8
Same-information judge	139	13	21	17	91.4	86.9

The refreshed analyzer admits 2/30 insufficient cases (6.7%); its precision error is 2/158 (1.3%), a different denominator.

Table 46: Admission on the common 300-question panel across three seeds. Counts sum the seeds; precision/recall average their separate rates. References are fully resolved. EDGE, MATPO, and MAGRPO use the same frozen EDGE analyzer on each policy's outputs; no positive gate admits every correct output.

Policy	Correct	Supported	TP	FP	Prec.	Recall
EDGE	571	480	467	6	98.7	97.3
MATPO	543	406	367	25	93.6	90.4
MAGRPO	517	327	278	37	88.3	85.0
No positive gate	561	380	380	181	67.7	100.0

These policy-output comparisons combine evidence quality with diagnostic difficulty. Table 45 instead holds EDGE's outputs fixed to isolate analyzer quality. Correctness does not guarantee positive eligibility: bypassing the gate admits all 181 insufficient correct answers in the no-gate panel. A separate enriched set of 40 correct uncited outputs and 40 correct partial-AND outputs tests rare boundaries, with 40/40 and 38/40 correctly rejected. These enrichment rates are not pooled with the uniformly sampled panel.

On the three EDGE seeds, exact action-set agreement is 172/190, 170/189, and 175/192, averaging 90.5%. Record intersections are 350/346/352, predicted-record totals 366/364/369, and reference totals 370/368/374, giving mean delivery precision/recall 95.4/94.2% (95.6/94.6% for the first seed). With one action per role, a nonempty delivered chain projects to all three actions; otherwise its projection is empty. Action intersections are therefore 507/501/513, predicted totals 528/522/531, and reference totals 540/537/546, giving 96.2/93.7% precision/recall. All record and action references in this panel are resolved. Delivery is scored before projection because agreement on all three owners can conceal an incorrect witness record.

Positive-advantage training audit. A separate batch of 300 training trajectories contains 250 supported, 42 insufficient, and 8 unresolved references. Full EDGE admits 244, 2, and 0 respectively: 246/300 coverage (82.0%), 244/250 supported recall (97.6%), and 2/42 insufficient admission (4.8%). A no-gate shadow rule admits all 300, isolating credit selection on fixed outputs. The trained no-gate policy is compared in Table 25. The two false admissions have $H = 1, H^* = 0$, distinct from implementation violations: full EDGE assigns no positive task credit when its own $H = 0$. Correct answers in all-correct groups have zero task advantage and contribute to LHD only when a confirmed break survives sufficient-path filtering.

F.3 Controlled Evidence Interventions

Figure 3 summarizes paired contrasts and OR redundancy; Table 47 retains all five primary conditions on 120 common questions whose pre-answer contexts contain a faithful sufficient path for every method, checked without replayed answers or final citation use. Each method retains its own context and frozen Reasoner from the prespecified seed 13. Probabilities therefore condition on evidence availability rather than estimating full-test accuracy. At the fixed pre-answer point, all five conditions use 16 samples with common decoding settings and recorded seeds: unmodified replay (*Original* in tables and figures), critical removal, matched irrelevant removal, critical-content replacement, and sufficient-path retention. Unmodified replay supplies the baseline even for a logged correct answer; shared seed IDs do not imply identical continuations across altered inputs.

Table 47: Correct-answer probabilities under content interventions (percentages). The 120 common eligible questions have 16 samples per condition. Remove/replace acts on critical content and its duplicates; retain keeps one complete sufficient path. An OR necessity test blocks all available paths. Panel (b) retains the 80 questions with a correct stored answer under all four policies, using 1,280 fresh responses per condition. Subset eligibility never uses these fresh draws.

