AUDITING ROLE COLLISIONS IN MULTIMODAL MEMORY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal memory evaluation assesses how retrieved evidence supports question answering. However, scores also depend on how this evidence is presented and how responses are generated and scored. We audit a public VisualMem answering path and identify missing explicit role and option bindings, a condition we term a role collision. We introduce role-preserving serialization (RPS) to express these bindings while keeping retrieval outputs fixed. Across four model stacks and all 58 eligible image-choice questions with balanced answer positions, RPS improves accuracy by up to 34.9 percentage points over source-derived serialization, with no gain on Molmo2. To examine how correct bindings affect scores, we separate mapping placement from role wording in a paired 2 × 2 factorial design. On Qwen3.5-9B, raising the output-token limit from 128 to 512 reduces the adjacent-over-global advantages from 25.86 to 5.60 points under neutral wording and from 19.83 to 8.62 under semantic wording; only the neutral reduction passes the expanded Holm correction. By contrast, Gemma and InternVL retain the same answer labels across 64/128/256/512-token runs, while semantic wording lowers InternVL's global-placement accuracy by 13.79 points. For scoring validation, independent human extraction agrees with the frozen parser on 193/200 archived Qwen outputs. Finally, serialization changes retrieval-method contrasts on 696 questions, while Retrieval-only remains best overall. On 120 MuirBench reference-injection questions, RPS exceeds flat injection by 6.67–27.50 points; these shifts and gains do not pass expanded Holm corrections. Together, these findings show why bindings, output-token limits, and answer-extraction rules must be specified before attributing score differences to retrieval quality.

1 INTRODUCTION

Personal-memory assistants retrieve information from earlier interactions to answer later questions. Recent systems study how to organize memories, retrieve relevant evidence, and reason over stored content (Chhikara et al., 2025; Wang and Chen, 2025; Nguyen et al., 2026; Long et al., 2026). Their evaluations commonly compare final answers. However, retrieved evidence must first be converted into a model-facing request. A measured score can therefore reflect both the memory system and the interface used to present its evidence, generate a response, and extract an answer.

This distinction matters when reference images and image-valued answer choices share one request. With one reference followed by four candidates labeled A–D, choice A occupies the second image position. Additional references change this offset. As Figure 1(a) illustrates, the dataset can define answer identities correctly while the serializer omits their explicit mapping. We term this omission of task-role and option bindings a **role collision**. It describes an observable interface condition, not an inferred internal confusion between references and answers.

Correcting this omission requires distinguishing valid binding from how that binding is expressed. Prior work examines multi-image position sensitivity and image separation (Tian et al., 2025; Lee et al., 2026). Answer-extraction studies further show that scores depend on how generated text becomes an answer label (Wang et al., 2024; Molfese et al., 2025; Yu et al., 2025). Thus, a gain after adding binding text does not by itself demonstrate better memory use. Evaluation must therefore also consider mapping placement, output-token limits, and answer extraction.

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

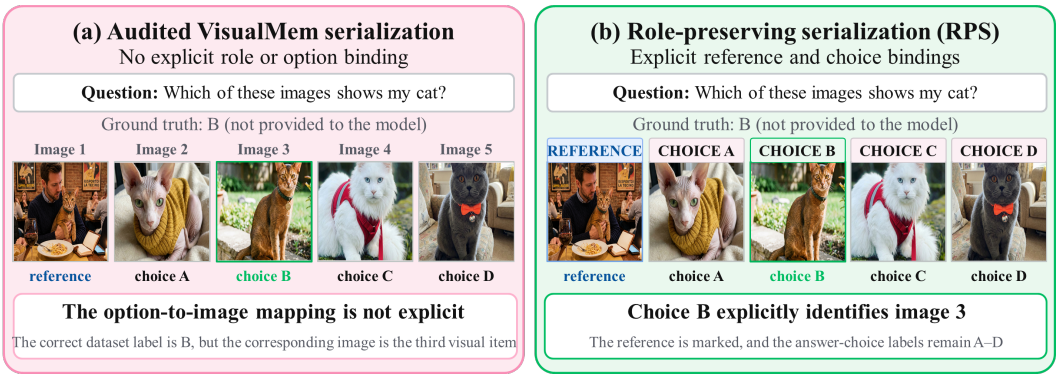

Figure 1: Missing role–option bindings and RPS. Both requests contain one reference and four answer images in the same order. (a) The audited serialization omits explicit bindings. (b) RPS identifies the reference and binds A–D to their images. The correct candidate remains the third image. This illustrates request construction, not an observed prediction or guaranteed answer.

Motivated by this distinction, we audit a pinned public VisualMem answering path (Nguyen et al., 2026). **Role-preserving serialization (RPS)** inserts text that distinguishes references from candidates and binds answer labels to images, without changing retrieval outputs or image order (Figure 1(b)). A **paired 2 × 2 factorial design** then compares mapping placement (global vs. adjacent) and role wording (neutral vs. semantic) on all 58 eligible image-choice questions. Four cyclic option orders place each correct candidate once at A–D.

The results reveal model-dependent effects: Qwen's placement advantages decrease across independently generated 64/128/256/512-token runs, whereas Gemma and InternVL retain their labels across the same four limits. Molmo2 does not reproduce the archived RPS gain. Scoring checks, retrieval comparisons, and external injection delimit these findings without assuming that binding repair guarantees useful memory.

The main contributions of this paper are summarized as follows.

- **We formalize a binding omission** and evaluate RPS while preserving retrieved content and ordered images.
- **We separate mapping placement and role wording** on 58 questions under four label-balanced option orders.
- **We test placement–budget interactions across models**, verifying whether native processors preserve each intervention.
- **We assess scoring and downstream consequences** with a 200-output human audit, four-model retrieval comparisons, and an external reference-injection cohort.

2 Related Work

2.1 Multimodal Memory and Evaluation

Memory systems organize earlier evidence for later questions (Chhikara et al., 2025; Wang and Chen, 2025), while M3-Agent combines persistent multimodal memory and reasoning (Long et al., 2026). VisualMem, MemEye, Mem-Gallery, and MemLens provide complementary evaluation settings (Nguyen et al., 2026; Guo et al., 2026; Bei et al., 2026; Ren et al., 2026). Rather than optimize retrieval, we ask whether the answering interface preserves image roles and identities.

2.2 Multi-Image Position, Separation, and Role Binding

SoFA modifies attention to address multi-image position bias (Tian et al., 2025); Lee et al. (2026) modify image delimiters. Structural perturbations test conversational roles and image placement (Shayegani et al., 2026). VisioMath permutes textual mappings while retaining image order and remapping gold answers, and studies visual–textual anchors (Li et al., 2026). We compare correct mapping placement and role wording, then test budget and retrieval consequences. Our roles

distinguish references, queries, and candidates, not conversational authority; our wrong-binding control leaves gold unchanged.

2.3 Answer Extraction and Evaluation Reliability

Generated answers can differ from first-token scores (Wang et al., 2024), and extraction policies can change multiple-choice performance (Molfese et al., 2025; Yu et al., 2025). We connect these issues to input binding and output-token limits, retaining unparsed outputs in accuracy denominators and reporting parse coverage. Independent human extraction separately checks the frozen parser.

3 Method

We audit evaluation validity using RPS under fixed retrieval. Figure 2 summarizes the pipeline and paired design: the binding requirement (Section 3.1), placement and wording of correct mappings (Section 3.2), and structural versus processor-level verification (Section 3.3).

3.1 Source Interface and Role-Preserving Serialization

For question i, let R_i denote the ordered retrieved reference images, Q_i an optional query image, and V_i the ordered candidate images. The audited VisualMem adapter emits the following source-content sequence:

$$S_i^{\text{off}} = \left[\text{instruction}, R_i^{(\leq 10)}, Q_i, V_i, \text{question/context}\right]. \quad (1)$$

The ten-image cap applies only to references; absent query blocks are omitted. The sequence preserves image order but lacks explicit reference and option identities. As Figure 1 illustrates, sequence position cannot substitute for answer identity.

Let L_i be valid option labels and $b_i: L_i \rightarrow V_i$ their intended mapping. RPS exposes b_i and distinguishes reference, query, and candidate roles. The archived implementation inserts REFERENCE ONLY, QUESTION IMAGE, or ANSWER CHOICE A before each corresponding image while retaining assets and order; its executable name remains typed. Adjacency is one implementation, not the validity requirement: correct global mappings, native identifiers, or structured fields can also preserve bindings.

3.2 Paired Mapping Placement and Role Wording

The paired 2 × 2 factorial design crosses mapping placement with role wording, yielding global-neutral (GN), global-semantic (GS), adjacent-neutral (AN), and adjacent-semantic (AS) (Figure 2(b)). Global placement lists mappings after all images and before the question/context; adjacent placement puts each mapping before its image. Within each wording level and question–order pair, statements, ordered images, and correct assignments remain fixed.

Neutral statements use Image 1 | ITEM R1. or Image 2 | ITEM A. Semantic wording replaces ITEM with REFERENCE, QUESTION, or ANSWER. Both retain R/Q/A–D identifiers, group wrappers, query, and output contract; neutral wording therefore retains role cues. The additional unlabeled anchor retains image order and output contract but omits groups and bindings. It lies outside the 2 × 2 design and is not archived Official.

Placement jointly changes distance, direction, absolute position, and recency; wording is not token-length-matched. These comparisons isolate neither attention mechanisms nor universally necessary role words.

3.3 Structural and Input Verification

ROLECHECK tests metadata and an adjacent-label policy, not understanding; valid global mappings may fail it. Of eleven inspected interfaces, three meet our scope: VisualMem lacks bindings, while M3Exam and Multimodal RewardBench 2 expose identities (Huang et al., 2026; Hu et al., 2025). This ledger does not estimate prevalence.

Source equality does not establish identical processor outputs. We record generation-input keys, shapes, dtypes, and hashes: visual tensors match across conditions, complete inputs match across budgets within each model, and Qwen, Gemma, and InternVL preserve the intended placement.

Molmo2 puts images before mapping labels, so its factorial placement effect is unavailable (NA), not zero. These checks do not fill historical recording gaps.

