ECR: EVIDENCE-CERTIFIED REVISION FOR SELF-CORRECTING LONG-VIDEO AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video agents increasingly answer questions by actively searching for task-relevant evidence. Once an answer has been formed, a separate decision arises: when does new evidence justify revising it? Evidence that supports an alternative answer does not necessarily refute the current one. We introduce ECR (Evidence-Certified Revision), a training-free framework for post-answer revision. ECR keeps the current answer as an anchor and treats a new answer as a challenger. It separately evaluates challenger support, anchor refutation, and evidence admissibility, records these judgments in a structured certificate, and applies a fixed reason-conditioned policy to retain the anchor, switch to the challenger, or escalate selected residual conflicts to blind pairwise arbitration that does not reveal anchor identity. On Video-MME Long, the full evaluation pipeline improves a frozen AVP agent from 52.11% to 62.33% accuracy while preserving 94.9% of initially correct answers. On a 655-question exact-replay cohort with frozen base executions and proposal outputs, ECR reaches 65.50% accuracy with 99 fixes and 13 breaks, compared with 65.19%, 141 fixes, and 57 breaks under unguarded proposal adoption. Under matched evidence access and nearly matched inference budgets, ECR-K2 reaches 55.29% accuracy versus 52.16% for a compute-matched symmetric selector, with more fixes and fewer breaks. Across five heterogeneous base agents on a shared 128-question cohort, ECR yields positive point estimates in all five settings. These results support treating the decision to revise as distinct from the process that produces complementary evidence.

1 INTRODUCTION

Long-video agents answer questions by searching for task-relevant evidence. Recent systems improve this process through iterative planning and reflection (Wang et al., 2026c; Li et al., 2026), adaptive temporal search (Ye et al., 2025b; Pan et al., 2026), and hierarchical memory (Yin et al., 2026). These methods improve how an agent acquires and organizes evidence before committing to an answer. Evidence outside the base trajectory or gathered to examine another option may support a different answer, the *challenger*. The remaining question is whether this evidence provides sufficient grounds to revise the current answer.

Revision can fix a wrong answer but also break one already correct (Huang et al., 2024). In a controlled replay of 655 Video-MME questions, unguarded proposal adoption produces 141 fixes and 57 breaks (Section 4.2). Fixes outnumber breaks, so final accuracy rises, but 16.6% of initially correct answers are lost. Reliable self-correction therefore requires deciding when new evidence justifies changing an answer already formed.

The difficulty is that support for a challenger differs from evidence against the current answer. Figure 1 shows an example. The question asks who takes the dog home at the end of the video. The current answer, the man, is supported by broader evidence from the base trajectory. The challenger, the woman, is supported by a local segment in which she walks with the dog. That segment makes the woman plausible, but it does not establish that the man failed to take the dog home. In general, $\mathrm{Support}(P, E_P) \not\Rightarrow \mathrm{Refute}(A, E_P)$, where A is the current answer, P is the challenger, and E_P is the complementary evidence packet used to examine P. A revision decision must therefore separately evaluate whether E_P supports P and provides counterevidence against A.

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: **Motivation and overview of Evidence-Certified Revision (ECR). Left:** the current answer A is the anchor, and the challenger P must meet a revision-specific evidential criterion. **Right:** P has local support but no valid counterevidence against A, so ECR retains A.

This motivates treating post-answer revision as an *asymmetric decision*. The current answer A serves as the *anchor* and is retained unless new evidence provides sufficient grounds for revision. A symmetric comparison asks which candidate is better supported, whereas revision asks whether the current answer should be overturned. A revision improves correctness only when it replaces a wrong anchor with a correct challenger. Support for the challenger without counterevidence against the anchor cannot establish that the anchor is wrong, and missing evidence does not constitute refutation. Residual disagreements follow reason-conditioned routing and usually lead to retention or direct comparison, which should not reveal which candidate is the original anchor.

Based on this formulation, we introduce ECR (Evidence-Certified Revision), a training-free framework for post-answer revision, outlined on the left of Figure 1. ECR receives the anchor with its evidence and a challenger, and forms a provenance-preserving complementary evidence packet for evaluating the challenger. Qualification evaluates challenger support, anchor refutation, and whether the temporal and logical scope of the complementary evidence is sufficient for the claim. A structured certificate records these relations and their reason. A fixed reason-conditioned policy then retains the anchor, switches to the challenger, or sends selected residual conflicts to blind pairwise arbitration. The certificate records what the evidence establishes, while the policy determines which revision action it permits.

Our experiments evaluate both the full evaluation pipeline and its revision mechanism. On the full 900-question Video-MME Long split (Fu et al., 2025), the full evaluation pipeline raises a frozen AVP agent (Wang et al., 2026c) from 52.11% to 62.33% accuracy while preserving 94.9% of initially correct answers. On exact replay, ECR reaches 65.50% accuracy with 99 fixes and 13 breaks, compared with 65.19%, 141 fixes, and 57 breaks under unguarded proposal adoption. Under matched evidence access and nearly identical inference budgets, ECR reaches 55.29% accuracy versus 52.16% for a compute-matched symmetric selector, although the paired accuracy difference remains statistically unresolved. The same frozen revision semantics yield positive point estimates across all five base agents on a shared 128-question cohort, including two whose challenger accuracy is below their base accuracy. Together, these results support a revision-specific criterion for deciding when complementary evidence warrants changing an existing answer.

In summary, our contributions are:

(1) Post-answer revision as an asymmetric decision. We formulate post-answer revision as deciding whether new evidence provides sufficient grounds to replace the current answer, and separate support for a challenger from refutation of the anchor (Section 3.1).

(2) Evidence-Certified Revision. We introduce a training-free framework that preserves evidence provenance, records revision-relevant evidence relations in a structured certificate, and sends only selected residual conflicts to blind arbitration (Sections 3.2 to 3.5).

(3) A controlled study of correction and preservation. We evaluate the full evaluation pipeline on Video-MME Long, use exact replay, ablation, and matched controls to characterize correction and preservation, and test the same frozen revision semantics across additional benchmarks and base agents (Section 4).

2 Related Work

Agentic long-video understanding. Earlier long-video systems reason over fixed or compressed visual contexts (Li et al., 2024; Ren et al., 2024). Recent agents acquire evidence on demand through adaptive temporal search and frame selection (Wang et al., 2024; 2025; Ye et al., 2025b; Pan et al., 2026; Ding et al., 2026; Lin et al., 2026; Qiu et al., 2026b), or through iterative reasoning with planning and memory (Yang et al., 2025; Wang et al., 2026b; Zhang et al., 2026b; Wang et al., 2026c; Li et al., 2026; Yin et al., 2026; Yeo et al., 2026). Recent work also separates evidence acquisition, verification, and provenance-aware aggregation (Zhang et al., 2026a). These methods improve where an agent looks and how it organizes acquired evidence. ECR begins after the base agent has committed to an answer and leaves the base trajectory unchanged. It asks whether complementary evidence provides sufficient grounds to revise that answer, allowing the same revision semantics to operate after different upstream search policies.

Verification and self-correction. Self-correction methods improve an initial output through iterative refinement, reflection, critique, or sampling and voting (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2024; Wang et al., 2022). Long-video systems also assess evidence sufficiency, verify candidate answers, revise accumulated memory, or arbitrate competing hypotheses (Qiu et al., 2026a; Wang et al., 2026a; Chen et al., 2026; Pan et al., 2026; Liu et al., 2026; Yan et al., 2026; Zhao et al., 2026). These approaches address whether evidence is sufficient, which candidate is better supported, or how an intermediate state should be updated. Self-correction can still turn initially correct answers into wrong ones (Huang et al., 2024). ECR isolates a narrower post-answer decision: whether complementary evidence provides sufficient grounds to overturn an already committed answer. It separately evaluates support for the challenger and counterevidence against the anchor.

Belief revision and judge bias. Classical belief revision studies how an agent should minimally change a belief set when new information arrives (Alchourrón et al., 1985). Work on language models examines how models update, retain, or withhold beliefs under conflicting information, including burden-of-proof protocols (Alqithami, 2026), over-revision (Wilie et al., 2024; Sadhu et al., 2026), and abstention under uncertainty (Zhang et al., 2024). ECR applies a related burden-of-proof principle to the narrower setting of revising a single committed answer while preserving the provenance of the evidence on both sides. Model judges can exhibit position, preference, and self-evaluation biases (Ye et al., 2025a; Panickssery et al., 2024; Wataoka et al., 2024). For residual conflicts, ECR removes explicit anchor identity from the verifier input.

3 Method

ECR separates what new evidence establishes from what action that evidence permits. Figure 2 organizes the procedure into three steps. Qualification evaluates complementary evidence relative to the anchor and challenger (Section 3.2). A certificate records the result and its reason (Section 3.3). A fixed policy maps the certificate to a revision action, with selected residual conflicts sent to blind arbitration (Sections 3.4 and 3.5).

