TRUE RECORD, WRONG SUBJECT: DIAGNOSING VISUAL MEMORY MISBINDING IN PERSONALIZED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A personal record can be true without applying to the subject shown in a query. We study this visual memory misbinding through a controlled diagnosis of record applicability and stage-specific decisions. A record-level contract separates subject presence (P), image–record assignment (E), and answer support (S). RecordAuth-Diag contains 3,690 cases with matched assignment interventions and distinct support, presence, and ambiguity controls. Memory-bank removal and nonce relabeling establish dependence on supplied records among cases with unsafe outputs before authorization. On 560 localized DAVIS cases, an ATRA-based context guard increases correct record answers from 170 to 228 while reducing unsafe record use from 149 to 5. Filtering full-bank answers by the same retained event set preserves only 144 correct answers. Context selection therefore changes which record answers the generator produces, beyond withholding an existing answer. Retaining the correct target does not ensure that generation uses it, and exact field execution can expose an invalid selected association. A lower rate of unsafe record use in model outputs does not necessarily imply a less contaminated selected context. These interactions motivate a structured-query reference pipeline separating record selection, field execution, and candidate acceptance. Under unchanged checks, the post-hoc Field comparison yields more correct answers and more unsafe releases than checked MLLM readout. A locked transfer to 310 VOT-derived cases reproduces this trade-off: 180 versus 152 correct answers and two versus zero unsafe releases. Code and reproducibility materials are available at https://anonymous.4open.science/r/visual-memory-misbinding-1E03.

1 INTRODUCTION

Contextual visual personalization connects a new image with a user's visual history and associated personal facts (Alaluf et al., 2024; Nguyen et al., 2024; Hao et al., 2024; Liu et al., 2025; Oh et al., 2026). Answering a personal query requires both retrieving a relevant memory and establishing that its record applies to the current subject. A fact can be correct for one person yet inapplicable to another. Record applicability is therefore a separate decision from visual relevance or factual correctness.

In Figure 1, Bob's clean bank supports "London," while Alice's later "Kyoto" record does not apply to him. Changing an image–record assignment can make Alice's record appear usable for Bob without changing its content or timestamp. The resulting error need not invent a fact; it can reuse a supplied fact for the wrong subject. We call this cross-subject record use *visual memory misbinding*.

We formulate *record-level authorization* through three conditions: the query identifies a unique subject represented in the bank (P), the cited image–record assignment is valid for that subject (E), and an active, available record supports the requested field under the task's recency rule (S). Authorization is record-level: selecting the right group does not validate every card in it. We measure *unsafe record use* against a fixed record-attribution set, alongside correctness under the original task target.

Our contribution is a controlled diagnosis of record applicability across selection, generation, and support checking. RecordAuth-Diag contains 3,690 cases separating assignment interventions from

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.

support, presence, and ambiguity controls. Reciprocal swaps preserve record text and the memory-image multiset while changing their assignment; memory-bank removal and nonce relabeling test dependence on supplied records. We implement these decisions with frozen same-MLLM verifiers and the Anchor-Triadic Relational Authorizer (ATRA). The ATRA context guard intersects learned and deterministic anchor–triad selections when their groups agree, supplying only the retained records to the unchanged generator.

Record selection changes which answers can be produced, but selecting the correct support does not ensure that generation uses it. Our readout comparison finds old-record errors despite retained targets, while exact timestamp execution can expose invalid newer associations. Support checking introduces a third decision: its reference set determines whether valid fallbacks are accepted. We separate these stages and evaluate their composition for structured queries.

Improved selection does not supply every authorization observation. Four of the guard's five residual unsafe cases involve subjects absent from the bank. A relative ranking can favor one candidate even when the queried subject is absent. We characterize this boundary for translation-invariant relative evidence and separately evaluate supplied subject-equivalent tokens as an additional presence observation.

Contributions. **(1) Controlled diagnosis:** RecordAuth-Diag combines assignment, support, presence, and ambiguity interventions with memory-bank removal and relabeling tests of supplied-record dependence. **(2) Record-level decisions:** the $P/E/S$ contract connects these conditions to explicit card verification, an ATRA-based context guard, and optional candidate checks. **(3) Stage-resolved evidence:** a selector-by-readout comparison and target-retention audit separate loss of task support from field-execution errors. Fixed-candidate checks identify the safety–utility effect of support-set consistency. A prospectively locked VOT transfer evaluates the resulting checked field-execution pipeline for structured queries.

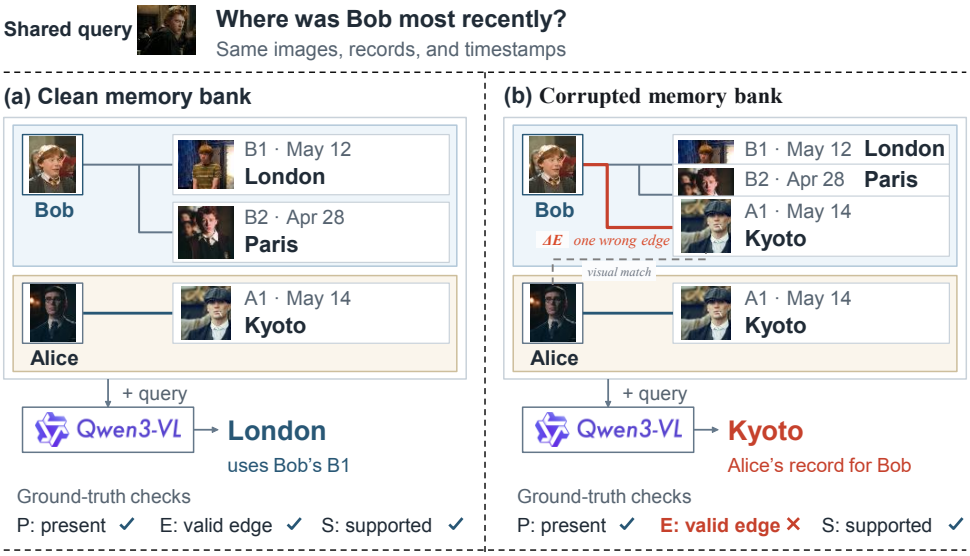

Figure 1: Visual memory misbinding. The query and record contents are unchanged; an incorrect image–record edge invalidates E while P and S remain valid.

2 Related Work

Contextualized visual personalization. MyVLM, Yo'LLaVA, RAP, and related systems evaluate recognition and use of user-specific concepts or accumulated context (Alaluf et al., 2024; Nguyen et al., 2024; Hao et al., 2024; Liu et al., 2025). MMPB evaluates personalization capability at larger scale (Kim et al., 2025). CoViP formalizes contextualized visual personalization and uses personalized captioning as a core capability (Oh et al., 2026); its LSD, LAR, and ITR diagnostics require visual recognition and contextual retrieval. RecordAuth-Diag asks a complementary question: whether a supplied record remains licensed after its visual assignment changes.

Memory authority and provenance. Access control, provenance graphs, typed tool returns, and multi-principal memory gates distinguish information availability from authority (Liu et al., 2024; Wang et al., 2026a;b; Delattre et al., 2026; Ren et al., 2026; Xu et al., 2026a). AuthMem-Bench holds a focal claim and downstream task fixed while changing source authority (Zhan et al., 2026). Our assignment interventions instead test whether record content applies to the current visual subject. The cited authority systems use machine-readable principals or sources; here visual evidence participates in the record-assignment decision. Supplied subject metadata is evaluated separately as an additional observation.

Misbinding and attribution. OmniHalluc-L tests audiovisual evidence attached to the wrong speaker, time, or modality (Dong et al., 2026). MultiBind, InstaBind-Lite, and attribute-misbinding attacks examine cross-subject transfer in generation or perception (Tian et al., 2026; Xu et al., 2026b; Fu et al., 2025). Our focus is the use of an explicit external record for a visual subject. Winoground, SugarCrepe, and AMELI provide related controlled tests of compositional structure and entity–attribute matching (Thrush et al., 2022; Hsieh et al., 2023; Yao et al., 2023). Correct composition or linking alone does not establish permission to use the linked record.

Selective prediction and grounding. Selective prediction studies abstention and risk–coverage trade-offs (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019; Chandu et al., 2025); grounding and hallucination evaluations ask whether a claim has supporting evidence (Yuksekgonul et al., 2022; Li et al., 2023; Guan et al., 2023; Rodriguez & Rohrbach, 2026). Record authorization additionally checks the record–subject relation, either before generation or for a cited answer. Safety and utility must be reported together, since always withholding records removes exposure without providing personalization.

3 Record Authorization: Decision and Identifiability

3.1 Decision object and behavioral endpoint

Let q be an image–question query and $\mathcal{B} = \{G_j\}_{j=1}^{J}$ a memory bank. Group $G_j = (a_j, M_j)$ has an enrollment anchor a_j representing a registered subject s_j, and cards $M_j = \{m_{jk}\}_{k=1}^{K_j}$ pairing images with records. The index j describes where a card is supplied, not whether that assignment is valid. Let $O^\star(m, j)$ denote the valid image–record assignment relation specified by the diagnostic provenance and intervention.

For an answer y citing m_{jk}, P_j means that the query has a unique subject matching s_j; $E_{jk} = O^\star(m_{jk}, j)$ means that the supplied assignment is valid; and $S_{jk}(q, y; R)$ checks the requested field and task constraints against reference set R. For latest-location queries, it requires an active, image-backed record matching y and no equally recent or newer competing event in R. The ideal reference $\mathcal{U}_j^\star$ contains the usable, assignment-valid events in group j; E checks the cited card's own assignment. The contract is

$$Y(q, m_{jk}, y) = P_j(q, \mathcal{B}) \wedge E_{jk}(\mathcal{B}) \wedge S_{jk}(q, y; \mathcal{U}_j^\star). \tag{1}$$

The construction fixes provenance and expected answers independently of model outputs. Reassigning an image can change E even though a card remains stored in group j. Withholding a view or marking a record stale instead changes its usable support. Appendix E specifies the expected action for each intervention.

The contract specifies the desired decision; observations determine how it can be estimated. Anchors and masks supply visual evidence for a referent, while ATRA predicts joint group/card actions. A supplied intended-subject token adds an observation unavailable to the visual guard.

For measurement, let $\mathcal{A}_i$ be the diagnostic's fixed *attribution-eligible event set*. It includes usable records assigned to the queried subject under $O^\star$, or either visible subject in an ambiguous query, without requiring the latest answer or a unique intended referent. Let $\mathcal{C}_i(y) = \{m_{jk} : Y(q_i, m_{jk}, y) = 1\}$ be the *contract-valid citation set* under the original query. These sets serve different decisions: an older event may belong to $\mathcal{A}_i$ but not support the latest-location answer; ambiguity leaves $\mathcal{C}_i(y)$ empty for every record answer.

Unauthorized card exposure occurs when the selector admits any event outside $\mathcal{A}_i$. Unsafe record use occurs when an output matches at least one supplied record but none in $\mathcal{A}_i$. Task correctness is evaluated against the unchanged expected answer, including NULL targets. Appendix A.5 decomposes attribution-eligible task errors and unmatched nonempty outputs. Gate coverage, supplied-record release and response coverage have distinct denominators. Withholding an answer cannot undo prior context exposure.

3.2 IDENTIFIABILITY BY AUTHORIZATION TYPE

An immutable anchor makes a supplied group a visual referent, but it does not make the queried subject identifiable when the subject is absent from the bank. Without an immutable referent, opaque group names and their associated observations can be permuted together, producing the same input under a different ownership assignment. RecordAuth-Diag therefore supplies one anchor per group, while treating that anchor as visual evidence rather than authenticated provenance.

Proposition 1 (Relative-evidence boundary)**.** *Let $v = (v_1, \ldots, v_J)$ be candidate evidence scores, and let a bank-relative authorizer depend only on a translation-invariant statistic $T(v)$. For a present world with scores v and an absent world with scores $v + c\mathbf{1}$, the release distributions coincide. A uniform absent-world false-release constraint α therefore limits recall in the paired present world to at most α.*

Rankings and pairwise margins satisfy this invariance. The result applies to policies whose observations are translation-invariant rankings or margins. ATRA also receives absolute similarities and metadata, so the proposition does not cover its full input space or an independently calibrated absolute presence channel.

An authenticated channel supplies a separate, conditional sufficiency endpoint.

Proposition 2 (Authenticated-provenance sufficiency)**.** *Suppose z_q authenticates a unique intended query subject, z_g authenticates the registered subject, and z_r authenticates the cited image–record association. The last credential binds the current image, record contents and subject as one integrity-protected object; substitutions invalidate it. If credential validation succeeds, equality of their certified subjects,* $\mathrm{subj}(z_q) = \mathrm{subj}(z_g) = \mathrm{subj}(z_r)$*, together with exact support over $\mathcal{U}_j^\star$ implies $P_j = E_{jk} = S_{jk} = 1$.*

Association integrity is part of credential validation. The empirical token control reads supplied subject metadata and tests an additional-observation endpoint; it does not implement credential validation (Appendix A).

4 TYPED RECORD AUTHORIZATION

Record use has three interfaces: selection returns a retained event set K, readout produces a candidate, and an optional acceptance check releases it or returns NULL. Figure 2 maps ATRA–triad selection to (a–c), Field and frozen-MLLM readouts of the same K to (d), and separate candidate checks to (e). The executable support test uses $K \cap M_j$ as its reference; the ideal contract uses $\mathcal{U}_j^\star$.

ATRA context guard. ATRA maps 24 saved visual, validity, position, and size features to a joint group/NULL decision and card-level actions. A group classification loss and weighted card supervision train these actions on the attribution-eligible event set. At inference, we ensemble the ATRA decisions and the deterministic anchor–triad decisions separately. If their selected groups agree, the guard intersects their retained event IDs; a group mismatch or empty intersection returns NULL. The original ensemble votes on latest-field answers and takes the first winning run's card set (Appendix I). This task-specific selection produces one K shared by both readouts. We denote the selection rule by ATRA $\cap$ triad in selector comparisons and call its use in readout pipelines the ATRA context guard.

Relevance selection control. Group relevance averages query–anchor similarity and the mean top-two usable query–card similarities. The best separated group supplies thresholded top-k cards. Thresholds and k maximize development accuracy minus twice exposure, subject to a clean-accuracy retention floor (Appendix A.5.1).

Readout and acceptance. MLLM readout uses frozen Qwen3-VL-8B on K. For the latest-location schema, Field returns the unique latest active, image-backed location under parsed timestamps; an empty usable set or latest-timestamp tie returns NULL. It changes execution without consulting evaluation targets or changing K. A candidate check requires a unique event trace, support in the path's event set, and localized visual evidence. The optional dual-path MLLM control is evaluated separately from the Field/MLLM readout comparison.

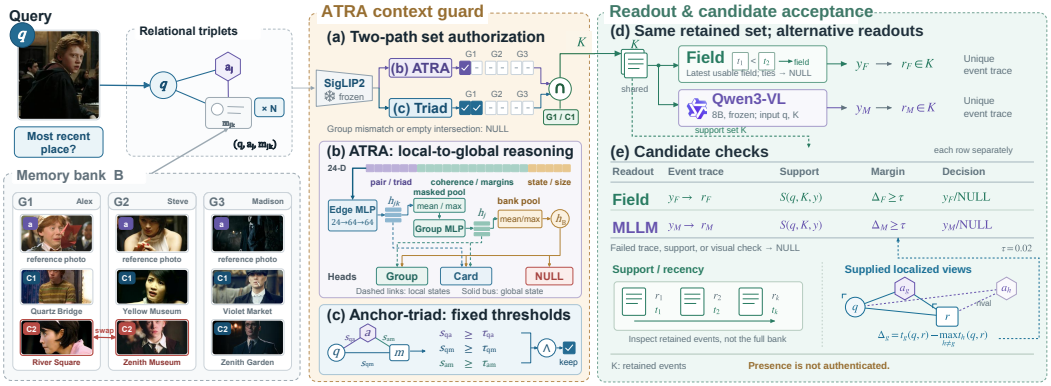

Figure 2: Record selection, alternative readouts, and candidate acceptance. ATRA–triad agreement yields K; Field and frozen-MLLM candidates use the same K and are checked separately for event trace, support, and localized edge evidence.

Counterfactual edge evidence. For query image q, cited record image r, and enrollment anchor a_g, the weakest of three visual relations scores the claimed assignment:

$$t_g(q, r) = \min\{\mathrm{sim}(q, r), \mathrm{sim}(q, a_g), \mathrm{sim}(r, a_g)\}, \tag{2}$$

$$\Delta_g(q, r) = t_g(q, r) - \max_{h \neq g} t_h(q, r). \tag{3}$$

The margin contrasts the claimed group with its strongest alternative. Conditional separation assumptions are stated in Appendix A.

Candidate paths and observations. Candidates from K use margin threshold 0.02. The dual-path MLLM control checks the guarded answer first, then a full-bank candidate at 0.075 if needed. Checks use frozen SigLIP2 and supplied localized views; missing evidence returns NULL. The Field composition retains the guard-path rule.

Alternative verification interface. The frozen Qwen/CoViP controls in Table 2 route to a group and verify its cards in separate calls; a one-pass control predicts both jointly. They see queries, anchors, card images and status, with record locations and task targets hidden. Formats and costs are in Appendix Q.

5 Experimental Design

RecordAuth-Diag contains 369 matched parents, 3,690 cases and 87 identities from LSD and Yo'LLaVA, spanning ten assignment, support, presence and ambiguity conditions. Public images are paired with synthetic records and timestamps; anchors and card status inform selection, localized views inform candidate checks, and provenance supplies supervision and evaluation labels. Authorizers use five identity folds and five seeds. DAVIS uses 100 development and 56 confirmation tracks; some source videos overlap. Its correct-record recall uses 392 positives, clean accuracy 56 cases, and aggregate unsafe rates all 560 cases. Paired intervals resample identities or tracks, with source-video sensitivity. Data, frozen generation settings and validation-selected thresholds are detailed in Appendices E, F and X.

6 RESULTS

6.1 DOES MISBINDING USE A SUPPLIED RECORD?

Assignment interventions induce unauthorized record use across all evaluated raw interfaces (Table 1). Reciprocal swaps preserve record text and the memory-image multiset; one-way replacement and duplication alter different input invariants. Anchored whole-group swaps remain a separate stress test (Appendix E).

Within each frozen raw-unsafe population, removing the memory bank preserves none of the exact unsafe answers, while displayed-nonce copying ranges from 38.10% to 92.78% in four attribution runs. These tests establish dependence on supplied records.

Table 1: Raw diagnosis and record dependence on RecordAuth-Diag. Local unsafe averages three assignment interventions; brackets are 95% identity-cluster intervals. n_{unsafe} pools all ten conditions and is the nonce-copy denominator. †: 368 parents (369 otherwise); –: not reported.

Model	Clean recall (%)	Local unsafe (%) [CI]	n_{unsafe}	Nonce copy (%)
Qwen3-VL-8B	48.10	22.49 [19.34, 25.90]	880	86.59
Phi-3.5-Vision	5.56	31.07 [25.84, 36.78]	1,258	60.33
CoViP (Qwen)	49.05	26.02 [23.25, 29.02]	970	92.78
RAP-Phi3-mini†	27.45	19.66 [16.29, 23.78]	1,105	38.10
Gemma-3-4B-IT	25.75	63.69 [59.74, 67.84]	–	–

6.2 FROM GROUP ROUTING TO RECORD-LEVEL DECISIONS

Explicit card decisions change the exposure–recall trade-off beyond group routing (Table 2). Staged verification reduces Qwen/CoViP exposure from 43.63/44.09% to 3.06/4.61%, while positive recall falls from 86.26/86.91% to 60.90/66.63%. One-pass prediction occupies another operating point. These same-MLLM verifier actions have different call costs and measure neither generated-answer accuracy nor ATRA's learned architecture. End-to-end outcomes and costs appear in Appendix Q.

Table 2: Same-MLLM verifier actions on 3,690 RecordAuth-Diag cases (%).

Model	Verifier interface	Card exposure ↓	Positive recall ↑	NULL specificity ↑
Qwen	Group / NULL	43.63	**86.26**	69.11
	One-pass group + cards	19.86	57.92	83.11
	Staged group + cards	**3.06**	60.90	**94.13**
CoViP	Group / NULL	44.09	**86.91**	68.74
	One-pass group + cards	21.46	65.04	80.31
	Staged group + cards	**4.61**	66.63	**90.79**

6.3 INPUT SENSITIVITY AND MATCHED SELECTORS

Each occlusion zeros one input block, keeping all 25 fold–seed checkpoints, training-fold transformations, thresholds and the ensemble rule fixed. For standardized continuous features, zero represents the training mean. Setting unstandardized metadata or cardinality features to zero may produce inputs outside the support of the training distribution. The validity/position block uses serialized card position, not timestamp rank (Appendix H).

Figure 3 presents safety and utility together. Query–card and anchor occlusion reduce correct recall by 47.00 and 32.79 points; their lower exposure accompanies lost personalization. Peer occlusion instead increases exposure by 1.71 points while coverage and recall fall by 2.63 and 1.97 points. Related information remains in derived features, so block occlusion measures input sensitivity rather than complete removal of an information channel.

Table 3 separates relational inputs from aggregation capacity. Flat-MLP matches ATRA's 24 features, folds, seeds and validation selection, with 55,773 versus 55,427 parameters, but fixes raw-feature

Table 3: Native selector operating points on 3,690 held-out RecordAuth-Diag cases (%).