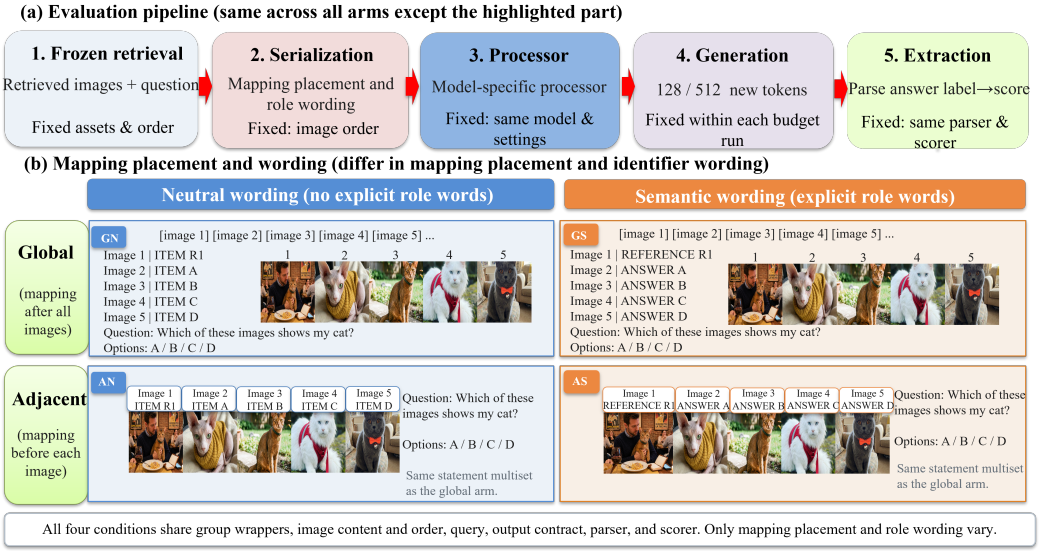

Figure 2: Evaluation pipeline and paired serialization design. GN/GS and AN/AS contrast global and adjacent placement under neutral/semantic wording. Correct assignments, mapping statements within each wording level, and ordered images remain fixed. The additional anchor (not shown) omits groups and labels. The diagram shows the 128/512 endpoints used for confirmatory tests; Qwen, Gemma, and InternVL also have independently generated 64/256-token runs reported descriptively in Figure 3. Molmo2 does not preserve adjacency. Prompts are schematic.

4 Experiments

We test binding, placement and wording, and output-token limits. Extraction checks scoring; retrieval comparisons and external injection assess consequences and scope (Table 1). Supplement A–K retains archived evidence, L describes the budget and external extension, and M completes binding and retrieval coverage.

4.1 Experimental Setup

Data and source. We pin VisualMem's answering source to 3fa332b1 and data to 1f4950f4 (Nguyen et al., 2026). Official denotes Equation (1), not the proprietary endpoint or its scores. The 696 questions span ten synthetic personas: 58 text-query/image-choice and 638 text-choice questions. F58 includes all 58 eligible tasks with fixed Full retrieval. H57 remains separate: among 57 shared tasks, 53 contexts and 42 reference-image sequences differ.

Table 1: Experimental coverage and units. Q/G/M/I denote Qwen/Gemma/Molmo2/InternVL. Binding and retrieval comparisons use 128 tokens; Molmo2 is excluded from the adjacency-preserving factorial. Repeated option orders, models, and budgets add no independent base questions.

Study	Units / protocol	Models and coverage
Counterbalanced binding	58 bases × 4 orders; 232/model; 128 tokens	Four models; Q/G/M archived, InternVL new
Paired factorial	58 × 4 orders × 5 arms / budget	Qwen, Gemma, InternVL 64/128/256/512
Processor check	Native templates and actual inputs	Molmo adjacency not preserved (NA)
Human extraction	200 archived endpoint outputs; 20/cell	Qwen only; separate archived sample
Retrieval comparison	696 tasks × 3 methods; 128 tokens	Four models; G/I archived, Q/M new
External injection	120 new questions; 5 tasks; 512	Qwen, Gemma, Molmo, InternVL

Controls and models. Paired serializers fix checkpoints, ordered images, context, decoding, and scoring. Four cyclic orders update candidates and gold together; each factorial matrix has 1,160

generations. Qwen3.5-9B, Gemma 3-4B, and InternVL3.5-8B-HF use separately generated 64/128/256/512-token runs, with budgets matched within model on RTX 4090 (Qwen Team, 2026; Gemma Team, 2025; Wang et al., 2025b). Molmo2-8B contributes binding, retrieval, and external evidence (Clark et al., 2026). New InternVL binding and Qwen/Molmo2 retrieval complete legacy 128-token coverage. Stack comparisons isolate neither architecture nor independent language lineages.

Scoring and estimands. The frozen parser accepts explicit answer formats, not isolated letters. Unparsed successful generations count as incorrect in the fixed denominator; technical failures remain separate. With $A_{a,b}$ denoting accuracy for condition a at output-token limit b, the placement effects are

$$D_1(b) = A_{\mathrm{AN},b} - A_{\mathrm{GN},b}, \qquad D_2(b) = A_{\mathrm{AS},b} - A_{\mathrm{GS},b}, \tag{2}$$

The wording effects and paired budget changes are

$$D_3(b) = A_{\mathrm{GS},b} - A_{\mathrm{GN},b}, \quad D_4(b) = A_{\mathrm{AS},b} - A_{\mathrm{AN},b}, \quad \Psi_k = D_k(512) - D_k(128). \tag{3}$$

Equation (3) measures placement–budget interactions for $k = 1,2$ and wording–budget interactions for $k = 3,4$. Effects use percentage points (pp). We average four orders within each base and weight bases equally. Pointwise 95% intervals use 50,000 whole-persona bootstrap draws; exact sign flips enumerate 1,024 assignments under symmetric cluster effects. Expanded Holm families retain 16 simple-effect slots per endpoint and eight each for placement and wording interactions, including unsupported Molmo2 slots. Intervals do not invert these tests. Intermediate budgets are descriptive; archived corrections remain separate (Supplement L.3).

4.2 Effects of Restoring the Binding

Correct versus missing or wrong binding. In the archived native-order F58 comparison, Qwen improves from 11/58 under Official to 34/58 under RPS, with 25 improvements and two regressions ($p = 5.65 \times 10^{-6}$; persona interval [25.0, 55.6] pp). Reassigning the same binding strings to wrong images reduces the score to 2/58. Correct assignment matters beyond added text, without identifying internal image selection.

Table 2: Counterbalanced binding at 128 tokens, correct/232; change is RPS−Official. Qwen/Gemma/Molmo2 are archived; InternVL is new. Intervals cluster by persona; raw sign-flip and Holm4 p are separate. Pale red/bold marks larger counts, not significance.

Model	Official	RPS	Change (pp)	95% interval	Raw p	Holm4 p
Qwen3.5-9B	54	**135**	+34.9	[25.0, 45.0]	0.00195	0.00781
Gemma 3-4B	63	**106**	+18.5	[10.3, 27.1]	0.00586	0.01172
Molmo2-8B	**63**	60	-1.3	[-5.0, 2.6]	0.6875	0.6875
InternVL3.5-8B-HF	65	**124**	+25.4	[17.4, 33.2]	0.00391	0.01172

Answer-position control. Qwen, Gemma, and InternVL retain positive RPS effects after balancing the correct label across A–D; all three pass Holm4 (Table 2). Native B/C concentration is therefore not their sole explanation. Molmo2 retains a small negative difference. These joint serialization effects differ from the factorial placement effects, and four orders do not establish invariance to all 24 permutations.

Parsing and components. An archived H57 answer-first comparison parses 55/57 outputs per condition while correct counts rise from 3 to 35. Equal coverage cannot explain the whole difference, although parsed subsets may differ. Adding neutral identities beyond grouping is unresolved for Gemma but passes correction for InternVL, motivating the factorial rather than assuming semantic words always help.

4.3 Mapping Placement, Role Wording, and Generation Budget

To test whether correct-map placement depends on the output-token limit, we compare complete within-model matrices. Table 3 reports the 128/512 confirmatory endpoints; Figure 3 shows independently generated Qwen/Gemma/InternVL runs at all four limits. Qwen improves in every condition as the limit increases, but its global mappings gain more than adjacent mappings.

Table 3: Factorial endpoints at 128/512 tokens. C;P denotes correct;parsed, each out of 232. Qwen, Gemma, and InternVL use their matched four-budget executions. Molmo2 adjacency is NA, not zero. Pale red/bold marks maximal C across the five conditions within each model/budget, including ties.

Arm	Qwen 128	Qwen 512	Gemma 128	Gemma 512	InternVL 128	InternVL 512
Anchor	39;100	84;198	65;232	65;232	53;224	53;224
GN	69;101	138;219	64;232	64;232	102;224	102;224
GS	99;160	132;222	63;232	63;232	70;230	70;230
AN	129;184	151;222	**80;232**	**80;232**	**116;232**	**116;232**
AS	**145;219**	**152;230**	76;232	76;232	111;232	111;232

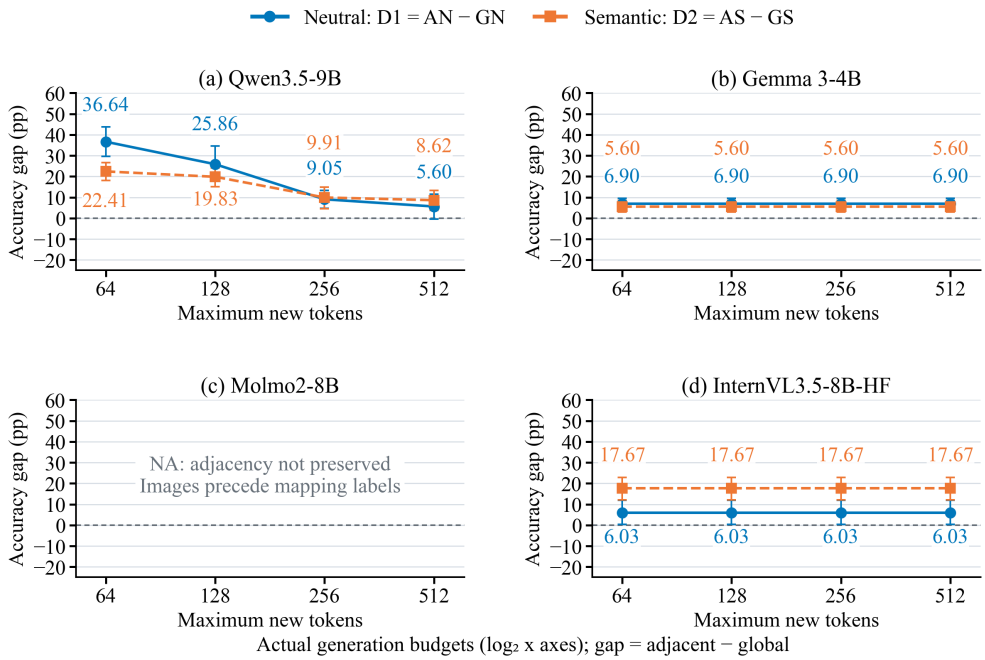

Figure 3: Placement effects with pointwise 95% persona intervals. Qwen, Gemma, and InternVL each use actual 64/128/256/512-token runs; budgets are matched within model, and all recorded runs used RTX 4090. Lines connect observations, not interpolated experiments; Molmo2 adjacency is NA. Only the 128/512 endpoint effects enter Holm16; their placement–budget interactions enter separate Holm8 families, while 64/256 are descriptive.