3.1 Problem Setup and ECR Interface

ECR starts from an answer that the base agent has already committed to. Consider a long video $\mathcal{V}$ and a question q with answer space $\mathcal{Y}$. A base long-video agent searches the video, commits to an answer, and produces the anchor state

$$B_A = (A, E_A, T_A), \tag{1}$$

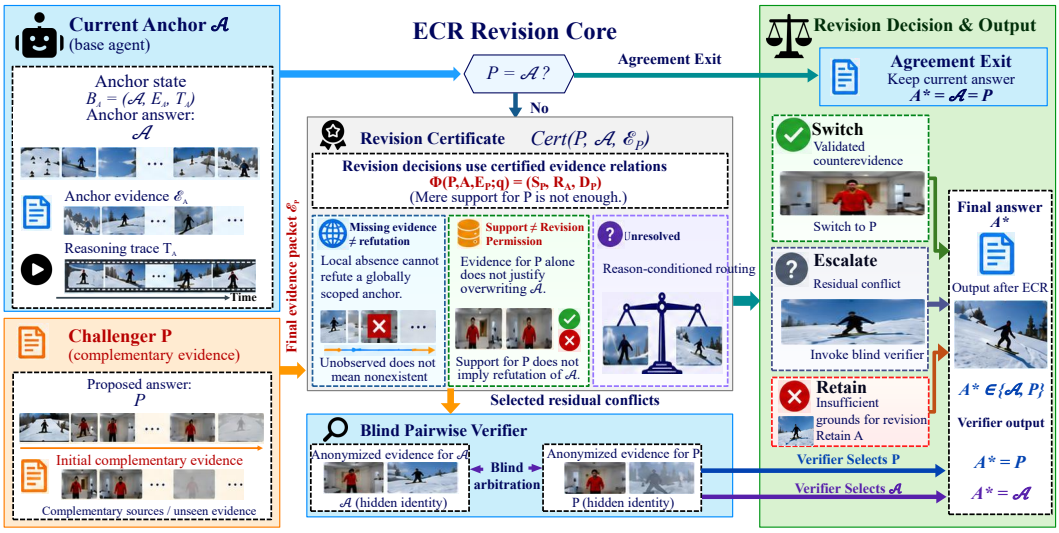


Figure 2: **Overview of Evidence-Certified Revision (ECR).** ECR starts from a current anchor A supported by the base trajectory and a challenger P supplied by an upstream proposal stage. A complementary evidence packet E_P is formed from evidence available for examining the challenger. ECR preserves the provenance of anchor and complementary evidence and evaluates whether E_P provides sufficient grounds to revise A. A structured revision certificate records this relation, and a fixed reason-conditioned policy maps it to RETAIN, SWITCH, or ESCALATE, with selected residual conflicts sent to blind arbitration.

where $A \in \mathcal{Y}$ is the anchor, E_A is its supporting evidence, and T_A is the reasoning trace of the base trajectory. ECR leaves this trajectory unchanged. Figure 2 shows the anchor state and challenger input.

An upstream proposal stage supplies a challenger P and initial complementary evidence. Before evidence-relation judgments, qualification forms the final packet E_P from evidence available for examining P. It may combine proposal-stage evidence with observations acquired during qualification. Proposal mechanisms may differ across base agents, while the qualification and revision rules of ECR remain fixed.

A symmetric selector treats A and P symmetrically in evidence selection and evaluation. ECR preserves their asymmetric roles and asks whether E_P provides sufficient grounds to change A. We write the revision operator as

$$A^* = \Gamma(B_A; P, E_P, q), \tag{2}$$

where Γ determines whether the anchor should change. A is the committed answer, while P is a challenger whose evidence must justify revision. If no valid challenger is produced or if $P = A$, ECR returns A; the latter is the *agreement exit*. Only a *live disagreement*, $P \neq A$, reaches evidence qualification.

3.2 PROVENANCE-PRESERVING EVIDENCE QUALIFICATION

Qualification records what the complementary evidence establishes about the two candidates. For a live disagreement, a qualifier receives (q, A, P, E_P) under a fixed instruction and returns three discrete judgments over the final complementary evidence packet E_P, which may include proposal-stage or qualification-stage evidence. The first,

$$S_P = \text{Support}(P, E_P), \tag{3}$$

asks whether E_P supports the challenger. The second,

$$R_A = \text{Refute}(A, E_P), \tag{4}$$

asks whether the same evidence provides counterevidence against the anchor. These judgments remain separate because support for P does not imply refutation of A. E_P is judged separately from E_A, which remains available for later arbitration.

The third judgment,

$$D_P = \mathrm{Adm}(E_P, q), \tag{5}$$

asks whether the temporal and logical scope of the evidence is sufficient for the claim it establishes. Local non-observation cannot establish global absence, so missing evidence is not refutation. Temporal-order claims also require evidence that reliably establishes the relevant ordering. These rules constrain which observations may serve as valid counterevidence.

Together, the three judgments form the evidence profile

$$\Phi(P, A, E_P; q) = (S_P, R_A, D_P). \tag{6}$$

Φ preserves challenger support, anchor counterevidence, and evidence admissibility as distinct relations that a single plausibility score would collapse. In the frozen configuration, each live disagreement uses $K = 2$ complementary evidence items. Section 4.4 evaluates this procedure under matched evidence access and nearly matched inference budgets.

3.3 Structured Revision Certificate

A certificate converts the evidence profile into an input for the revision policy. ECR maps the profile in Eq. (6) to

$$C = \mathrm{Cert}(\Phi) = (z, \rho), \tag{7}$$

where

$$z \in \{\textsc{Valid}, \textsc{Refuted}, \textsc{Unresolved}\} \tag{8}$$

is a coarse evidential state and ρ records its reason. Both are retained because the same coarse state can permit different downstream actions.

A VALID certificate indicates an admissible relation relevant to revision. Direct anchor counterevidence is the clearest case, while admissible support for an alternative that logically excludes the anchor may also be VALID. A REFUTED certificate means that a reason-specific qualification rule rejects the stated revision basis, for example support for P without valid counterevidence against A, or a negative claim based on unobserved video. The label concerns the revision claim and does not imply that P is false. UNRESOLVED covers evidence relations not resolved by the direct VALID or REFUTED conditions and therefore requiring reason-specific routing. The certificate records an evidential judgment, not a formal guarantee that the selected answer is correct.

Appendix A specifies the qualification interface, structured output, parser behavior, and reason vocabulary. Figure 4 illustrates one certificate through downstream routing.

3.4 Asymmetric Revision Policy

The revision policy acts on the state and reason recorded in the certificate. A fixed, non-learned policy maps each certificate to

$$a = G(C) \in \{\textsc{Retain}, \textsc{Switch}, \textsc{Escalate}\}. \tag{9}$$

Because G reads both z and ρ, routing is reason-conditioned rather than state-only. Validated anchor counterevidence is the principal route to SWITCH. Reasons that do not authorize revision retain the anchor, while selected residual conflicts are sent to ESCALATE.

Certificate state and revision action are distinct. In the frozen map, explicit counterevidence switches, mutually exclusive support retains, and residual reasons retain or escalate, with one reason-specific unresolved route mapping directly to SWITCH. Appendix B gives the complete deterministic map and routing outcomes.

The final answer is

$$A^* = \begin{cases} P, & G(C) = \textsc{Switch}, \\ V(A, P, E_A, E_P), & G(C) = \textsc{Escalate}, \\ A, & G(C) = \textsc{Retain}. \end{cases} \tag{10}$$

The certificate records what the evidence establishes, while G determines the permitted revision action, making each revision traceable to its evidential basis.

Table 1: **Protocol-aware positioning on Video-MME Long.** Published rows use other protocols and serve only as context. Only the last two rows share our protocol. † marks an adaptive average frame count.

Method	Inference / Visual Budget	Training	Long Acc. (%)
VideoAgent (Wang et al., 2024)	87† frames	–	49.0
VideoTree (Wang et al., 2025)	128 frames	–	54.2
T + GPT-4o (Ye et al., 2025b)	32 frames	–	59.3
TimeSearch-R-7B (Pan et al., 2026)	31.9† frames	✓	53.3
VideoZoomer-7B (Ding et al., 2026)	up to 128 frames	✓	55.8
VideoMind-7B (Liu et al., 2026)	agentic reasoning	✓	49.2
AVP + Qwen3-VL-Plus (Wang et al., 2026c)	adaptive agent	–	52.11
AVP + ECR (ours)	adaptive agent + revision	–	**62.33**

Table 2: **Paired revision performance on Video-MME.** Full900 is the primary evaluation. Exact replay provides a controlled item-level comparison between guarded and unguarded revision.

Method	Acc.	Fix/Brk	BM	Harm.	Prec.
Full900: primary paired evaluation					
Anchor	469/900 (52.11)	–	–	–	–
ECR	**561/900 (62.33)**	116/24	94.9%	2.67%	.829
Exact replay: 655 questions					
Anchor	343/655 (52.37)	–	–	–	–
Unguarded proposal	427/655 (65.19)	141/57	83.38%	8.70%	.712
ECR	**429/655 (65.50)**	**99/13**	**96.21%**	**1.98%**	**.884**

3.5 BLIND ARBITRATION FOR RESIDUAL CONFLICTS

Arbitration resolves selected residual conflicts without revealing anchor identity. When $G(C) = \text{ESCALATE}$, ECR passes two anonymized candidate–evidence pairs, $(\widetilde{Y}_1, \widetilde{E}_1)$ and $(\widetilde{Y}_2, \widetilde{E}_2)$, to a pairwise verifier V. The verifier is not told which candidate is the original anchor. This removes explicit anchor identity but does not guarantee invariance to candidate order or other indirect source cues (Ye et al., 2025a; Panickssery et al., 2024; Wataoka et al., 2024).