Encoder	Selector	Acc. ↑	Unsafe exposure ↓	Gate coverage	Positive recall ↑
SigLIP2	Anchor-triad	74.01	11.46	72.33	79.09
	Flat-MLP	**81.25**	6.91	67.40	**81.15**
	ATRA	80.46	6.53	65.75	79.40
	ATRA ∩ triad	77.24	**1.90**	53.25	69.14
DINOv2	Anchor-triad	56.31	28.21	86.40	**69.53**
	Flat-MLP	65.34	11.63	59.11	60.98
	ATRA	68.29	9.73	60.62	65.31
	ATRA ∩ triad	**68.83**	**3.14**	48.37	58.65

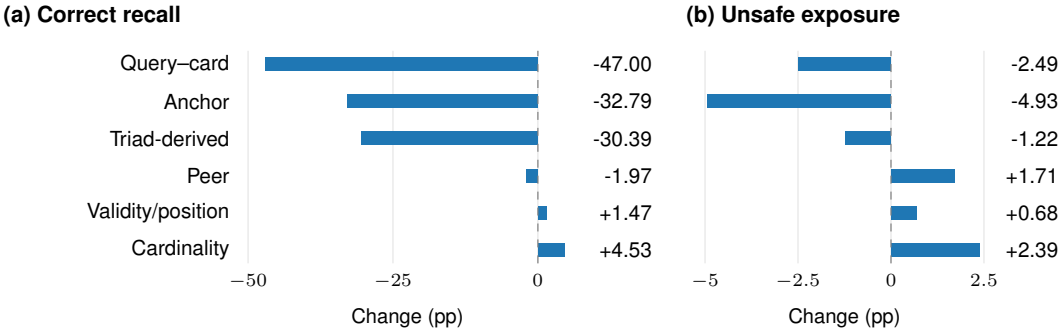

Figure 3: Frozen input-block occlusion on 3,690 held-out cases. Changes are relative to ATRA: 79.40% correct positive-task recall and 6.53% unauthorized card exposure.

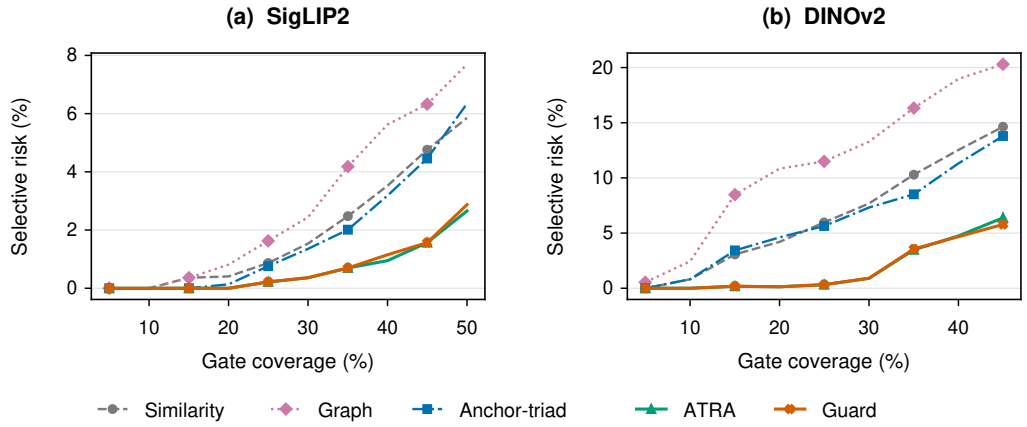

Figure 4: Common-ranking retention on RecordAuth-Diag (3,690 held-out cases). Risk is unauthorized card exposures divided by retained cases. Guard denotes ATRA ∩ triad. Only native authorizations are retained; no selector is recalibrated.

mean/max pooling. On SigLIP2, both paired accuracy and exposure intervals contain zero; on DINOv2, both favor ATRA (Table 52). The structural benefit is therefore encoder-dependent.

Figure 4 compares risk under a common ranking; Figure 6 reports cost sensitivity.

6.4 Selection and readout expose different errors

Table 4 crosses selectors and readouts with matched observations, anchors and frozen MLLM settings. The development-selected policies differ in retained sets and coverage (Appendix A.5.1).

At fixed MLLM readout, relevance yields 260 correct and 40 unsafe answers, versus 228 and 5 for the guard. At fixed Field readout, correct counts are 304 and 303, but unsafe counts are 119 and 10. The

latter guard-minus-relevance differences are -0.26 recall points (95% track interval $[-8.67, 7.40]$) and -19.46 unsafe points ($[-23.04, -15.89]$).

At fixed relevance selection, Field raises correct answers by 44 but unsafe by 79, leaving task-invalid record answers nearly unchanged (156 to 155). At fixed guard selection, it raises correct answers by 75 and reduces task-invalid answers from 83 to 12, while unsafe rises from 5 to 10. Thus generation can avoid executing invalid selected records as well as misread valid ones. The sets themselves are contaminated in 119 relevance cases and 10 guard cases, measured directly from $K_i \setminus \mathcal{A}_i$. Full paired and source-video intervals appear in Appendix A.5.2. Table 5 also exposes an order effect: guard–MLLM correctness falls from 36 to 15 on reordering, with unchanged retained content in all 56 pairs (Appendix A.5).

Table 4: Selectors and readouts on DAVIS-560. Exposed K contains any attribution-ineligible card. MLLM rows use 560 generation calls; Field rows use none.

Selector	Readout	Exposed K /560	Mean $\lvert K\rvert$	Correct ↑ /392	Unsafe ↓ /560	Other record errors ↓
Relevance	MLLM	119	3.56	**260**	40	116
ATRA guard	MLLM	**10**	2.39	228	**5**	**78**
Relevance	Field	119	3.56	**304**	119	36
ATRA guard	Field	**10**	2.39	303	**10**	**2**

Table 5: Condition-level correct and unsafe record counts on DAVIS-560 (56 cases per condition and arm). Dashes indicate NULL-target controls with no positive-record target.

	Correct records ↑		Unsafe records ↓	
Condition	Full bank	Guard + MLLM	Full bank	Guard + MLLM
Assignment interventions				
Reciprocal swap	12	**37**	27	**0**
One-way replacement	29	**37**	4	**0**
Cross-group duplicate	33	**34**	10	**0**
Support interventions				
Stale latest record	16	**36**	28	**0**
Missing latest view	21	**33**	19	**0**
Unchanged content				
Clean	**38**	36	1	**0**
Clean, reordered	**21**	15	1	**0**
Distinct NULL-target controls				
Target absent	–	–	6	**4**
Whole-group swap	–	–	52	**1**
Multi-subject ambiguous	–	–	1	**0**
Total (560 cases)	170	**228**	149	**5**

6.5 Retained support, execution, and acceptance

Generating from K yields 228 correct answers; filtering full-bank answers by the same K retains 144 (Table 61). Selection changes generation rather than only withholding an existing answer.

The target-retention audit locates the guard's field errors. Of 392 positive tasks, 305 retain a target event and 87 receive empty contexts. Every correct output occurs in the retained-target group, giving the empirical decomposition

$$\frac{228}{392} = \frac{305}{392} \times \frac{228}{305} \quad \text{(MLLM)}, \qquad \frac{303}{392} = \frac{305}{392} \times \frac{303}{305} \quad \text{(Field)}.$$

The same 77.81% support retention combines with 74.75% versus 99.34% conditional readout correctness. All 77 old-record errors retain the target. Field corrects them but converts two correct answers to unsafe use by choosing invalid newer edges. Paired readout intervals are reported in Appendix Z.

On fixed guard MLLM candidates, alignment rejects 78 other errors, two unsafe and nine correct answers, reducing invalid releases $83 \rightarrow 3$. Replacing legacy full-group recency with retained-set recency excludes rejected newer records, restoring 68 valid fallbacks ($151 \rightarrow 219$ correct), with the same three unsafe cases and unchanged tracing, localization and margin rules (Table 61).

Structured-query reference pipeline. The reference pipeline is ATRA guard $\rightarrow$ Field $\rightarrow$ check. On DAVIS, replacing MLLM readout with Field adds 70 correct and three unsafe releases under unchanged checks (Table 6).

The margin rejects 14 correct and four unsafe candidates; two ambiguous-query candidates lack localized evidence. Field already enforces the same-K support rule, so all 313 candidates reaching that check pass. Its additional screening comes from visual evidence (Appendix A.5.2).

Table 6: Record-answer counts and task-invalid fraction among releases. Other denotes task-invalid matches outside unsafe use.

Policy	Released	Correct	Unsafe	Other	Invalid / released (%)
DAVIS-560: 392 positive tasks; diagnostic composition					
Full bank + MLLM	454	170	149	135	62.56
Guard + MLLM + check	222	219	**3**	**0**	**1.35**
Guard + Field + check	295	**289**	6	**0**	2.03
VOT-310: 217 positive tasks; prospectively locked transfer					
Full bank + MLLM	254	91	83	80	64.17
Guard + MLLM + check	153	152	**0**	**1**	**0.65**
Guard + Field + check	184	**180**	2	2	2.17

6.6 Locked source transfer and residual errors

We lock the full pipeline before evaluating VOT2021 (Kristan et al., 2021), a source unused in development. The existing crop/mask criteria retain 31 of 60 sequences, yielding 310 cases. Checkpoints, ensemble and thresholds are unchanged; Appendix Z gives the source-overlap audit.

On VOT-310, checked Field gains 28 correct answers over checked MLLM: 12.90 recall points ($[4.61, 22.58]$), with two additional unsafe releases: 0.65 points ($[0.00, 1.61]$); total task-invalid releases rise from one to four. Checking Field removes four unsafe and two other errors, retaining all 180 correct answers (Table 6).

Residual errors comprise two absent-subject releases and two older answers after selection deletes the true latest event. The latter pass the same-K check, directly separating local consistency from original-task support. Presence evidence and support preservation are complementary requirements.

7 Scope and Limitations

Controlled interventions establish inducibility; natural prevalence remains unmeasured. Qwen positive recall changes from 37.05% to 35.08% on RecordAuth-Diag and 57.62% to 58.10% on DreamBooth; MyVLM clean accuracy falls from 51.26% to 25.21% (Tables 39 and 19). Utility gains are source-dependent. The Field composition is diagnosed on DAVIS and prospectively tested on VOT, with supplied anchors and localizations. DAVIS is not video-disjoint; free-form evaluation requires human validation. Appendix X retains the full scope.

8 Conclusion

RecordAuth-Diag separates factual correctness from record applicability through controlled interventions. The $P/E/S$ contract and ATRA-based pipeline connect record selection, field execution and candidate acceptance. Retention and readout audits show that generation can ignore correct support, while exact execution can expose invalid associations. Locked VOT transfer reproduces the checked-Field utility gain and its risk trade-off. Residual target absence and omitted latest records identify distinct requirements for presence evidence and support preservation.

AI USE STATEMENT

Generative AI tools assisted with hypothesis refinement, experimental-design feedback, the conceptual framework, formulation and checking of mathematical claims and proofs, method and analysis implementation, synthetic-metadata construction and cleaning, result interpretation, translation, literature discovery and summarization, manuscript editing, reference and table formatting, and schematic revision. They were not sources of empirical measurements or substitutes for model inference or human annotation; the authors retained responsibility for those checks. They verified citations against source records, regenerated reported measurements from saved predictions, reviewed the proofs, tested AI-assisted code and artifacts, and take responsibility for the final content.

ETHICS STATEMENT

This study uses public research datasets and synthetic event metadata; no new human-subject data were collected. Public availability and the evaluated diagnostic contracts do not confer consent, legal identity, or permission to collect personal information. Deployment would require authenticated enrollment and updates, data minimization, access control, deletion, demographic evaluation, and legal review. The masks and subject metadata used in controlled analyses are supplied observations rather than identities inferred from appearance.

REPRODUCIBILITY STATEMENT

The anonymous supplementary artifact provides versioned data, code, predictions and analyses. Appendix Y maps the scientific results to their reproduction inputs and records retained failures and pending human annotation. The artifact records oracle eligibility, native output metrics, paired relevance errors, the capacity-matched selector, and the post-hoc context-consistency diagnostic, all evaluated on the stated frozen predictions.

REFERENCES

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries, 2024. URL https://arxiv.org/abs/2403.14599.

Khyathi Raghavi Chandu, Linjie Li, Anas Awadalla, Ximing Lu, Jae Sung Park, Jack Hessel, Lijuan Wang, and Yejin Choi. Certainlyuncertain: A benchmark and metric for multimodal epistemic and aleatoric awareness. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cQ25MQQSNI.

Blaise Delattre, Cong Wang, and Yang Cao. CAGE: Certified authorization under typed-return uncertainty for tool-using agents. *arXiv preprint arXiv:2607.29190*, 2026.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. OmniHalluc-L: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination, 2026. URL https://arxiv.org/abs/2606.03614.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Junming Fu, Jishen Zeng, Yi Jiang, Peiyu Zhuang, Baoying Chen, Siyu Lu, and Jianquan Yang. Unveiling the attribute misbinding threat in identity-preserving models, 2025. URL https://arxiv.org/abs/2512.15818.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2023. URL https://arxiv.org/abs/2310.14566.

Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Rap: Retrieval-augmented personalization for multimodal large language models, 2024. URL https://arxiv.org/abs/2410.13360.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. URL https://arxiv.org/abs/2306.14610.

Jaeik Kim, Woojin Kim, Woohyeon Park, and Jaeyoung Do. MMPB: It's time for multi-modal personalization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0872. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/24d7018563a6e5b9ee2031d5eed5d79b-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, Ondrej Drbohlav, Jani Käpylä, Gustav Häger, Song Yan, Jinyu Yang, Zhongqun Zhang, and Gustavo Fernández. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2711–2738, October 2021.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2305.10355.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. SudoLM: Learning access control of parametric knowledge with authorization alignment, 2024. URL https://arxiv.org/abs/2410.14676.

Shi Liu, Weijie Su, Xizhou Zhu, Wenhai Wang, and Jifeng Dai. Comemo: Lvlms need image context with image memory, 2025. URL https://arxiv.org/abs/2506.06279.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant, 2024. URL https://arxiv.org/abs/2406.09400.

Yeongtak Oh, Sangwon Yu, Junsung Park, Han Cheol Moon, Jisoo Mok, and Sungroh Yoon. Contextualized visual personalization in vision-language models, 2026. URL https://arxiv.org/abs/2602.03454.

Zhe Ren, Yibo Yang, Yimeng Chen, Zijun Zhao, Benshuo Fu, Zhihao Shu, Bingjie Zhang, Yangyang Xu, Dandan Guo, and Shuicheng Yan. GateMem: Benchmarking memory governance in multi-principal shared-memory agents, 2026. URL https://arxiv.org/abs/2606.18829.

Hector G. Rodriguez and Marcus Rohrbach. SIEVES: Selective prediction generalizes through visual evidence scoring. *arXiv preprint arXiv:2604.25855*, 2026.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

Wenqing Tian, Hanyi Mao, Zhaocheng Liu, Lihua Zhang, Qiang Liu, Jian Wu, and Liang Wang. MultiBind: A benchmark for attribute misbinding in multi-subject generation, 2026. URL https://arxiv.org/abs/2603.21937.

Peiran Wang, Ying Li, and Yuan Tian. Aligning provenance with authorization: A dual-graph defense for LLM agents. *arXiv preprint arXiv:2605.26497*, 2026a.

Yiqi Wang, Zihao Yan, Jiaqi Zhang, Zhangkai Wu, Mingkai Zheng, Zequn Sun, Yanming Zhu, and Taotao Cai. MAP-Graph: Provenance-aware shared memory for multi-agent workflows. *arXiv preprint arXiv:2608.10509*, 2026b.

Jinghan Xu, Yiyong Xiao, Wanru Shao, Hankai Liu, and Xinjin Li. Memory provenance laundering in LLM agents: A non-amplification firewall for persistent memory, 2026a. URL https://arxiv.org/abs/2607.29167.

Yuanzhi Xu, Qian Gao, Jun Fan, Guohui Ding, Zhenyu Yang, Yuteng Xiao, and Sixue Lin. Diagnosing dense same-class attribute misbinding in large vision-language models, 2026b. URL https://arxiv.org/abs/2608.16805.

Barry Menglong Yao, Sijia Wang, Yu Chen, Qifan Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. Ameli: Enhancing multimodal entity linking with fine-grained attributes, 2023. URL https://arxiv.org/abs/2305.14725.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2022. URL https://arxiv.org/abs/2210.01936.

Qiuyang Zhan, Rui Zhang, Sheng Guo, Lepeng Zhao, and Zhuotao Liu. When memory becomes authority: Benchmarking authority collapse at the memory consolidation boundary, 2026. URL https://arxiv.org/abs/2608.01679.

A TYPED IDENTIFIABILITY PROOFS AND AUTHORIZATION PROTOCOL

The appendix follows verification questions: proofs and observation assumptions (A.1–A.4 and D); source construction and endpoints (E–G); authorizer training and controls (H–M); transfer and robustness (N–S); measurement and reproduction (T–Z). A.5–A.9 report output, localized-confirmation and fixed-candidate audits.

A.1 RELATIVE-EVIDENCE NON-IDENTIFIABILITY

Let a bank-relative policy be $a(T(v))$, where $v \in \mathbb{R}^J$ is a vector of candidate evidence scores and $T(v + c\mathbf{1}) = T(v)$ for every constant c. Consider a present world with score vector v and an absent world with vector $v + c\mathbf{1}$. The policy receives the same value of T in both worlds and hence has the same release distribution. If false release in the absent world is bounded by α uniformly over this world class, release in the paired present world is also at most α. Rankings and pairwise margins satisfy the stated invariance. The construction does not apply to selectors with an independently calibrated absolute presence observation.

A.2 TYPED BOTTLENECK BOUND

The typed conjunction also admits a standard bottleneck bound. For predicate type $\ell \in \{P, E, S\}$, let $\mathbb{P}^+$ denote an all-valid observable distribution and $\mathbb{P}^-_\ell$ a matched intervention that flips only predicate ℓ. If gate a_ℓ has false acceptance at most α_ℓ and release requires every gate, then

$$\Pr_{\mathbb{P}^+}(\text{release}) \le \min_\ell \left\{ \alpha_\ell + \mathrm{TV}(\mathbb{P}^+, \mathbb{P}^-_\ell) \right\}. \tag{4}$$

Equation 4 states a distributional bottleneck for the specified matched observables. The experiments report finite-sample authorization and answer outcomes under named observation channels; no evaluated policy receives a numerical risk certificate from this bound. The derivation below bounds valid release under the stated false-acceptance constraints.

For one predicate type $\ell \in \{P, E, S\}$, let $\mathbb{P}^+$ denote the common all-valid observable distribution and $\mathbb{P}^-_\ell$ a matched distribution that flips only predicate ℓ while retaining the same candidate indices. Final release implies acceptance by gate a_ℓ, so

$$\Pr_{\mathbb{P}^+}(\text{release}) \le \mathbb{E}_{\mathbb{P}^+}[a_\ell] \tag{5}$$

$$\le \mathbb{E}_{\mathbb{P}^-_\ell}[a_\ell] + \mathrm{TV}(\mathbb{P}^+, \mathbb{P}^-_\ell) \tag{6}$$

$$\le \alpha_\ell + \mathrm{TV}(\mathbb{P}^+, \mathbb{P}^-_\ell). \tag{7}$$

Taking the minimum over ℓ gives Eq. 4. The middle inequality is the bounded-function characterization of total variation.

A.3 Conditional edge separation

Assume that every traced authorized candidate satisfies $\Delta_g \ge \delta_+$ and every traced misbound candidate satisfies $\Delta_g \le \delta_-$, with $\delta_- < \delta_+$. For any $\tau \in (\delta_-, \delta_+]$, thresholding the margin accepts all candidates in the first set and none in the second. If the record trace and support witness are exact, every accepted present-identity candidate has a valid edge and support. This guarantee does not include target presence; an absent query may still have a positive best-group margin.

A.4 Authenticated-provenance sufficiency

The query credential identifies one intended subject and validates its enrollment in the group registry. The cited credential authenticates the complete image–record association, including the current image, record contents and subject. Verification must detect substitution of either the image or record, even when an unchanged identity string is replayed. Equality of validated subjects then establishes P_j and E_{jk}. The exact support witness establishes S_{jk} over task-valid events, completing the conjunction. Typed false acceptance is zero under these premises. Identity-label equality alone does not supply association integrity.

The query token is read from `query_identity`. Group and image tokens use `source_identity` and `image_source_identity`, respectively. The latter describes the currently attached image rather than only the text record's original owner. The benchmark's image interventions update that field with the image. The oracle checks these metadata values and a unique candidate's active-view, field and full-bank recency constraints. It neither authenticates image bytes nor validates cryptographic credentials. A test replacing only the image path while leaving metadata unchanged is accepted, explicitly separating this oracle from the proposition's integrity-protected observation. The metadata control also supplies an intended subject for ambiguous queries, so its observation differs from the original visual task. Original NULL targets remain unchanged in scoring.