Placement–budget interaction. Qwen's neutral-wording position advantage falls from 25.86 to 5.60 pp, and its semantic-wording advantage falls from 19.83 to 8.62 pp. The corresponding changes are −20.26 pp (Holm8 p=.03125) and −11.21 pp (p=.109375). Only the neutral reduction passes the expanded correction. Neither remaining placement effect at 512 tokens passes Holm16 (p=1.0 and p=.28125). The estimates describe attenuation, not disappearance or equivalence. Original Qwen-only tests from the earlier runtime are retained separately in Supplement K.

Wording and model dependence. Gemma's placement effects remain +6.90/+5.60 pp and InternVL's +6.03/+17.67 pp across all four limits. Every primary answer label remains unchanged, giving zero observed budget interactions without establishing population equivalence or a universally sufficient budget. Qwen's positive 128-token wording estimates do not pass expanded Holm16. Gemma's wording estimates are small and negative. Semantic wording lowers InternVL's global-placement accuracy by 13.79 pp at both endpoints (Holm16 $p = .0312$ each); its adjacent effect is −2.16 pp and unresolved. No wording–budget interaction passes Holm8. Correct semantic labels therefore do not necessarily improve measured accuracy.

4.4 Answer Availability and Extraction Reliability

When answers become available. All 1,160 comparable Qwen token sequences match as prefixes for every tested budget pair. In the 512-token runs, complete literal FINAL=<A>–FINAL=<D> fields first appear after token 128 in 99/120/62/38/11 Anchor/GN/GS/AN/AS outputs. Fewer late declarations under adjacent placement are consistent with an answer-availability contribution, not causal mediation: extra tokens also permit computation, and explicit answers can be wrong.

Extraction-rule sensitivity. Six frozen rules cover all 13,920 factorial outputs (Supplement S45). Qwen's neutral placement effect spans 25.86–27.16/5.60–6.47 pp at 128/512 tokens, with a 20.26–21.12 pp reduction. InternVL's strict-closed rule rejects 314 primary labels per budget, including 125 correct answers; its neutral effect spans +2.59–+6.03 pp and global wording remains negative. Gemma's labels agree across rules. Gemma and InternVL retain zero budget interactions under every rule. The primary scorer remains unchanged; prefix diagnostics are separate from budget runs.

Qwen's 512-token limit is not established as sufficient. Sixty unparsed outputs reach it with unknown stop causes, while twelve below-cap Anchor outputs select E/F. Ten responses contain conflicting declarations; two unclosed fields are accepted by the parser. These risks remain separate from the random human sample, and E/F alone does not establish internal reference selection.

Table 4: Independent human extraction on the fixed 200-output archived endpoint sample. Proportion intervals are Wilson 95% intervals; the κ interval uses 10,000 paired-output bootstrap resamples. Blue rows mark parser validation. Extraction agreement is not task accuracy.

Quantity	Estimate	95% interval
Two-rater label agreement	194/200 (.970)	[.9361, .9862]
Unweighted Cohen's κ	.961971	[.929761, .987367]
Conflict-flag agreement	198/200 (.990)	[.9643, .9973]
Parser–adjudicated label agreement	193/200 (.965)	[.9295, .9829]
Parser–human correctness agreement	199/200 (.995)	[.9722, .9991]
Human-extracted task accuracy	96/200 (.480)	[.4118, .5490]
Frozen-parser task accuracy	95/200 (.475)	[.4069, .5440]

Independent human check. The unchanged Qwen sample contains twenty outputs in each of the ten archived endpoint budget–condition cells (128/512 × five conditions), selected by hash without using response text, correctness, parsing, or length. Two annotators independently extract A/B/C/D/NONE/AMBIGUOUS and flag explanation conflicts, blind to gold, parser predictions, and conditions. After both passes are locked, a third annotator resolves six label and two distinct conflict-flag disagreements.

The parser agrees with adjudicated extraction on 193/200 labels (Table 4). Seven differences change correctness for one output: human extraction gives 96 correct answers, the parser 95. This supports extraction on the archived endpoint sample, not the fresh matched-runtime generations. Twenty outputs per endpoint cell cannot re-test the interactions; Supplement K.9–K.12 retains full linkage.

4.5 Consequences for Retrieval-Method Comparisons

We next ask whether serialization changes comparisons between retrieval methods. Naive, Retrieval-only, and Full provide fixed evidence for the same 696 questions. Each method is re-answered under Official and RPS at 128 tokens, using archived Gemma/InternVL and new Qwen/Molmo2 runs. Table 5 reports actual correct counts. For Full and Naive, we combine all four matched correctness outcomes:

$$\Gamma = \frac{1}{N}\sum_i \left[\left(Y_{i,F,\mathrm{RPS}} - Y_{i,N,\mathrm{RPS}}\right) - \left(Y_{i,F,\mathrm{off}} - Y_{i,N,\mathrm{off}}\right)\right]. \quad (4)$$

Figure 4 also reports Retrieval-only−Naive and Full−Retrieval-only, which sum to Full−Naive and are not independent replications. Supplement S51 supplies All/Image/Text correct and parsed counts; S52 reports all 36 contrasts and correction families.

Observed comparison shifts. Full–Naive shifts on All/Image are +0.29/+18.97 pp for Qwen, +1.58/+18.97 for Gemma, −0.14/+1.72 for Molmo2, and +3.30/+31.03 for InternVL. None of the eight primary or sixteen secondary interactions passes Holm correction. Retrieval-only remains highest overall for every model and serializer, although image-choice rankings can differ. A changed method difference does not necessarily change its winner.

Binding repair does not ensure useful references. Archived full-input versus choice-only comparisons favor full input for Gemma, remain unresolved for Qwen, and favor choice-only for Molmo2. Input validity and memory utility are separate questions.

Table 5: Correct counts at 128 tokens for all 696 questions and 58 image-choice questions. Pale red/bold marks the maximum across three methods within each model, serializer, and subset, including ties, not significance.

Model / method	Official /696	RPS /696	Official /58	RPS /58
Qwen · Naive	276	293	**15**	26
Qwen · Retrieval-only	**289**	**298**	**15**	23
Qwen · Full	266	285	10	**32**
Gemma · Naive	178	177	**6**	5
Gemma · Retrieval-only	**193**	**192**	5	4
Gemma · Full	178	188	5	**15**
Molmo · Naive	215	215	**31**	**29**
Molmo · Retrieval-only	**234**	**227**	25	16
Molmo · Full	200	199	16	15
InternVL · Naive	186	186	**26**	32
InternVL · Retrieval-only	**197**	**203**	22	33
InternVL · Full	167	190	10	**34**

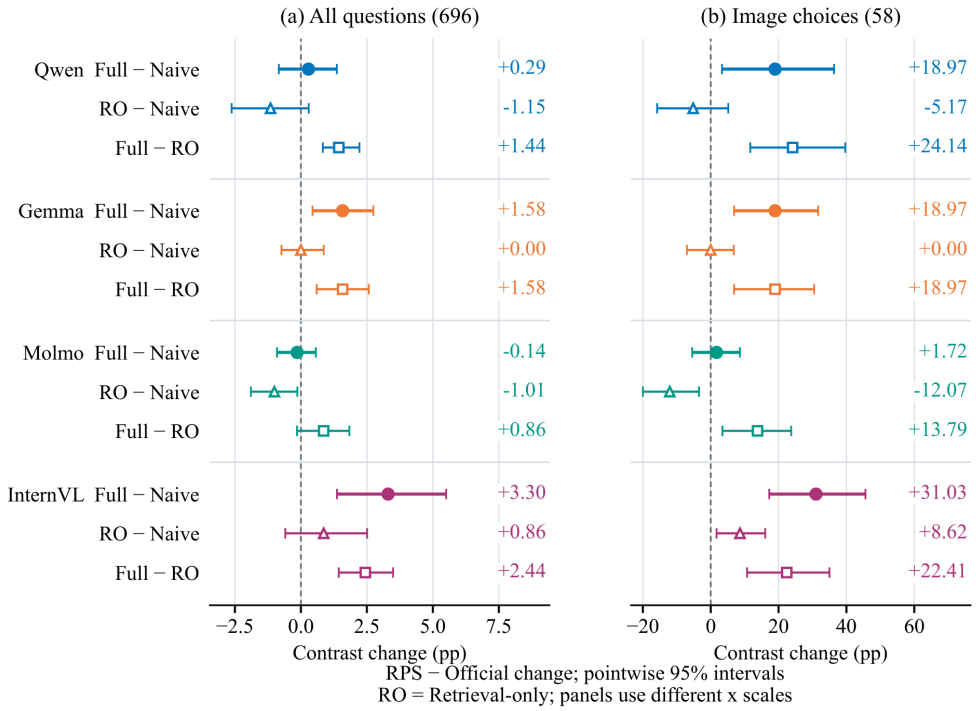

Figure 4: All 24 RPS−Official changes in method contrasts at 128 tokens. All (696) and Image (58) panels use different x scales; RO denotes Retrieval-only. Pointwise 95% persona-cluster intervals do not establish sign-flip significance. The algebraically dependent contrasts use primary Holm8 and secondary Holm16 families (Supplement S52).

4.6 Cross-Source Tests and Diagnostic Boundaries

Constructed reference injection. We test 120 frozen MuirBench questions across five tasks (Wang et al., 2025a). Flat and RPS share two reference distractors and target images; No-ref retains task-image identities without distractors. RPS exceeds Flat by 6.67–27.50 pp at 512 tokens (Table 6), but all Holm4 p values are .25. Five clusters impose minimum raw two-sided $p = .0625$; positive bootstrap intervals do not establish significance.

Table 6: MuirBench correct;parsed (C;P), N=120, five task clusters, 512 tokens. Intervals resample whole tasks; no test passes Holm4. No-ref is descriptive. Pale red/bold marks maximal C across conditions. The archived 40-question/32-token study remains separate.

Model	No-ref C;P	Flat C;P	RPS C;P	RPS - Flat, pp [95% CI]
Qwen3.5-9B	**53;115**	18;92	51;119	+27.50 [11.67, 43.33]
Gemma 3-4B	**51;120**	29;120	41;120	+10.00 [0.83, 20.83]
Molmo2-8B	**26;120**	11;119	19;120	+6.67 [1.67, 14.17]
InternVL3.5-8B-HF	23;37	7;29	**27;29**	+16.67 [9.17, 25.83]

Qwen's parsed count rises from 92 to 119. InternVL parses 29 outputs per injected condition, but the sets differ; 265/360 outputs across its conditions are unparsed at the cap. RPS remains below No-ref for three models. This constructed injection does not establish general memory utility or another native interface defect; Molmo2 uses its transformed template.

Archived diagnostics. A 74-question Gemma text-choice control scores 37 per serializer, with one win and one harm. Qwen improves from 13/240 to 240/240 synthetic answers as the limit rises from 128 to 512. These query-image fixtures differ from F58 and do not certify natural-task competence.