Arbitration is the only stage that treats the candidates symmetrically and is invoked only for designated residual conflicts. Section 4.4 compares this selective routing with exhaustive arbitration. The verifier remains fallible, as Figure 6(e) illustrates.

4 EXPERIMENTS

The experiments ask four questions: whether ECR improves a frozen agent while preserving correct answers (Section 4.2); which components control fixes and breaks (Section 4.3); whether the observed advantage remains under matched evidence access and nearly matched inference cost (Section 4.4); and whether the same frozen revision rules transfer across benchmarks and base agents (Section 4.5). The first uses the full benchmark; the ablation fixes the cohort; matched controls fix candidate pairs and evidence access; transfer freezes the revision rules across new inputs.

4.1 EXPERIMENTAL SETUP

Benchmarks and cohorts. We use four long-video benchmarks: Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), EgoSchema (Mangalam et al., 2023), and LongVideoBench (Wu et al., 2024). Full900 is the 900-question Video-MME long-video split, with one frozen base execution supplying all anchors. The 655-question *exact-replay* cohort reuses stored base executions and proposal outputs, so all revision policies share the same frozen base and proposal states. The fixed `ECR-SCOPE-256` cohort serves the component ablation. Matched controls use 255 fixed *revision-live* cases from Video-MME (147), MLVU (75), and EgoSchema (33). Benchmark transfer uses 128-question subsets of MLVU, EgoSchema, and LongVideoBench.

Base agents and baselines. The primary base agent is AVP (Wang et al., 2026c) with Qwen3-VL-Plus (Bai et al., 2025). A frozen proposal stage supplies the challenger and initial complementary

Table 3: **Component ablation on `ECR-SCOPE-256`.** All arms use the same frozen cohort and backbone. p is the exact McNemar test against A0.

Arm	Acc. (/256)	Δ	Fix/Brk	Prec.	Harm.	p
A0 Anchor only	167 (65.23%)	–	–	–	–	–
A1 Unconditional proposal	168 (65.62%)	+0.39	34/33	.507	12.89%	1.000
A2 + certificate & gate	176 (68.75%)	+3.52	11/2	.846	0.78%	0.0225
A3 + selective verifier (full ECR)	**180 (70.31%)**	**+5.08**	19/6	.760	2.34%	0.0146

evidence to ECR, which runs after the AVP trajectory without changing its search policy. The cross-agent study adds TimeChat-7B (Ren et al., 2024), VideoMind-2B (Liu et al., 2026), LensWalk (Li et al., 2026), and LongVideo-R1 (Qiu et al., 2026b) on a shared 128-question cohort, using the same frozen revision rules for all five agents. *Unguarded proposal adoption* accepts every valid challenger and measures revision without the ECR decision rule. Self-consistency (SC@3) (Wang et al., 2022) serves as a sampling baseline. Matched controls compare ECR-K2 with random selection and a symmetric selector that ranks evidence for distinguishing anonymized A and P; Symmetric-K2+ adds generic review computation. All arms share candidate pairs, the evidence pool, $K = 2$, and the blind verifier.

Metrics. Because revision can fix or break an answer, we report accuracy, wrong→correct *fixes*, correct→wrong *breaks*, belief update $\mathrm{BU} = \Pr(A^* = \text{correct} \mid A = \text{wrong})$, belief maintenance $\mathrm{BM} = \Pr(A^* = \text{correct} \mid A = \text{correct})$, harmful-flip rate (breaks over all questions), and correction precision (fixes over fixes plus breaks). Paired comparisons use exact McNemar tests (McNemar, 1947) and paired bootstrap confidence intervals (Koehn, 2004). All main experiments use the same frozen ECR configuration. The base-agent comparison measures the end-to-end pipeline; exact replay fixes base executions and proposal outputs; the ablation fixes the cohort and backbone; matched controls fix candidate pairs, evidence access, and the blind verifier.

4.2 Main Results

Table 1 places our result among published systems. Their protocols differ from ours and serve only as context. Our claims rest on the paired comparisons in Table 2, where the anchor and ECR share one frozen base execution.

The full evaluation pipeline improves accuracy while preserving most correct anchors. On Full900, the full evaluation pipeline raises accuracy by 10.22 points, from 52.11% to 62.33% (McNemar $p = 1.15 \times 10^{-15}$, paired-bootstrap CI95 $[+7.78, +12.78]$). It fixes 116 of 431 wrong anchors ($\mathrm{BU} = 26.9\%$) and preserves 445 of 469 correct ones ($\mathrm{BM} = 94.9\%$), giving a 2.67% harmful-flip rate. Frozen-subset and independent re-execution checks are also positive (Appendices E and H).

Guarded revision preserves the accuracy gain while removing most breaks. On exact replay, unguarded proposal adoption reaches 65.19% with 141 fixes, 57 breaks, and an 8.70% harmful-flip rate. ECR reaches 65.50% with 99 fixes and 13 breaks, raising BM from 83.38% to 96.21% and lowering harmful flips to 1.98%. Unguarded adoption already raises accuracy because fixes substantially outnumber breaks. ECR preserves 44 additional correct anchors while retaining the overall accuracy gain.

Repeated sampling yields a much smaller gain. On the same cohort, SC@3 raises accuracy by 1.83 points to 54.20%, whereas ECR gains 13.13 points. Additional self-consistency settings are reported in Appendix D.3.

4.3 Ablation: What Makes Revision Reliable?

Exact replay shows that the revision rule controls harmful changes but does not identify the responsible component. Table 3 and Figure 3(a) therefore add components one at a time on `ECR-SCOPE-256`.

Most observed break reduction appears after adding the certificate and policy. Unguarded proposal adoption (A1) gives 34 fixes and 33 breaks, so accuracy changes little. Adding the certificate and revision policy (A2) gives 11 fixes, 2 breaks, and raises correction precision from .507 to .846.

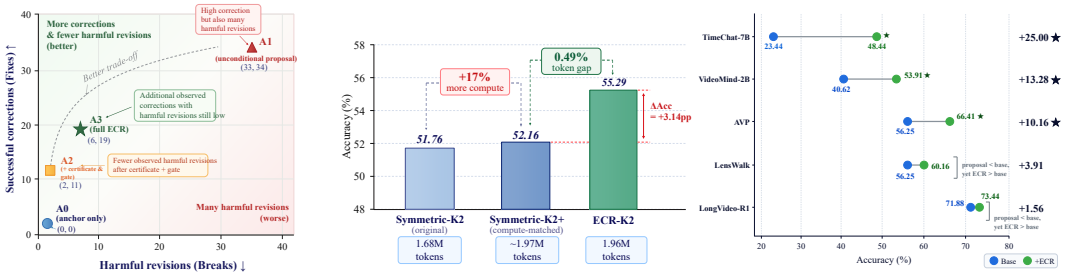


Figure 3: **Mechanism and portability of ECR. (a)** The observed break count drops most after adding the certificate and revision policy, while the full system shows additional fixes after arbitration. **(b)** At a 0.49% token gap, ECR-K2 reaches 55.29% and the symmetric selector reaches 52.16%. **(c)** The same frozen revision rules yield positive accuracy changes across all five base agents.

Table 4: **Matched evidence access and compute on 255 revision-live cases.** All arms share the candidate pairs, evidence pool, $K = 2$, and blind verifier. Symmetric-K2+ is nearly compute-matched to ECR-K2.

Method	Acc.	Fix/Brk	BM	Prec.	Harm.
Random-K2	42.35%	9/17	85.34%	.346	6.67%
Symmetric-K2	51.76%	27/11	90.52%	.711	4.31%
Compute-Matched Symmetric-K2+	52.16%	29/12	89.66%	.707	4.71%
ECR-K2	**55.29%**	**33/8**	**93.10%**	**.805**	**3.14%**

A2 is significantly better than the anchor ($p = 0.0225$), while its difference from A1 is statistically unresolved ($p = 0.341$).

Blind arbitration is associated with additional observed fixes. Adding blind arbitration for residual conflicts (A3) gives 19 fixes and 6 breaks and reaches 70.31%, 5.08 points above the anchor ($p = 0.0146$). The A2-to-A3 difference is smaller and statistically unresolved ($p \approx 0.388$). The largest observed break reduction appears at A2, while the full system shows additional fixes at A3. SCOPE-256 is a fixed component-comparison cohort and does not replace full-benchmark evaluation (Appendix D).

4.4 Matched Evidence and Compute

The ablation arms differ in evidence and computation, so they cannot separate the effect of revision-specific qualification from these resources. The controls below match evidence access and nearly match inference cost.

Extra generic compute does not close the observed gap. Table 4 and Figure 3(b) compare four arms on 255 revision-live cases. A generic symmetric criterion improves over random selection from 42.35% to 51.76%. ECR-K2 uses 17.0% more tokens than Symmetric-K2, so Symmetric-K2+ receives extra generic review until the gap falls to 0.49%. It reaches 52.16%, while ECR-K2 reaches 55.29% with more fixes (33 vs. 29), fewer breaks (8 vs. 12), and higher correction precision (.805 vs. .707). The 3.14-point difference remains statistically unresolved (McNemar $p = 0.1849$, CI95 $[-0.78, +7.06]$). This provides little evidence that generic inference compute alone explains the observed gap, but does not establish a causal effect of the qualification criterion.