A.5 Attribution, task validity, and stage-resolved output audit

The construction defines several sets before inference. Visible-subject records $\mathcal{V}_i$ are records assigned to subjects visible in the original query. Attribution-eligible events $\mathcal{A}_i$ additionally satisfy the construction's assignment and usability rules; ambiguous queries retain both visible subjects. The task-correct event set $\mathcal{T}_i$ consists of events supporting the unchanged positive target and is empty for NULL targets. The contract-valid set $\mathcal{C}_i(y)$ requires the full conjunction in Eq. 1 for the particular answer y. An older valid record can lie in $\mathcal{A}_i$ without lying in $\mathcal{T}_i$ or supporting the latest-location answer in $\mathcal{C}_i(y)$. For an ambiguous query, $\mathcal{V}_i$ and $\mathcal{A}_i$ can be nonempty while $\mathcal{T}_i$ and every record-answer $\mathcal{C}_i(y)$ are empty. The stored field `authorized_memory_ids` denotes $\mathcal{A}_i$.

Ideal and executable support references. Let $U(m)$ indicate active status, an available image and a recorded query field. The ideal comparison set is $\mathcal{U}^\star_j = \{m \in M_j : U(m) \wedge O^\star(m, j)\}$. Writing $\ell(m)$ and $t(m)$ for location and timestamp, the evaluated schema uses

$$S(q, m, y; R) = U(m) \wedge [y = \ell(m)] \wedge \big[\forall e \in R \setminus \{m\},\ t(e) < t(m)\big].$$

Assignment remains a separate condition. In Figure 1, the invalid newer record can pass this field/time test, while $E = 0$ prevents a valid citation; the older valid fallback is not superseded by an invalid

competitor. Reconstructing this conjunction reproduces all task-correct citation sets on 3,690 held-out RecordAuth-Diag and 560 DAVIS cases.

The executable check substitutes the selected group's usable K for $\mathcal{U}_j^\star$. For a group whose $\mathcal{U}_j^\star$ has a unique latest event $m^\star$, the retained reference selects the same event if it retains $m^\star$ and every other usable retained event has an earlier timestamp. If selection deletes the true latest event, an older candidate can be locally consistent with K but task-incorrect. Conversely, a full-bank reference can reject a valid fallback because an invalid newer record remains in the bank. The retention audit measures the former possibility; the fixed-candidate comparison isolates the latter.

Decision or measurement	Reference set	Consequence
Gate exposure	$\mathcal{A}_i$	Any admitted event outside the set is exposed
Unsafe record use	$\mathcal{A}_i$	No parsed matching event belongs to the set
Correct record answer	Original target / $\mathcal{T}_i$	NULL tasks cannot yield correct record answers
Typed release	$\mathcal{C}_i(y)$	Unique referent, valid edge, and task support are all required

A CPU-only audit rejoins all 560 saved outputs per arm with the original manifest and task targets. Table 7 partitions outputs without changing the parser or unsafe labels. The guard's 78 attribution-eligible errors comprise 77 answers citing an older record than the valid target and one answer to an ambiguous NULL-target query. Each guard record answer has a unique event trace. Hence 83/311 record releases are task-invalid (26.69%), compared with 5/311 unsafe record releases (1.61%). The original aggregate unsafe endpoint remains 5/560 (0.89%). These denominators describe different decisions.

Table 7: Mutually exclusive output categories on DAVIS-560. Counts use the original parser, attribution set and task target.

Output category	Full bank	Guard
Correct record answer	170	228
Unsafe record answer	149	5
Attribution-eligible older-record error	107	77
Attribution-eligible answer to a NULL task	27	1
Other attribution-eligible record error	1	0
Unmatched nonempty error	0	8
Abstention	106	241
Total	560	560

The remaining full-bank error has two parser matches, one attribution-eligible and one ineligible, and is non-exact. The frozen unsafe-use rule treats any attribution-eligible match as sufficient to avoid the unsafe label. We retain this case explicitly rather than assigning it an unobserved unique citation.

Ordering after fixed selection. All 56 clean/reordered pairs preserve the retained event IDs and their image, record, timestamp, status and subject contents, as well as query image, question and expected answer. The ordered event sequence changes in 44 pairs. The guard has 36 correct clean answers and 15 correct reordered answers: 22 change from correct to incorrect, one changes from incorrect to correct, 14 remain correct and 19 remain incorrect. Predictions change in 27 pairs. The net correctness loss therefore occurs after the unchanged selected content is serialized. This paired result locates the failure at the generation stage on these cases; it does not identify a neural representation responsible for it.

Support preservation versus set agreement. Rejoining the saved legacy and context-aligned check decisions finds 68 newly accepted answers, all correct under the original targets; the aligned check retains 219 correct and three unsafe answers. Two executable schema-level controls clarify the relevant condition. Both hold the candidate older answer and its retained context fixed. If the excluded newer record has an invalid image assignment, the old answer is a valid fallback. If that newer record is valid, the same context-local recency test still accepts the old answer, which is now

task-incorrect. These are logical counterexamples tested on synthetic records, not additional model evaluations.

Excluded newer event	Full-bank recency	Retained-set recency	Original task
Invalid assignment	Reject old	Accept old	Correct
Valid, mistakenly removed	Reject old	Accept old	Incorrect

A support check must preserve the task-valid comparison set, rather than merely reuse the generator's subset. The observed recovery diagnoses the legacy check's false rejection; its post-hoc design measures that specific repair.

A.5.1 Target retention, field execution, and same-stage relevance

Evaluation-only support sets. For each positive task, $\mathcal{T}_i$ contains the attribution-eligible events whose complete location field equals the original target. It is empty for NULL tasks. These sets are joined to the saved K_i after selection and never enter the readout or test-time relevance selector. Among 392 positive cases, 305 retain a target event and 87 receive empty contexts. All 228 correct generated record answers occur in the first group. All 77 older-record errors also occur there: none is explained by deletion of the target support.

Table 8: Target-support retention and readout on positive DAVIS tasks. Targets are used only for evaluation. Rates use the denominator in each row.

Audit quantity	Count / denominator	Rate (%)
Target retained by guard	305/392	77.81
Target absent from guard K_i	87/392	22.19
MLLM correct given retained target	228/305	74.75
Field readout correct given retained target	303/305	99.34
Older-record errors with target retained	77/77	100.00
Older-record errors with target deleted	0/77	0.00

Outcome-independent field execution. The readout receives only the predicted group and K_i together with its record fields. It filters for active status, an available image, and a recorded location; it returns the unique latest location under parsed ISO timestamps. An empty usable set or a tie at the latest timestamp returns NULL. Ties reject even if the tied fields agree, so serialized order cannot determine the answer. No query identity, image-source identity, target or attribution label enters this decision. All 1,337 selected guard cards already satisfy the usability rule; the comparison changes readout rather than card eligibility.

The deterministic policy corrects the 77 older answers. Two reciprocal/one-way assignment cases move in the opposite direction: an invalid newer edge remains in K_i, and timestamp execution selects it over the valid older answer. There are 303 correct answers overall and 10 unsafe releases, seven on absent targets, two on these assignment cases, and one on a whole-group swap. Two ambiguous queries yield other record errors. Clean and reordered correctness are both 44/56, compared with 36/56 and 15/56 for generation. The control resolves field recency within a selected bank; it inherits invalid edges and unresolved subject presence from selection.

Applying the same field rule to relevance-selected contexts completes the selector-by-readout comparison without further inference or parameter changes. It yields 459 record releases: 304 correct, 119 unsafe and 36 other record errors. The guard's corresponding counts are 315, 303, 10 and 2. At fixed field execution, the net differences are 109 fewer unsafe and 34 fewer other record errors, with one fewer correct answer. These are paired benchmark counts, not an equivalence claim about population utility. The relevance generator's lower unsafe count (40 versus 119 under direct readout) also shows that generation can withhold or avoid invalid selected records rather than only introduce errors.

Candidate-check transitions. The saved acceptance logs verify that the context-aligned guard is a filter of the same guard candidates. It rejects nine correct record answers, two unsafe answers, and all 78 other record errors. Thus the change from 83 to 3 task-invalid record releases is an observed case transition, not only a difference between rounded aggregate rates. Relative to the legacy check,

the aligned check accepts 68 additional correct candidates; the same three unsafe candidates remain. Unmatched nonempty errors are rejected by the unique-trace requirement and are reported separately from record-matching errors.

Development-locked relevance selection. Both selectors observe the same full-frame SigLIP2-base cache used by the frozen guard, with one supplied enrollment anchor per group and the same original DAVIS crops. Instance-mask evidence belongs to the separate candidate checks; it is not added to either pre-generation selector. The relevance score is

$$r_j = \tfrac{1}{2}\left[\mathrm{sim}(q, a_j) + \mathrm{meanTop2}_{m \in U_j}\, \mathrm{sim}(q, m)\right],$$

where U_j contains active, image-backed records with a recorded location. A group with no usable card receives a rejecting score. The highest-scoring group is selected if its score and margin over the next group pass thresholds. Within it, the selector retains up to k highest query–card scores above a card threshold. Group-score ties use opaque subject IDs; card-score ties use event IDs. Selection reads neither source identities nor test labels.

A 400-configuration grid uses group/card thresholds $\{-1, 0.2, 0.4, 0.6, 0.8\}$, group margins $\{0, 0.02, 0.05, 0.1\}$ and $k \in \{1, 2, 4, 8\}$. On the 1,000 DAVIS development cases, each configuration is scored by exact deterministic-readout accuracy minus twice unauthorized card exposure, subject to retaining at least 95% of the best clean accuracy. Ties prefer higher accuracy, lower exposure, fewer cards, then the first lexicographic parameter tuple. The selected thresholds are $(r, \text{margin}, \text{card}, k) = (0.6, 0.02, 0.6, 8)$, with 91/100 clean correct, 662/1,000 overall correct, and 202/1,000 exposed on development. All parameters were fixed before new test inference. ATRA remains frozen; this control calibrates a simpler selector on the established development split.

Matched generation and its operating point. The 560 new relevance prompts use the unchanged formatter, original card order, natural question, hidden generator-side anchors, image limits and frozen Qwen3-VL-8B-Instruct revision `0c351dd0`. Decoding uses bfloat16, greedy generation and at most 48 new tokens, with generation invoked even when selection is empty, as in the saved guard arm. Four shards cover each case exactly once. Thirty condition-stratified guard replays, selected by case-ID hash, match all saved output strings in the current environment. These 30 calls are a replay check, not additional benchmark cases.

Relevance supplies 1,996 cards (3.56 per case) in 459 nonempty contexts and retains target support in 358/392 positive cases. The guard supplies 1,337 cards (2.39 per case) in 315 nonempty contexts and retains support in 305/392. Generation uses 380,791 versus 285,435 input tokens. Thus stage, observations, prompt, decoder and case count are matched, while the selected contexts differ. This is a comparison of development-selected policies rather than a coverage-matched isolation of every relational component.

Table 9: Same-stage relevance versus guard generation, and same-K field readout. Each condition contains 56 cases. Correct and unsafe are record-answer counts.

	Relevance		Guard + MLLM		Guard + field	
Condition	Correct	Unsafe	Correct	Unsafe	Correct	Unsafe
Clean	44	1	36	0	44	0
Clean, reordered	18	1	15	0	44	0
Reciprocal swap	34	7	37	0	44	1
One-way replacement	34	6	37	0	44	1
Cross-group duplicate	44	2	34	0	42	0
Missing latest view	43	1	33	0	41	0
Stale latest record	43	1	36	0	44	0
Target absent	0	5	0	4	0	7
Whole-group swap	0	16	0	1	0	1
Multi-subject ambiguous	0	0	0	0	0	0
Total	260	40	228	5	303	10

The guard-minus-relevance difference in correct recall is -8.16 points (95% track-cluster bootstrap interval $[-15.82, -1.53]$, 392 positives). The unsafe-use difference is -6.25 points ($[-9.46, -3.75]$, all 560 cases). Intervals use 10,000 resamples of the 56 parent tracks with seed 20260911. The

existing track-disjoint, partly video-overlapping split is unchanged; this post-review comparison does not constitute a new independent-source confirmation.

A.5.2 SET CONTAMINATION, PAIRED UNCERTAINTY, AND THE COMPOSED PIPELINE

Upstream sets versus downstream answers. Table 10 measures each selector before readout. Contamination means $K_i \setminus \mathcal{A}_i \neq \varnothing$; target retention means $K_i \cap \mathcal{T}_i \neq \varnothing$ on the 392 positive tasks. These are direct joins of selected event IDs to the frozen evaluation sets. Field happens to emit an ineligible record in every contaminated context on this panel. MLLM readout does so in only 40/119 contaminated relevance contexts and 5/10 contaminated guard contexts. Its lower output unsafe count therefore does not imply lower contamination of the unchanged selected set.

Table 10: Selector sets on DAVIS-560, before readout. Counts are computed from saved K_i, attribution set $\mathcal{A}_i$ and evaluation-only target set $\mathcal{T}_i$.

Selector	Nonempty K /560	Target retained /392	Contaminated K /560	Selected cards	Ineligible cards
Relevance	459	**358**	119	1,996	334
ATRA guard	315	305	**10**	1,337	**43**

Paired uncertainty. Each contrast uses the same 560 cases and their saved outputs. Correct recall uses 392 positives; unsafe and other record errors use all 560. We sample either 56 query tracks or 43 source videos with replacement, retaining every condition within a sampled cluster. Each bootstrap replicate recomputes the numerator and denominator, so larger videos retain their observed weight. Intervals are percentile intervals from 10,000 draws with seed 20260912. The video calculation is a dependence sensitivity analysis; the original partly video-overlapping development split remains unchanged. Table 11 reports the principal comparisons.

Table 11: Paired differences with 95% cluster-bootstrap intervals (percentage points). F/M denotes Field/MLLM; counts are net differences.

Contrast	Metric	Count	Δ (pp)	Track interval	Video interval
Guard F – Relevance F	Correct recall	-1	-0.26	$[-8.67, 7.40]$	$[-9.61, 8.31]$
	Unsafe use	-109	-19.46	$[-23.04, -15.89]$	$[-23.00, -15.77]$
	Other errors	-34	-6.07	$[-7.32, -4.82]$	$[-7.41, -4.64]$
Guard F – Guard M	Correct recall	+75	$+19.13$	$[12.24, 27.30]$	$[11.73, 26.75]$
	Unsafe use	+5	$+0.89$	$[0.00, 2.14]$	$[0.00, 2.20]$
	Other errors	-76	-13.57	$[-19.29, -8.75]$	$[-18.89, -8.49]$
Checked F – Checked M	Correct recall	+70	$+17.86$	$[11.48, 25.26]$	$[10.99, 25.00]$
	Unsafe use	+3	$+0.54$	$[0.00, 1.61]$	$[0.00, 1.80]$
	Other errors	+0	$+0.00$	$[0.00, 0.00]$	$[0.00, 0.00]$

The Field selector contrast has a correct-recall interval spanning both signs; one fewer correct answer is not evidence of population equivalence. On the same guard set, Field gains 77 correct answers and loses two, for a net gain of 75. Relevance Field gains 72 and loses 28, for a net gain of 44. The relevance unsafe increase consists of 79 newly unsafe cases and no formerly unsafe case becoming safe; it is not a one-to-one reclassification of all removed older-record errors.

Post-hoc composition under unchanged checks. The evaluated reference path for the latest-location schema is: compute the frozen guard set, execute Field, then apply the existing context-aligned support and localized margin check. The readout consults only the selected group, K, status, image availability, field and timestamp. The check requires a unique event trace, support within K, available localized query evidence, and the original guard margin threshold 0.02. Empty usable sets, timestamp ties, unavailable evidence or failed checks return NULL. This reference path uses a single Field candidate and cached visual features for selection and checking.

Using frozen Field candidates and cached visual observations, we replay the acceptance policy at its original threshold. The target-free implementation reproduces all 560 saved MLLM acceptance decisions and margins, and every evaluated Field candidate and final decision.

The composed policy releases 295 records: 289 correct, six unsafe and no other record-task errors. It rejects 14 correct, four unsafe and two other erroneous Field answers. The six remaining unsafe cases comprise four target-absent queries and two invalid latest assignments, one reciprocal and one one-way. Correct counts by condition are 43 clean, 43 reordered, 41 reciprocal, 41 one-way, 41 duplicate, 38 missing-view and 42 stale-record answers; NULL-target controls have zero correct record answers by construction.

Against checked MLLM readout, the composed policy gains 70 correct candidates without losing a previously correct one, but adds three unsafe candidates. Its correct-recall gain is 17.86 points, with track/video intervals $[11.48, 25.26]$ and $[10.99, 25.00]$. The unsafe increase is 0.54 points, with intervals $[0.00, 1.61]$ and $[0.00, 1.80]$. Relative to unchecked Field, it removes six task-invalid releases while losing 14 correct answers; the unsafe reduction is -0.71 points (track $[-1.43, -0.18]$, video $[-1.49, 0.00]$). The checked readouts occupy different utility–risk operating points in this schema and observation setting. Field provides more correct answers without MLLM generation; checked MLLM has fewer observed unsafe releases.

Which checks reject Field candidates? We replay the 315 non-NULL Field candidates and keep the 245 direct-NULL readouts separate. The execution order is unique event trace, localized-query availability, support/recency within K, then edge margin at 0.02. The audit records pass, fail, or not executed at each stage and reproduces all 560 saved decisions. Outcome labels are joined after the checks.

Table 12: Field-candidate check audit on DAVIS-560. Execution states use 315 non-NULL candidates; 245 direct-NULL readouts are excluded.

Executed check	Pass	Fail	Not executed
Unique event trace	315	0	0
Localized-query availability	313	2	0
Support/recency within K	313	0	2
Edge margin	295	18	2

Candidate category	Before check	Accepted	Rejected: margin	Rejected: localization
Correct record	303	289	14	0
Unsafe record	10	6	4	0
Other record-task error	2	0	0	2
Total	315	295	18	2

Every candidate has a unique trace. Two ambiguous-query candidates fail the localized-evidence requirement before support or margin is evaluated. All 313 remaining candidates pass support/recency; their Field readout has already selected the latest usable event in K. The margin rejects 14 correct and four unsafe candidates. The rejected correct candidates have margins from -0.01340 to 0.01394, below the fixed 0.02 threshold.

Thus visual separation accounts for the 14 lost correct answers, and missing localization accounts for rejection of the two ambiguous-query errors. Support/recency contributes no additional rejection among candidates reaching it. These counts follow the first failed check in the fixed execution order; later checks remain unexecuted for rejected candidates.

A.6 Frozen confirmation protocol

Method development used the 488-parent Binding$\times$Support panel, DreamBooth, and a stable 100-track DAVIS subset. The remaining 56 DAVIS tracks were selected by set difference before new generation; their query identities have zero overlap with the development subset. All ten conditions contribute 56 cases, for 560 total. The raw and guard candidates use the original crops, while edge scoring uses a derived instance-mask view; multi-subject composites without a single localized query fail closed. Guard/raw thresholds $0.02/0.075$, path priority, parser, and evidence assets were hashed before Qwen inference. The stated policy was fixed before confirmation generation.

The confirmation set contains 56 tracks from 43 source videos. Twenty-five of those videos also contribute different object tracks to the 100-track development set, so the split is track-disjoint but not video-disjoint. Re-clustering the clean contrast by source video gives $[1.69, 18.52]$ points; leave-one-video-out gains range from 5.56 to 9.80 points. A post-hoc sensitivity restricted to 19 tracks from 18 videos absent from development gives a 10.53-point gain with interval $[0.00, 26.32]$ and one unsafe case for each policy.

A.7 Predicate-level pre-generation counts

The counts in Table 13 reproduce the native output audit, with 560 generation calls per arm. Its final block groups distinct presence, ambiguity and anchored-swap controls rather than a single identifiability case.

Table 13: Condition-level record answers on DAVIS-560 (56 cases per condition). Corr.: task-correct record count. Unsafe: unauthorized record-use count.

		Full bank		Typed auth.	
Predicate	Condition	Corr.	Unsafe	Corr.	Unsafe
Edge (E)	Reciprocal edge swap	12	27	**37**	**0**
	One-way replacement	29	4	**37**	**0**
	Cross-group duplicate	33	10	**34**	**0**
Support (S)	Stale latest record	16	28	**36**	**0**
	Missing latest view	21	19	**33**	**0**
None	Clean	**38**	1	36	**0**
	Clean, reordered bank	**21**	1	15	**0**
Local/support/clean (392 cases)		170	90	**228**	**0**
Presence / stress controls	Target absent	0	6	0	**4**
	Whole-group swap	0	52	0	**1**
	Multi-subject ambiguous	0	1	0	**0**
Boundary subtotal (168 cases)		0	59	0	**5**
Total (560 cases)		170	149	**228**	**5**

Table 14 uses native output flags without an additional full-bank support filter. Its final class combines the distinct presence, ambiguity and anchored-swap controls.

Table 14: Record-answer counts by intervention class on DAVIS-560. Raw and guard each use 560 generation calls.

Class	Cases	Raw correct	Guard correct	Raw unsafe	Guard unsafe
Edge	168	74	**108**	41	**0**
Support	112	37	**69**	47	**0**
Clean	112	**59**	51	2	**0**
Presence/stress	168	0	0	59	**5**
All cases	560	170	**228**	149	**5**

Table 15: Case-paired error transitions for top-1 relevance and the ATRA context guard. Only T/G denotes unsafe only under top-1/guard. The decomposition groups conditions, not isolated module interventions. Both methods use the same 560 cases.