(a) Full 120-question panel

Method	Original	−Irrelevant	−Critical	Replace	Retain path
EDGE	88.5	87.6	54.2	52.8	87.8
MATPO	83.8	82.9	64.0	61.5	81.8
MAGRPO	79.6	78.9	66.3	63.9	77.5
No positive gate	85.4	84.5	64.7	62.5	83.6

(b) Common logged-correct subset (80 questions)

Method	Original	−Irrelevant	−Critical	Retain
EDGE	89.2	88.4	53.3	88.8
MATPO	84.6	83.6	63.0	82.2
MAGRPO	80.0	79.5	65.4	78.1
No positive gate	86.0	85.0	64.0	84.2

Removal deletes content and duplicate disclosures throughout the accessible context, not merely citation IDs. Irrelevant controls match record count and length. Replacement substitutes unrelated content matched in format and modality, including every duplicate disclosure of the targeted facts. AND removal targets an indispensable record; OR necessity removes a set intersecting every available sufficient path. Removing one OR alternative while retaining another is a redundancy control. Retention keeps one complete path and applies only when one exists. Question, options, rule context, policy, and frozen reference answer stay fixed.

Paired accuracy and insufficient-evidence response changes use program-clustered intervals on common eligible questions, retaining unsuccessful outputs. Critical versus irrelevant removal tests dependence on the targeted content; retention versus unmodified replay tests the effect of discarding other content. These local Reasoner contrasts complement the matched training-gate ablation.

EDGE's insufficient-evidence response rate rises by 31.6 points, accounting for most of its 34.3-point accuracy decline after critical removal. All methods use the same response criterion and parser, with refusals scored incorrect. The accuracy decline thus includes recognition of missing evidence. Table 48b shows positive adjusted drops for both AND necessity and OR path blocking.

On the full 120-question panel, the adjusted drops for EDGE, MATPO, MAGRPO, and no positive gate are 33.4, 18.9, 12.6, and 19.8 points; their 95% paired program-bootstrap intervals are [26.8,39.9], [12.4,25.3], [6.9,18.5], and [13.2,26.3]. The corresponding retention-minus-unmodified-replay contrasts are -0.7, -2.0, -2.1, and -1.8 points, with intervals [-2.8,1.4], [-4.6,0.6], [-4.9,0.7], and [-4.3,0.7]. Each of 10,000 resamples retains all paired conditions and repetitions within a source program. These are pointwise intervals for local intervention effects; a contrast of method effects requires its own paired interval. Figure 3a plots critical minus irrelevant removal, so its estimates and interval endpoints reverse the sign of the positive adjusted drops above. The retention intervals include zero but do not by themselves establish equivalence. Critical-content replacement remains in Table 47: EDGE falls to 52.8%, 35.7 points below unmodified replay. Un-

Table 48: Response and content effects on the 120-question intervention panel. (a) Insufficient-evidence response rates (%); these responses score incorrect. (b) Structure-specific contrasts (percentage points), with 60 AND and 60 OR questions. Adjusted drop is irrelevant-removal minus critical-removal accuracy; retention is retained-path minus unmodified-replay accuracy. Each condition uses 16 draws per question.

(a) Insufficient-evidence responses

Method	Original replay	Critical removal
EDGE	4.2	35.8
MATPO	4.6	16.7
MAGRPO	3.8	8.8
No positive gate	3.1	12.1

(b) Effects by support structure

Method	AND drop	OR drop	AND retention	OR retention
EDGE	37.9	28.9	-1.1	-0.2
MATPO	22.0	15.8	-2.1	-1.9
MAGRPO	14.5	10.7	-2.0	-2.2
No positive gate	21.6	18.0	-1.8	-1.9

like the matched removal contrast, replacement also changes the supplied content, so this response alone does not isolate dependence on correct evidence.

OR redundancy control. Forty of the 60 OR questions contain two faithful paths for every method. Removing one path changes EDGE's accuracy by -0.3 points; blocking all paths changes it by -33.3 points (Table 49). Only the latter removes sufficient evidence, whereas an incomplete AND path cannot suffice. Figure 3b plots the subset means; the table reports uncertainty calculated on the subset itself.

Table 49: OR redundancy control on 40 common questions, 16 draws per condition (640 responses). The unmodified and all-path-blocked counts are subsets of the primary OR panel; one-path removal is the additional condition. Top: condition percentages. Bottom: paired changes and uncertainty ranges in percentage points.