5 Discussion and Limitations

The answering interface belongs to the measurement specification. RPS supplies bindings, the output-token limit controls generation, and extraction converts responses into scores. A larger budget does not supply missing bindings; structural checks do not validate the parser. Retrieval methods are the upstream comparators. Evaluation should retain processor inputs, report fixed-denominator accuracy and parse coverage, and directly test method-contrast changes.

The natural factorial covers 58 tasks, ten synthetic personas, and one native interface. Models, budgets, and orders add no independent tasks; the repository ledger does not estimate prevalence. Coupled positional changes and extra-token computation/reporting prevent identifying a shared attention mechanism or universally sufficient budget. Human validation covers 200 archived Qwen endpoint outputs and does not validate the fresh four-budget generations. The external cohort has five task clusters and constructed distractors; exact-image checks cannot exclude all related sources or pretraining overlap. Historical provenance gaps remain explicit. This result-informed extension is neither preregistered nor an independent reproduction.

6 Conclusion

In this paper, we investigate how the answering interface affects multimodal memory evaluation under fixed retrieval. RPS makes missing roles and option identities explicit without changing image content or order. Paired comparisons distinguish binding from its placement and wording, and reveal model-dependent responses to output-token limits. Independent human extraction supports the frozen parser on the archived sample. Retrieval and external comparisons show observed shifts, but their expanded tests remain unresolved. Together, these findings support specifying bindings, processor inputs, output-token limits, and answer-extraction rules before attributing score differences to retrieval quality.

ETHICS STATEMENT

This study audits public evaluation interfaces and uses VisualMem's synthetic personas and released MuirBench data. It does not allege misconduct by benchmark authors or reproduce their proprietary-model scores. Figures 1–2 illustrate input structure and are not experimental observations. Two independent human annotators extract explicit answers, and a third adjudicates disagreements. The responsible author confirms independence, blinding, and compliance with applicable consent and ethics requirements. Names, contacts, consent records, and free-form notes remain private. Approved de-identified labels, evidence spans, adjudications, and scientific linkage are retained in the anonymous human-evidence materials. Third-party assets remain subject to their original access and redistribution conditions.

REPRODUCIBILITY STATEMENT

Supplementary Sections A–M document pinned interfaces, retrieval configurations, serialization, scoring, and uncertainty. The anonymous companion includes a compact executable ROLECHECK checker and synthetic example, the unchanged 200-output human linkage and approved anonymous evidence, and a portable derived-label statistical replay. This replay recomputes numerical results; it does not regenerate model responses or re-extract raw text. Sections L–M document matched four-budget execution identities, CPU reanalysis, completed legacy coverage, and exact original and Molmo-compatible inference runners with dependency and artifact hashes. These are not a self-contained inference release: frozen contexts, images, and historical dependencies must be obtained separately. Full author-side records retain input and token hashes. Historical gaps remain explicit, and same-team software checks are not independent human reproduction. Further releases remain subject to third-party rights and human privacy.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript preparation, specifically language polishing and editing of the text, as well as with organizing the code repository and assembling the final submission materials. The authors have reviewed all AI-assisted work and take full responsibility for the final text, code, results, and disclosures.

REFERENCES

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-Gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. arXiv:2601.03515, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. arXiv:2504.19413, 2025. doi:10.48550/arXiv.2504.19413.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv:2601.10611, 2026.

Gemma Team. Gemma 3 technical report. arXiv:2503.19786, 2025.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, and Ruixiang Tang. MemEye: A visual-centric evaluation framework for multimodal agent memory. arXiv:2605.15128, 2026.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal RewardBench 2: Evaluating omni reward models for interleaved text and image. arXiv:2512.16899, 2025.

Zhengjun Huang, Wenxuan Liu, Zhoujin Tian, Wei Chen, Junle Chen, Yuqian Wu, Fangyuan Zhang, Qintian Guo, and Xiaofang Zhou. M3Exam: Benchmarking multimodal memory for realistic user-agent interactions. arXiv:2606.07402, 2026.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In International Conference on Learning Representations, 2026.

Can Li, Ying Liu, Ting Zhang, Mei Wang, and Hua Huang. VisioMath: Benchmarking figure-based mathematical reasoning in LMMs. In *International Conference on Learning Representations*, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In International Conference on Learning Representations, 2026.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 18477–18494, 2025. doi:10.18653/v1/2025.findings-acl.950.

Viet Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal visual memory from explicit and implicit evidence. arXiv:2605.28806, 2026.

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

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, Zhengnan Li, Yangqiu Song, Ginny Wong, and Simon See. MemLens: Benchmarking multimodal long-term memory in large vision-language models. arXiv:2605.14906, 2026.

Erfan Shayegani, G. M. Shahariar, Sara Abdali, Lei Yu, Nael Abu-Ghazaleh, and Yue Dong. Misaligned roles, misplaced images: Structural input perturbations expose multimodal alignment blind spots. In International Conference on Learning Representations, 2026.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10599–10609, 2025.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A comprehensive benchmark for robust multi-image understanding. In International Conference on Learning Representations, 2025a.

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

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 7407–7416, 2024. doi:10.18653/v1/2024.findings-acl.441.

Yu Wang and Xi Chen. MIRIX: Multi-agent memory system for LLM-based agents. arXiv:2507.07957, 2025. doi:10.48550/arXiv.2507.07957.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In International Conference on Learning Representations, 2025.

SUPPLEMENTARY MATERIAL

CONTENTS

Click a section title or page number to jump to that section.
Subsections are also available in the PDF bookmark panel.

SECTION		SUPPLEMENT PAGE	PDF PAGE
A	SOURCE PROVENANCE AND AUDIT SCOPE	1	13
B	DATASET CENSUS AND RETRIEVAL CONFIGURATIONS	2	14
C	SERIALIZATION CONTRACTS AND EXTRACTION	3	15
D	PAIRED EFFECTS AND COMPONENT EVIDENCE	4	16
E	COUNTERBALANCED LABELS AND POSITION DIAGNOSTICS	5	17
F	RETRIEVAL COMPARISONS AND DIRECT INTERACTIONS	5	17
G	CONSTRUCTED MUIRBENCH REFERENCE INJECTION	6	18
H	MODEL-STACK BOUNDARIES AND HISTORICAL SCREENS	7	19
I	REPRODUCIBILITY, RELEASE SCOPE, AND REMAINING GAPS	8	20
J	ARCHIVED PAIRED 2 × 2 FACTORIAL DESIGN	9	21
K	COMPLETE PAIRED F58 BUDGET FOLLOW-UP	13	25
L	EXPANDED EXPERIMENTS AND ANALYSIS	21	33
M	LEGACY PROTOCOL COVERAGE COMPLETION	28	40
	REFERENCES	32	44

Supplement page refers to the original printed page number.
PDF page refers to the page position in this combined file.

SUPPLEMENTARY MATERIAL
AUDITING ROLE COLLISIONS IN MULTIMODAL MEMORY EVALUATION

Anonymous authors
Paper under double-blind review

This supplement follows the main paper from source construction to scored outcomes. Sections A–C establish the audited interface, dataset configurations, and serialization contracts. Sections D–H retain archived comparisons and their scope, while I distinguishes numerical replay from inference reproduction. Sections J–K preserve the original paired 2 × 2 factorial, Qwen budget follow-up, and 200-output human extraction audit. Section L adds matched four-budget runs for Qwen, Gemma, and InternVL, together with a separate four-model external cohort; M completes legacy 128-token InternVL binding and Qwen/Molmo2 retrieval. Historical test families remain distinct from the expanded corrections in L and M. Throughout, role collision denotes an interface condition and role-preserving serialization (RPS) denotes the binding intervention. Executable identifiers containing typed remain unchanged. Colors retain the source captions' comparison-specific meanings.

A SOURCE PROVENANCE AND AUDIT SCOPE

A.1 THE PINNED ANSWERING PATH

We begin with the public VisualMem request builder rather than reconstructing a prompt from model errors (Nguyen et al., 2026). The answering source is fixed at revision 3fa332b13a88f89b7ed81f2c7117db9acc81ebd5. Table S1 maps its relevant operations in evaluation/responses.py: retrieved images are appended first, followed by question and choice images, and finally the complete question/context prompt. This ordering locates the omitted explicit role and option mapping. It does not reveal how a model internally interprets the images.

Table S1: Source-to-request map in the pinned VisualMem answering file.

Line	Function or block	Observed behavior
63	Response parser	Source answer-extraction entry point
84	Question builder	Constructs question and choice text
139	Answering entry	Starts request construction
157	Memory-image loop	Appends retrieved images first
164	Question-image loop	Appends question and choice images
171	Prompt append	Appends the complete prompt last

To make the source identification reproducible, we retain both newline conventions. The Windows CRLF checkout has SHA-256 0d6504e0489d945656040ad77929906ea8af32f6491d341388ad7f9c2530b7d3; the canonical LF Git blob has SHA-256 9b4a3617299dc783d193c611cbe38727d5cc169e5647f8f14168e23d09481b7e. Replacing CRLF with LF makes the retained checkout identical to that blob at the audited revisions. The author-side reconciliation record preserves both digests.

Throughout the paper, **Official** names this source-derived content order. The original path uses a proprietary Gemini endpoint, whereas our experiments use open-weight model stacks and their own processors. We also apply a common strict parser instead of the source's permissive extraction policy. Thus, the comparison reproduces the request construction, not the original endpoint, preprocessing, scoring denominator, or published closed-model results.

A.2 APPLICABLE INTERFACES AND EVIDENCE LEVELS

We extend the source check to eleven keyword-discovered public interfaces, using a search cutoff of 2026-08-24. An interface enters the applicable set when reference or query images coexist with competing image-valued answers in one native request. Three interfaces satisfy this rule.

VisualMem omits the audited binding, while M3Exam and Multimodal RewardBench 2 (MMRB2) expose identity conventions. Table S2 reports the inspected adapters, not every task within each benchmark. Because the ledger is not a probability sample, neither 1/3 nor 1/11 estimates failure prevalence.

Table S2: Complete interface ledger. "No" means outside the selected native audit scope, not benchmark failure.

Interface	Revision	Applies	Observed binding or scope
VisualMem	ed40dc4	Yes	Explicit role/choice binding absent
M3Exam gallery	1dbe104	Yes	Adjacent filename
MMRB2 judge	7ff0552	Yes	Prompt/A/B markers
MemEye	0358e70	No	History/query turns
Mem-Gallery	a93959e	No	Memory/query flow
M3Eval	d11ffb0	No	Video and text choices
MemLens	77f3ab9	No	Source turns
MuirBench	840b85f	No	Native placeholders; injection tested separately
MMSI-Bench	13e58a2	No	Scenes and text choices
MULTI	aa40c5b	No	Question/knowledge images
ATM-Bench	ef4e5df	No	Open answers and evidence IDs

Although the ledger lists a different VisualMem repository revision, the relevant file has the same audited digest. Historical traces record asset paths, task roles, labels, indices, and digests. These fields support provenance checks but are not automatically model-visible: only emitted text and image blocks enter the request. Accordingly, the historical checks establish supplied-asset and content-order invariance. The newer generation-boundary checks in Sections J–K separately test actual input tensors; they do not retroactively establish crop, tensor, or embedding equality for historical runs.