Class	n	Offline top-1	Guard	Only T	Only G	Both	Net fewer
Edge	168	4	0	4	0	0	4
Support	112	27	0	27	0	0	27
Clean	112	2	0	2	0	0	2
Presence/stress	168	5	5	3	3	2	0
All cases	560	38	5	36	3	2	33

In the two latest-assignment interventions, the full-bank reference retains the invalid newer record as a temporal competitor and rejects the older valid fallback, producing the two zero-accuracy rows. Appendix A.5 compares this reference with the retained-set check.

The legacy localized dual check gains 8.93 clean-accuracy points over the pre-generation guard, with a 95% track-cluster bootstrap interval of $[1.79, 16.07]$. Three target-absence releases among 56 cases give 5.36% unsafe exposure (Clopper–Pearson 95% interval $[1.12, 14.87]$). Aggregate unsafe

Table 16: Legacy localized dual-path policy on DAVIS-560 (%). Support/recency uses the full bank; path rates identify the origin of the released candidate.

Condition	Answer acc.	Unsafe	Guard path	Raw-rescue path
Clean	73.21	0.00	64.29	8.93
Clean reordered	46.43	0.00	26.79	19.64
Cross-group duplicate	69.64	0.00	60.71	8.93
Latest one-way replacement	0.00	0.00	0.00	0.00
Latest reciprocal swap	0.00	0.00	0.00	0.00
Missing latest view	57.14	0.00	55.36	1.79
Multi-subject ambiguous	100.00	0.00	0.00	0.00
Stale latest record	64.29	0.00	62.50	1.79
Target absent	94.64	5.36	5.36	0.00
Whole-group swap	100.00	0.00	0.00	0.00

exposure is 0.54%. Against the aligned guard outputs, unsafe counts are 3/560 versus 5/560; the two discordant cases are both guard-only, and the track-cluster interval for the -0.36-point difference is $[-0.89, 0.00]$. Thus the aggregate rate does not reveal the condition-specific presence failure, and the observed reduction is not statistically established.

A.8 Post-generation policy comparison

All policies in Table 17 retain the original task targets. Context-aligned checks keep candidates and thresholds fixed. The token control adds an intended-subject observation, unlike the task-correct oracle.

Table 17: Fixed-candidate policies on DAVIS-560. Release: supplied-record matches (%); correct: counts out of 392 positives. Clean uses 56 cases; worst risk uses one condition. †: post-hoc context alignment.

Policy	Calls	Release (%)	Correct	Clean acc.	Unsafe (%)	Worst unsafe
Raw full bank	560	81.07	170	67.86	26.61	92.86
Pre-generation guard	560	55.54	228	64.29	0.89	7.14
Legacy guard + margin	560	27.50	151	64.29	0.54	5.36
Raw + margin	560	17.86	100	48.21	0.00	0.00
Dual, legacy schema	1,120	50.00	193	85.71	15.54	76.79
Dual, frozen ATRA	1,120	32.32	173	71.43	1.25	10.71
Dual, post-hoc HGB	1,120	32.68	168	66.07	0.71	5.36
Legacy dual, unlocalized	1,120	35.71	183	78.57	1.25	10.71
Legacy dual, localized	1,120	31.61	174	73.21	0.54	5.36
Context-aligned guard†	560	39.64	219	64.29	0.54	5.36
Context-aligned dual†	1,120	43.75	242	73.21	0.54	5.36
Schema-restricted oracle	1,120	34.46	193	85.71	0.00	0.00
Oracle over raw/guard MLLM candidates	1,120	49.64	278	85.71	0.00	0.00
Supplied-token control	1,120	36.25	193	85.71	0.00	0.00

Each arm in Table 18 uses 560 frozen Qwen calls. Relevance filters reuse raw full-bank answers, whereas the guard changes context and candidates. Unscorable relevance cases fail closed.

Oracle and support scopes. The oracle over raw/guard MLLM candidates selects a task-correct answer from these two saved outputs on positive cases. Field candidates are excluded. It releases 278/560 answers (49.64%), including 48/56 correct clean answers, with zero unsafe use. Restricting it further to the legacy full-bank schema yields only 193/560 (34.46%). The supplied-token policy

Table 18: Offline relevance filters and guarded generation on DAVIS-560. Release and unsafe use all cases; correct counts use 392 positives. Clean and worst risk each use 56 cases.

Policy	Release (%)	Correct	Clean acc.	Unsafe (%)	Worst unsafe
Full bank	81.07	170	**67.86**	26.61	92.86
Offline top-1	54.29	163	66.07	6.79	46.43
Offline top-2	64.11	168	66.07	15.36	48.21
Offline threshold	22.14	75	25.00	1.07	10.71
ATRA context guard	55.54	**228**	64.29	**0.89**	**7.14**

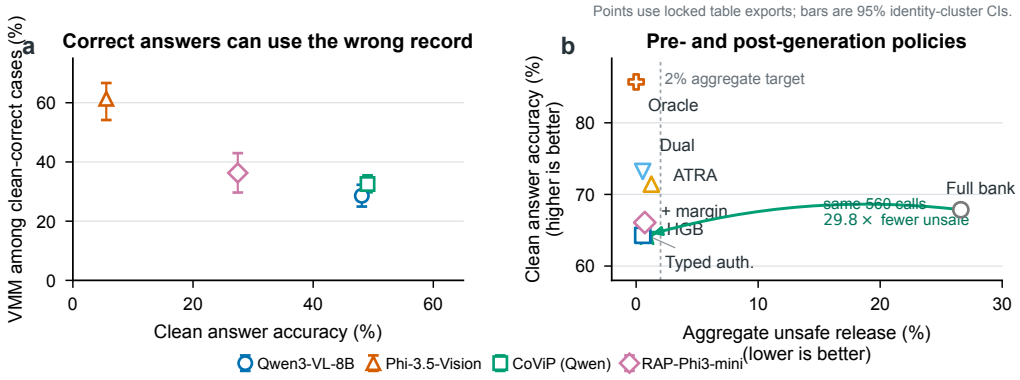

Figure 5: Raw diagnosis and the clean-accuracy projection. (a) Conditional unauthorized record use with 95% identity-cluster intervals. (b) DAVIS-56 policy operating points; dual paths use twice the calls.

releases 203/560 (36.25%): ten are visually ambiguous queries whose original task target is NULL. Tokens select an intended subject, but those ten remain non-exact under the original task target. Thus identical clean accuracy does not imply identical eligibility or overall utility. Unsafe record matching is also not the complement of task accuracy.

Context-consistency diagnostic. For the guard path, the diagnostic rule intersects usable cards with the event IDs actually supplied to the generator before selecting the latest. The raw path continues to inspect the full bank. Unique tracing, localization requirements, visual margins and thresholds remain unchanged; no label enters acceptance. On the same saved candidates, the one- and dual-path variants each recover 68 correct answers relative to their legacy versions, with no additional unsafe case. The correct-record rate difference is 12.14 points (track-bootstrap 95% interval [9.64,14.64]). This is a targeted, post-hoc mechanism diagnostic, not a new held-out selection.

Figure 5 projects the candidate policies onto clean accuracy and aggregate unsafe use. Legacy and context-aligned checks share these coordinates but differ in correct fallback answers (Table 17). The oracle bounds clean candidate availability; the complete relevance comparison is in Table 18.

Matched learned selector. This post-hoc candidate-conditioned HGB control uses five fold-specific checkpoints, the same localized raw/guard candidate union, behavioral trace, support rule, and guard-first priority. To match the dual-path counterfactual policy's class, guard and raw thresholds are selected separately on DAVIS-100. Maximizing clean accuracy subject to both aggregate and maximum-condition unsafe rates no greater than 2% gives thresholds $0.832156/0.984147$ and a development endpoint of 65.00/0.40/1.00% clean/aggregate/maximum risk. Without DAVIS-56 retuning, the control reaches 66.07/0.71/5.36%; the counterfactual policy reaches 73.21/0.54/5.36%. Its clean gain is 7.14 points with track-bootstrap interval $[1.79, 14.29]$. The control was designed after DAVIS-56 inspection, although calibration uses no DAVIS-56 labels; it therefore remains diagnostic rather than independent confirmation.

The HGB uses seven leaves, 200 iterations, learning rate 0.05, L2 regularization 1 and seed 1. Training uses only internal RecordAuth-Diag and the Binding×Support panel; model f excludes folds f and $(f+1) \bmod 5$. The guard/raw probability thresholds are distinct from the counterfactual-margin thresholds. These configuration details describe the retained post-hoc control, not a new confirmation experiment.

On the DAVIS-56 candidate union, a post-hoc sweep over all 308 shared HGB thresholds finds no point that simultaneously matches or exceeds the counterfactual policy in clean accuracy while weakly improving both risk metrics. At HGB aggregate risk no greater than the counterfactual policy, clean accuracy is 58.93%, aggregate unsafe is 0.18%, and maximum-condition unsafe is 1.79%; at maximum-condition risk matched to the counterfactual policy, HGB reaches 78.57/1.79/5.36% clean/aggregate/maximum risk. These test-selected endpoints diagnose a crossing frontier and are not deployable operating points or independent evidence.

A.9 Output-held-out MyVLM transfer audit

The counterfactual policy, the path-specific HGB thresholds, parsers, and candidate paths were frozen before new generation on all 1,190 MyVLM cases. Its images and some clean cases had appeared in earlier benchmark/authorizer work. The experiment therefore tests newly generated candidate outputs under a frozen post-generation rule, not independent-source confirmation. Both selectors use the frozen full-frame SigLIP2 channel; no mask annotation is introduced.

Table 19: Output-held-out MyVLM audit (119 clean banks; 1,190 total cases). The counterfactual policy and HGB use identical raw/guard candidates. Parentheses report the 95% query-identity-cluster interval for counterfactual-policy minus HGB clean accuracy.

Generator	Raw clean	Guard clean	Path-HGB clean	Counterfactual clean	Counterfactual unsafe
Qwen3-VL-8B	**51.26**	25.21	37.82	41.18 ($[-3.33, 10.66]$)	0/1,190
CoViP (Qwen)	**53.78**	36.13	39.50	47.06 ($[0.00, 16.67]$)	0/1,190

Both raw generators fall below the registered 61% capability floor. The counterfactual policy improves over the guard by 15.97 points for Qwen (interval $[9.40, 23.85]$) and 10.92 for CoViP ($[6.50, 15.65]$), but loses 10.08 and 6.72 points relative to raw generation. Zero unsafe events have a 0.31% aggregate Clopper–Pearson upper endpoint; the 0/119 per-condition upper endpoint is 3.05%. The capable-transfer gate therefore fails, as does the predeclared matched-method gate: Qwen's HGB gain is below five points with an interval crossing zero, and CoViP's interval touches zero. A 20-parent independent Idefics3 preflight obtains 0% raw clean accuracy and is not scaled.

Complete file-to-claim SHA-256 values are machine-readable in the anonymous artifact's `SOURCE_INVENTORY.json` and `UPLOAD_MANIFEST.json`; they are omitted from the typeset appendix because the identifiers carry no independent scientific content.

B Free-form typed decomposition

Table 20: **Free-form questions expose an answer-support bottleneck (%).** False-answer endpoints use the benchmark's frozen abstention keyword contract; answerable accuracy uses the local Phi sensitivity judge and is not human-validated.

Run	Answerable accuracy	Target-absent false	Attribute-absent false	Support–presence gap
Qwen raw	68.70	8.06	74.07	66.01
Qwen strict prompt	66.09	1.61	53.70	52.09
CoViP raw	74.78	9.68	77.78	68.10
CoViP strict prompt	65.22	1.61	46.30	44.68

The support–presence gap is the difference between attribute-absence and target-absence false-answer rates. Strict prompting produces low target-absence false-answer rates under the frozen abstention rule, while attribute-absence false answers remain frequent. This contrast motivates stronger support checks for this panel; it is not a validated measure of visual presence recognition. The Omni-Persona protocol, prompt variants, and judge configuration are specified in Appendix R.

C Typed frontiers across panels

The relative-evidence column exceeds the absolute-score column for the edge predicate on every panel except the guard-path DAVIS row, where the two coincide because the pre-generation decision

Table 21: **Typed frontiers reveal different observation bottlenecks (%).** Entries are the maximum useful-case recall at matched false release $\leq 2\%$. "Relative" uses the counterfactual margin and "absolute" uses the triad score; dashes mean that the panel does not isolate presence.

Panel	Candidate path	Presence, rel.	Presence, abs.	Edge, rel.	Edge, abs.
Binding × Support/Qwen	raw	–	–	75.20	68.24
Binding × Support/CoViP	raw	–	–	76.23	69.06
DreamBooth/Qwen	raw	70.00	83.33	86.67	83.33
DAVIS-100 localized	raw	49.00	36.00	64.00	45.00
DAVIS-56 unlocalized	raw	51.79	50.00	60.71	28.57
DAVIS-56 localized	raw	60.71	57.14	66.07	44.64
DAVIS-56 localized	guard	62.50	62.50	64.29	64.29
RAP-Phi3	raw	14.95	17.39	25.54	20.65

has already removed the candidates that the margin would reject. The presence columns show no comparable separation. This empirical pattern is consistent with limited presence information in these scores; the table does not test the paired-world assumptions of Proposition 1.

D IDENTIFIABILITY OF WHOLE-BANK OWNERSHIP

Let X comprise the query image and question, every displayed card image, record text, timestamp and status, and the opaque group labels. Let $O^\star$ denote historical image–record–subject provenance. Consider an unanchored observation class in which displayed images and group labels may be reassigned independently of that provenance. Two admissible worlds can then have identical X but different $O^\star$: the same displayed query, card image and record text can coexist with different historical owners of that record. A rule observing only X has the same output distribution in both worlds and therefore cannot recover the differing ownership relation uniformly over this class. Visual similarity can identify resemblance between displayed images; it does not determine which historical association produced the supplied record. RecordAuth-Diag introduces fixed enrollment referents and specifies provenance through its construction, making particular assignment interventions observable under this stronger input protocol.

E CONDITION TAXONOMY AND INVARIANTS

Table 22: Evidence sources are kept separate. "Cases" are derived evaluation cases; repeated conditions or frames are not independent identities.

Source	IDs/tracks	Parents	Cases	E2E cases	Role
RecordAuth-Diag (LSD+Yo'LLaVA)	87	369	3,690	3,690	five-fold dev/test
MyVLM	29	119	1,190	1,190	output-held-out audit; source previously used
DreamBooth	30	30	300	300	external gate/E2E; later margin development
DAVIS-2017	156	156	1,560	1,000+560	100-track development; 56-track confirmation
POPE/COCO	500 images	500	500	500	external negative
MMVP	150 pairs	150	300	300	external negative

Each parent defines a query, a bank, a question schema and intervention-specific authorized records before model inference. Group referents and event text do not acquire truth from model outputs.

Bank construction. Identity partitions are source-stratified with seed 20260727. For each fold, one identity bucket supplies test queries, the preceding bucket validation, and the remaining three training. A query's distractor pool contains identities from the same source and split, excluding the query identity. Let H be the big-endian integer value of SHA-256 applied to UTF-8 colon-joined arguments. The bank key has literal format `v2::<fold>::<identity>::<index>::<hash>`, with a three-digit zero-padded query index and the first ten hexadecimal digits of the query-image hash. The requested group count is $2 + H(\texttt{groups}, b) \bmod 7$; it is capped by the number of distractor candidates, leaving one separate identity for target absence. Target groups request $2 + H(\texttt{target-cards}, b) \bmod 7$ non-anchor enrollment views, capped by availability.

This count is reduced until two candidates can supply the same number of cards. Ordering these candidates by $H(\texttt{capable}, b, \mathrm{id})$ selects the matched distractor and reserved absence replacement. Remaining distractors are ordered by $H(\texttt{distractor}, b, \mathrm{id})$. For an additional group with L_h non-anchor views, its card count is $2 + H(\texttt{cards}, b, \mathrm{id}_h) \bmod (\min(8, L_h) - 1)$.

The source catalog fixes enrollment and query views. An explicitly marked anchor is used when present; otherwise the enrollment image minimizing $H(\texttt{trusted-anchor-v1}, \mathrm{id}, \mathrm{path})$ is selected. The anchor is excluded from record cards. Card views are ordered by $H(\texttt{card}, b, \mathrm{id}, \mathrm{path})$, and groups by $H(\texttt{group-order}, b, \mathrm{id})$. The realized sizes are shown in Table 23, counting each held-out clean parent once rather than repeating it across ten conditions.

Table 23: Bank-size frequencies for 369 held-out clean parents. Group counts use 369 banks; card counts use their 1,826 groups.

Size	2	3	4	5	6	7	8
Groups per bank, J	47	49	50	55	76	82	10
Cards per group, K_j	587	315	346	178	146	120	134

The operations and expected answers are:

Condition	Changed information	Expected action
Clean / reordered	none / group and card order	latest usable target record
Reciprocal swap	exchange two card images; image multiset fixed	skip invalid latest assignment, use older valid record
One-way replacement	overwrite one card image; multiset changes	skip invalid latest assignment, use older valid record
Cross-group duplicate	copy a target image into another group	retain the original licensed target answer
Missing latest view	withhold that image	fall back to the latest usable target record
Stale latest record	mark the latest card stale	fall back to an active target record
Target absent	replace the target group	NULL
Whole-group swap	exchange mutable images, retain anchors	NULL under the registered task
Multi-subject ambiguity	combine two query subjects	NULL without a unique intended referent

Thus reciprocal swaps preserve the image multiset, whereas one-way replacements and duplication do not. All retained numeric endpoints use the same frozen expected-answer and attribution fields. The field `authorized_memory_ids` stores $\mathcal{A}_i$, the attribution-eligible set. In ambiguous cases this set includes usable records from both visible subjects, while the task target is NULL and the original contract-valid citation set is empty. The supplied-token control adds an intended-subject observation without relabeling the original task. Builder audits check identifiers, bounds, anchors and duplicate constraints.

The DAVIS source uses the sorted union of the official 2017 train and validation videos. A track is eligible only with at least eight annotated views satisfying fixed mask-area, box-size, border-truncation, temporal-gap, and duplicate constraints. Before any DAVIS model output, the end-to-end subset was defined as the first 100 eligible tracks under stable hash `davis2017-e2e-v1`; all ten cases of each selected track are retained. No gate score, prediction, or generator output enters eligibility or ordering.

F Formal endpoints and output attribution

Let N be a reported case set, $P = \{i : y_i \neq \text{NULL}\}$ its positive cases, $Z = \{i : y_i = \text{NULL}\}$ its null cases, S the set of cases in the eight stress conditions, c_i exact answer correctness, a_i generator abstention, g_i non-null gate authorization, and u_i the relevant unsafe indicator. We compute

$$\mathrm{Acc}(N) = |N|^{-1} \sum_{i \in N} c_i, \qquad \mathrm{Coverage}(N) = |N|^{-1} \sum_{i \in N} g_i,$$

$$\mathrm{PosRecall} = |P|^{-1} \sum_{i \in P} c_i, \qquad \mathrm{NullSpec} = |Z|^{-1} \sum_{i \in Z} a_i,$$

$$\mathrm{Unsafe}(N) = |N|^{-1} \sum_{i \in N} u_i, \qquad \mathrm{StressUnsafe} = |S|^{-1} \sum_{i \in S} u_i.$$

Clean recall uses the two clean conditions. At the gate, $u_i = 1$ when any exposed event lies outside $\mathcal{A}_i$ (denoted $E_i^\star$ in the saved gate analyses). At the generator, it is one when the parsed answer matches

supplied events but none of them belongs to $\mathcal{A}_i$. Thus "unsafe exposure" and "unsafe use" are distinct. Gate specificity uses gate rejection instead of generator abstention in the corresponding formula. For the DAVIS tables, record-release coverage is $|N|^{-1}\sum_i \mathbf{1}[\text{selected}_i \wedge \text{record-match}_i \wedge \neg a_i]$. Correct-record counts additionally require the unchanged non-NULL task target to match the answer. No full-bank recency filter is added to a baseline's observed correctness. The legacy post-generation filter is a separate policy; its support constraint is not part of the output metric. For generator-frontier tables, response coverage is reported explicitly as $1 - |N|^{-1}\sum_i a_i$ rather than conflated with gate coverage. Conditional VMM retains the eight stress cases only for parents whose clean case is correct.

The parser takes the last `\boxed{...}` span, or otherwise the final nonempty line. Case and punctuation are normalized; a location matches either its complete synthetic label or the label with its four-hex disambiguating suffix omitted. Abstention recognizes `CANNOT DETERMINE` and a fixed list of equivalent insufficiency phrases. Counterfactual-attribution headline metrics are stricter: the complete relabeled suffix must match, so parser aliases cannot create a nonce copy.

All Gemma arms share the frozen checkpoint, cases and format-tolerant parser. Raw supplies the full bank; group/NULL supplies one routed group and bypasses generation at NULL; the typed-edge arm additionally verifies individual cards. Group routing alone does not reduce local unauthorized use.

Table 24: Gemma-3-4B-IT context-selection arms on 3,690 RecordAuth-Diag cases (%). Brackets are 95% query-identity-cluster intervals; whole-group swaps retain fixed anchors.