Method	Original	One path removed	All paths blocked
EDGE	88.9	88.6	55.6
MATPO	84.7	84.2	64.7
MAGRPO	78.9	78.4	67.2
No positive gate	85.8	85.3	65.0

Paired contrasts and uncertainty ranges

Method	Δ one path [range]	Δ all paths [range]
EDGE	-0.3 $[-1.6, 0.9]$	-33.3 $[-39.1, -27.5]$
MATPO	-0.5 $[-2.0, 1.1]$	-20.0 $[-25.6, -14.4]$
MAGRPO	-0.5 $[-2.1, 1.2]$	-11.7 $[-16.5, -6.9]$
No positive gate	-0.5 $[-2.0, 1.1]$	-20.8 $[-26.3, -15.3]$

Pointwise 95% intervals use 10,000 paired source-program resamples, retaining all subset questions and their 16 draws per condition, with seed 13 fixed. The contrasts support selective dependence; inclusion of zero does not establish equivalence. The logged-correct subset (Table 47b) retains the direction, complementing the full panel and matched training comparison in Section 4.4.

Four methods, five primary conditions, and the 40-question one-path control require $4(120 \times 5 + 40)16 = 40{,}960$ Reasoner calls; subset analyses reuse these draws. Annotation time, eligibility, and calls are logged separately from attribution replay.

G Qualitative Cases and Training Walkthrough

G.1 Representative Calibration Trajectories

These calibration trajectories illustrate successful diagnosis, a rule-use comparison, and two attribution boundaries. Case numbers identify entries in the calibration set. They show the observed handoffs and required corrections. Quantitative repair validation is reported separately in Appendix E.1.

Case 3 L_3 **break** **C1 element identification**

Question. Which jump starts the second half? *Reference:* triple flip. *System:* triple Lutz.
Trajectory and diagnosis. The planner requests the takeoff window with pose. Perception acquires the correct window but reports a back outside edge, while the video shows an inside edge. The reasoner follows that record. Thus L_1 and L_2 pass and L_3 fails, assigning A_M rather than A_R.
Hindsight to A_M. Correct the reported takeoff edge from outside to inside.

Case 5 **EDGE versus MATPO** **C4 rule-grounded judgment**

Question. Which call applies to the late triple Lutz? *Reference:* $<$ in the official protocol.
Trajectories. Both systems obtain an equivalent landing record: blade contact about 140° short of full rotation. MATPO answers q; EDGE answers $<$. Under ISU Communication No. 2474 (2022/23), q denotes a quarter turn short with full base value; more than a quarter but less than half receives $<$ and 80% base value. MATPO misapplies the rule after obtaining the evidence. Such a rule-use error belongs to L_4 and would route the threshold to the Reasoner as hindsight. This pattern occurs in 23 of the 40 C4 questions answered correctly by EDGE and incorrectly by MATPO; the case contrast alone does not isolate the cause of the systems' difference.

Case 6 **Analyzer omission miss at** L_3

Question (C2). Does the free foot touch the ice on the triple Salchow landing? *Reference:* yes. *System:* no.
Trajectory. Perception reports "landing on the right back outside edge; free foot low to the ice", omitting the brief touch. The reasoner answers no. The analyzer accepts the noncontradictory description and predicts L_4, blaming A_R.
Correct diagnosis. With L_1/L_2 passing, omission of the decisive fact fails L_3 completeness and assigns A_M. Hindsight must repair the perception description.
Evidence boundary. A strict-refusal Reasoner lacks the touch fact in this description and must abstain. The analyzer output is a false pass of L_3, not the result prescribed by the completeness predicate. Table 38 counts link errors, without isolating omissions.

Case 7 **Cross-agent breaks and joint repair**

Question (C3). How many jumps occur between the flying spin and step sequence? *Reference:* two. *System:* three.
Trajectory and diagnosis. The planner's request stops before the second required boundary; perception misdates the first. The two records fail at L_1 and L_3, respectively, selecting A_P and A_M.
Repair scope. Repairing only one agent can leave the other required record unavailable. Joint intervention tests whether both records are restored (Appendix E.3); the observed breaks alone identify neither a unique best single repair nor optimal credit weights.