Selective escalation yields a more favorable observed operating point than exhaustive arbitration. With the same qualified evidence, the frozen policy reaches 53.4% on the 268 exact-replay live disagreements, compared with 50.8% under exhaustive blind arbitration. Selective escalation gives more fixes (99 vs. 94), fewer breaks (13 vs. 15), and about 189 rather than 268 logical adjudications. This is one favorable observed operating point and does not establish policy optimality (Appendix C).

4.5 Transfer and Portability

Benchmark transfer is positive but small. We apply the same frozen rules to 128-question subsets of three additional benchmarks. ECR improves MLVU and EgoSchema by 1.56 points each and

Table 5: **Transfer of the same frozen ECR revision rules across long-video benchmarks.** No dataset-specific revision rule is introduced.

Dataset	N	Anchor	ECR	Δ	Fix/Brk
Video-MME Long	900	52.11%	**62.33%**	**+10.22**	116/24
MLVU	128	82.03%	83.59%	+1.56	3/1
EgoSchema	128	78.91%	80.47%	+1.56	3/1
LongVideoBench	128	71.88%	74.22%	+2.34	6/3

Table 6: **Cross-agent portability of ECR on a shared 128-question cohort.** The same frozen ECR revision semantics are used across all five base agents. Reported McNemar p-values are unadjusted.

Base agent	Base	Prop.	+ECR	Δ	Fix/Brk	Prec.	p
TimeChat-7B	23.44%	50.78%	**48.44%**	+25.00	37/5	.881	4.4×10^{-7}
VideoMind-2B	40.62%	56.25%	**53.91%**	+13.28	23/6	.793	0.0023
AVP	56.25%	60.94%	**66.41%**	+10.16	18/5	.783	0.0106
LensWalk	56.25%	46.88%	**60.16%**	+3.91	11/6	.647	0.3323
LongVideo-R1	71.88%	54.69%	**73.44%**	+1.56	7/5	.583	0.7744

LongVideoBench by 2.34 points. None is individually significant ($p = 0.625$, 0.625, and 0.508), so we interpret them as directionally consistent rather than conclusive benchmark-level improvements.

ECR yields positive point estimates across all five base agents. Table 6 and Figure 3(c) show positive point estimates across all five base agents, from 1.56 to 25.00 points, with three unadjusted McNemar $p < 0.05$. The Prop. column reports challenger accuracy. It falls below base accuracy for LensWalk and LongVideo-R1, yet ECR still produces a net gain. For TimeChat and VideoMind, challengers are substantially stronger than the anchors and also exceed final ECR accuracy (50.78% vs. 48.44% and 56.25% vs. 53.91%). These cases show that conservative revision can trade some proposal accuracy for protection against harmful changes when proposals are substantially stronger than anchors.

Challenger quality changes the role of revision. When unguarded proposals are strong, as in exact replay, ECR mainly protects correct anchors while preserving the accuracy gain. When fixes and breaks are balanced, as on SCOPE-256, certification and routing improve the correction–preservation profile. ECR can also remain net positive when aggregate challenger accuracy is below base accuracy, as for LensWalk and LongVideo-R1. Across these settings, challenger accuracy alone does not determine whether selective post-answer revision is useful.

5 Conclusion and Limitations

Reliable post-answer revision depends on what new evidence establishes about the current answer. A challenger can be supported without the anchor being refuted, and ECR builds revision on this distinction. On Video-MME Long, the full evaluation pipeline improves a frozen AVP agent by 10.22 points while preserving 94.9% of initially correct answers. On exact replay, ECR reduces breaks from 57 to 13 with comparable final accuracy to unguarded proposal adoption. Under matched evidence access and nearly matched inference cost, ECR-K2 shows a 3.14-point observed advantage over a symmetric selector, but the paired difference remains statistically unresolved. These results support separating revision decisions from the process that produces additional evidence.

Limitations. Full900 measures the end-to-end effect of proposal generation and ECR revision, not a same-budget causal effect of the revision rule alone. The 255-case matched difference remains statistically unresolved, while transfer uses 128-question subsets with small gains. ECR also depends on available revision evidence and model-based qualification, whose individual judgments are not separately validated against human annotations. Blind arbitration remains fallible. Broader evaluation on full transfer benchmarks and additional agent designs is needed to establish the scope of these findings.

Ethics Statement

This work studies post-answer revision for long-video agents using established benchmarks and does not involve human-subject intervention or the collection of new personal data. ECR can reduce harmful answer changes but does not guarantee correctness, and the revised outputs may still contain errors. Accordingly, the method should not be treated as a substitute for human oversight in high-stakes applications.

Reproducibility Statement

We document the protocol definitions, evaluation settings, and implementation details used for the reported results in the paper and appendix, including the frozen revision policy, exact-replay analyses, evidence-access- and compute-matched controls, cross-agent portability evaluations, and stability analyses. All reported paired analyses are derived from retained per-question execution records. Code, prompts, evaluation scripts, and evaluation manifests will be released with the paper.

AI Use Statement

Large language and vision-language models are integral components of the studied system. The reported experiments use multiple long-video agents and reasoning backbones, including Qwen3-VL-Plus (Bai et al., 2025) and GPT-5.5 (OpenAI, 2026), and an LLM is used for blind pairwise arbitration as specified in Sections 3 and 4. These model calls are part of the experimental method, and their token usage is reported for the matched-compute analysis.

AI-assisted tools were also used during manuscript preparation for language editing, code assistance, and research workflow support. The authors reviewed and verified the manuscript, experimental records, analyses, and reported conclusions, and take full responsibility for the contents of the paper.

References

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530, 1985.

Saad Alqithami. Preregistered belief revision contracts. *arXiv preprint arXiv:2604.15558*, 2026.

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

Shuimu Chen, Yuteng Chen, Yuanshen Guan, Zebang Cheng, Zeyu Zhang, Shengqian Qin, Bin Xia, Jiaran Li, Wenming Yang, and Fei Ma. Reflect-r1: Evidence-driven reflection for self-correction in long video understanding. In *European Conference on Computer Vision*, pp. 463–480. Springer, 2026.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 388–395, 2004.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. *arXiv preprint arXiv:2603.24558*, 2026.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European conference on computer vision*, pp. 323–340. Springer, 2024.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*, 2026.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 57481–57506, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

OpenAI. GPT-5.5 System Card. OpenAI System Card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/. Published April 23, 2026.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8376–8405, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. Videoseal: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *arXiv preprint arXiv:2605.12571*, 2026a.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026b.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Jayanta Sadhu, Sayem Shahad, and Kenneth Marino. Derelab: Probing defeasible reasoning and confirmation bias in llms with a generative benchmark. *arXiv preprint arXiv:2608.30413*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. *arXiv preprint arXiv:2603.04977*, 2026a.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10467–10475, 2026b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026c.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. Reveal: A rubric-guided agent for explicit evidence sufficiency verification in long-video question answering. *arXiv preprint arXiv:2608.08612*, 2026.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. Vca: Video curious agent for long video understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20168–20179. IEEE, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025a.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025b.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. Gcagent: Long-video understanding via schematic and narrative episodic memory. *IEEE Transactions on Multimedia*, 2026.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24042–24051, 2026.

Can Zhang, Baofeng Zhang, Xiaotian Han, Junyuan Shang, Yuchen Ding, Shuohuan Wang, Dianhai Yu, and Ruirui Li. From intent to evidence: Policy-steered multi-strategy retrieval for long-video agents. *arXiv preprint arXiv:2608.31005*, 2026a.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 140958–140985, 2026b.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, 2024.

Xinkui Zhao, Enbo Chen, Yifan Zhang, Chang Liu, Guanjie Cheng, Naibo Wang, and Yueshen Xu. Visage: Constructing self-correcting memories for long-form video understanding. *arXiv preprint arXiv:2607.28678*, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

A Operational Qualification and Certificate Construction

This appendix gives the operational details behind Section 3. It specifies the revision interface, evidence qualification, certificate construction, parser behavior, and blind arbitration used by ECR.

A.1 Revision Interface

For each question q, the base agent produces the frozen anchor state

$$B_A = (A, E_A, T_A),$$

where A is the anchor, E_A the evidence supporting it, and T_A the reasoning trace. ECR does not modify this trajectory.

An upstream proposal stage returns a challenger P with initial complementary evidence. Before the evidence-relation judgments are made, the qualification stage forms the final complementary evidence packet E_P from the available evidence for examining P. The packet may include proposal-stage evidence and observations acquired during qualification. The proposal mechanism may differ across base agents, while the qualification and revision rules of ECR remain fixed. For the primary AVP evaluation and benchmark transfer with AVP, the proposal configuration is also frozen. In cross-agent evaluation, proposal generation follows each base agent's interface while the ECR revision semantics remain unchanged. The resulting revision input is

$$(B_A; P, E_P, q).$$

A proposal is valid when P parses to the task answer space and its evidence references resolve to provenance-preserving observations. Otherwise, ECR takes the no-valid-challenger exit and returns A. If $P = A$, it takes the agreement exit and also returns A. Only a live disagreement, $P \neq A$, reaches qualification. In the frozen configuration, the evidence-relation judgments use a packet of $K = 2$ selected complementary evidence items whose identifiers and timestamps preserve their provenance.

This interface also marks a limit of the method. ECR can judge only the evidence that reaches the revision stage, so it cannot recover discriminative evidence that the revision pipeline never acquires.