B Dataset census and retrieval configurations

The VisualMem archive is fixed at data revision 1f4950f402d6facc3a6534029f78ed99f16432be. It contains 696 questions from ten synthetic personas, with the query/answer roles shown in Table S3. All 58 image-choice questions have text queries and no query image. Eligibility follows observable answer-value types, not model correctness. The complete manifest covers 276 distinct assets and has SHA-256 cdb3eb0e37b6c72b856364d7e72125a84bc6c147358a73270f6d674c8048802d.

Table S3: Observable query and answer roles in the pinned archive.

Query type	Text choices	Image choices	Total
Text query	631	58	689
Image query	7	0	7
Total	638	58	696

B.1 H57 is not F58 minus one prediction

The two archived image-choice configurations use different retrieval outputs: vmem-vmem_svg4 for **H57** and vmem-xm_qwen4b_svg for **F58**. H57 excludes persona 3/event 45/question 1 before answer generation, even though its candidate assets are available. Three text-choice questions in an earlier "central 60" list are also excluded from image-choice denominators. F58 restores the full eligible census, but this is not its only change. Among the 57 shared tasks, retrieved context differs in 53 cases and the ordered reference-image list differs in 42. Queries, candidates, and gold labels remain the same.

Table S4: Shared task identities across retrieval configurations. Each count covers the 57 shared questions. This is not a same-input repeatability comparison.

Model	H57 Official / RPS	F58 shared Official / RPS	Net gains H57 / F58
Qwen3.5-9B	6 / 28	11 / 33	22 / 22
Gemma 3-4B	4 / 17	5 / 15	13 / 10
Molmo2-8B	8 / 7	15 / 14	−1 / −1

Table S4 therefore compares shared task identities across retrieval configurations, not identical requests. In particular, Gemma's 17/57 and 15/58 RPS counts are not a one-question extension: F58 already scores 15/57 on the shared tasks. Within each configuration, however, paired serializers use its same fixed retrieval output. All source-derived, group, neutral-ID, and RPS builders cap retrieved images at ten. The observed maxima for Naive, Retrieval-only, and Full on 696 questions are six, six, and four, respectively. Candidates do not consume this reference cap.

Internal plans and digest sidecars preserve the sequence of original runs, result-aware reanalyses, and targeted follow-ups. The archived correction dated 2026-09-09 reconciles inputs and analysis definitions without adding model generations, human labels, or an untouched test set. Sections J–K specify protocols fixed before their respective follow-up outputs. These records are internal freezes, not external preregistrations; completing the census does not remove adaptation to earlier observations.

C SERIALIZATION CONTRACTS AND EXTRACTION

Our binding requirement is an unambiguous option-to-image mapping together with a recoverable distinction between evidence and selectable answers. Adjacent text is one implementation, not the definition of validity. A correct global index map, native image identifier, or structured field can express the same correspondence. ROLECHECK tests the narrower adjacent-label policy. Section J.1 compares correct global and adjacent mappings, whereas Table S5 describes the earlier component recipes.

Table S5: Archived serializer recipes. Equal label text applies only to correct RPS versus permuted RPS.

Recipe	References	Image order	Visible IDs / wording
Official	Retained	Source	No adjacent IDs
Group boundaries	Retained	Source	Generic groups
Neutral IDs	Retained	Source	Correct A–D; generic groups
RPS (typed)	Retained	Source	Correct IDs; explicit roles
Permuted RPS	Retained	Source	Wrong assignment; same RPS text
RPS choice-only	Removed	Choice order	Correct IDs; explicit roles

C.1 WHAT EACH CONTROL CHANGES

Each archived control targets a different part of the request. The grouping recipe adds BEGIN/END GROUP 1 and BEGIN/END GROUP 2 around reference and question/choice images. Neutral IDs then add ITEM A–ITEM D before the candidates. RPS uses per-image reference and answer labels. Thus, neutral IDs versus grouping tests the added candidate identities, but RPS versus neutral IDs also changes the grouping template and is not a pure role-word ablation. The permuted control rotates the complete preceding text-block assignments. It retains the ordered images and exact label-text multiset while changing which image receives each label; text order and tokenization need not remain identical.

The output contract is also configuration-specific. Here, generation budget means the output-token limit, which is distinct from actual response length. Archived JSON runs allow 128 new tokens, and H57 answer-first runs have a documented output-token limit of 32. Retained call-level records do not establish the effective output-token limit for archived F58 answer-first runs, so we do not infer it from a filename or the later paired experiment. The frozen parser accepts supported explicit JSON or answer formats, including FINAL=C and FINAL=<C>, rather than guessing from the first isolated option letter. Unparsed successful outputs remain incorrect in the fixed denominator. E is valid only for MuirBench questions with five choices.

Table S6: Qwen parse and correct counts, reported as parsed / correct out of n. FINAL-32 is the H57 answer-first contract. The F58 answer-first budget is unknown; a dash marks an unrun arm.

Configuration	Output	Official	RPS	RPS choice-only
H57, n = 57	FINAL-32	55 / 3	55 / 35	56 / 30
F58, n = 58	JSON-128	28 / 11	45 / 34	44 / 30
F58, n = 58	Answer-first	52 / 5	56 / 38	–

Table S6 separates correct counts from parse coverage. In the F58 JSON pair, 24 outputs parse under both serializers; within this subset, RPS has nine wins and two harms, with sixteen additional wins outside it. For F58 answer-first, fifty outputs parse in both arms, with 31 RPS wins and one harm; three further wins lie outside this subset. Because membership depends on both outputs, these are descriptive overlap checks rather than unbiased estimates of visual reasoning. H57 answer-first choice-only coverage is 56/57, or 98.2%.

D PAIRED EFFECTS AND COMPONENT EVIDENCE

We define a **win** as a task correct only under the candidate serializer and a **harm** as one correct only under its reference. Their paired accuracy difference is $(W - H)/n$. Archived analyses use 20,000 persona-bootstrap draws with seed 20260822; explicitly identified reanalyses use 50,000 PCG64 draws with seed 20260909. Exact persona sign flips assume symmetric cluster effects rather than randomized treatment assignment. With ten synthetic personas, percentile intervals and exact sign-flip tests need not cross a decision threshold together.

Table S7: Complete F58 Qwen paired comparisons. Effects and persona intervals are percentage points. The last row has unknown effective-budget provenance.

Candidate − reference	Correct pair	W / H	Effect	McNemar p	Persona 95% CI
RPS − Official	34 / 11	25 / 2	+39.7	5.65e−6	[25.0, 55.6]
Permuted − RPS	2 / 34	1 / 33	−55.2	4.07e−9	[−67.9, −43.3]
Choice-only − RPS	30 / 34	6 / 10	−6.9	.454	[−19.0, 6.8]
Answer-first RPS − Official	38 / 5	34 / 1	+56.9	2.10e−9	[46.7, 68.4]

D.1 COMPONENT COMPARISONS AND THEIR LIMITS

For the complete F58 JSON comparison, the RPS–Official effect remains positive after removing any one persona, ranging from 35.8 to 44.2 pp; its persona sign-flip p is .003906. Correct binding also outperforms the equal-text wrong-binding control. By contrast, the full-input versus choice-only comparison does not establish a general memory benefit for Qwen. Table S7 therefore supports a distinction between repairing the representation and demonstrating the usefulness of the retained references.

Table S8: Archived source-order component counts. H57 and F58 have different retrieval inputs. A dash denotes no reported arm.

Model / set	Official	Group	Neutral IDs	RPS	Permuted
Gemma / H57	4	6	12	17	–
Molmo2 / H57	8	8	8	7	–
InternVL / F58	10	20	35	34	11

Table S9: Successive component-recipe contrasts. Holm correction uses the three comparisons within each checkpoint. RPS − IDs also changes the wrapper.

Model	Contrast	Wins	Harms	Raw p	Holm p
Gemma	Group − Official	2	0	.500	.500
Gemma	IDs − Group	8	2	.1094	.3281
Gemma	RPS − IDs	6	1	.125	.3281
InternVL	Group − Official	13	3	.0213	.0425
InternVL	IDs − Group	20	5	.00408	.0122
InternVL	RPS − IDs	5	6	1.000	1.000

The Gemma component results illustrate why the comparison must match the claim. Neutral IDs versus Official gives ten wins and two harms (p = .0386), but this comparison also includes grouping. The incremental contrast is IDs versus grouping: eight wins and two harms, with p = .109375 before correction. Thus, the complete configurations improve over Official, while the additional contribution of identities beyond grouping is not statistically resolved (Tables S8–S9).

InternVL3.5-8B-HF combines an InternViT-300M encoder with a Qwen3 language component, totaling approximately 8.5B parameters (Wang et al., 2025b). IDs versus Official yields 31 wins and six harms (p = 4.13e−5; archived persona interval [27.1, 55.9] pp), and the IDs-versus-grouping contrast passes its stated correction. RPS versus IDs instead gives five wins and six harms (p = 1; interval [−10.7, 8.6]). Correct RPS also exceeds permuted binding by 31 wins

against eight harms (p = 2.94e−4; interval [28.1, 51.7]), with 55/58 parses in both arms. These results add a multimodal implementation, not a wholly independent language-model family.

Molmo2 provides a retained non-replication. On H57, Official, grouping, and neutral IDs each score 8/57, while RPS scores 7/57. On F58, Official scores 16/58 and RPS scores 15/58, with no RPS wins and one harm. The archived evidence does not establish which implementation factor causes this difference.

E COUNTERBALANCED LABELS AND POSITION DIAGNOSTICS

To remove native correct-label imbalance, each F58 question is evaluated in four cyclic option orders, updating candidates and gold labels together. Every correct candidate appears once at A–D, so a constant-letter strategy scores 25% across the 232 variants. The variants remain repeated measurements of 58 tasks. We average their paired differences within each base before resampling bases or personas. This balances answer position without enumerating all 24 permutations. The archived option-order study in Table S10 uses a different template and output contract from the paired five-condition experiment in Section J.

Table S10: Archived binding comparison with four option orders. Pooled counts are out of 232; each gold-position count is out of 58.

Model	Serializer	Pooled	Gold A	Gold B	Gold C	Gold D
Qwen	Official	54	19	11	10	14
Qwen	RPS	135	30	30	40	35
Gemma	Official	63	43	8	2	10
Gemma	RPS	106	46	23	14	23
Molmo2	Official	63	13	21	14	15
Molmo2	RPS	60	17	19	11	13

The counterbalanced results retain a positive RPS effect for Qwen and Gemma. Qwen improves by 34.9 pp, with persona interval [25.0, 45.0], base interval [24.6, 45.7], and sign-flip p = .00195. Gemma improves by 18.5 pp, with respective intervals [10.3, 27.1] and [10.3, 27.2], and p = .00586. Both improve at each gold position, although Gemma retains a strong gold-A advantage. Molmo2 instead changes by −1.3 pp, with persona interval [−5.0, 2.6], base interval [−4.7, 2.2], and p = .6875. These reanalyses use 50,000 draws and seed 20260909.