G.2 Worked Routing Examples

The following algebraic examples illustrate the routing rules. Each trajectory has one Planner, Perception, and Reasoner action, so $M_g = 3$; $w_{\max} = 2$, and ϵ is omitted for readability. Sets are shown as action projections after path filtering. For an eight-trajectory all-incorrect group, let four trajectories have neither format flag and four have both. All task rewards and advantages are zero, so $w = (0, 0, 0)$ and $A = (0, 0, 0)$ throughout. A Planner break retained by the path filter still supplies a Planner LHD target. Format flags do not affect task rewards, group statistics, or advantage signs.

For a mixed group, repeat each of the four cases below twice. Four answers are correct and four are incorrect, so $\mu_q = 0.5$, $\sigma_q = 0.5$, and task advantages are $+1$ and -1.

Table 50: Algebraic routing examples. Tuples use (P, M, R); $A = \hat{A}w$. (a) F records schema/citation-format flags and is not a reward. The unknown case has no confirmed break; the uncited success has passing upstream links. (b) OR alternatives after sufficient-path selection. The unresolved case has unknown L_3 on r_1 and an L_1 failure on r_2, so $H = 0$ means unconfirmed support. Diagnostic and training columns show action projections.

(a) Mixed-reward group

Case	R_{task}, F	$\hat{A}$	$\Phi^{\text{train}}_{\text{act}}$	Ψ, H	w	A	LHD
Wrong, Planner break	$0, (1, 1)$	-1	$\{P\}$	$\emptyset, 0$	$(2, 0, 0)$	$(-2, 0, 0)$	P
Wrong, unresolved	$0, (1, 1)$	-1	$\emptyset$	$\emptyset, 0$	$(0, 0, 0)$	$(0, 0, 0)$	–
Correct, no citation	$1, (1, 0)$	$+1$	$\{R\}$	$\emptyset, 0$	$(0, 0, 0)$	$(0, 0, 0)$	R
Correct, sufficient	$1, (1, 1)$	$+1$	$\emptyset$	$\{P, M, R\}, 1$	$(1, 1, 1)$	$(1, 1, 1)$	–

(b) Alternative OR paths

Case	$\hat{A}$	H, Ψ	Diag. actions	Train actions	w	LHD
Complete path	$+1$	$1, \{P, M, R\}$	$\{P\}$	$\emptyset$	$(1, 1, 1)$	–
Complete path	0	$1, \{P, M, R\}$	$\{P\}$	$\emptyset$	$(0, 0, 0)$	–
Use failure	-1	$0, \emptyset$	$\{P, R\}$	$\{R\}$	$(0, 0, 2)$	R
Use failure	$+1$	$0, \emptyset$	$\{P, R\}$	$\{R\}$	$(0, 0, 0)$	R
Use failure	0	$0, \emptyset$	$\{P, R\}$	$\{R\}$	$(0, 0, 0)$	R
Unresolved r_1	-1	$0, \emptyset$	$\{P\}$	$\emptyset$	$(0, 0, 0)$	–

The Planner-only weight is capped from 3 to 2. The unresolved trajectory receives no attributed task update. The uncited correct answer receives no positive task credit, but its confirmed use failure still gives the Reasoner hindsight. Repeating that case in an all-correct group also leaves its LHD target intact. With one action per role, a complete positive path gives $\Psi = \{P, M, R\}$ and weights $(1, 1, 1)$: positive routing filters support eligibility, while negative routing supplies the nonuniform action weights. For a conjunction with only some completed records, Ψ can be nonempty while $H = 0$; positive task weights are still zero. Format flags, answer correctness, support sufficiency, and attribution uncertainty therefore have distinct roles.

OR alternatives and filtered targets. Let $E^*(q) = \{r_1, r_2\}$ with two singleton sufficient paths. In the complete case, r_1 passes all four links and r_2 is unrequested, so $\Phi^{\text{diag}} = \{(1, P, t_P, r_2)\}$, $D = \{r_1\}$, and $H = 1$. The selector gives $S^\dagger = \emptyset$ and $\Phi^{\text{train}} = \emptyset$, while $\Psi = \{P, M, R\}$ still permits positive task credit. Appendix G.3 traces this complete-path case through concrete agent inputs, outputs, and training weights. In the use-failure case, r_1 passes L_1–L_3 and fails L_4, whereas r_2 fails L_1. Then $H = 0$, $\Psi = \emptyset$, and $\Phi^{\text{diag}} = \{(4, R, t_R, r_1), (1, P, t_P, r_2)\}$. Since $p_\tau(r_1) = 3 > p_\tau(r_2) = 0$, $S^\dagger = \{r_1\}$ and $\Phi^{\text{train}} = \{(4, R, t_R, r_1)\}$.