A.2 Evidence Packet Formation

Let E_P^{init} denote the initial complementary evidence returned by the upstream proposal stage. The qualification stage builds a provenance- preserving candidate pool from E_P^{init} and, when the initial packet does not provide sufficient temporal or logical coverage, one targeted acquisition round. The acquisition request is conditioned on (q, A, P) and the current complementary evidence, but does not expose E_A. It uses the same video-access interface as the base system and returns up to four additional timestamped observations.

All candidate observations keep their source identifiers and timestamps. In the primary AVP pipeline, Qwen3-VL-Plus ranks the pool for usefulness in distinguishing P from A, evidence admissibility, and redundancy. The selector then keeps exactly $K = 2$ items, preferring non-overlapping

temporal windows when scores are tied. Remaining ties are resolved by timestamp and evidence identifier. The selector sees q, A, and P, but never the gold answer. Qualification-stage acquisition is limited to one retrieval round and one ranking call per live disagreement.

Table 7: Primary AVP implementation used to form and judge the revision packet. Cross-agent proposal generation follows the corresponding base-agent interface, while the ECR stages remain fixed.

Stage	Model	Visible candidates	Decoding
Proposal	Qwen3-VL-Plus	base trajectory	frozen upstream rule
Packet acquisition / ranking	Qwen3-VL-Plus	A, P	greedy, $T = 0$
Qualification / certificate	Qwen3-VL-Plus	A, P	greedy, $T = 0$
Blind arbitration	Qwen3-VL-Plus	anonymous pair	greedy, $T = 0$

A.3 Evidence-Relation Qualification

For a live disagreement, the qualifier judges the complementary evidence E_P through three discrete relations,

$$S_P = \mathrm{Support}(P, E_P), \qquad R_A = \mathrm{Refute}(A, E_P), \qquad D_P = \mathrm{Adm}(E_P, q).$$

S_P records whether the new evidence supports the challenger. R_A records whether the same evidence contradicts the anchor. D_P records whether the temporal and logical scope of the evidence is sufficient for the claim it is used to establish. Together they form

$$\Phi(P, A, E_P; q) = (S_P, R_A, D_P).$$

The qualifier judges E_P against A and P without pooling it with E_A. The anchor evidence stays available in case the conflict is later escalated to blind arbitration. Table 8 summarizes the three relations.

Table 8: Operational interpretation of the three evidence-relation components.

Relation	Question evaluated	Role in revision
S_P	Does E_P support challenger P?	Challenger support
R_A	Does E_P disconfirm anchor A?	Anchor counterevidence
D_P	Is the observed scope sufficient for the claim?	Evidence admissibility

Two rules constrain admissibility. Local non-observation does not establish global absence. A claim about temporal order does not count as counterevidence when the evidence cannot establish that order reliably. Appendix F reports how often these rules act in the frozen replay.

A.4 Structured Revision Certificate

The evidence profile is converted into

$$C = \mathrm{Cert}(\Phi) = (z, \rho),$$

where

$$z \in \{\textsc{Valid}, \textsc{Refuted}, \textsc{Unresolved}\}$$

summarizes the evidential state and ρ keeps the reason that produced it. Section 3.3 defines the three states. The revision policy reads both z and ρ because the state alone does not fix the action. For example, admissible support for an answer that logically excludes the anchor yields a VALID state, and the frozen policy still retains the anchor for this reason (Appendix B). A REFUTED label concerns the revision claim and does not imply that P is false.

Operationally, the qualifier returns (S_P, R_A, D_P, z, ρ) jointly in one structured response. Equation (7) is therefore compact notation for constructing a structured certificate from the evidence judgment, rather than a claim that (z, ρ) is recovered by a deterministic truth table from the three ternary fields alone. The parser deterministically checks that the state and reason are compatible with the reported evidence relations and rejects illegal combinations. Table 9 summarizes these consistency conditions.

A.5 Qualification Prompt and Structured Output

Qualification runs only on live disagreements. The qualifier receives the question with its answer choices, the anchor, the challenger, and the complementary evidence E_P. It does not receive the anchor evidence. For presentation, the normalized instruction template is:

```
You are evaluating whether newly acquired evidence justifies
revising an already-formed answer.

Question:
{question}

Answer choices:
{choices}

Current anchor:
{anchor}

Challenger:
{challenger}

Complementary evidence:
{challenger_evidence}

Evaluate the complementary evidence along three dimensions:

1. SUPPORT:
   Does the evidence support the challenger?

2. ANCHOR_REFUTATION:
   Does the same evidence provide valid counterevidence against
   the current anchor?

3. ADMISSIBILITY:
   Is the temporal and logical scope of the observed evidence
   sufficient for the claim being made?

Follow these rules:
- Support for the challenger is not by itself evidence against
  the anchor.
- Missing observation is not evidence of absence.
- Local evidence cannot establish a global negative claim unless
  its coverage is sufficient.
- Claims about temporal order require evidence that reliably
  establishes the relevant order.
- Judge only what the supplied evidence establishes.

Return exactly one JSON object and no additional text.
```

The qualifier returns one structured object:

```
{
  "support": "yes | no | uncertain",
  "anchor_refutation": "yes | no | uncertain",
  "admissibility": "yes | no | uncertain",
  "certificate_state": "valid | refuted | unresolved",
  "reason": "<reason_code>",
  "rationale": "<single concise sentence>"
}
```

Table 9: Paper-normalized reason vocabulary and canonical consistency conditions used by the ECR certificate. The labels preserve the frozen routing semantics and should not be read as literal raw-log strings.

Reason code	State	Canonical evidence relation	Interpretation
`A_explicit_counterevidence`	VALID	admissible anchor counterevidence	E_P directly contradicts A.
`B_mutually_exclusive_support`	VALID	admissible support for an incompatible P	P is supported and cannot jointly hold with A.
`U_keep`	UNRESOLVED	incomplete or non-decisive evidence	No revision route is justified.
`U_reason_retain`	UNRESOLVED	reason-specific retain condition	The frozen map conservatively retains A.
`U_reason_switch`	UNRESOLVED	reason-specific switch condition	The frozen map directly switches to P.
`U_residual_escalate`	UNRESOLVED	admissible residual two-candidate conflict	The conflict is sent to arbitration.
`R_rollback`	REFUTED	stated revision basis rejected	The anchor is retained.
`R_residual_escalate`	REFUTED	direct revision rejected but conflict remains	The conflict is sent to arbitration.

The free-text rationale is kept for audit only. The revision policy ignores it and acts only on the parsed certificate state and reason code.

A.6 NORMALIZED REASON VOCABULARY

For presentation, Table 9 normalizes the implementation-specific reason strings into a compact vocabulary. The labels preserve the certificate state and routing distinction used by the frozen policy and should not be read as literal raw-log strings. The certificate state remains separate from the action.

Two UNRESOLVED labels, `U_reason_retain` and `U_reason_switch`, denote reason-specific routes in the frozen policy. Their actions are determined by the reason code, and they are not treated as additional certificate states. The certificate state alone therefore does not determine the action.

A.7 CERTIFICATE PACKET AND PARSER BEHAVIOR

Each qualified evidence packet contains the fields in Table 10. Evidence identifiers and timestamps record the provenance of every observation used in revision.

Table 10: Fields retained in a qualified ECR certificate packet.

Field	Content
`question_id`	evaluation item identifier
`question`	question text and answer choices
`anchor`	committed answer A
`challenger`	proposed answer P
`evidence_ids`	selected complementary evidence identifiers
`timestamps`	temporal locations of selected evidence
`evidence`	textualized or serialized evidence content
`support`	parsed S_P judgment
`anchor_refutation`	parsed R_A judgment
`admissibility`	parsed D_P judgment
`certificate_state`	z
`reason`	normalized reason code ρ

The parser normalizes the output deterministically before the policy acts on it. It normalizes whitespace and capitalization and checks that each categorical field belongs to its declared vocabulary. A packet is accepted only when the certificate state and reason code form a legal pair in Table 9. The rationale field may be empty because the policy does not use it.

If a required categorical field is missing, malformed, or outside its vocabulary, the packet becomes

$$(\textsc{Unresolved}, \texttt{U_keep}),$$

which retains the anchor. A parser failure therefore cannot authorize a revision. Invalid evidence identifiers and broken provenance links are handled in the same way. Questions without a valid challenger and agreement exits are resolved before qualification and never reach the parser.

The revision policy is

$$a = G(C) \in \{\textsc{Retain}, \textsc{Switch}, \textsc{Escalate}\}.$$

Some frozen logs contain an implementation tag named `rollback`. It is not a fourth action. It marks a reason that leads to RETAIN, so the final answer stays the anchor.

A.8 BLIND ARBITRATION INTERFACE

When $G(C) = \textsc{Escalate}$, the verifier receives the question and two anonymized candidate–evidence packets,

$$(\widetilde{Y}_1, \widetilde{E}_1), \qquad (\widetilde{Y}_2, \widetilde{E}_2).$$

The packets expose the candidate answers and their supporting evidence but do not identify which candidate is the original anchor. Certificate states, routing reasons, and anchor/challenger labels are not included in the arbitration input. The verifier selects the candidate better supported by the supplied evidence, and the selected anonymous label is mapped back to the corresponding answer only after the comparison.