E.1 HISTORICAL DIAGNOSTICS USE A DIFFERENT BASELINE

Earlier positional screens use a text-first reconstruction rather than the pinned Official order. On H57, accuracy across four option orders is 15.8% for this reconstruction and 48.7% for RPS. Placing references before the choices, between B/C, or after the choices gives reconstruction accuracies of 8.8%, 40.4%, and 56.1%, compared with 52.6%, 61.4%, and 57.9% for RPS. The narrowing gap when candidates occupy the first four visual positions motivates the later positional analysis. These observations are not pooled with the source-order results.

The native H57 gold distribution is A/B/C/D = 0/24/32/1. Always selecting C therefore scores 56.1%, and the single D case cannot support a meaningful stratum-level claim. Native absolute accuracy alone is not strong evidence of visual reasoning. The balanced F58 design tests whether B/C concentration alone explains the serializer difference, without claiming invariance to arbitrary image order or prompt paraphrase.

F RETRIEVAL COMPARISONS AND DIRECT INTERACTIONS

To assess consequences for retrieval-method comparisons, the analysis holds three released outputs fixed: vmem-xm_qwen4b_naive, vmem-vmem_svg_ro, and vmem-xm_qwen4b_svg, named **Naive**, **Retrieval-only**, and **Full**. Each contains the same 696 task identities, questions, candidates, and gold labels; retrieved context differs by method. Their union contains 810 distinct assets. Within each model and method, Official and RPS share preprocessing, greedy decoding, scoring, and output-token limit. We re-answer these outputs rather than retrain retrieval or reproduce a closed-model leaderboard.

Table S11: Complete retrieval consequence counts. The final two columns isolate the 58 image-choice questions. Pale red/bold marks the largest correct count within the stated comparison.

Model	Retrieval	Official /696	RPS /696	Official /58	RPS /58

Model	Retrieval	Official /696	RPS /696	Official /58	RPS /58
Gemma	Naive	178	177	**6**	5
Gemma	Retrieval-only	**193**	**192**	5	4
Gemma	Full	178	188	5	**15**
InternVL	Naive	186	186	**26**	32
InternVL	Retrieval-only	**197**	**203**	22	33
InternVL	Full	167	190	10	**34**

Table S11 shows that serialization changes some method comparisons while Retrieval-only remains highest overall. Gemma's Full and Naive counts tie under Official. Under RPS, Full has eighteen wins and seven harms over Naive (raw p = .0433; three-comparison Holm p = .1299), leaving its corrected advantage unresolved. InternVL moves from a nineteen-answer deficit to a four-answer advantage. Its Official pair has 15/34 wins/harms (raw p = .00940; Holm p = .0188), and its RPS pair has 21/17 (p = .627). Overall Kendall rank agreement is .816 for Gemma and .333 for InternVL; the corresponding image-choice values are 0 and −1. These rankings are descriptive.

A change in statistical significance is not itself a test of a changed method contrast. We therefore combine all four correctness outcomes. With F/N denoting Full/Naive and T/O denoting RPS/Official, the per-question change and its average are

$$d_i = \left(Y_{i,F,T} - Y_{i,N,T}\right) - \left(Y_{i,F,O} - Y_{i,N,O}\right), \quad \Gamma = \frac{1}{n}\sum_i d_i. \tag{S1}$$

Table S12: Direct full-minus-naive serializer interaction. Effects and intervals are percentage points. These are post-result reanalyses with ten persona clusters.

Model	Question subset	Interaction	Persona 95% CI	Sign-flip p
Gemma	All 696	+1.580	[0.440, 2.744]	.0625
Gemma	Image-choice 58	+18.966	[6.897, 31.667]	.0625
Gemma	Text-choice 638	0.000	[−0.472, 0.450]	1.000
InternVL	All 696	+3.305	[1.366, 5.506]	.015625
InternVL	Image-choice 58	+31.034	[17.241, 45.614]	.0078125
InternVL	Text-choice 638	+0.784	[−0.328, 1.974]	.375

F.1 Uncertainty and text-choice changes

For the direct interaction, all four outcomes stay together during resampling. Table S12 uses 50,000 whole-persona draws with seed 20260909. Gemma's percentile interval excludes zero while its exact sign-flip p is .0625, and both summaries are retained. InternVL's positive interaction likewise does not establish that RPS Full outperforms RPS Naive: a change in a contrast and the final contrast are different quantities. The ten synthetic personas limit population-level interpretation.

Text-choice totals can remain unchanged even when predictions change. Gemma's Naive, Retrieval-only, and Full counts are respectively 172/172, 188/188, and 173/173 under Official/RPS. Nevertheless, Full changes one label without changing correctness, while Naive changes seven labels and two correctness indicators. InternVL Full changes from 157 to 156 text-choice answers correct, with twelve changed labels and three correctness changes; Naive changes 25 labels and twelve correctness indicators. The larger observed interaction occurs on image-choice questions, but these records do not support general text-choice invariance.

G Constructed MuirBench reference injection

We use MuirBench to construct a second input setting, not to claim a defect in its native interface (Wang et al., 2025a). The dataset revision is 4c393cffc985c77d28de3b9045e2e5186920df80, and the evaluation repository revision is 840b85fe. Selection ranks SHA-256 of 20260823|idx, retains the member with idx < counterpart_idx, and requires an exact placeholder mapping with at least one query image and two image-valued choices. Five tasks contribute eight examples each; the selected set does not cover all twelve benchmark tasks.

For each selected question, we inject two reference images from other tasks. Their canonical PNG digests differ from the target and from each other. Flat and RPS injection use the same sequence, [reference 1, reference 2, target images], and differ in the visible bindings. The no-reference

control retains explicit RPS answer identities. The references are therefore constructed distractors rather than naturally retrieved personal memories.

Table S13: MuirBench injection. All counts are out of 40. W/H, tests, and effects compare RPS with flat injection; intervals cluster five tasks.

Model	No-ref.	Flat	RPS	W / H	McNemar p	95% CI (pp)
Qwen	15	3	20	17 / 0	1.53e−5	[25.0, 60.0]
Gemma	23	9	22	15 / 2	.00235	[15.0, 50.0]
Molmo2	7	2	4	2 / 0	.500	[0.0, 10.0]

Table S14: Qwen injection by task, eight examples per row.

Task	No-ref.	Flat	RPS	Wins	Harms
Attribute Similarity	0	0	1	1	0
Diagram Understanding	6	2	7	5	0
Difference Spotting	1	0	2	2	0
Geographic Understanding	2	0	4	4	0
Visual Retrieval	6	1	6	5	0

Tables S13–S14 report the complete injection results. Qwen has seventeen RPS-only wins and no flat-only wins, but five nonzero task clusters impose a minimum two-sided sign-flip p of .0625. Gemma improves over flat injection, not over the no-reference condition. Molmo2's two-answer gain is unresolved. These outcomes support sensitivity to the constructed interface manipulation without establishing improved memory use or the prevalence of native omissions.

H Model-stack boundaries and historical screens

We retain the complete-census Gemma and Molmo2 results in Table S15 to show the limits of the archived intervention. Every row has full strict-parse coverage, so differences in correct counts cannot be attributed to differences in coverage. The effective generation budgets of the archived answer-first conditions remain unknown.

Table S15: Archived complete-census counts, correct out of 58; all outputs parse.

Condition	Gemma 3-4B	Molmo2-8B
Official JSON-128	5	16
RPS JSON-128	15	15
RPS choice-only JSON-128	4	23
Permuted RPS JSON-128	4	16
Official answer-first	8	31
RPS answer-first	18	30

The reference-image comparison differs across models. Gemma's full RPS condition exceeds choice-only by eleven wins and no harms ($p = .000977$; persona interval [8.5, 30.0] pp). Qwen's corresponding comparison remains unresolved. Molmo2 instead favors choice-only, with ten wins against two harms ($p = .0386$). Its answer-first prompt improves both serializers without reversing their one-answer difference. Thus, a benefit from explicit binding and a benefit from references are separate, model-dependent findings.

The historical Gemma template study gives 17/57, 13/57, and 13/57 correct for explicit, concise, and XML RPS, compared with 4/57 for Official. Concise and XML each give eleven wins and two harms over Official ($p = .0225$). Relative to explicit RPS, their win/harm counts are 0/4 ($p = .125$) and 2/6 ($p = .289$). A separate exhaustive control covers 74 eligible text-choice questions with references. Both serializers score 37/74, with one win and one harm, 72 identical labels, and only five identical raw strings. These controls do not establish template equivalence or invariant generations.

Table S16: Earlier architecture/size screens on 57 historical questions. Accuracy is a percentage. Reconstruction is text-first, not the pinned Official baseline.

Checkpoint	Reconstruction	RPS full	RPS choice-only
Qwen3.5-2B	8.8	35.1	35.1
Qwen3.5-4B	7.0	66.7	61.4
Qwen3.5-9B	8.8	52.6	68.4
LLaVA-OneVision-Qwen2-7B	15.8	24.6	45.6

Checkpoint	Reconstruction	RPS full	RPS choice-only
SmolVLM2-2.2B	0.0	0.0	0.0

Table S16 keeps the earlier architecture and size screens separate because they use the text-first reconstruction. LLaVA's RPS–reconstruction contrast has eight wins and three harms (p = .227). An Idefics3-8B gate reaches 1/12 in its better condition and stops, so no complete result is claimed. A Gemma 3-12B download fails before inference; InternVL3.5-8B-HF is the planned fallback. Consequently, these records do not supply evidence above 9B. The older 512-token reconstruction gives full/choice-only accuracies of 70.2/71.9%, with two full wins and three harms (p = 1). It is not pooled with the matched F58 budget study.

I REPRODUCIBILITY, RELEASE SCOPE, AND REMAINING GAPS

The submission includes an anonymous code attachment, and we commit to publishing our implementation and analysis code. Its documented evidence scope is numerical, not a redistribution of the upstream benchmark. Third-party code, images, questions, retrieved context, and raw generations remain subject to the relevant redistribution conditions. Human identities, contact details, consent records, and free-form notes remain private. The following sections explain which claims each verification route can support.

I.1 WHAT EACH VERIFICATION LAYER ESTABLISHES

We distinguish three verification levels. File and summary verification checks hashes, schemas, and registered values, establishing consistency of the delivered records. Row-level recomputation derives statistics from de-identified predictions, labels, and grouping keys without rerunning a model. Inference reproduction additionally requires source assets, checkpoints, processors, environment, and generation configuration. A successful summary check alone is neither a new inference run nor independent verification of data collection.