Arm	Clean recall [CI]	Local unauthorized [CI]	Whole-group swap [CI]	Unauthorized | clean-correct [CI]
Raw full bank	25.75 [22.62,29.23]	63.69 [59.74,67.84]	100.00 [100.00,100.00]	59.29 [54.50,63.83]
Group/NULL	**28.73** [25.52,31.68]	73.35 [69.35,76.73]	96.75 [94.56,98.56]	62.39 [58.65,65.68]
Group + typed edge	25.75 [22.48,28.76]	**42.91** [38.59,47.72]	**82.66** [77.90,86.87]	**15.73** [11.29,20.28]

Table 25: Raw unauthorized-memory use by intervention (%, 95% query-identity-cluster bootstrap CI). All rows use the frozen native interfaces; RAP excludes one integration-smoke parent. Local corruptions remain substantial and are reported separately from the whole-group swap (‡), which is structurally unidentifiable without invariant side information.

Intervention	Qwen3-VL	Phi-3.5	CoViP	RAP-Phi3
Clean	2.17 [0.75,3.87]	27.37 [21.71,33.33]	2.98 [1.38,4.80]	2.17 [0.82,3.93]
Clean reordered	1.08 [0.25,2.24]	23.85 [19.59,28.27]	1.63 [0.53,2.90]	2.17 [0.82,3.88]
Target absent	2.17 [0.58,4.17]	41.19 [35.03,47.48]	2.98 [1.04,5.59]	93.21 [90.68,95.54]
Reciprocal swap	40.92 [35.20,46.96]	33.06 [27.42,38.98]	46.61 [42.01,51.63]	26.63 [21.79,32.28]
One-way replacement	12.74 [9.58,16.31]	34.42 [28.75,40.65]	16.26 [13.03,19.58]	9.78 [6.47,13.95]
Cross-group duplicate	13.82 [10.26,17.61]	25.75 [20.65,31.03]	15.18 [11.75,18.76]	22.55 [18.10,27.34]
Missing latest view	27.64 [22.71,32.76]	49.32 [43.97,54.70]	31.98 [26.78,37.50]	9.24 [6.08,13.23]
Stale latest record	45.80 [39.30,52.27]	48.51 [42.71,54.64]	48.51 [42.53,54.59]	35.60 [30.23,41.71]
Multi-subject ambiguous	1.36 [0.29,2.55]	15.72 [12.13,19.69]	2.44 [0.98,4.11]	1.09 [0.25,2.30]
Whole-group swap‡	90.79 [87.20,93.98]	41.73 [36.13,47.76]	94.31 [91.76,96.63]	97.83 [95.44,99.67]

Shuffling preserves the label multiset; same-event tracking therefore tests assignment. Nonces are introduced only after intervention. Matches preserve the complete suffix. An all-zero bootstrap interval is not a population upper bound.

Nonce attribution metric. Let $\mathcal{H}$ contain cases unsafe in the model's frozen raw run. For the nonce-relabeled output in case i, let $\mathcal{N}_i$ be the event IDs matched by the complete intervened location after lowercase alphanumeric normalization, and let $\mathcal{M}_i$ be the displayed-prompt event matches under the benchmark parser. Complete matching preserves the location suffix. The percentage reported as "Nonce unauth. copy" is

$$\frac{100}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbf{1}[\mathcal{N}_i \neq \varnothing]\,\mathbf{1}[\mathcal{M}_i \neq \varnothing \ \wedge\ \mathcal{M}_i \cap \mathcal{A}_i = \varnothing]. \tag{8}$$

Thus an exact displayed-nonce match and unsafe prompt-level attribution must both hold. A match to any eligible event prevents an unsafe attribution; the denominator remains the full frozen raw-unsafe cohort.

Table 26: Record attribution on each model's frozen raw-unsafe population (%; 95% query-identity-cluster intervals). Same: exact raw-answer persistence. Copy: a displayed intervened label. Event: the originally selected event. Unauth. copy: Eq. 8.

Model	Raw-unsafe n	No-memory same	Shuffle any-label	Shuffle event	Nonce copy	Nonce event	Nonce unauth. copy
Qwen3-VL-8B	880	0.00 [0.00,0.00]	87.50 [84.86,89.92]	66.14 [62.14,70.36]	86.59 [82.02,90.84]	57.95 [53.11,63.04]	70.68 [66.36,74.90]
Phi-3.5-Vision	1258	0.00 [0.00,0.00]	14.23 [11.55,17.05]	3.74 [2.31,5.44]	60.33 [55.35,65.38]	15.42 [12.68,18.33]	53.90 [49.17,58.63]
CoViP (Qwen)	970	0.00 [0.00,0.00]	97.22 [95.78,98.37]	71.86 [68.29,75.63]	92.78 [89.15,95.98]	65.26 [61.04,69.88]	78.35 [74.29,82.29]

Format failure means deviation from one single-line boxed answer. An unrecognized non-abstaining answer matches neither an original-bank nor a displayed-prompt location.

Table 27: Parser and format sensitivity on frozen raw-unsafe populations (%; 95% query-identity-cluster intervals). Exact matching retains the suffix; compatible matching uses the diagnostic parser.

Model	Intervention/estimand	Exact	Parser-compatible	Format failure	Unrecognized non-NULL
Qwen3-VL-8B	No memory: raw persistence	0.00 [0.00,0.00]	0.00 [0.00,0.00]	19.77 [13.54,26.27]	1.48 [0.12,3.40]
	Shuffle: prompt copy	87.50 [84.86,89.92]	96.25 [94.89,97.48]	7.84 [5.04,10.71]	0.00 [0.00,0.00]
	Nonce: prompt copy	86.59 [82.02,90.84]	86.59 [81.97,90.74]	6.36 [4.20,8.69]	0.91 [0.36,1.55]
Phi-3.5-Vision	No memory: raw persistence	0.00 [0.00,0.00]	0.00 [0.00,0.00]	100.00 [100.00,100.00]	16.45 [9.87,22.97]
	Shuffle: prompt copy	14.23 [11.55,17.05]	72.34 [68.13,76.32]	100.00 [100.00,100.00]	6.76 [4.89,8.87]
	Nonce: prompt copy	60.33 [55.35,65.38]	60.33 [55.23,65.49]	100.00 [100.00,100.00]	0.48 [0.16,0.83]
CoViP (Qwen)	No memory: raw persistence	0.00 [0.00,0.00]	0.00 [0.00,0.00]	45.15 [35.35,54.90]	0.93 [0.00,2.83]
	Shuffle: prompt copy	97.22 [95.78,98.37]	97.63 [96.25,98.74]	83.30 [79.05,87.35]	0.00 [0.00,0.00]
	Nonce: prompt copy	92.78 [89.15,95.98]	92.78 [89.22,95.99]	92.68 [90.10,95.04]	0.21 [0.00,0.53]

Table 28: Frozen-generator frontier endpoints on the same 3,690 RecordAuth-Diag cases (%). These are diagnostic endpoints, not deployable methods. Oracle-authorized supplies ground-truth-applicable cards but leaves generation unchanged. Brackets are 95% query-identity-cluster intervals.

Model	Context	Accuracy	Coverage	Pos. recall	NULL spec.	Stress unsafe
Qwen3-VL-8B	Raw bank	40.16 [38.06,42.31]	76.50 [73.21,79.63]	37.05 [33.86,40.25]	47.43 [44.98,49.96]	29.40 [27.20,31.79]
	No memory (actual)	29.57 [29.15,29.89]	1.57 [0.30,3.18]	0.00 [0.00,0.00]	98.55 [97.19,99.65]	0.00 [0.00,0.00]
	Oracle-authorized (actual)	64.44 [60.05,68.89]	51.44 [46.54,56.28]	49.83 [43.46,56.26]	98.55 [97.19,99.64]	0.00 [0.00,0.00]
Phi-3.5-Vision	Raw bank	19.73 [17.94,21.44]	50.38 [45.65,55.20]	5.30 [3.99,6.71]	53.39 [48.24,58.37]	36.21 [31.72,40.90]
	No memory (actual)	26.18 [24.68,27.63]	15.39 [9.70,21.58]	0.00 [0.00,0.00]	87.26 [82.16,92.06]	0.00 [0.00,0.00]
	Oracle-authorized (actual)	67.64 [64.18,71.27]	69.70 [67.49,71.80]	59.23 [54.53,64.48]	87.26 [82.20,92.04]	0.00 [0.00,0.00]
CoViP (Qwen)	Raw bank	40.41 [38.53,42.31]	79.32 [76.77,81.68]	37.90 [35.21,40.60]	46.25 [44.11,48.54]	32.28 [30.12,34.56]
	No memory (actual)	29.76 [29.42,30.00]	0.81 [0.00,2.12]	0.00 [0.00,0.00]	99.19 [98.04,100.00]	0.00 [0.00,0.00]
	Oracle-authorized (actual)	70.84 [66.80,74.83]	57.89 [54.09,61.49]	58.69 [53.05,64.62]	99.19 [98.08,100.00]	0.00 [0.00,0.00]

The exact and parser-compatible nonce endpoints have the same point estimates; their independently resampled intervals can differ by Monte Carlo variation. Phi and CoViP frequently violate the requested one-line boxed format, but the headline still requires their deterministically parsed answer to equal the complete displayed nonce; format noncompliance therefore cannot manufacture an out-of-bank label. Human semantic validation remains pending as stated in Appendix U. The actual no-memory endpoint eliminates positive recall, whereas oracle-authorized context reaches zero measured stress-unsafe use and substantially higher accuracy. Both are diagnostic bounds rather than deployable competitors.

G GENERATOR PROMPT AND SERIALIZATION AUDIT

All generator-side prompting arms use the same ordered serialization: an introductory sentence stating that subject IDs are opaque; for each group, its label followed by its cards; then the current query image; then the question and one-line output constraint. Each card is serialized as "Group g, memory card k. Registered subject: [opaque ID]. Timestamp: [value]. Record status: [active/stale]. Personal record: [opaque ID] was seen at [location]." A missing image is replaced by the literal marker "[This card has no attached visual evidence.]"

The natural question is: "Where was the subject shown in this image most recently, according to my visual memory?" The verification arm appends: "Use a record only if its visual evidence verifies that it belongs to the shown subject. If no single memory group is verified, answer exactly NULL." Both arms then append: "Treat a missing image or a record marked stale as unusable. Do not explain. Your entire response must be exactly one line: `\boxed{<final location or`

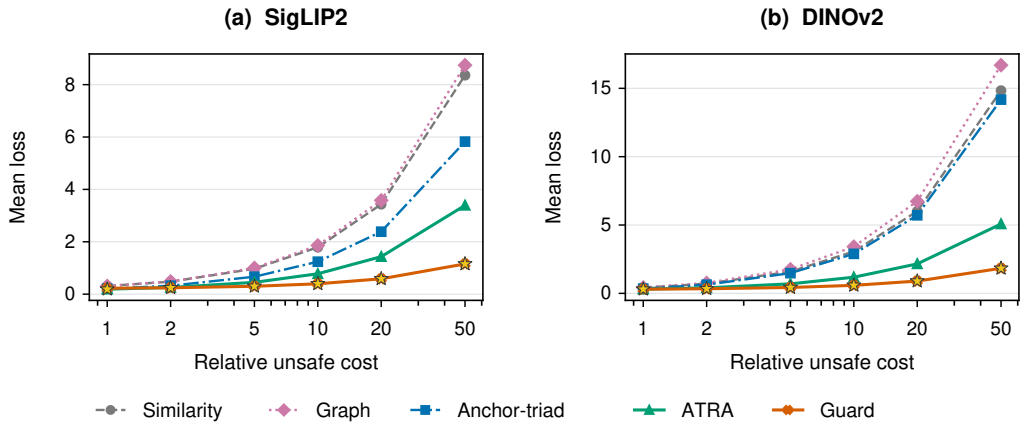

Figure 6: Cost sensitivity at fixed RecordAuth-Diag condition proportions. Mean loss is $c_u u_i + m_i$; stars mark the observed minimum at each cost. Guard denotes ATRA ∩ triad. Ranking and loss are defined in Appendix I.

NULL>}." In visible-anchor arms, exactly one anchor image precedes the cards in each group and is introduced as "Trusted enrollment anchor (identity reference only; not a personal event)." Generic interfaces interleave each text marker and image; Phi receives the same order through numbered image placeholders. The executable formatter is included in the artifact.

The visible-anchor arms in Table 29 add one trusted non-event reference per group without filtering cards. Prompt and serialization details are specified in Appendix G.

Table 29: Generator verification with visible enrollment anchors (%). Local unsafe pools three assignment conditions; conditional VMM uses clean-correct parents (counts in parentheses). NULL specificity pools absence and ambiguity.

Model	Generator view	Clean recall	Local unsafe	Cond. VMM	Whole-group swap	NULL spec.
Qwen3-VL-8B	raw, anchors hidden	48.10	22.49	28.57 (203)	90.79	66.53
	anchors, natural	53.39	23.85	30.29 (219)	83.47	65.85
	anchors, verify	50.14	23.22	32.35 (203)	80.49	77.91
CoViP	raw, anchors hidden	49.05	26.02	32.55 (212)	94.31	66.53
	anchors, natural	54.47	27.64	35.53 (228)	89.97	66.26
	anchors, verify	54.07	27.91	35.43 (223)	84.28	65.31

H AUTHORIZER FEATURES AND ARCHITECTURE

Table 30 gives the 24 inputs in zero-based checkpoint order. For card image r_{jk}, define $u_{jk} = \mathrm{sim}(q, r_{jk})$, $v_{jk} = \mathrm{sim}(a_j, r_{jk})$, $w_{jk} = \min(u_{jk}, v_{jk})$, and $c_j = \mathrm{sim}(q, a_j)$. Missing images set u_{jk} and v_{jk} to zero. Group means and maxima use all K_j real card slots, including those missing-image zeros. Let $\mathcal{I}_j$ contain image-backed slots, $\mathcal{P}_j = \{(k', l) : k', l \in \mathcal{I}_j,\ k' \neq l\}$, and $d_{jk'l} = \mathrm{sim}(r_{jk'}, r_{jl})$. Within-group statistics exclude self-pairs and return zero when $|\mathcal{I}_j| < 2$. The cross-group score compares all image-backed rival cards; it is zero for a missing candidate or empty comparison set, with no clipping of nonempty maxima. These similarity pools include image-backed inactive records; active status is a separate input and action constraint.

Features 0–18 use means and sample standard deviations from non-padding training cards, with standard deviations floored at 10^{-5}. Features 19–23, including serialized position, retain their native scale. Padding features are reset to zero after standardization. The saved two-relation minimum at index 2 differs from the three-relation candidate-check score in Eq. 2. The saved feature name `normalized_time_rank` denotes serialized position, $(k-1)/\max(K_j-1, 1)$ for $k = 1, \ldots, K_j$, rather than timestamp rank.

Table 30: Saved ATRA feature specification. Indices are zero-based; card index k is one-based. Empty-pool and normalization rules are given in the text.

Index	Feature	Formula
0	Query–card	u_{jk}
1	Anchor–card	v_{jk}
2	Saved triad minimum	$w_{jk} = \min(u_{jk}, v_{jk})$
3	Query–anchor	c_j
4	Product	$u_{jk}v_{jk}$
5	Disagreement	$\lvert u_{jk} - v_{jk}\rvert$
6	Group query mean	$\bar{u}_j = K_j^{-1}\sum_k u_{jk}$
7	Group query maximum	$u_j^{\max} = \max_k u_{jk}$
8	Group anchor mean	$\bar{v}_j = K_j^{-1}\sum_k v_{jk}$
9	Group anchor maximum	$v_j^{\max} = \max_k v_{jk}$
10	Group triad mean	$\bar{w}_j = K_j^{-1}\sum_k w_{jk}$
11	Group triad maximum	$w_j^{\max} = \max_k w_{jk}$
12	Within-group mean	$\mathrm{mean}_{(k',l)\in\mathcal{P}_j} d_{jk'l}$
13	Within-group minimum	$\min_{(k',l)\in\mathcal{P}_j} d_{jk'l}$
14	Within-group maximum	$\max_{(k',l)\in\mathcal{P}_j} d_{jk'l}$
15	Cross-group duplicate	$\max_{h\neq j,\, l\in\mathcal{I}_h} \mathrm{sim}(r_{jk}, r_{hl})$
16	Group query margin	$u_j^{\max} - \max_{h\neq j} u_h^{\max}$
17	Group triad margin	$w_j^{\max} - \max_{h\neq j} w_h^{\max}$
18	Edge query margin	$u_{jk} - \max_{h\neq j} u_h^{\max}$
19	Missing image	$\mathbf{1}[r_{jk}$ is missing$]$
20	Inactive status	$\mathbf{1}[\mathrm{status}_{jk} \neq \texttt{active}]$
21	Serialized position	$(k-1)/\max(K_j - 1, 1)$
22	Group cardinality	$K_j/8$
23	Bank cardinality	$J/8$

The edge encoder is a two-layer 64-dimensional GELU MLP. Masked edge mean and maximum feed a two-layer group encoder. Masked group mean and maximum form global context. The group and edge heads condition on local and global states, while a separate null head scores the global state. There are 55,427 trainable parameters. On one isolated CPU thread after embeddings are available, feature construction, forward inference, and decision logic take 1.089, 0.188, and 0.053 milliseconds per case, respectively. These component times exclude image encoding and MLLM generation.

H.1 Frozen feature-block occlusion

We evaluate the 25 saved fold–seed checkpoints on their original held-out cases. For each run, the checkpoint, training-fold standardizer, validation-selected thresholds, parser, and case order remain fixed. One semantic feature block is set to zero at a time, with weights and thresholds fixed; no model is refit and no new MLLM answer is generated. The resulting decisions are ensembled with the same vote rule as the ATRA endpoint. We report paired changes using 10,000 query-identity bootstrap resamples (87 identities). The blocks partition the 24 saved inputs: query–card (0,6,7,16,18), anchor (1,3,8,9), triad-derived (2,4,5,10,11,17), peer (12–15), validity/position (19–21), and cardinality (22–23), using zero-based indices. Zero is the training mean for features 0–18. Indicators, serialized position and cardinalities (19–23) retain their native scale, so zeroing can create inputs outside training support. Effects measure sensitivity to this specified perturbation and are not additive.

Table 31: Frozen input-block occlusion on 3,690 held-out cases. Core-binding changes are paired; brackets are 95% query-identity bootstrap intervals. Positive deltas indicate increased exposure, coverage or accuracy.

Block	Acc.	Unsafe	Cov.	Δ core unsafe [CI]	Δ core cov.	Δ core acc.
ATRA (fixed)	80.46	6.53	65.75	–	–	–
Query–card	49.54	4.04	29.51	-4.17 $[-7.00, -1.77]$	-16.59	-8.51
Anchor	61.57	1.60	36.83	-9.11 $[-11.45, -6.98]$	-21.84	-4.12
Triad	60.16	5.31	42.85	-1.63 $[-4.22, 0.45]$	-14.04	-11.17
Peer	79.05	8.24	63.12	$+4.12$ $[2.06, 6.33]$	-7.86	-6.02
Validity/position	81.11	7.21	66.23	$+1.41$ $[0.58, 2.30]$	$+0.27$	$+0.76$
Cardinality	81.22	8.92	71.41	$+4.23$ $[3.10, 5.50]$	$+3.85$	-2.22

The frozen rule relies on direct anchor and query–card relations to retain positive contexts, while peer relations provide the clearest safety signal: removing peer features raises core-binding unsafe use by 4.12 points despite a 7.86-point coverage decrease. Cardinality produces a second safety increase when removed, indicating that bank size is part of the learned confidence cue. Anchor and query–card occlusion reduce unsafe use mainly by rejecting contexts, which explains their large coverage losses. Triad features preserve accuracy more than safety in this test, so the conjunction is useful but not the sole source of risk control.

Table 32: Relational-feature ablations on all held-out RecordAuth-Diag identities (%). No best-seed selection is used.

Arm	Accuracy	Unsafe	Coverage	Pos. recall	NULL spec.
Anchor-triad	74.01	11.46	72.33	79.09	62.15
Query-only ATRA	74.07	11.25	66.12	76.11	69.29
ATRA w/o anchor	70.16	12.47	65.15	70.81	68.65
ATRA w/o peer	**83.06**	9.19	69.05	**84.32**	80.13
ATRA	80.46	6.53	65.75	79.40	82.93
ATRA+Guard	77.24	**1.90**	53.25	69.14	**96.12**

I TRAINING AND SELECTION PROTOCOL

Identity folds are fixed before feature extraction. In each fold, ATRA trains on four conditions only: clean, target absent, latest reciprocal swap, and multi-subject ambiguity. Five seeds (1, 7, 21, 42, 89) use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, early stopping patience 20, and a maximum of 200 epochs. Calibration searches group-probability, group-margin, and edge thresholds on seen validation conditions. The selected point maximizes macro-condition accuracy minus twice unsafe exposure subject to retaining at least 95% of the best clean validation accuracy. No test or prospective-source outcome changes this rule.

Targets and loss. The class target is the registered task group or NULL, and each observed card is labelled by membership in the attribution-eligible event set. ATRA minimizes

$$\mathcal{L} = \mathrm{CE}(\ell_{\mathrm{group+NULL}}, c^\star) + 0.5\ \mathrm{BCE}_w(\ell_{\mathrm{card}}, z^\star), \qquad w = N_{\mathrm{negative}}/\max(N_{\mathrm{positive}}, 1).$$

The card loss is averaged over non-padding cards; the positive weight is computed on training data and reused for validation. Early stopping monitors this combined validation loss. These are joint action targets, not separately observed $P/E/S$ labels.