Candidate 1/2 assignment is deterministically randomized from the question identifier before the verifier input is constructed. Both packets use the same field order and per-item evidence token cap, and the anonymous-to-original mapping is restored only after parsing the winner. These controls remove explicit anchor identity but do not establish invariance to all indirect source cues.

For presentation, the arbitration instruction is normalized as follows:

```
You are comparing two answer candidates using only the supplied
question and evidence.

Question:
{question}

Answer choices:
{choices}

Candidate 1:
{candidate_1}
Evidence for Candidate 1:
{evidence_1}

Candidate 2:
{candidate_2}
Evidence for Candidate 2:
{evidence_2}

Select the candidate better supported by the supplied evidence.
Do not infer or use which candidate was the original answer.
Return exactly one JSON object and no additional text.
```

The verifier returns one structured object:

```
{
  "winner": "candidate_1 | candidate_2",
  "rationale": "<single concise sentence>"
}
```

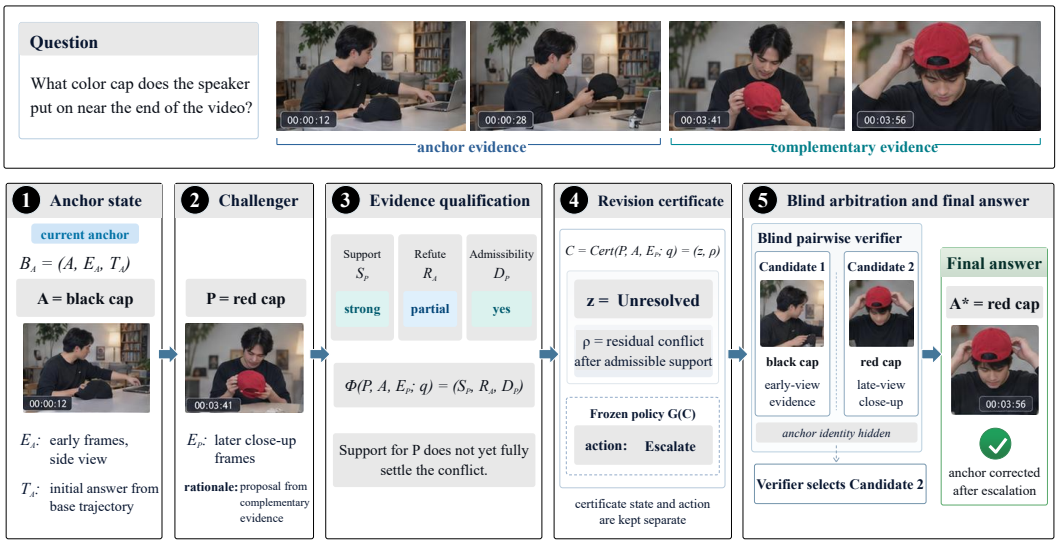


Figure 4: **Operational trace of ECR on a representative frozen revision-live case.** ECR keeps the provenance of the anchor and challenger evidence, judges support, refutation, and admissibility, and maps the resulting certificate to an action. Here the residual conflict is escalated to blind pairwise arbitration, which selects the challenger and fixes the answer.

The rationale is retained for audit and is not used by routing. The parsed winner is mapped back to the corresponding answer. An output that cannot be parsed as one of the two anonymous candidate labels does not authorize a revision and retains the anchor. Figure 4 traces one escalated case through qualification, routing, and arbitration.

B Revision Policy and Certificate Decomposition

This appendix decomposes the 655-question exact-replay cohort and its 268 live disagreements, the only questions on which revision can act. All analyses reuse the same frozen base executions and proposal outputs. Policy comparisons on the 268 live disagreements therefore use identical frozen anchor–challenger pairs.

B.1 Exact-Replay Cohort Decomposition

Table 11: Decomposition of the 655-question exact-replay cohort. Questions without a live disagreement retain the anchor by construction, so all fixes and breaks occur among the 268 live disagreements.

Cohort	N	Anchor correct	ECR correct	Fixed	Broken
Non-live exits	387	286 (73.90%)	286 (73.90%)	0	0
Live disagreements	268	57 (21.27%)	143 (53.36%)	99	13
All exact replay	655	343 (52.37%)	429 (65.50%)	99	13

Table 11 splits the cohort by whether revision can act. The 387 questions without a live disagreement are 59 cases without a valid challenger and 328 agreement exits, and all of them retain the anchor (Figure 5). Revision acts only on the 268 live disagreements, where the anchor is correct on just 57 questions. ECR raises accuracy on this subset to $143/268$ (53.36%), with 99 fixes and 13 breaks.

Over the whole cohort, ECR raises accuracy from $343/655$ (52.37%) to $429/655$ (65.50%), a gain of 13.13 points. Its correction precision is

$$\frac{99}{99+13} = 0.884,$$

and its belief maintenance is

$$\frac{343 - 13}{343} = 96.21\%.$$

On the live disagreements alone, $\mathrm{BU} = 99/211 = .469$ and $\mathrm{BM} = 44/57 = .772$.

B.2 Certificate-State Semantics

Table 12 shows how the frozen policy treats each certificate state of a live disagreement.

Table 12: Certificate states of live disagreements and their frozen actions. The action depends on the reason, and the coarse state alone does not determine it.

Verdict / reason	Frozen routing outcome
VALID: `A_explicit_counterevidence`	SWITCH
VALID: `B_mutually_exclusive_support`	RETAIN
UNRESOLVED	reason-conditioned RETAIN/SWITCH/ESCALATE
REFUTED	reason-conditioned RETAIN/ESCALATE

Explicit counterevidence against the anchor is the clearest case for direct revision and leads to SWITCH. Mutually exclusive support establishes that the challenger and the anchor cannot both hold, but the frozen policy does not switch on this reason and retains the anchor (Appendix H). UNRESOLVED and REFUTED certificates lead to different actions depending on their reason, which is why the state alone cannot serve as an action label.

B.3 Frozen Reason-to-Action Map

Table 13 gives the complete deterministic map. Once (z, ρ) has been parsed, no learned rule is involved.

The map was frozen before benchmark-scale evaluation, including the two reason-specific UNRESOLVED routes. Table 14 reports the route frequencies and changed-answer outcomes on the 268 live disagreements.

The route counts sum to 136 RETAIN, 59 SWITCH, and 73 ESCALATE. The changed routes yield 99 fixes, 13 breaks, and 20 wrong-to-wrong changes, matching the aggregate replay results in Table 11. Figure 5 visualizes the same flow.

C Revision Policy Controls

We compare three policies on the same 268 live disagreements using identical frozen anchor–challenger pairs (Table 15). ECR and exhaustive arbitration use the same qualified evidence packets. Unguarded proposal adoption accepts every challenger in this live-disagreement cohort.

Logical-adjudication accounting. We use *logical adjudication* as an analysis-level count of explicit candidate-resolution decisions rather than model calls. At the certificate level, one adjudication is counted when qualification makes an explicit revision determination: a VALID or REFUTED certificate, or a reason-specific UNRESOLVED route with a direct action. Unresolved routing without a direct action does not itself count as an adjudication. Each blind pairwise arbitration then contributes one additional adjudication when the verifier resolves an escalated conflict.

Under the frozen exact-replay routing, this gives

$$47_{\text{VALID}} + 41_{\text{REFUTED}} + 28_{\text{reason-specific unresolved}} + 73_{\text{blind arbitration}} = 189$$

logical adjudications for ECR. The 28 reason-specific unresolved decisions comprise 2 `U_reason_retain` and 26 `U_reason_switch` cases. The 73 arbitration decisions comprise 62 `U_residual_escalate` and 11 `R_residual_escalate` cases. Thus an `R_residual_escalate` case contributes once when the direct revision basis is rejected and once when the residual conflict is resolved by arbitration, whereas an `U_residual_escalate` case contributes only the arbitration decision. Exhaustive arbitration instead performs one pairwise

Table 13: Normalized reason-to-action map of the frozen ECR revision policy.

Certificate state	Normalized reason	Action
VALID	`A_explicit_counterevidence`	SWITCH
VALID	`B_mutually_exclusive_support`	RETAIN
UNRESOLVED	`U_keep`	RETAIN
UNRESOLVED	`U_reason_retain`	RETAIN
UNRESOLVED	`U_reason_switch`	SWITCH
UNRESOLVED	`U_residual_escalate`	ESCALATE
REFUTED	`R_rollback`	RETAIN
REFUTED	`R_residual_escalate`	ESCALATE

Table 14: Reason-level routing frequency and changed-answer outcomes on the 268 exact-replay live disagreements. Retain routes do not change the answer.