The archived paths here and in Section K.12 are relative to archived_studies/ in the anonymous companion. Its data/budget_contrasts_v1.csv, data/budget_interactions_v1.csv, data/prefix_effects_v1.csv, and data/endpoint_audit_summary_v1.json retain the earlier registered quantities and endpoint reconciliation. The file data/human_scoring_linkage_v1.jsonl joins 200 de-identified records to condition, parser, and gold fields; data/human_scoring_aggregate_reference_v1.json contains their aggregate reference. These historical files remain byte-identical. New statistics and their declared families are separate in expanded_studies/.

The recommended entry point is python -B verify_bundle.py from the anonymous companion root, using Python 3.11 or newer and NumPy 2.x. It runs archived verify_all.py in an external working copy, then runs the expanded numerical replay and checks that packaged bytes are unchanged. The historical verifier writes its own results directory, so it should not be run directly inside the preserved distribution. No CUDA, model weights, API keys, network access, or benchmark installation is needed after installing the numerical dependency. The root README specifies commands and fresh output locations.

I.2 A RUNNABLE SCOPED CONTRACT

The synthetic request data/synthetic_rolecheck_example.json contains a schema version, expected choices, an instruction, visible labels, and image-block metadata. Running python -m src.rolecheck data/synthetic_rolecheck_example.json returns status 0 for a pass and 2 for a contract failure. Under this policy, each image must immediately follow its matching visible label. A valid global mapping can fail this adjacency-specific check. Conversely, a pass establishes neither equality of processor tensors nor semantic use by the model.

I.3 HISTORICAL PROVENANCE IS NOT REPAIRED RETROSPECTIVELY

The author-side workspace preserves saved generations, manifests, execution metadata, and historical analyzers. Some archived runs do not link every checkpoint file unambiguously to a complete execution-code revision. Nearby manifests cannot supply this missing historical evidence. The complete archived runs record Qwen revision c2022362, Molmo2 e28fa285, an InternVL mirror revision 3974c115, and a Gemma public-mirror snapshot identified by digests. Unresolved fields remain unresolved; the newer checks in Sections J–K do not certify older executions.

The older 456-output H57 annotation packet is separate from the frozen 200-output F58 human audit. Only the latter supports the reported human statistics. Software checks record structure and arithmetic, while human identity, independence, blinding, and consent rely on the authors' attestation. The numerical release does not independently reproduce the collection of human labels.

J Archived Paired 2 × 2 Factorial Design

J.1 Four factorial conditions and an additional anchor

The paired 2 × 2 factorial design crosses mapping placement (global vs. adjacent) with role wording (neutral vs. semantic). It asks whether adjacent placement helps and whether semantic wording adds value at a fixed placement. It uses F58, separately from the H57 retrieval configuration. Qwen3.5-9B and Gemma 3-4B each process 58 bases, four cyclic option orders, and five conditions, yielding 1,160 generations per checkpoint. These option orders jointly update candidate positions, displayed labels, and gold labels while preserving the query, context, and reference images. Within each option order, conditions use the same retrieved evidence and ordered images. The anchor is an additional control, not a fifth factorial combination.

The frozen identifiers are official_anchor (Anchor), global_neutral (GN), global_semantic (GS), adjacent_neutral (AN), and adjacent_semantic (AS). All four mapped conditions share BEGIN/END GROUP 1 and BEGIN/END GROUP 2. Neutral statements use forms such as Image k | ITEM R1. and Image k | ITEM A. Semantic statements replace ITEM with REFERENCE, QUESTION, or ANSWER. The prompt defines k over the complete image sequence. Adjacent placement puts each statement immediately before its image; global placement collects the same statements after END GROUP 2 and before the final question/context.

For each question and option order, the exact mapping-statement multiset and its correct assignments stay fixed within each wording level. The placement contrast jointly changes distance, before/after direction, absolute position, and recency; it does not isolate an internal attention mechanism. Neutral IDs also retain structural cues through R/Q/A–D and the group wrappers. Role wording is not constrained to equal tokenizer length. The anchor shares the answer-first output contract but omits the added groups and labels, and therefore differs from the archived Official execution.

J.2 Generation and actual-input evidence

The shared instruction lists all four permitted first-line answers, FINAL=<A>, FINAL=<B>, FINAL=<C>, and FINAL=<D>, and allows a brief explanation. No single letter serves as the only format example. At generation, the runner checks the output-token limit, max_new_tokens=128 and do_sample=False, with seed 20260909. It retains effective arguments, generated token IDs and counts, precision, processor settings, and checkpoint digests. Greedy decoding is not assumed to guarantee bitwise repeatability.

Raw decoded text is saved before the frozen _strip_thinking transformation. The scored text and strict parse_response prediction are retained separately, with no manual answer repair or outcome-dependent fallback. An unparsed successful call counts as wrong; a technical failure is recorded separately and prevents a claim of complete execution. Source-byte and decoded-RGB hashes identify the images, while tensor keys, shapes, dtypes, and hashes identify the actual generation inputs. Image-bearing tensors must agree across conditions; text tensors may differ when labels move. These checks establish the input boundary, not the model's understanding of it.

J.3 Protocol and execution discipline

The protocol, manifests, code, and tests are frozen internally before this experiment's outputs, after earlier outcomes inform the design. They are not an external preregistration or an untouched test set. Append-only events and validated resume keys prevent duplicate cases. Technical pilots remain in separate files and outside formal counts. A three-hour per-invocation loop allowance is checked between calls, while monitoring handles hung calls. Repairs are versioned, and neither accuracy nor statistical significance determines stopping or replacement.

J.4 PAIRED ANALYSIS AND 128-TOKEN RESULTS

The four primary contrasts compare adjacent minus global placement under each wording level and semantic minus neutral wording at each placement. We first average the signed differences across the four option orders within each base, so all 58 bases receive equal weight. Pointwise uncertainty uses 50,000 whole-persona PCG64 bootstrap draws with seed 20260909 and the ratio-of-sums estimator specified in Section K.3. Exact two-sided sign-flip sensitivities enumerate all 1,024 persona signs. Four simple effects form one Holm family per checkpoint. The placement-by-wording interaction has a separate two-checkpoint family.

Table S17: Paired F58 2 × 2 factorial design at 128 tokens. Counts are out of 232 dependent variants per arm. The anchor shares the new answer contract, not the archived Official execution. Pale red/bold marks the largest correct count within the stated comparison.

Arm	Qwen correct	Qwen parsed	Gemma correct	Gemma parsed
Anchor	35	100	66	232
GN	71	101	61	232
GS	104	158	60	232
AN	130	187	**78**	232
AS	**145**	218	**78**	232

Table S18: Factorial simple effects in percentage points. W/H are descriptive paired counts among dependent variants. Holm4 corrects four persona tests per checkpoint. Blue marks effects supported by the declared corrected tests.

Model	Effect	W / H	Estimate	Pointwise 95% CI	Holm4 p
Qwen	Placement, neutral	69 / 10	**+25.43**	[16.67, 34.05]	**.0117**
Qwen	Placement, semantic	52 / 11	**+17.67**	[13.36, 21.25]	**.00781**
Qwen	Wording, global	40 / 7	**+14.22**	[6.36, 22.77]	**.0234**
Qwen	Wording, adjacent	23 / 8	**+6.47**	[3.33, 9.65]	**.0234**
Gemma	Placement, neutral	23 / 6	**+7.33**	[5.00, 9.58]	**.0117**
Gemma	Placement, semantic	21 / 3	**+7.76**	[5.91, 9.75]	**.00781**
Gemma	Wording, global	4 / 5	−0.43	[−2.63, 1.69]	1.0000
Gemma	Wording, adjacent	8 / 8	+0.00	[−3.33, 3.07]	1.0000

Both checkpoints complete all five conditions, totaling 2,320 generations. Table S17 shows complete parsing for Gemma but substantial coverage differences for Qwen. Qwen's fixed-denominator effects therefore reflect both task correctness and compliance with the specified output contract and budget. Anchor comparisons also change groups, IDs, and added text; they do not isolate binding alone. Table S18 retains every declared effect, including negative and unresolved comparisons.

J.5 INTERACTIONS, DELETION SENSITIVITY, AND TOKEN EVIDENCE

Table S19: Placement-by-wording interaction: (AS − AN) − (GS − GN), in percentage points. Holm2 corrects the two model-level tests separately from Table S18.

Model	Estimate	Pointwise 95% CI	Holm2 p
Qwen3.5-9B	−7.76	[−15.18, −1.25]	.1953
Gemma 3-4B	+0.43	[−2.59, 3.45]	1.0000

Leave-one-persona-out estimates test sensitivity to any single cluster. For Qwen, their ranges are [23.08, 28.37] and [16.83, 19.23] pp for placement, and [11.79, 16.35] and [5.29, 7.69] for wording. Gemma's corresponding ranges are [6.73, 8.17], [7.21, 8.17], [−0.96, 0.48], and [−0.96, 0.96]. The placement-by-wording interaction ranges are [−9.62, −5.66] for Qwen and [−0.48, 1.44] for Gemma. These are descriptive deletion ranges, not confidence intervals. Qwen's negative pointwise interaction interval does not supersede its unresolved corrected test in Table S19.

Jointly parsed subsets provide a second descriptive view. Qwen's neutral placement pair contains 100 such outputs, with eight wins and nine harms; its semantic pair contains 158, with seventeen wins and eleven harms. Global wording has 95 jointly parsed cases, with four wins and one harm, while adjacent wording has 186, with eight wins and seven harms. Membership depends on both

outcomes. These subsets distinguish total-score changes from changes among already parsed outputs, but neither replace fixed-denominator effects nor identify parsing as a causal mediator.

Identical text multisets need not produce identical token sequences. Moving labels adjacently adds 3, 4, 5, or 6 input tokens in 168, 40, 20, or four cases, with the same distribution across wordings and checkpoints. Qwen's semantic wording adds one token per image-label statement; full-input increases of 4, 5, 6, 7, or 8 occur in 24, 144, 40, 20, or four cases. Gemma instead has identical observed label-token vectors and total lengths in all 232 wording pairs. Qwen's wording contrast is therefore not a length-controlled test of a semantic mechanism.

The 128-token cap is reached in 165/166/131/102/78 Qwen outputs for Anchor/GN/GS/AN/AS. Among unparsed outputs, 130/132, 131/131, 74/74, 45/45, and 14/14 reach the cap; parsed cap-hit counts are 35/35/57/57/64. Cap saturation is neither equivalent to parse failure nor proof that extension yields a correct answer. Gemma reaches the cap in 5/12/1/9/3 cases while retaining complete parsing. These observations motivate the separately frozen complete Qwen extension in Section K.

J.6 Program-verifiable diagnostic construction

The synthetic suite separates matching and answer reporting from natural-task complexity. It contains sixty bases: twenty exact image matches, twenty within-panel ordinal selections, and twenty query-image matches with reference distractors. Four cyclic option orders yield 240 variants per checkpoint. High-contrast 256-pixel geometric icons use seed 20260909 and contain no rendered answer letters. Exact matches reuse target pixels, and ordinal panels reuse candidate tiles. The two distractors include one nongold duplicate and one additional image. Gold labels follow checked construction rules rather than model judgments. Every fixture has an explicit query image, unlike F58, so these tasks do not measure a natural-task perceptual ceiling or beneficial memory use.