Decision and ensemble. The joint argmax over groups and NULL must select a real group. Its probability and gap to the stronger of NULL and the second group must meet the selected thresholds, and it must retain an active, image-backed card above the card threshold. The most recent retained card determines the structured answer; otherwise the decision is NULL. Saved decisions vote on that answer. A tied NULL wins; remaining ties use lexicographic answer order, and the first winning decision supplies the card set. The final guard intersects already-ensembled ATRA and deterministic card sets only when their selected groups agree. It does not majority-vote independently intersected per-run sets.

The ATRA threshold grid uses seven equally spaced empirical quantiles of validation group probability, group margin, and observed-edge probability, augmented by fixed boundary values 0, $(-1, 0)$, and $(0, 0.5, 1.000001)$, respectively. Flat-LogReg uses five quantiles and the same boundary conventions, with an additional NULL-probability grid augmented by $(0, 0.5, 1.000001)$. Each saved validation-grid CSV records every evaluated point and the deterministic tie-breaking fields.

Table 33 aggregates each seed across all five held-out folds, rather than treating folds or cases as independent training repetitions. The final ensemble improves over the seed mean on both encoders, but no best-seed result is used in any headline table.

Common ranking and cost curves. The matched-coverage analysis only removes native authorizations. For each card, its common score is the mean of query–card, anchor–card and query–anchor cosines. A group's score is the mean of its two largest available card scores (or its only score).

Table 33: ATRA training-seed stability (%). Each seed pools all five held-out identity folds; SD is across five complete fold replicates. The final column is the preregistered five-seed decision ensemble, not the best seed.

Encoder	Metric	Seed mean ± SD	Range	Ensemble
SigLIP2-base	Accuracy	78.09 ± 1.14	76.83–79.59	80.46
SigLIP2-base	Unsafe	8.09 ± 0.88	6.72–9.13	6.53
SigLIP2-base	Positive recall	77.89 ± 2.37	75.34–80.87	79.40
SigLIP2-base	NULL specificity	78.55 ± 1.95	76.60–81.12	82.93
DINOv2-base	Accuracy	67.20 ± 0.66	66.61–68.05	68.29
DINOv2-base	Unsafe	10.56 ± 1.35	8.62–11.76	9.73
DINOv2-base	Positive recall	64.20 ± 0.97	63.14–65.70	65.31
DINOv2-base	NULL specificity	74.20 ± 2.67	70.01–77.15	75.25

Confidence is the selected group's gap to its strongest competitor plus half the weakest retained-card score. All methods use this unfitted ranking; descending order retains round(coverage $\times N$) cases, with stable input order for ties. Selective unsafe risk divides unsafe exposures by retained cases, not by the full population. The cost curves use all 3,690 held-out RecordAuth-Diag cases with their fixed empirical condition proportions. Per-case loss is $c_u u_i + m_i$, where $m_i = 1$ only for a positive task without correct personalization and without unsafe exposure; unsafe cases are not counted twice. Figure 6 is therefore conditional on this benchmark composition, not deployment prevalence.

J GENERALIZATION TO HELD-OUT INTERVENTION FAMILIES

Table 34: Condition-level generalization on held-out fold identities (%). Seen/held-out denotes participation in ATRA training and threshold selection, not knowledge during feature design.

Encoder	Mechanisms	Method	n	Accuracy	Unsafe	Coverage
SigLIP2-base	seen	Similarity	1476	62.94	12.20	56.98
SigLIP2-base	seen	Anchor-triad	1476	65.38	14.02	64.43
SigLIP2-base	seen	ATRA	1476	**90.04**	3.32	47.83
SigLIP2-base	seen	ATRA+Guard	1476	84.08	**2.30**	39.30
SigLIP2-base	unseen	Similarity	2214	64.72	19.24	84.69
SigLIP2-base	unseen	Anchor-triad	2214	**79.77**	9.76	77.60
SigLIP2-base	unseen	ATRA	2214	74.07	8.67	77.69
SigLIP2-base	unseen	ATRA+Guard	2214	72.67	**1.63**	62.56
DINOv2-base	seen	Similarity	1476	41.26	30.49	77.10
DINOv2-base	seen	Anchor-triad	1476	41.06	32.45	85.37
DINOv2-base	seen	ATRA	1476	**79.81**	5.42	43.43
DINOv2-base	seen	ATRA+Guard	1476	78.12	**3.52**	37.06
DINOv2-base	unseen	Similarity	2214	57.32	28.73	85.05
DINOv2-base	unseen	Anchor-triad	2214	**66.49**	25.38	87.08
DINOv2-base	unseen	ATRA	2214	60.61	12.60	72.09
DINOv2-base	unseen	ATRA+Guard	2214	62.65	**2.89**	55.92

All rows in Table 34 use held-out fold identities; "seen" and "held out" refer only to optimization and threshold selection; all families were known during feature design. On the six held-out mechanisms with SigLIP2, ATRA changes anchor-triad accuracy/unsafe exposure from 79.77/9.76% to 74.07/8.67%, while the guard reaches 72.67/1.63%. With DINOv2 the corresponding values are 66.49/25.38%, 60.61/12.60%, and 62.65/2.89%. Thus relational authorization transfers most clearly in safety; it does not establish accuracy dominance over the deterministic control.

K MATCHED BASELINES

All deterministic thresholds are selected from the same fold validation identities. Similarity ranks groups by query–card evidence. Medoid uses the card with greatest within-group support. Graph joins cards above pairwise consistency thresholds and scores coherent components. Anchor-only scores anchor–card consistency. Anchor-triad jointly thresholds query–anchor, query–card, and anchor–card evidence and returns NULL when no candidate satisfies the rule. The appendix tables in the final evidence export include every deterministic method, validation grid, and both encoders; Table 37 reports the full gate comparison.

Table 35: Capacity-matched fixed-aggregation control on all 3,690 held-out cases (%). Flat-MLP matches ATRA's 24 features, folds, seeds, training, and validation-only selection, but fixes raw-feature mean/max pooling. Counts are trainable parameters.

Encoder	Method	Params	Acc. ↑	Unsafe ↓	Cov.	Pos. recall ↑	NULL spec. ↑
SigLIP2	Anchor-triad	0	74.01	11.46	72.33	79.09	62.15
	ATRA	55,427	80.46	**6.53**	65.75	79.40	**82.93**
	Flat-MLP	55,773	**81.25**	6.91	67.40	**81.15**	81.48
DINOv2	Anchor-triad	0	56.31	28.21	86.40	**69.53**	25.47
	ATRA	55,427	**68.29**	**9.73**	60.62	65.31	75.25
	Flat-MLP	55,773	65.34	11.63	59.11	60.98	**75.52**

Table 36: External method isolation at 30% matched population coverage (%). Entries are Anchor-triad/ATRA. A common frozen label-free ranking only removes native exposures; native coverage records the unrecoverable utility boundary.

Source	Encoder	Native cov.	Unsafe@30	Error@30	Pos. recall@30
DreamBooth	SigLIP2	80.67/71.67	0.00/0.00	0.00/0.00	42.86/42.86
DreamBooth	DINOv2	91.33/72.33	2.22/2.22	4.44/3.33	40.95/41.43
DAVIS-2017	SigLIP2	77.18/71.15	2.56/0.43	2.99/0.64	41.58/42.58
DAVIS-2017	DINOv2	95.64/69.42	5.13/0.85	9.19/2.35	38.92/41.85

This Flat-LogReg control matches inputs, folds, conditions and selection objective, but deliberately does not match parameter capacity. Its intervals resample the 87 query identities.

L External-source gate results

On prospective DreamBooth, SigLIP2 anchor-triad/ATRA/guard accuracy is 84.00/93.67/95.00% and unsafe exposure is 7.67/4.00/0.67%. DINOv2 reaches 68.33/82.00/88.33% accuracy and 21.67/10.33/3.00% unsafe exposure. On all 156 DAVIS tracks, SigLIP2 reaches 69.74/83.59/84.62% accuracy and 16.99/9.62/2.12% unsafe exposure; DINOv2 reaches 59.04/75.58/79.81% and 31.28/12.24/4.62%. Video-cluster sensitivity and person/non-person breakdowns are retained in the artifact registry.

M Paired end-to-end attribution tests

All-case unsafe use, NULL specificity and paired contrasts accompany the four-arm results in the supplementary artifact; Table 40 reports the paired attribution contrasts.

Table 37: Complete deterministic, flat learned, set learned, and guarded authorizer comparison on all held-out RecordAuth-Diag identities (%).

Encoder	Method	Accuracy	Unsafe	Coverage	Pos. recall	NULL spec.
SigLIP2	Similarity	64.01	16.42	73.60	72.71	43.72
	Medoid	61.17	19.81	74.25	69.45	41.82
	Graph	63.82	17.21	73.33	72.51	43.54
	Anchor-only	56.34	31.84	80.14	62.14	42.82
	Anchor-triad	74.01	11.46	72.33	79.09	62.15
	Flat-LogReg	72.28	17.53	83.12	**82.66**	48.06
	ATRA	**80.46**	6.53	65.75	79.40	82.93
	ATRA+Guard	77.24	**1.90**	53.25	69.14	**96.12**
DINOv2	Similarity	50.89	29.43	81.87	63.07	22.49
	Medoid	47.37	35.20	85.66	60.05	17.80
	Graph	49.00	33.20	85.04	62.37	17.80
	Anchor-only	46.53	39.46	89.95	59.43	16.44
	Anchor-triad	56.31	28.21	86.40	**69.53**	25.47
	Flat-LogReg	53.93	30.73	88.05	67.98	21.14
	ATRA	68.29	9.73	60.62	65.31	75.25
	ATRA+Guard	**68.83**	**3.14**	48.37	58.65	**92.59**

Table 38: ATRA minus Flat-LogReg (percentage points; 95% query-identity intervals). Positive accuracy/recall and negative unsafe exposure favor ATRA.

Encoder	Metric	ATRA − Flat-LogReg [95% CI]
SigLIP2-base	Accuracy	+8.18 [+5.32, +10.94]
SigLIP2-base	Unsafe	-11.00 [-13.34, -8.68]
SigLIP2-base	Positive recall	-3.25 [-7.39, +0.81]
DINOv2-base	Accuracy	+14.36 [+11.20, +17.53]
DINOv2-base	Unsafe	-21.00 [-23.98, -18.11]
DINOv2-base	Positive recall	-2.67 [-6.39, +1.07]

Table 39: Matched four-arm generation on the same cases and frozen decoder (%). Stress unsafe measures unauthorized use on non-clean conditions; positive recall measures correct personalization.

Source	Model	Metric	Raw	Anchor-triad	ATRA	ATRA+Guard
RecordAuth-Diag	Qwen3-VL-8B	Accuracy	40.16	54.66	**55.93**	53.74
		Stress unsafe	29.40	1.15	2.57	**0.64**
		Pos. recall	37.05	39.88	**40.34**	35.08
RecordAuth-Diag	Phi-3.5-Vision	Accuracy	19.73	47.94	**54.55**	52.28
		Stress unsafe	36.21	11.86	7.38	**2.00**
		Pos. recall	5.30	44.87	**46.65**	38.75
RecordAuth-Diag	CoViP	Accuracy	40.41	59.51	**60.70**	57.97
		Stress unsafe	32.28	2.07	3.39	**0.85**
		Pos. recall	37.90	47.27	**47.74**	41.00
DreamBooth	Qwen3-VL-8B	Accuracy	50.33	67.00	69.33	**70.67**
		Stress unsafe	38.33	**0.00**	4.17	**0.00**
		Pos. recall	57.62	58.57	**60.95**	58.10
DreamBooth	Phi-3.5-Vision	Accuracy	8.00	51.67	59.00	**60.00**
		Stress unsafe	47.92	8.75	4.17	**0.00**
		Pos. recall	2.86	54.76	**57.62**	55.71
DreamBooth	CoViP	Accuracy	51.00	73.33	76.33	**77.67**
		Stress unsafe	39.58	2.50	5.00	**0.83**
		Pos. recall	59.05	69.52	**70.95**	68.10
DAVIS-2017	Qwen3-VL-8B	Accuracy	44.30	65.60	69.10	**71.00**
		Stress unsafe	34.75	6.00	9.88	**1.88**
		Pos. recall	44.00	61.29	**65.57**	60.86
DAVIS-2017	Phi-3.5-Vision	Accuracy	21.30	41.40	50.10	**51.80**
		Stress unsafe	35.50	15.38	10.25	**2.12**
		Pos. recall	7.86	38.29	**42.43**	39.14

The track- and video-cluster intervals agree on every DAVIS safety contrast. The guard's accuracy advantage over ATRA is not significant under either bootstrap, whereas its unsafe-use reduction is. The DreamBooth Qwen anchor-triad equality and CoViP uncertainty are retained in the same table rather than summarized as universal success.

N Independent personalized-pipeline confirmation

RAP-Phi3-mini is independently authored and post-trained to consume retrieved personal concepts. We pin checkpoint revision `0264b32f`, RAP source revision `3c0c9c2e`, and OpenAI CLIP ViT-L/14-336 revision `ce19dc91`. Each RecordAuth-Diag card becomes one RAP record: its image is the retrieval key, while opaque subject, location, timestamp, and status form `Name/Info`. The query image is first, followed by deterministic full-frame cosine top-2 records in RAP's released multi-image syntax. Full-frame retrieval replaces RAP's detector crops because the diagnostic views are already single-subject images; we do not call this an exact deployment reproduction.

The adapter was debugged on 30 validation parents. An initial over-explicit refusal prompt produced 100% abstention and is retained as a failed run. The locked native answering prompt was then

Table 40: Paired end-to-end attribution contrasts in percentage points. RecordAuth-Diag intervals resample query identities; DreamBooth resamples causal parents; DAVIS resamples tracks and additionally reports a case-weighted source-video sensitivity. McNemar p is a case-level diagnostic, not clustered inference.

Source	Model	Contrast	Metric	Δ [95% CI]	Cluster	Video CI	p
RecordAuth-Diag	Qwen3-VL-8B	Guard–Anchor	Accuracy	-0.92 [-2.49, +0.54]	identity	–	0.03
	Qwen3-VL-8B	Guard–Anchor	Stress unsafe	-0.51 [-0.86, -0.21]	identity	–	6.1e-05
	Qwen3-VL-8B	Guard–ATRA	Accuracy	-2.20 [-3.50, -1.05]	identity	–	7.44e-09
	Qwen3-VL-8B	Guard–ATRA	Stress unsafe	-1.93 [-2.65, -1.28]	identity	–	1.39e-17
	Phi-3.5-Vision	Guard–Anchor	Accuracy	+4.34 [+2.48, +6.37]	identity	–	3.77e-12
	Phi-3.5-Vision	Guard–Anchor	Stress unsafe	-9.86 [-13.00, -7.32]	identity	–	5.03e-88
	Phi-3.5-Vision	Guard–ATRA	Accuracy	-2.28 [-4.21, -0.66]	identity	–	4.38e-06
	Phi-3.5-Vision	Guard–ATRA	Stress unsafe	-5.39 [-6.58, -4.32]	identity	–	2.74e-48
	CoViP	Guard–Anchor	Accuracy	-1.54 [-3.51, +0.25]	identity	–	0.00131
	CoViP	Guard–Anchor	Stress unsafe	-1.22 [-2.14, -0.51]	identity	–	2.91e-11
	CoViP	Guard–ATRA	Accuracy	-2.74 [-4.33, -1.29]	identity	–	1.57e-10
	CoViP	Guard–ATRA	Stress unsafe	-2.54 [-3.42, -1.73]	identity	–	5.29e-23
DreamBooth	Qwen3-VL-8B	Guard–Anchor	Accuracy	+3.67 [+1.67, +5.67]	parent	–	0.00342
	Qwen3-VL-8B	Guard–Anchor	Stress unsafe	+0.00 [+0.00, +0.00]	parent	–	1
	Qwen3-VL-8B	Guard–ATRA	Accuracy	+1.33 [-1.67, +4.33]	parent	–	0.454
	Qwen3-VL-8B	Guard–ATRA	Stress unsafe	-4.17 [-6.25, -2.08]	parent	–	0.00195
	Phi-3.5-Vision	Guard–Anchor	Accuracy	+8.33 [+4.33, +12.33]	parent	–	1.62e-06
	Phi-3.5-Vision	Guard–Anchor	Stress unsafe	-8.75 [-14.58, -3.75]	parent	–	9.54e-07
	Phi-3.5-Vision	Guard–ATRA	Accuracy	+1.00 [-1.67, +3.33]	parent	–	0.549
	Phi-3.5-Vision	Guard–ATRA	Stress unsafe	-4.17 [-6.25, -2.08]	parent	–	0.00195
	CoViP	Guard–Anchor	Accuracy	+4.33 [+1.33, +7.33]	parent	–	0.00443
	CoViP	Guard–Anchor	Stress unsafe	-1.67 [-5.00, +0.00]	parent	–	0.125
	CoViP	Guard–ATRA	Accuracy	+1.33 [-1.67, +4.33]	parent	–	0.454
	CoViP	Guard–ATRA	Stress unsafe	-4.17 [-6.25, -2.08]	parent	–	0.00195
DAVIS-2017	Qwen3-VL-8B	Guard–Anchor	Accuracy	+5.40 [+2.80, +8.00]	track	[+2.72, +8.02]	1.91e-07
	Qwen3-VL-8B	Guard–Anchor	Stress unsafe	-4.12 [-6.38, -2.12]	track	[-6.31, -2.13]	2.33e-10
	Qwen3-VL-8B	Guard–ATRA	Accuracy	+1.90 [-0.80, +4.30]	track	[-0.74, +4.34]	0.0503
	Qwen3-VL-8B	Guard–ATRA	Stress unsafe	-8.00 [-10.38, -5.88]	track	[-12.04, -5.23]	1.08e-19
	Phi-3.5-Vision	Guard–Anchor	Accuracy	+10.40 [+8.20, +12.60]	track	[+7.96, +12.90]	2.87e-21
	Phi-3.5-Vision	Guard–Anchor	Stress unsafe	-13.25 [-16.50, -10.12]	track	[-16.88, -10.00]	2.47e-32
	Phi-3.5-Vision	Guard–ATRA	Accuracy	+1.70 [-0.30, +3.50]	track	[-0.40, +3.60]	0.043
	Phi-3.5-Vision	Guard–ATRA	Stress unsafe	-8.12 [-10.62, -6.00]	track	[-11.79, -5.29]	5.42e-20

evaluated once on test. Because one clean test parent had already been inspected during an integration smoke test, its complete ten-condition parent is excluded, leaving 3,680 cases, 368 parents, and 86 query identities. No test-conditioned prompt, top-k, threshold, or parser change followed.

Table 41: Independent RAP-Phi3 transfer (%, 95% query-identity-cluster CI; 3,680 cases, 368 causal parents). The frozen guard sees the same cases and can only filter cards before native CLIP top-2 retrieval.

Arm	Clean recall	Stress unsafe	NULL specificity	All accuracy	Positive recall
Raw	**27.45** [22.15,33.33]	36.99 [34.87,39.51]	0.00 [0.00,0.00]	22.12 [19.78,24.68]	**31.60** [28.33,35.17]
Strict guard	19.97 [15.08,25.07]	**1.56** [0.49,3.50]	**96.20** [93.48,98.05]	**47.58** [44.17,51.11]	26.75 [22.31,31.61]
Guard − raw	−7.47 [−10.71,−4.69]	−35.43 [−38.13,−32.84]	+96.20 [93.56,98.06]	+25.46 [23.18,27.64]	–

Table 42: RAP generator attribution (%, 95% query-identity-cluster CI). The primary population is frozen before each context intervention.

Population / intervention	Endpoint	Rate [CI]
Raw unsafe ($n = 1{,}105$), no memory	Same raw prediction	0.00 [0.00,0.00]
Raw unsafe, nonce relabel	Exact displayed-nonce copy	38.10 [34.06,42.80]
Raw unsafe, nonce relabel	Same-event exact tracking	26.97 [23.29,30.97]
Raw unsafe, record shuffle	Same-event tracking	66.61 [63.05,70.04]
Clean correct ($n = 101$), nonce relabel	Exact displayed-nonce copy	31.68 [23.26,40.45]

The guard's reduction in unsafe use is paired with a significant clean-recall cost. Exact McNemar tests give $p < 10^{-14}$ for clean recall and $p < 10^{-294}$ for stress unsafe use; identity-cluster intervals, not these case-level tests, remain primary. Nonce and shuffle hold visual keys and their retrieval scores

fixed while altering record text; removal exposes no card. Thus nonce and same-event endpoints cannot be explained by an unchanged visual match alone.

O ROLE-LEVEL SPLIT AND DEPENDENCE AUDIT

The fold constructor assigns an identity to one split before any causal cases are expanded. We nevertheless audit the serialized manifests, because a query-only audit would miss leakage through distractor, anchor, or card roles. For every record we enumerate query identity/image, every candidate-group identity, trusted-anchor image, and memory-card image. Table 43 reports the resulting all-role sets.

Table 43: Serialized all-role split audit. Counts are train/validation/test. Overlap columns count identities or images appearing in more than one split within a fold; "viol." counts records containing a role assigned outside their query split.