Reason	N	Action	Fix	Break	W→W
`A_explicit_counterevidence`	33	Switch	28	2	3
`B_mutually_exclusive_support`	14	Retain	0	0	0
`U_keep`	90	Retain	0	0	0
`U_reason_retain`	2	Retain	0	0	0
`U_reason_switch`	26	Switch	17	2	7
`U_residual_escalate`	62	Escalate	46	7	9
`R_rollback`	30	Retain	0	0	0
`R_residual_escalate`	11	Escalate	8	2	1
Total	268	–	99	13	20

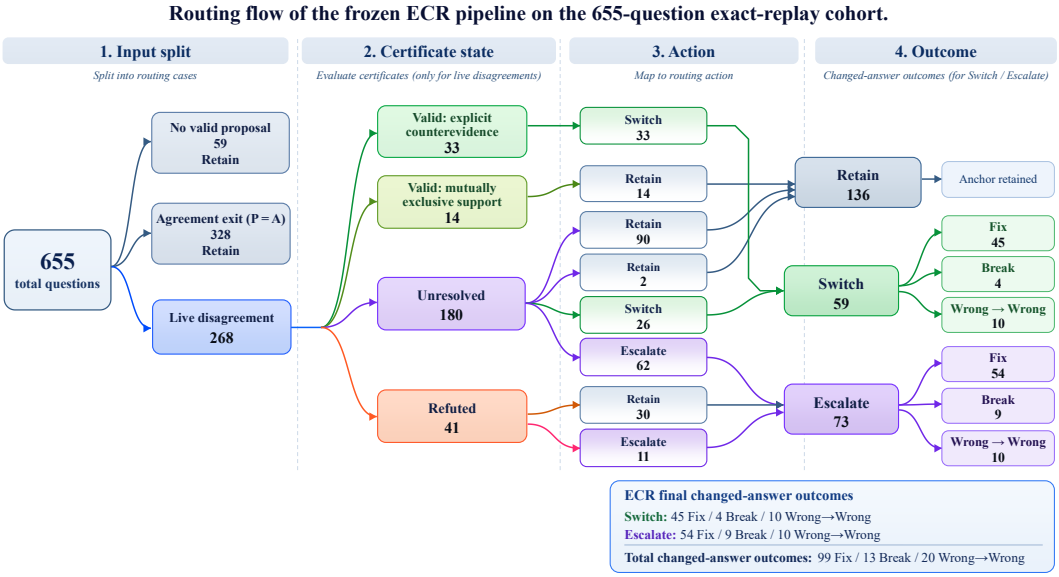


Figure 5: **Flow of the frozen ECR revision policy on the exact-replay cohort.** The flow separates early exits from live disagreements and traces each certificate state to its frozen action and outcome. States and actions appear as separate layers because the action depends on the reason.

adjudication on each of the 268 live disagreements. This accounting measures policy-level decision events and should not be interpreted as model calls or token cost.

With the same packets, selective escalation reaches 53.4% accuracy and exhaustive arbitration reaches 50.8%. Selective escalation gives 99 fixes and 13 breaks against 94 fixes and 15 breaks, with correction precision .884 against .862, and uses 189 logical adjudications compared with 268 under exhaustive arbitration.

Table 15: Policy comparison on the 268 live disagreements. *Adjud.* counts explicit candidate-resolution decisions under the accounting defined above. A case may contribute more than one such decision. The count is distinct from the number of ESCALATE cases and does not measure model calls or token cost.

Policy	Acc	Fixed	Broken	BU	BM	Prec.	Adjud.
Anchor	.213	0	0	.000	1.000	–	0
ECR (full)	**.534**	**99**	**13**	**.469**	**.772**	**.884**	189
Exhaustive arbitration	.508	94	15	.446	.737	.862	268
Unguarded proposal adoption	.526	141	57	.668	.000	.712	0

Unguarded adoption reaches 52.6% by changing many more answers and breaks all 57 correct anchors in this subset. ECR gives the highest observed accuracy at this operating point while preserving more correct anchors. This comparison does not establish policy optimality.

D Ablation and Matched Controls

D.1 Supplementary Diagnostics on ECR-SCOPE-256

Belief metrics. In the setting of Table 3, adopting every challenger (A1) gives $\mathrm{BU} = 38.20\%$ and $\mathrm{BM} = 80.24\%$. Adding the certificate and revision policy (A2) raises BM to 98.80%, and the full method (A3) keeps it at 96.41%.

Basic precondition control. Checking only provenance, proposal validity, and anchor legality gives 27 fixes and 28 breaks, reaching $166/256$ (64.84%) with correction precision 0.4909. These checks leave fixes and breaks nearly balanced, while the full qualification arm shows a stronger observed shift toward keeping correct anchors.

Qualification-isolation cohort. The 85 qualification-isolation cases are live disagreements defined by the protocol. Four of them inherit an unparsable anchor output from the frozen base trajectory and remain in the cohort. The archived raw condition contains only proposal-stage evidence, while the final E_P packets may also contain evidence acquired during qualification, which accounts for about 28% of their evidence items. The complete raw evidence pool was not archived, so this comparison matches evidence count and budget but not the full set of available evidence. We therefore read it as a comparison of qualification pipelines, which does not isolate the effect of certificate-based selection alone.

D.2 Matched Evidence Access and Compute

This control uses 255 fixed, unique revision-live cases from Video-MME (147), MLVU (75), and EgoSchema (33). All arms share the candidate pair (A, P), the evidence pool, exactly $K = 2$ selected evidence items, and the blind verifier (Table 16).

Random-K2 samples two evidence items with a fixed per-item seed. Symmetric-K2 anonymizes the two candidates and ranks the shared evidence pool for how useful each item is for deciding which candidate is better supported. The selector sees the question, the two anonymous candidate labels, evidence identifiers, timestamps, and evidence content. It does not receive anchor identity, certificate states, or ECR routing reasons. Ties are resolved by timestamp and evidence identifier, and the top two items are passed to the same blind verifier used by ECR-K2.

For presentation, the symmetric selection instruction is normalized as:

```
Given a question, two anonymous answer candidates, and a shared
evidence pool, select exactly two evidence items that are most
useful for deciding which candidate is better supported.

Treat the candidates symmetrically.
Do not infer which candidate is the original answer.
Prefer evidence that directly distinguishes the candidates and
```

```
avoid redundant items.

Return only the two selected evidence identifiers.
```

Symmetric-K2+ uses the same selected items and adds one generic review pass that summarizes the evidence for each anonymous candidate before the common blind verifier. It does not use support/refutation labels, certificate states, or anchor-preservation rules. The symmetric selector and review use Qwen3-VL-Plus with greedy decoding ($T = 0$). ECR-K2 uses 17.0% more total tokens than Symmetric-K2 (1,960,356 vs. 1,675,518), so the additional review brings Symmetric-K2+ to about 1.97M total tokens.

Table 16: Matched control of evidence access and compute on 255 revision-live cases. Symmetric-K2+ and ECR-K2 use nearly identical total inference budgets.

Method	Acc.	Fix/Brk	BU	BM	Prec.	Harm.
Random-K2	108/255 (42.35%)	9/17	6.47%	85.34%	.346	6.67%
Symmetric-K2	132/255 (51.76%)	27/11	19.42%	90.52%	.711	4.31%
Compute-Matched Symmetric-K2+	133/255 (52.16%)	29/12	20.86%	89.66%	.707	4.71%
ECR-K2	**141/255 (55.29%)**	**33/8**	**23.74%**	**93.10%**	**.805**	**3.14%**

The remaining token gap to ECR-K2 is only 0.49%. Against Symmetric-K2+, ECR-K2 gains 3.14 points, with four more fixes and four fewer breaks.

The paired outcomes are 123 cases where both arms are correct, 104 where both are wrong, 18 where only ECR-K2 is correct, and 10 where only Symmetric-K2+ is correct. The 3.14-point accuracy difference remains statistically unresolved (exact McNemar $p = 0.1849$, paired-bootstrap CI95 $[-0.78, +7.06]$ points). The observed correction and preservation metrics are higher for ECR-K2, but the comparison does not establish a causal effect of the qualification criterion.

D.3 Self-Consistency Head-to-Head

Each self-consistency arm runs the same base agent several times on the 655-question exact-replay cohort with the same frames and prompts (Table 17). The final answer is the majority vote over legal outputs, ties go to the earliest sample, and sample 1 serves as the anchor.

Table 17: Self-consistency comparison on the 655-question exact-replay cohort. McNemar p-values compare each arm with the shared anchor.

Arm	Acc	Δ vs base	Fixed/Broken	Prec.	Harmful	McNemar p
Base	343/655	–	–	–	–	–
SC@2	347 (52.98%)	+0.61	4/0	1.000	0	0.125
SC@3	355 (54.20%)	+1.83	29/17	0.630	2.60%	0.104
ECR-v2E	**429 (65.50%)**	**+13.13**	**99/13**	**0.884**	**1.98%**	1.50×10^{-17}

SC@3 reaches 54.20%, a gain of 1.83 points (CI95 $[-0.15, +3.82]$), and ECR reaches 65.50%, a gain of 13.13 points, with 99 fixes and 13 breaks. SC@3 therefore yields a substantially smaller observed gain than ECR on this cohort.

E Transfer and Portability

The certificate semantics, revision policy, and packet size $K = 2$ remain frozen across these evaluations. For benchmark transfer with AVP, the proposal configuration is also frozen. In the cross-agent study, proposal generation follows each base agent's interface while the ECR revision rules remain unchanged. These experiments test portability and do not isolate a causal effect of dataset or architecture.

E.1 TRANSFER ACROSS BENCHMARKS

Table 18 reports the full transfer statistics.

Table 18: Transfer under frozen ECR revision semantics.