At $\hat{A} = -1$, use-failure credit is $A = (0, 0, -2)$; the unfiltered union would give $A = (-1.5, 0, -1.5)$. At zero advantage, the selected Reasoner hindsight remains eligible. For a tie with both OR records failing L_3, the lower fixed ID selects one record's facts for S_M, even though both records have owner M. Reordering the reference list preserves this choice. For AND with r_1 complete and necessary r_2 unrequested, the selector retains the Planner break on r_2; at negative unit advantage it gives $w = (2, 0, 0)$. Thus the same missing record is a training target for AND and an unused alternative for an already completed OR path.

G.3 From Agent Outputs to Training Targets

Figure 8 gives an illustrative numerical example of the complete-path OR case in Table 50. The dialogue, timestamps, and reward group instantiate the routing rules; the frames provide a schematic view of the handoffs. The question asks whether the skater steps out after landing the second jump. Given the question and a program view sampled at 2 fps, the Planner requests the landing interval from 61.20 to 62.80 s in RGB. The tool returns 48 frames at 30 fps, sampling the start-inclusive, end-exclusive interval. Perception receives the request and these frames, and produces record `e1`: initial contact occurs at 61.60 s, followed by an extra supporting step at 62.00 s. The Reasoner receives the question and `e1`, answers yes, and cites that record. Acquisition is a tool execution, so the trajectory contains three trainable role actions.

During training, the analyzer additionally receives the source and reference annotations. The sufficient sets are $\mathcal{S}_q = \{\{r_1^*\}, \{r_2^*\}\}$: either the original landing view or an alternative replay is sufficient. The reference-window overlay in the figure belongs to this audit; it is not an additional Planner input. Record `e1` matches r_1^* and passes all four links, giving $D = \{r_1^*\}$ and $H = 1$. The unrequested replay r_2^* fails L_1; its downstream links are blocked rather than counted as further failures. Thus the record-level diagnosis contains $(1, A_P, t_P, r_2^*)$, where t_P indexes the Planner turn. Because a complete OR path already exists, Equation (17) gives $S^\dagger = \emptyset$ and $\Phi^{\text{train}} = \Phi^{\text{train}}_{\text{act}} = \emptyset$. The replay remains visible in the diagnostic log without becoming a corrective target.

For the eight-trajectory reward group shown in the figure, four answers are correct and four are incorrect. The displayed trajectory has reward 1, so Equation (3) gives

$$\mu_q = \sigma_q = 0.5, \qquad \hat{A} = \frac{1 - 0.5}{0.5 + 10^{-6}} \approx 0.999998.$$

Its positive credit set is the witnessed delivery chain $\mathcal{C} = \Psi = \{(A_P, t_P), (A_M, t_M), (A_R, t_R)\}$. With $M_g = N_g = 3$ and $w_{\max} = 2$, each action has weight 1 and routed advantage approximately 0.999998. BRPO therefore reinforces all three actions that delivered and used the sufficient evidence, despite the unused replay's Planner diagnosis.

LHD uses the filtered break set, which is empty here, giving role masks $(m_P, m_M, m_R) = (0, 0, 0)$. No corrective context is constructed for the unused replay. If the same supported trajectory instead belongs to an all-correct group, its task advantage and task weights are also zero; the complete path still leaves all LHD masks zero. KL regularization and updates from other trajectories remain as specified in Appendix C.3. This example separates a record-level failure from a task-level need for correction while retaining positive credit for supported reasoning.