Fold	Cases	Identities	Images	ID/image overlap	Viol.
0	2320/700/670	53/16/18	755/226/218	0/0	0
1	2300/670/720	51/18/18	746/218/235	0/0	0
2	2190/720/780	51/18/18	711/235/253	0/0	0
3	2090/780/820	52/18/17	679/253/267	0/0	0
4	2170/820/700	54/17/16	706/267/226	0/0	0

Each fold contains the same 87 identities and 1,199 images overall; no group-role identity is absent from the query-role assignment. Repeated use remains substantial within a split (up to 528 group/anchor and 2,648 card occurrences per identity), so query-identity resampling captures the main ownership cluster but not every shared-distractor dependency. The audit therefore removes a leakage explanation without turning the existing intervals into full learning-procedure uncertainty.

As a complementary diagnostic, we delete each identity together with every case in which it appears as query, candidate group, anchor, or card, then recompute the principal guard–anchor and guard–ATRA accuracy/unsafe contrasts. Across Qwen and CoViP, all leave-one-identity-out estimates retain the full-data sign (zero sign reversals). This tests leverage by repeated non-query roles; it is not a new split, retraining run, or confidence interval over the fold-construction procedure.

P FROZEN REPRESENTATION-TO-POLICY ANALYSES

The public LSD diagnostic contains 50 identity clusters with ten contexts each. We serialize four conditions per row: authorized original, gray-image missing mapping, absent-persona query, and a derangement that preserves all real images and dialogues but permutes their binding. Five fixed folds use 30/10/10 identities for train/ validation/test. Linear logistic probes select regularization and threshold on source validation identities only. The primary feature is the final prompt-token state at layer 24; controls combine null likelihood and image-similarity scalars. Split-crossing absent-persona rows are excluded, leaving 1,936 cases for the all-invalid comparison.

Cross-checkpoint rows fit and calibrate only on the source checkpoint. The target-fold summary distinguishes ranking transfer from portable calibration.

Table 44: Frozen representation audit. Within-checkpoint values are out-of-fold AUROC with 95% identity-cluster intervals. Macro: target-fold mean±SD. BA: balanced accuracy at the source validation threshold.

Train/evaluate	Task	Hidden pooled [CI]	Scalar pooled [CI]	Target macro	Source-thr. BA	Leave-mispair-out
Base/Base	all invalid	.949 [.926,.968]	.802 [.770,.831]	–	.872	.599
CoViP/CoViP	all invalid	.966 [.944,.983]	.810 [.788,.830]	–	.908	.646
Base/CoViP	all invalid	.896 [.864,.931]	–	.960±.019	.686	–
CoViP/Base	all invalid	.876 [.834,.921]	–	.948±.030	.698	–
Base/CoViP	mispair	.871 [.839,.907]	–	.958±.015	.777	–
CoViP/Base	mispair	.820 [.780,.869]	–	.924±.020	.732	–

The separate within-checkpoint mispair analysis gives AUROCs of .926/.964 for Base/CoViP and .608/.603 for the scalar control. Table 44 instead lists all-invalid within-checkpoint results and

cross-checkpoint mispair transfer; these are different comparisons. A layer sweep rises from chance at layer 0 to its maximum near layer 24; paired hidden scores rank original above derangement in 97.8/99.6% of rows. Cross-checkpoint transfer is meaningful because the checkpoints share architecture and aligned initialization, but it is not cross-architecture evidence. The gap between pooled and fold-macro AUROC and the moderate transferred BA show that ranking direction is more stable than global calibration.

The simulated gate uses only out-of-fold readouts. It retains 82/83 and 115/121 correct originals while leaving 50 and 44 invalid false attributions (3.48/3.06%). This estimates recoverable policy headroom; it does not validate deployment under shift. The leave-mispair-out column trains on missing and absent-persona negatives and then evaluates a never-trained derangement. Its weak .599/.646 result is retained. Finally, shifting the discriminative layer-24 direction by -5 and -10 in a preliminary prefill activation-steering experiment on three cases changes no outputs. The tested direction transfers weakly to unseen derangements and does not alter these three generations at the tested layer and scales.

Q STRUCTURED SAME-MLLM VERIFIERS AND REFERENCE ROBUSTNESS

The staged verifier is frozen before all RecordAuth-Diag inference. The group prompt presents the query, every enrollment referent, every image-backed card, and active/missing/stale status, but hides record locations and expected answers. It requires exactly one `GROUP_k` token or `NULL`. The optional edge prompt then presents only the selected group's referent, query, and labelled cards and requires an exact `KEEP: CARD_i,...` list or `NULL`. Any parse failure maps to operational NULL. Both prompts use greedy decoding and the frozen Qwen3-VL or CoViP checkpoint; neither receives RecordAuth-Diag labels, demonstrations, training, or threshold selection.

The later one-pass control receives exactly the full-bank evidence of the group prompt but must emit either `GROUP_k; KEEP: CARD_i,...` or `NULL` in one call. Group and card indices must be in-range, unique, and sorted; every other output fails closed. Its prompt, parser, two checkpoint revisions, and all 3,690 case IDs were hash-frozen before output generation. It is an information-matched structured control, not a compute-matched one: Qwen/CoViP use 11.25M input tokens and 3,583/3,517 verifier seconds, versus 13.37/13.38M and 3,421/3,440 seconds for group plus edge. The joint output is longer, so fewer input tokens do not imply lower measured latency.

Every decision file covers the same 3,690 test cases exactly. The unchanged natural answer generator receives only the retained cards; a group/edge rejection invokes a hard bypass without generation. Table 45 reports the resulting record answers. Staged card verification reduces unsafe record answers from 517 to 71 for Qwen and from 689 to 125 for CoViP, while correct record answers decrease from 792 to 708 and from 936 to 826. Other task-invalid record answers remain substantial after card verification.

Table 45: End-to-end outcomes after same-MLLM verification on 3,690 RecordAuth-Diag cases (2,583 positive tasks), with unchanged generation and hard NULL bypass. Other denotes attribution-eligible task errors; unmatched outputs are non-NULL.

Model	Interface	Released	Correct	Unsafe	Other	Unmatched
Qwen	Group/NULL	1882	**792**	517	573	6
	One-pass	1213	502	287	**424**	9
	Staged	1454	708	**71**	675	**1**
CoViP	Group/NULL	2198	**936**	689	**573**	**1**
	One-pass	1578	585	417	576	7
	Staged	1730	826	**125**	779	**1**

Table 46 separates verification and generation costs. The group/card rows sum the full-bank routing and edge-verification calls. Generation costs include zero-cost bypass cases. Times sum the recorded per-call durations rather than measuring the parallel job's elapsed time. Peak verifier memory is 18.13 GiB on an RTX 4090.

Anchor-reference robustness. This separate audit uses frozen ATRA, not the preceding same-MLLM verifier. Each of 156 DAVIS tracks contributes two eligible temporal views unused by the

Table 46: Recorded same-MLLM verification and generation costs on 3,690 cases. Tokens are input tokens in millions; seconds are summed call durations.

Model	Interface	Verifier tokens (M)	Generation tokens (M)	Generation calls	Verifier seconds	Generation seconds
Qwen	Group/NULL	11.253	2.283	2,595	2294.0	1591.6
	One-pass	11.250	0.970	1,696	3582.7	972.8
	Staged	13.367	0.913	1,643	3421.5	955.0
CoViP	Group/NULL	11.253	2.300	2,617	2273.3	1425.0
	One-pass	11.250	1.030	1,907	3517.5	1018.4
	Staged	13.384	1.003	1,833	3440.4	938.3

original anchor, cards, or query (312 assets), plus blurred and center-occluded transforms of its original anchor (312 assets; 624 total). Exact and perceptual-hash audits find no forbidden temporal-view overlap. The wrong view for each group is the frozen-encoder nearest different track in the same candidate pool. The 2-of-3 wrapper retains a group and each edge independently by majority vote; ties and empty results return NULL. The quorum is applied to the original validation-locked ATRA outputs without retraining or threshold changes. Primary intervals resample tracks; a DAVIS-video sensitivity gives the same direction. With three clean views, SigLIP2/DINOv2 reach 83.78/74.10% accuracy and 9.55/11.92% unsafe exposure, on the three-clean-view condition. One-wrong and two-wrong decisions are retained in the reference-corruption artifact. The two-wrong construction tests failure of the majority-clean assumption; no unshown numerical comparison is implied here.

R OMNI-PERSONA VISUAL PROTOCOL AND COMPLETE ROUTING RESULTS

We pin Omni-Persona code commit `818b0804` and dataset revision `a5f99514`. All 231 visual items and 860 referenced assets pass path and hash audits. Each item interleaves four persona contexts; its 115 answerable, 54 attribute-absent, and 62 target-absent questions collapse to 172 unique source/query clusters. Because the released visual item provides one context reference per persona, the same reference is used as enrollment anchor and one-card group. This passes structural mapping but removes anchor–card diversity: external ATRA performance is a hard schema-shift test, not another RecordAuth-Diag edge-corruption result.

Prompt and local judgment. The strict arm appends a rule requiring one uniquely matching context image; otherwise it emits "I cannot determine that from the provided context." It does not change the images or their order. For answerable, non-abstaining outputs, the local Phi-3.5 judge receives the gold answer and prediction under the frozen correctness rubric, decodes greedily for at most eight tokens, and parses CORRECT/WRONG; invalid verdicts count as wrong. Unanswerable items are scored by the fixed abstention rule, without the judge. "Cal" is the fraction of all 231 items that are judged correct when answerable or correctly abstained when unanswerable. It is not a probability calibration score. Both this aggregate and answerable accuracy remain non-human-validated sensitivity measures.

Raw and strict arms use the released context order and deterministic decoding. Frozen SigLIP2 and ATRA use only non-Omni thresholds. The integrated verifier presents the four context images and query image to the same frozen MLLM, requires exactly one context index or NULL, and runs before the unchanged answer prompt. Table 47 reports exact, judge-free routing.

Table 47: Judge-free Omni routing with 95% source-sample-cluster intervals (%). T-abs. is target-absent NULL specificity; A-abs. is correct identity routing when the selected context lacks the queried attribute.

Router	Exact route [CI]	Unsafe [CI]	Pos. recall [CI]	T-abs. [CI]	A-abs. [CI]
Frozen ATRA	55.84 [47.64,63.95]	3.03 [0.43,6.67]	42.60 [33.54,51.76]	91.94 [80.00,100]	44.44 [30.77,58.14]
Qwen router	90.91 [85.90,95.32]	**1.30** [0.00,3.51]	89.35 [83.05,94.84]	**95.16** [87.04,100]	**90.74** [82.14,98.00]
CoViP router	**91.77** [86.96,95.95]	3.90 [0.88,7.69]	**93.49** [88.76,97.48]	87.10 [74.24,97.83]	**90.74** [82.14,98.04]

The integrated result is a strong architecture-equivalent baseline against the claim that a separate learned gate is necessary. It supports the decision decomposition: the frozen answer model can

verify identity when forced to emit an explicit routing action. It does not check whether the chosen biography states the requested attribute. Frozen SigLIP2 similarity reaches 71.43% exact routing and 10.82% unsafe routing. Order sensitivity remains nonzero. Four-order unanimous voting, added only after the single-order outputs were seen, lowers unsafe routing to 0.87/1.30% for Qwen/CoViP but also lowers positive recall to 81.66/89.35%; it is marked exploratory and is excluded from confirmatory claims.

Filtered generation invokes the released answer-oriented generator on the selected context. The NULL-bypass variant instead returns the fixed abstention without invoking the generator when routing rejects.

Table 48: Filtered generation versus operational NULL bypass (%). Cal uses the local Phi-3.5 sensitivity judge; target-absent false answers use the deterministic abstention rule.

Model	Verifier	Filtered Cal	Bypass Cal	Filtered false	Bypass false
Qwen	frozen ATRA	36.36	56.28	64.52	3.23
Qwen	integrated	47.19	64.07	64.52	3.23
CoViP	frozen ATRA	47.19	54.98	25.81	3.23
CoViP	integrated	54.55	60.61	29.03	6.45

The composer audits two invariants over all four bypass arms: every covered prediction is byte-identical to the filtered run, and every rejected prediction is exactly the fixed abstention. Thus the difference in Table 48 is caused by making NULL operational, not regeneration or prompt editing. The benchmark's official GPT-5.4-mini judge was unreachable from the execution environment; all free-form metrics are explicitly labeled as local sensitivity measurements. No AI judge is described as human validation.

S EXTERNAL ALL-NEGATIVE TRANSFER

Table 49: Actual ordinary-versus-authorized capability after an audited NULL decision (%). Values are base→authorized. Exact output agreement and exposed-card count audit the bypass rather than inferring preservation from a proxy.

Task	Model	n	Accuracy	Task-specific	Output agree.	Cards
MMVP	Qwen3-VL-8B	300	79.33→79.33	Strict-pair: 62.00→62.00	100.00	0
MMVP	Phi-3.5-Vision	300	66.33→66.33	Strict-pair: 38.67→38.67	100.00	0
POPE	Qwen3-VL-8B	9000	88.47→88.47	F1-yes: 87.75→87.75	100.00	0
POPE	Phi-3.5-Vision	9000	79.92→79.92	F1-yes: 75.60→75.60	100.00	0

The contamination tasks replace the native VQA question with a question about whether any personal memory applies to an independently sourced COCO/MMVP image. A stable template selection draws banks from the development registry and rejects exact/perceptual-hash image collisions. The source construction sets the applicable groups and authorized event set to empty, so the task target is NULL. Ordinary VQA preservation is a separate, known task mode: it withholds personal cards but retains the original benchmark question. Thus no-memory context selection and final-answer abstention are different actions; the experiment does not learn which mode a real request requires.

On 500 POPE/COCO cases, SigLIP2 similarity/graph/anchor-triad/ATRA/guard produce 7/8/7/0/0 false authorizations. DINOv2 produces 104/196/78/2/1. On 300 MMVP cases, the corresponding SigLIP2 counts are 3/3/3/0/0 and DINOv2 counts are 35/97/41/1/1. These are gate decisions only. They motivate but do not substitute for the registered generator-level contamination tests.

Table 50: Generator-level transfer on independent all-negative banks (%). Cells are raw→guard with guard-minus-raw parent-cluster 95% intervals. "Accuracy" is task-defined NULL correctness rather than ordinary VQA accuracy; abstention is reported separately so that reduced response coverage is not interpreted as improved personalization.

Source	Model	n	NULL acc. Δ [CI]	Unsafe Δ [CI]	Abstain
MMVP	Qwen3-VL-8B	300	100.00→100.00 (+0.00 [+0.00,+0.00])	0.00→0.00 (+0.00 [+0.00,+0.00])	100.00→100.00
MMVP	Phi-3.5-Vision	300	100.00→100.00 (+0.00 [+0.00,+0.00])	0.00→0.00 (+0.00 [+0.00,+0.00])	100.00→100.00
POPE/COCO	Qwen3-VL-8B	500	100.00→100.00 (+0.00 [+0.00,+0.00])	0.00→0.00 (+0.00 [+0.00,+0.00])	100.00→100.00
POPE/COCO	Phi-3.5-Vision	500	100.00→100.00 (+0.00 [+0.00,+0.00])	0.00→0.00 (+0.00 [+0.00,+0.00])	100.00→100.00

At generator level, both raw and guarded Qwen/Phi paths abstain on every custom POPE and MMVP contamination case. This is a retained null result: the prompts do not force either decoder to consume the irrelevant bank, so these sets demonstrate neither a repair gain nor positive personalization. The ordinary VQA runs in Table 49 answer the benchmark questions and serve the separate capability-preservation claim.

T FAILURE RETENTION AND HARDWARE AUDIT

The original strict-guard implementation intersected per-run decisions before majority voting. Because voting and intersection do not commute, the resulting ensemble could add an edge relative to the separately voted ATRA output. That output is archived and excluded. The corrected pipeline ensembles first and intersects once; audits across ten sources find zero added edges, index inconsistencies, or newly introduced unsafe cases. A registered 300-case rerun on A100 reproduces A800 outputs exactly for Qwen and Phi, raw and guarded: exact text, parsed predictions, answer flags, and unsafe flags all agree in 100% of cases.

The DAVIS-56 cost registry counts actual non-bypassed generation calls. Guard-only generation uses 285,435 input and 5,859 output tokens in 300.576 seconds; the raw path adds 2,032,209 input and 5,485 output tokens in 538.978 seconds. Serialized dual-path cost is therefore 2,317,644 input tokens, 11,344 output tokens, and 839.554 seconds. The counterfactual policy and path-HGB consume these same saved candidates. Timed CPU replays with saved SigLIP2 embeddings take 3.96 and 14.05 seconds, respectively, including process/model loading but excluding image embedding; these are audit runtimes, not online latency.

Table 51: A800-to-A100 deterministic reproducibility (%) on the registered 300-case subset.

Model / arm	n	Text agree.	Parsed agree.	Accuracy	Unsafe
Qwen / Raw	300	100.00	100.00	39.33→39.33	27.00→27.00
Qwen / Guard	300	100.00	100.00	49.00→49.00	1.67→1.67
Phi / Raw	300	100.00	100.00	21.00→21.00	22.67→22.67
Phi / Guard	300	100.00	100.00	41.33→41.33	5.67→5.67

U ADDITIONAL MEASUREMENT AND COMPARISON LIMITATIONS

The CoViP authors describe their downstream LSD diagnostic images as evaluation-only (Oh et al., 2026), but we cannot independently audit the checkpoint's complete training corpus; public image overlap in foundation-model pretraining also remains possible. Frozen base Qwen and Phi separate the CoViP-specific post-training effect, not generic pretraining exposure. RAP supplies an independent checkpoint lineage, but its complete training corpus and foundation-model image overlap are likewise not independently auditable. Synthetic nonce labels and paired within-parent changes reduce the value of memorizing a diagnostic answer, but do not make public-image provenance irrelevant.

Generator attribution is automatic and deliberately conservative: its headline requires suffix-preserving equality with a synthetic location displayed after an intervention, with the ordinary parser reported only as a sensitivity. This makes nonce copying causally auditable but does not validate semantic attribution in free-form personal narratives. A blinded 180-item packet has been specified for two human annotators; until those judgments are actually collected, we make no claim of human agreement or semantic-parser validity. No-memory and oracle-authorized runs are diagnostic endpoints, not usable deployment policies.

The capacity-matched Flat-MLP uses 55,773 parameters versus ATRA's 55,427. Reanalysis pairs all 3,690 held-out cases and resamples 87 query identities. Table 52 gives Flat-MLP minus ATRA differences. Both SigLIP2 intervals contain zero; this does not establish equivalence. The DINOv2 intervals favor ATRA for accuracy and unsafe exposure. Hierarchical aggregation is therefore an encoder-dependent design choice, not a demonstrated universal requirement.

Table 52: Capacity-matched paired differences (percentage points). 10,000 query-identity bootstrap replicates; fixed predictions and seeds. Positive accuracy favors Flat-MLP; positive unsafe favors ATRA.

Encoder	Endpoint	Flat-MLP−ATRA	95% CI
SigLIP2	Accuracy	+0.79	[-0.61,2.22]
SigLIP2	Unsafe exposure	+0.38	[-0.55,1.36]
DINOv2	Accuracy	-2.95	[-4.33,-1.54]
DINOv2	Unsafe exposure	+1.90	[0.78,3.11]

V PRESENCE-CHANNEL AND FREE-FORM CONTROLS

Target absence dominates the boundary: no supplied record belongs to the queried subject. The supplied-token control provides intended-subject metadata. It evaluates an additional-observation endpoint without implementing the association-integrity validation assumed in Proposition 2.

Table 53: Supplied-subject metadata control (%). Fixed candidates use query, registered-subject and attached-image identity metadata; zero observed unsafe releases are not population-risk bounds.

Panel	Observation channel	Clean accuracy	Aggregate unsafe	Worst-condition unsafe
DAVIS-56	legacy localized dual check	73.21	0.54	5.36
DAVIS-56	supplied subject metadata (oracle)	**85.71**	**0.00**	**0.00**
Support/Qwen–CoViP	supplied subject metadata (oracle)	87.30–88.93	0.00	0.00

On DAVIS, the token policy has 0/560 observed unsafe releases and 48/56 correct clean answers. It supplies identity-equivalent metadata rather than inferring credentials from images. Its eligibility differs from the raw/guard MLLM candidate oracle on visually ambiguous queries; the full-case audit in Appendix A.8 separates them. Proposition 1 does not exclude independently calibrated absolute presence signals.

The free-form Omni-Persona panel asks a different question. Strict prompting lowers target-absence false answers to 1.61%, while attribute-absence false answers remain 46.30–53.70% (Appendix B). These false-answer rates use a frozen abstention contract; answerable correctness is a separate local-judge sensitivity measure. The result motivates support checks beyond the location schema, not a claim of validated free-form semantic accuracy.

W RETAINED DETAILED ACTION AND REPLAY TABLES

On DAVIS-560, the guard increases correct records from 170 to 228 and reduces unsafe use from 149 to 5. Offline top-1 filtering has similar record release (54.29% versus 55.54%) but 163 correct and 38 unsafe answers (Table 18). The paired unsafe difference is -5.89 points (95% track-cluster interval $[-8.57, -3.57]$); 27 of the net 33 fewer unsafe cases occur in support conditions. Unsafe-only counts are 36 for top-1 and 3 for the guard, with 2 shared cases (Table 15). Each arm uses 560 saved Qwen calls, but offline filters reuse full-bank answers while the guard changes generation context. This is a condition-level policy contrast, not an isolated module effect.

Table 54: Same-MLLM verifier decisions on 3,690 cases (%). One-pass jointly predicts group and cards; staged verification checks cards after routing. Parse failures reject.