Dataset	N	Base	+ECR	Δ	Fix/Brk	Prec.	Harmful	p
Video-MME-900	900	52.11%	62.33%	+10.22	116/24	0.829	2.67%	1.15×10^{-15}
MLVU-128	128	82.03%	83.59%	+1.56	3/1	0.750	0.78%	0.625
EgoSchema-128	128	78.91%	80.47%	+1.56	3/1	0.750	0.78%	0.625
LongVideoBench-128	128	71.88%	74.22%	+2.34	6/3	0.667	2.34%	0.508

All three additional benchmarks have positive point estimates, 1.56 points for MLVU and EgoSchema and 2.34 points for LongVideoBench. None is statistically significant, so we read them as directionally consistent but limited evidence of transfer.

On the frozen Video-MME subsets `UNSEEN719` and `UNSEEN_STRICT`, ECR gains 10.85 and 10.67 points.

E.2 CROSS-AGENT PORTABILITY

Table 19 repeats the paired statistics behind Table 6 and Figure 3(c). All five base agents use the same frozen revision rules.

Table 19: Cross-agent portability on the shared 128-question cohort. McNemar p-values are unadjusted.

Base agent	Base	Prop.	+ECR	Δ	Fix/Brk	Prec.	p
TimeChat-7B	23.44%	50.78%	48.44%	+25.00	37/5	.881	4.4×10^{-7}
VideoMind-2B	40.62%	56.25%	53.91%	+13.28	23/6	.793	0.0023
AVP	56.25%	60.94%	66.41%	+10.16	18/5	.783	0.0106
LensWalk	56.25%	46.88%	60.16%	+3.91	11/6	.647	0.3323
LongVideo-R1	71.88%	54.69%	73.44%	+1.56	7/5	.583	0.7744

All five point estimates are positive, and three have unadjusted McNemar $p < 0.05$. For LensWalk and LongVideo-R1, aggregate challenger accuracy is below base accuracy, yet ECR still yields positive final point estimates. For TimeChat and VideoMind, challenger accuracy exceeds final ECR accuracy, showing that conservative revision does not always retain the full gain of a strong proposal branch (Section 4.5).

E.3 ADDITIONAL BACKBONE PORTABILITY

Table 20 runs the same frozen rules with two backbones on V48, a set of 48 questions.

Table 20: Additional backbone portability on V48 ($n = 48$ per backbone) using the same frozen revision semantics.

Backbone	Base	+ECR	Δ	Fixed/Broken	BU / BM	McNemar p
GPT-5.5	81.25%	85.42%	+4.17pp	3/1	33.3% / 97.4%	0.625
Qwen3-VL-Plus	52.08%	56.25%	+4.17pp	5/3	21.7% / 88.0%	0.727

Both point estimates are positive, and neither is statistically significant at $n = 48$. We treat this experiment as a small check across backbones and draw no conclusion about broad generalization across models.

F Coverage and Temporal Rules

The coverage rule and the temporal rule limit which observations may serve as counterevidence. We treat them as constraints on counterevidence and make no claim that they improve accuracy on their own.

On the 268 live disagreements of the frozen replay, removing either rule does not change any final answer. The `non_observation_is_not_absence` signal fires on 29 items, but it affects the downstream decision only once, after the conflict has already been settled.

The temporal rule returns `NO_OP` on 266 items, produces two UNRESOLVED outcomes, and never leads to a direct switch. We therefore make no separate accuracy claim for either rule. In the current system both rules only restrict which evidence relations can support revision.

G Qualitative Case Studies

The aggregate results do not show how qualification and the revision policy behave on individual questions. Figure 6 presents representative successes and failures that cover the main qualification outcomes and all three actions.

Support without anchor refutation. Figure 6(a) illustrates the support–refutation distinction. The complementary evidence supports the challenger but gives no sufficient counterevidence against the anchor, so the revision is not authorized and the anchor is retained.

Valid direct counterevidence. Figure 6(b) shows a direct switch. Admissible complementary evidence contradicts the anchor, so the certificate is VALID and the policy switches without blind arbitration.

Safe retention (604-3). In Figure 6(c), a documentary question, the anchor is the correct answer B and the challenger is C. The complementary evidence gives no valid counterevidence against the anchor, so the frozen policy retains B. Qualification thus prevents an unsupported revision from breaking a correct answer.

Successful residual arbitration (612-3). In Figure 6(d), a question about Goya's *Atropos*, the anchor is C and the challenger is B. Qualification leaves a residual conflict, so the case is escalated. The blind verifier selects B from the anonymized answer and evidence pairs, which fixes the answer.

Residual verifier failure (686-1). In Figure 6(e), a PC-building question, the anchor is the correct answer B and the challenger is A. The certificate remains UNRESOLVED, so the case is escalated. The verifier selects A, which breaks a correct answer. Hiding the anchor removes its identity from the verifier input but does not make arbitration free of errors.

H Protocol Freeze, Pre-Specification, and Stability

Frozen revision policy and development isolation. The deployed revision policy was selected on a 64-question development set drawn from a separate Video-MME development pool and excluded from every reported evaluation cohort. The development items were used to select the reason-to-action map, packet size $K = 2$, parser fallback, qualification instruction, and blind-arbitration instruction. A branch that would also switch on mutually exclusive support was evaluated during development and was not added to the deployed policy. No reported evaluation label was used to modify these choices after the freeze.

After this freeze, benchmark-scale runs may reuse stored base executions, proposal outputs, or qualified packets as specified by each control, but the certificate vocabulary, reason-to-action map, parser behavior, packet size, and arbitration instruction are unchanged.

Disagreement cohort. The 268 items used for policy comparison are defined by $A \neq P$ and a non-empty challenger, without reference to correctness. All policy comparisons on this cohort use the same frozen anchor–challenger pairs.

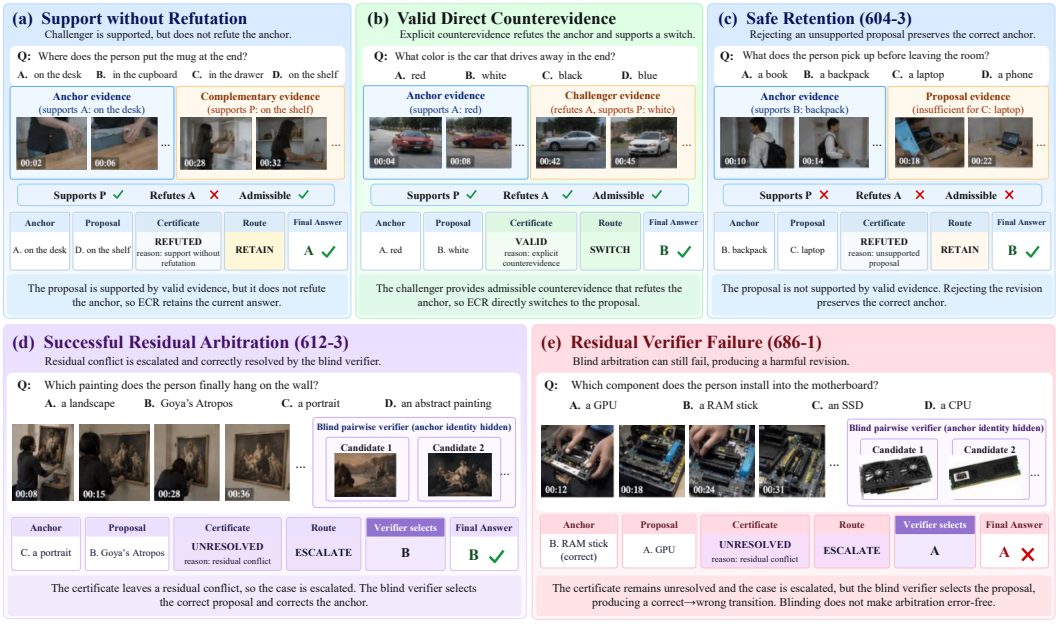


Figure 6: **Representative success and failure cases of ECR. (a)** Complementary evidence supports the challenger without providing sufficient counterevidence against the anchor, leading to RETAIN. **(b)** Valid direct counterevidence authorizes SWITCH without blind arbitration. **(c)** Qualification blocks an unsupported revision and preserves the correct anchor. **(d)** A residual conflict is escalated and successfully resolved by blind pairwise arbitration. **(e)** Blind arbitration selects the wrong challenger, illustrating a remaining failure mode.

Table 21: Development/evaluation separation used for the frozen policy. Overlap reports shared question identifiers with the 64-question development set.

Cohort	N	Role	DEV64 overlap
DEV64	64	policy selection	64
Full900	900	primary evaluation	0
Exact replay	655	replay control	0
ECR-SCOPE-256	256	component ablation	0
Matched control	255	matched comparison	0
Transfer subsets	3×128	portability	0

Stability protocol. Three end-to-end runs repeat challenger generation, qualification, and verification on a 222-question stratified subset. They give gains of 14.41, 10.36, and 10.81 points. Pairwise anchor agreement across base-agent samples is 68.2%.

Pre-specified policy controls. The exhaustive-arbitration analysis was specified before benchmark-scale evaluation and uses the same qualified evidence as the frozen policy. Under the adjudication accounting used in Appendix C, the frozen selective policy uses about 189 logical adjudications and exhaustive arbitration uses 268. These counts are distinct from the number of cases routed to ESCALATE and do not measure total model calls or token cost.

Implementation consistency. Within each paired comparison, model identities, prompts, parser rules, and routing semantics are held fixed except for the factor explicitly varied by the control. Paper-facing reason labels and instruction templates are normalized for presentation, while the frozen action semantics and reported outcomes are those used in evaluation.