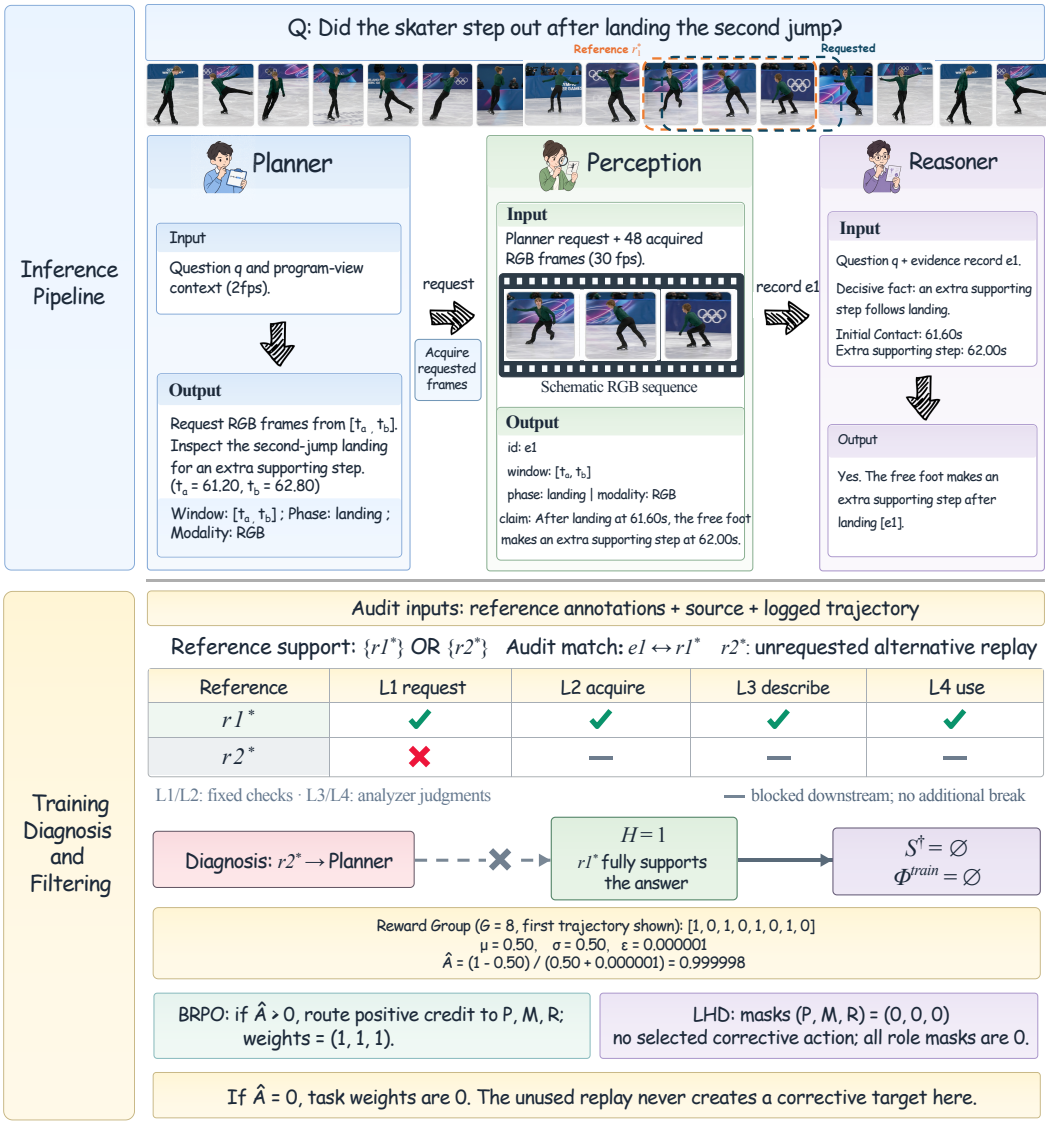

Figure 8: **A complete OR path from inference to training.** The upper panel shows the three agents' inputs and outputs, linked by the requested frames and evidence record `e1`. The lower panel audits both reference paths and applies sufficient-path selection: the completed path permits positive BRPO credit, while the unused alternative receives no corrective target. Weights and masks are ordered Planner, Perception, Reasoner.