Model	Interface	Group acc.	Group unsafe	Card exposure	Positive recall	NULL specificity
Qwen	group/NULL	**81.11**	9.95	43.63	**86.26**	69.11
	one-pass group+cards	65.47	5.42	19.86	57.92	83.11
	Staged group + cards	70.87	**1.90**	**3.06**	60.90	**94.13**
CoViP	group/NULL	**81.46**	10.08	44.09	**86.91**	68.74
	one-pass group+cards	69.62	6.15	21.46	65.04	80.31
	Staged group + cards	73.88	**3.04**	**4.61**	66.63	**90.79**

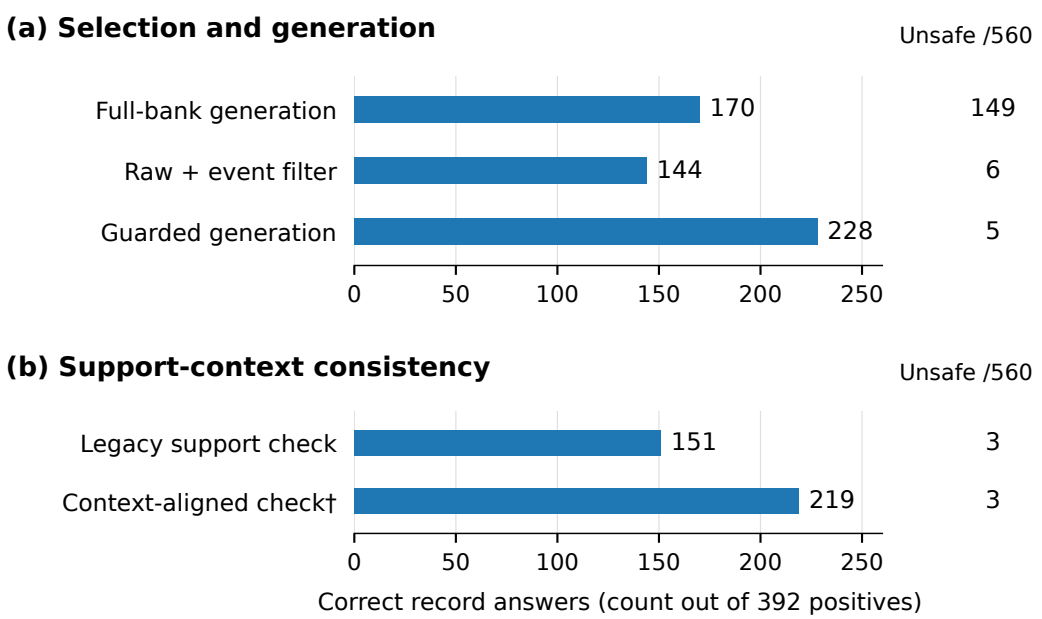

Figure 7: Fixed-candidate replay on DAVIS-560. Correct counts use 392 positives; unsafe counts use all 560 cases. †: post-hoc support alignment with candidates, thresholds and task targets fixed.

Table 55: Frozen input-block occlusion on 3,690 held-out cases (%). Correct recall uses positive tasks; no model is refit.

Occluded block	Accuracy	Unsafe	Coverage	Correct recall
None (ATRA)	80.46	6.53	65.75	79.40
Query–card	49.54	4.04	29.51	32.40
Anchor	61.57	1.60	36.83	46.61
Triad-derived	60.16	5.31	42.85	49.01
Peer	79.05	8.24	63.12	77.43
Validity/position	81.11	7.21	66.23	80.87
Cardinality	81.22	8.92	71.41	83.93

The retained-event set in Table 56 is the full authorizer selection, read independently of either answer. Changing the eligibility set cannot reverse the generator's prior exposure.

Table 56: Fixed-answer retention replay on DAVIS-560. Release, correct and unsafe are supplied-record counts; no candidate is regenerated.

Answer source	Eligible event set	Release	Correct	Unsafe
Full-bank generation	Full bank	454	170	149
Full-bank generation	ATRA-retained	246	144	6
Guarded generation	Full bank	311	**228**	**5**
Guarded generation	ATRA-retained	311	**228**	**5**

Post-generation trace and support. Optional checks reuse a fixed answer rather than selecting its input context. The location task requires the latest usable record. We first map the answer to one exact event occurrence, then require an active, image-backed card with the requested field and the latest timestamp among usable cards in the inspected group. The legacy check inspects the full bank, which can disagree with the filtered generation context. We also evaluate a post-hoc context-aligned check restricted to the guard's retained event IDs; raw-path support remains full-bank. Candidates and margin thresholds are unchanged. A missing or nonunique trace returns NULL. This is a schema-specific support test; free-form questions require a separate evaluation (Appendix B). Figure 2 summarizes ATRA–triad selection, the alternative Field and frozen-MLLM readouts of K, and their separate candidate checks.

X Detailed experimental protocol and scope

X.1 Experimental protocol

Sources and evidence roles. RecordAuth-Diag contains 369 matched parents, 3,690 cases and 87 identities derived from LSD and Yo'LLaVA with synthetic event metadata. Five identity folds separate the learned authorizers' training, validation and test roles. Raw-interface diagnostics evaluate Qwen3-VL-8B, CoViP, Phi-3.5-Vision, RAP and Gemma-3-4B-IT. Their differences in training and interfaces preclude an architecture ranking.

The localized study uses 156 DAVIS tracks. The first 100 support development; the remaining 56 contribute ten conditions each to frozen confirmation. The 56 tracks span 43 videos, of which 25 also contribute development tracks: confirmation is track-disjoint, not video-disjoint. DreamBooth contributes 30 subjects, and a 488-parent Binding$\times$Support panel supplies crossed binding/support controls. Appendix E gives the full source–stage map. Full-frame, mask-localized and token-supplied observations are evaluated separately.

Comparisons. The DAVIS guard-MLLM and full-bank MLLM arms each generate once per case. Checks of their outputs reuse those frozen answers; Field candidates are evaluated separately under the same check. We evaluate two relevance controls. Offline relevance filters reuse saved full-bank answers (Appendix W). The same-stage relevance selector chooses records before generation with parameters locked on DAVIS-100 and is crossed with MLLM generation and Field readout (Appendix A.5.1). The latter matches observations and stage; selected sets, coverage and input-token counts remain method-dependent. The candidate-conditioned HGB is a post-hoc diagnostic control designed after DAVIS-56 inspection, with development-only threshold calibration (Appendix A.8). The method and candidate paths were frozen before the MyVLM output-held-out audit, although that source had appeared in earlier authorizer work.

Endpoints and uncertainty. *Card exposure* counts an unauthorized record in the selected context. *Record release* counts a selected, non-abstaining answer that matches a supplied record. *Correct record answers* match the unchanged task target on positive cases; *unsafe use* matches supplied records with no attribution-eligible match. Other record-task errors include eligible matches that fail the original field or NULL target. We report correctness and release alongside that narrower endpoint, rather than treating unsafe use as every $P/E/S$ failure. Unmatched non-NULL strings count toward response coverage but not record release (Appendix F).

DAVIS rates use 560 cases, with 392 positive cases and 56 clean cases. RecordAuth-Diag clean recall uses both clean conditions. Gate-level positive recall and NULL specificity remain distinct from generator accuracy. Paired bootstrap intervals resample query identities or tracks; source-video resampling checks DAVIS dependence. The 19-track video-disjoint analysis and the context-aligned support variant are post-hoc diagnostics. None of these rates is a calibrated population-risk certificate.

Compute. The DAVIS-56 guard uses 560 MLLM calls, 0.285M input tokens, 300.6 s serialized generation and 16.76 GiB peak GPU memory. Full-bank generation uses the same 560 calls, 2.032M input tokens and 539.0 s. Thus the guard uses $7.1\times$ fewer input tokens, not $7.1\times$ less total computation. Dual-path policies consume both arms: 1,120 calls, 2.318M input tokens and 839.6 s, with 18.65 GiB peak memory. CPU replay with saved embeddings takes 4.0–14.1 s for 560 cases and excludes image encoding.

X.2 Comparison and transfer scope

Diagnostic scope. The interventions establish controlled inducibility and supplied-record dependence, not natural prevalence or realized harm. Public-image pretraining overlap and interface differences limit cross-model interpretation. Free-form judgments remain local sensitivity measurements until human annotations are collected.

Comparison scope. Offline relevance compares filtering a frozen full-bank answer with generation from guarded context. The same-stage relevance control instead selects before generation and is crossed with both readouts. Its development-selected operating point differs in coverage and context

size from the guard. The main feature-role analysis is frozen input-block occlusion: checkpoints, training-fold transformations, thresholds, and ensemble rules remain unchanged. Separately retrained feature ablations use validation-selected thresholds (Table 32). These analyses evaluate sensitivity of a fixed policy and performance after adaptation, respectively. For offline filtering, nineteen full-bank cases with scored memory answers lack a usable candidate-conditioned relevance score and fail closed: 17 lack a record image, one lacks a unique trace, and one lacks localized query evidence.

Observation and dependence. Local visual decisions condition on supplied anchors and, in localized rows, masks. Fixed anchors make the whole-group-swap control different from the unanchored symmetry argument. Relative-score limits hold in their stated observation class. DAVIS confirmation is not video-disjoint; the video-level and 19-track analyses are sensitivity checks.

Transfer. The evaluated interfaces exclude proprietary VLMs. MyVLM and Idefics3 do not pass the specified capability gates, and all-negative banks provide no generator-level repair gain. These outcomes delimit the supported settings. Record authorization is a machine decision contract, not identity authentication or human consent.

Y Reproducibility contract

Access and result mapping. The anonymous supplementary artifact versions inputs, predictions, code and analyses. Endpoints record expected case counts, unique keys, source manifests, and model/authorizer hashes. A table-to-artifact manifest links paper rows to canonical summaries and SHA-256 values; the exporter rejects incomplete outputs. Historical registries are immutable, and additive records preserve their original lookup identifiers independently of the appendix organization.

Models and inference. The frozen revisions are Qwen3-VL-8B-Instruct `0c351dd0`, CoViP-Qwen3-VL-8B-GSPO `2283a306`, Phi-3.5-Vision `12b77fb4`, Gemma-3-4B-IT `4b66b5f8`, SigLIP2-base-patch16-224 `75de2d55`, and DINOv2-base `f9e44c81`. RAP-Phi3-mini, RAP source and its CLIP retriever use `0264b32f`, `3c0c9c2e` and `ce19dc91`. Full revisions, licenses and metadata hashes are in the registry. Qwen/CoViP and Gemma use 3,136–100,352 pixels per image; Phi uses one crop with SDPA. Core generation is greedy with at most 48 new tokens; RAP uses 32. The post-generation SigLIP2 encoder uses L2-normalized 768-dimensional vision pooler outputs, 224^2 RGB input and channel mean/std 0.5. Localized DAVIS crops set masked-out pixels to RGB 127 before encoding.

Data, decisions and training. Manifests preserve each source identity, case, intervention, expected answer, authorized record set and observation channel. The package includes frozen embeddings, 25 checkpoints per encoder, validation grids, ensembled decisions and raw outputs. Appendix I gives the actual training targets, loss and decision order; Appendix Q specifies the training-free MLLM action formats. Serialization includes group/card indices, opaque subject labels, timestamps, active/stale state, record text and images or explicit missing-image markers.

Analyses and retained failures. The evidence map indexes raw diagnosis and nonce attribution, learned gates, same-MLLM verifiers, masked DAVIS confirmation, fixed-candidate controls, representation probes, reference corruption and free-form sensitivity tests, with candidate paths, resampling units, post-hoc status and no-new-call replays. The artifact retains failed prompts, incomplete attempts, hardware checks and the excluded pre-ensemble guard. All-role and source-video audits address distinct dependencies; failed official-judge access, local judgments and pending human annotations remain documented separately.

Analysis protocol. CPU-only reanalyses retain frozen predictions and cover oracle eligibility (full-case versus clean-only), paired relevance errors with track/video intervals, capacity-matched Flat-MLP intervals, native endpoints and post-hoc context consistency. Scripts, commands, per-case decisions, input hashes and test logs accompany the revision. The stage-control extension adds deterministic readout on saved selections and 560 new same-stage relevance generations, plus 30 guard replays for environment consistency. Relevance parameters are locked on DAVIS-100 development data before new test inference. No model is retrained; historical registries and sources

remain unchanged. The subsequent pipeline analysis substitutes Field candidates under the existing support/margin rule, adds track- and source-video-cluster paired intervals, and audits set membership. It uses no new MLLM calls and does not retune thresholds.

Z LOCKED PIPELINE TRANSFER AND IMPLEMENTATION AUDITS

Z.1 PROSPECTIVE NEW-SOURCE PROTOCOL

The complete Guard–Field–check composition was locked on 17 September 2026, before any VOT predictions, in a timestamped local protocol with SHA-256 hashes. This confirmation uses the official VOT-ST2021 image sequences and dense masks (Kristan et al., 2021), a source not previously used in method development. The protocol orders the 60 sequences by a fixed hash, allows at most 40 eligible sequences and requires at least 24. The existing geometric and temporal rules retain 31; all 29 exclusions have fewer than eight eligible temporal views. Each sequence supplies up to ten quantile-spread frames, with the earliest as anchor, the latest as query and intermediate frames as record views. Crops use the original 15% padding. The area, box-side and border constraints match the DAVIS construction. The ten task interventions and expected-answer rules are unchanged. Distractors are selected deterministically within the new source pool, rather than DAVIS semantic-category pools.

The image audit compares selected crops and their full source frames with 16,537 previously used images and DAVIS frames. There are no SHA-256 matches or perceptual hashes within Hamming distance four; the minimum observed distance is ten. Sequence identity is the object-track proxy. This audit establishes absence from the examined development assets, not an identity or pretraining certificate. All 310 task targets also agree with the explicit support-reference contract. Public-source images remain subject to their original licenses.

The protocol retains all 25 ATRA checkpoints, saved standardizers and thresholds, the answer-vote ensemble, the deterministic triad intersection, and the 0.02 candidate margin. The relevance control reuses the DAVIS-100 settings $(0.6, 0.02, 0.6, 8)$ for group score, group gap, card score and k. Full-crop features drive both selectors; neutral-background instance masks are used only by candidate checks. Frozen Qwen3-VL-8B uses the original greedy 48-token, hidden-anchor, natural-prompt setting. There are 930 new generated outputs across raw, guard and relevance contexts, plus 30 pre-existing DAVIS cases replayed for environment verification; all 30 predictions match. No threshold, checkpoint, exclusion rule or policy is selected using VOT outcomes.

Table 57: Complete locked VOT-310 policy counts (217 positive tasks). All arms share the same query cases; checks retain the original 0.02 threshold.

Policy	Released	Correct	Unsafe	Other	Invalid
Full bank + MLLM	254	91	83	80	163
Relevance + MLLM	229	155	18	56	74
Guard + MLLM	187	154	3	30	33
Relevance + Field	265	173	68	24	92
Guard + Field	190	**180**	6	4	10
Guard + MLLM + check	153	152	**0**	**1**	**1**
Guard + Field + check	184	**180**	2	2	4

Guard and relevance retain nonempty contexts in 190 and 265 cases, respectively. They retain true support in 181 and 201 of 217 positive tasks, while exposing an ineligible card in six and 68 contexts. Among the guard's retained-target cases, MLLM and Field answer correctly in 154/181 and 180/181 cases. The remaining 36 positive cases contain 34 empty contexts and two nonempty contexts with the true latest event deleted. Both latter cases pass the Field check and supply its two attribution-eligible task errors. The other two residual releases are subject-absent. Candidate checking rejects four unsafe Field candidates by margin and two ambiguous-query candidates by unavailable localization, losing no correct Field answer. The MLLM check rejects 28 candidates on local recency, three on margin and two on localization; abstention/nonunique trace accounts for 124 cases.

Intervals resample all ten conditions of each query sequence together. The same constructed bank pool supplies distractors across queries, so the intervals are conditional on those banks. The signed changes preserve the utility–risk trade-off; a zero observed unsafe count in 310 cases is not a zero-risk guarantee.

Table 58: Paired VOT differences with 95% query-sequence bootstrap intervals (percentage points; 10,000 draws). Correct recall uses positive tasks; unsafe and task-invalid releases use all cases.

Contrast	Correct recall	Unsafe	Task-invalid
Checked Field − checked MLLM	$12.90\ [4.61, 22.58]$	$0.65\ [0.00, 1.61]$	$0.97\ [0.00, 2.26]$
Checked Field − full bank	$41.01\ [25.81, 54.84]$	$-26.13\ [-31.29, -21.29]$	$-51.29\ [-57.42, -45.16]$
Checked Field − Field Guard − relevance, Field readout	$0.00\ [0.00, 0.00]$	$-1.29\ [-2.58, -0.32]$	$-1.94\ [-3.55, -0.65]$
	$3.23\ [-6.91, 11.98]$	$-20.00\ [-24.19, -15.48]$	$-26.45\ [-31.29, -21.61]$

Z.2 Frozen DAVIS margin intervention

The rival score in Eq. 3 includes query–rival-anchor similarity. That term can suppress a rival even when the candidate image matches its anchor. We remove that term from rival scores only, holding the claimed score, 315 Field candidates, selected sets, localization rules and 0.02 threshold fixed. Of 313 candidates reaching the margin stage, the original rule accepts 289 correct and six unsafe answers. The intervention accepts 284 correct and three unsafe answers. It changes three unsafe decisions (two assignment interventions on one track and one absent-subject case), together with five correct decisions. The three remaining unsafe acceptances are absent-subject cases. This diagnostic isolates a score dependency; the VOT confirmation retains the original rule.

Table 59: Fixed DAVIS candidates under a rival-score intervention. Localization and support are unchanged; no thresholds are retuned.

Rival score	Correct accepted	Unsafe accepted	Other accepted
$\max_{h \neq g} \min\{\mathrm{sim}(q, r), \mathrm{sim}(q, a_h), \mathrm{sim}(r, a_h)\}$	289	6	0
$\max_{h \neq g} \min\{\mathrm{sim}(q, r), \mathrm{sim}(r, a_h)\}$	284	3	0

Z.3 Answer-vote and run-order dependence

The saved ensemble votes on each run's latest-field answer and takes the first winning run's card set. On DAVIS-560, 37 cases have multiple ATRA card sets behind the same winning answer. Reversing run order changes the final guard set in two cases but changes neither Field nor checked-Field outcomes. Changing only the vote proxy from latest to earliest retained field changes 21 final sets and the resulting latest-field outputs (Table 60). Thus the implementation has an explicit task-schema dependency even though it operates before MLLM generation. Expected answers and attribution labels are used only for the audit, not for either vote.

Table 60: Frozen DAVIS ensemble interventions. All per-run groups, card actions, weights and thresholds are fixed. Outputs use the original latest-field task.

Ensemble rule	Changed K	Field correct	Field unsafe	Checked correct	Checked unsafe
Original latest-field vote	0	303	10	289	6
Reverse run order	2	303	10	289	6
Earliest-field vote proxy	21	293	21	280	5

Deterministic group selection. For each group, the anchor–triad selector averages its top two $\min\{\mathrm{sim}(q, m), \mathrm{sim}(a_j, m)\}$ values and takes the minimum with $\mathrm{sim}(q, a_j)$. It selects the highest-scoring group, breaking score ties by group index, and applies its saved score/gap thresholds. Only active, image-backed cards passing the saved relation threshold are retained. This defines the deterministic group entering the ensemble intersection.

Z.4 Complete DAVIS candidate-policy comparison

On DAVIS, Field-minus-MLLM correct recall changes by 19.13 points ($[12.24, 27.30]$) and unsafe use by 0.89 points ($[0.00, 2.14]$). After checking, Field-minus-MLLM changes are 17.86 recall

points ($[11.48, 25.26]$) and 0.54 unsafe points ($[0.00, 1.61]$). These 95% intervals use the original track-cluster resampling.

Table 61 compares retained-event filtering, legacy and context-aligned checks, dual paths and candidate-restricted oracles on the same frozen DAVIS predictions. Field outputs are outside the raw/guard MLLM oracle union. The legacy full-bank recency rule rejects valid older fallbacks when an invalid newer record remains present; context alignment changes the comparison set.

Table 61: Record releases on DAVIS-560. Other denotes task-invalid matches outside unsafe use. †: post-hoc context alignment; ‡: new Field candidates under unchanged checks.

Policy / observation	MLLM calls	Released /560	Correct /392	Unsafe	Other	Invalid / released (%)
Full-bank generation	560	454	170	149	135	62.56
Raw + retained-event filter	560	246	144	6	96	41.46
Guard + MLLM	560	311	228	5	78	26.69
Guard + MLLM + legacy check	560	154	151	**3**	**0**	1.95
Guard + MLLM + aligned check†	560	222	219	**3**	**0**	1.35
Guard → Field → check‡	0	295	**289**	6	**0**	2.03
Legacy localized dual	1,120	177	174	**3**	**0**	1.69
Context-aligned dual†	1,120	245	242	**3**	**0**	**1.22**
Supplied-token control	1,120	203	193	0	10	4.93
Oracle over raw/guard MLLM candidates	1,120	278	278	0	0	0.00

Reproduction. The prospective locks, construction scripts, source CRC/SHA-256 checks, per-case predictions and paired statistics are indexed in `artifacts/recordauth_pipeline_confirmation_20260917/artifact_manifest.json`.