J.7 Synthetic results under the shared 128-token budget

Table S20: Synthetic diagnostic counts. Variant denominators apply to correct/parsed; base denominators apply to all-four measures. Original-choice consistency requires four parsed selections of the same original candidate.

Model / task	Correct	Parsed	All 4 correct	Same original
Qwen: exact	9/80	9/80	0/20	0/20
Qwen: ordinal	2/80	2/80	0/20	0/20
Qwen: distractors	2/80	2/80	0/20	0/20
Qwen: all	13/240	13/240	0/60	0/60
Gemma: exact	32/80	80/80	0/20	0/20
Gemma: ordinal	34/80	80/80	0/20	0/20
Gemma: distractors	42/80	80/80	0/20	0/20
Gemma: all	108/240	240/240	0/60	0/60
Molmo2: exact	20/80	80/80	0/20	0/20
Molmo2: ordinal	38/80	80/80	0/20	0/20
Molmo2: distractors	32/80	80/80	0/20	0/20
Molmo2: all	90/240	240/240	0/60	0/60

Table S21: Synthetic accuracy (%) and pointwise 95% base-bootstrap intervals. All four option orders stay together; 50,000 draws use seed 20260909. Overall draws resample all 60 bases without task stratification.

Task	Qwen accuracy [CI]	Gemma accuracy [CI]	Molmo2 accuracy [CI]
Exact	11.25 [3.75, 21.25]	40.00 [35.00, 45.00]	25.00 [25.00, 25.00]
Ordinal	2.50 [0.00, 6.25]	42.50 [33.75, 51.25]	47.50 [40.00, 55.00]
Distractors	2.50 [0.00, 6.25]	52.50 [43.75, 61.25]	40.00 [33.75, 46.25]
All	5.42 [2.08, 9.17]	45.00 [40.42, 49.58]	37.50 [33.75, 41.67]

All 720 planned synthetic calls complete. At 128 tokens, Qwen has no base with four parsed outputs; fifty of sixty bases have none. Its zero all-four-correct count therefore does not represent sixty fully observed visual failures. All 227 unparsed responses and one parsed response reach the limit, and no unparsed string contains a FINAL field. The 38 differences between raw and scored

text are whitespace only. Fixed inspections show unfinished item-by-item explanations rather than evidence of an internal thinking mode.

Gemma and Molmo2 have complete parsing and never reach the output-token limit, with maximum lengths of 115 and 125 tokens. Their low accuracy and inconsistency across option orders are therefore not observed coverage failures. In one Gemma example, the response starts with FINAL=<A> but later identifies B in its explanation. The frozen score remains A; the explanation does not independently establish recognition. Molmo2's exact-match bootstrap interval is [25, 25] because each observed base has one correctly answered option order. This is an empirical-bootstrap property, not zero population uncertainty.

J.8 NATIVE-TEMPLATE BOUNDARY AND SEPARATE BUDGET SENSITIVITY

Source-level adjacency need not survive a native chat template. Molmo2's template moves numbered image placeholders to the front and joins the text blocks afterward. Its processor expands the placeholders in that order without restoring interleaving. A CPU Jinja replay confirms this arrangement on all 240 fixtures: 160 have five images and eighty have seven. The pinned chat_template.jinja and processing_molmo2.py at e28fa285 match runtime SHA-256 prefixes ec66862fda57dc1a and 61ef1329d8a02cae. This verifies layout, not new crops or attention, and does not identify the cause of wrong answers or explain the archived Molmo2 results retrospectively.

Qwen's missing declarations motivate a separate internally frozen budget sensitivity. We increase only max_new_tokens to 512 on every synthetic fixture, rather than selecting failed cases or replacing the original results. Checkpoints, BF16 precision, runtime, assets, image order, prompt, greedy decoding, seed, and scoring remain fixed. Each complete processor-tensor dictionary must match its 128-token counterpart. One technical pilot is excluded. The paired analysis resamples all sixty bases, or twenty within each task, retaining their four option orders and introducing no further significance family.

Table S22: Complete synthetic Qwen budget sensitivity. Every option-order variant remains in each denominator.

Task	Correct 128	Correct 512	Parsed 128	Parsed 512
All tasks	13/240	240/240	13/240	240/240
Exact	9/80	80/80	9/80	80/80
Ordinal	2/80	80/80	2/80	80/80
Distractors	2/80	80/80	2/80	80/80

Table S23: Paired 512-minus-128 changes in synthetic accuracy and parsing (percentage points). Both measures have identical changes and intervals. Intervals are pointwise whole-base percentiles.

Task	Parse change [95% CI]	Accuracy change [95% CI]
All tasks	94.58 [90.83, 97.92]	94.58 [90.83, 97.92]
Exact	88.75 [78.75, 96.25]	88.75 [78.75, 96.25]
Ordinal	97.50 [93.75, 100.00]	97.50 [93.75, 100.00]
Distractors	97.50 [93.75, 100.00]	97.50 [93.75, 100.00]

The 240 additional formal calls all complete, and every task reaches 80/80 correct. The overall comparison has 227 improvements and no losses in either parsing or correctness. All sixty bases become all-four-correct and original-choice-consistent. Processor dictionaries and comparable token prefixes match, including all 228 pairs that reach the original 128-token limit. Longer outputs have median length 201, range 36–282, and no 512-token cap hits. The run takes 25.80 wrapper minutes and 25.86 outer-call minutes, with peak allocation of 17.81 GiB. These simple-task results do not establish perfect population accuracy or F58 competence; longer generation allows both additional computation and later reporting.

J.9 CHECKPOINT PROVENANCE AND EXECUTION ACCOUNTING

Table S24: Original five-run execution records. Input lengths cover all calls. Wrapper time includes synchronized generation and decoding/ID handling; peak is allocated CUDA memory.

Run	Calls	Input tokens	Wrapper min	Peak GiB
Gemma factorial	1,160	1,182–3,933	45.04	8.92
Qwen factorial	1,160	1,597–4,534	78.76	18.74

Run	Calls	Input tokens	Wrapper min	Peak GiB
Gemma synthetic	240	1,513–2,049	5.63	8.79
Qwen synthetic	240	543–751	16.10	17.81
Molmo2 synthetic	240	2,274–3,118	6.81	17.78

The original runs use one RTX 5090, BF16, Python 3.12.3, PyTorch 2.12.1+cu130, and CUDA 13.0. Qwen and Gemma use Transformers 5.15.1; Molmo2 uses a compatible 4.57.1 environment. Runtime hashes and source receipts reconcile Qwen to revision c2022362 and Gemma to 093f9f38. Gemma's raw descriptive revision field remains unresolved rather than being retrospectively edited. Checkpoint names alone do not establish file identity.

Molmo2 revision e28fa285 supplies FP32 shards. To fit the 32-GB device, source tensors are deterministically cast to BF16. Each converted shard is reloaded and checked against that cast before acceptance. This is BF16 weight conversion, not source-byte equality or 4/8-bit quantization. The source and destination digests and per-shard checks remain in the author-side audit; model weights are excluded from the anonymous attachment.

Table S24 accounts for the original five-run study: 3,040 formal calls and thirteen excluded technical calls. The separate synthetic budget sensitivity adds 240 formal calls and one pilot. The subtotal before Section K is therefore 3,280 formal calls and fourteen excluded calls. Reused comparators are counted only once. This total covers the specified extension rather than every historical experiment.

Timing records distinguish the synchronized generation wrapper from the complete outer call. Outer-call sums, which also include processing, tracing, and scoring, are 50.88/84.63 minutes for the Gemma/Qwen factorials and 6.01/16.16/7.05 for their three synthetic runs. Both measures exclude loading, downloads, and inter-call persistence. Wall-clock spans may include loading and resume gaps. These observations describe the executions, not a throughput benchmark or billed rental time.

The internally frozen analyzer supplies the original statistics. Versioned export and verification programs link result snapshots to append-only events, manifests, source receipts, and code digests. Preserved analysis paths include experiments/analyze_binding_revision_0909.py and experiments/analyze_binding_release_0909.py. These identify author-side workflows, not additional files promised in the thin numerical attachment. Checking a saved execution does not create an independent second model run.

K Complete paired F58 budget follow-up

K.1 Scope and sole intended manipulation

The natural-task follow-up repeats every F58 Full-retrieval case: 58 bases, four cyclic option orders, and five conditions. It retains all 1,160 original 128-token outputs and adds 1,160 outputs at 512 tokens. No H57 task or synthetic fixture substitutes for this matrix, and wrong, unparsed, or capped cases are not selectively rerun. The fifth condition remains the shared answer-first anchor rather than historical Official.

K.2 Input, generation, and completion evidence

The checkpoint is Qwen3.5-9B revision c202236235762e1c871ad0ccb60c8ee5ba337b9a. All fourteen runtime files, including four weight shards, tokenizer, template, and processor configuration, match the original hashes. The environment records Python 3.12.3, PyTorch 2.12.1+cu130, Transformers 5.15.1, CUDA 13.0, and RTX 5090. The earlier record lacks the driver and physical GPU identity, so equality of those fields is not claimed.

The sole intended manipulation raises the output-token limit from 128 to 512. BF16, greedy do_sample=False, seed 20260909, question/context, image bytes and order, serializers, thinking configuration, EOS behavior, cleaner, and parser remain fixed. The adapter passes temperature 1.0 with sampling disabled. Although this differs from the requested API temperature 0.0, it does not activate sampling. We add no forced-letter decoding, image resizing, answer repair, or FINAL-based stopping rule.

Before generation, source-byte and decoded-RGB checks reconstruct each ordered image-role list. The complete tensor-key set, shapes, dtypes, and hashes must match the paired 128-token case, including text inputs, masks, visual pixels, and grids. Across conditions, visual content agrees

while text-dependent tensors may differ. New records also retain full input IDs, non-tensor arguments, and the library-prepared generation configuration. A configuration-only reconstruction at 128 tokens checks the intended parameter difference without inference. It does not fill older missing ID arrays or non-tensor logs.

Generated IDs and raw responses are saved durably before scoring. A hash-chained journal records task, budget, input digest, attempt, and the first protocol-conforming success. Wrong or unparsed successes are not retried; technical interruptions require a documented repair. Two excluded pilots use the shortest and longest original inputs, breaking ties lexically, while those tasks remain in the formal matrix. The six-hour loop allowance and 6.5-hour watchdog are technical safeguards, not significance-based stopping criteria.

Table S25: Qwen F58 budget matrices. Each cell is a count out of 232 with percentage in parentheses. Every original output remains a comparator. Pale red/bold marks the largest correct count within the stated comparison.

Arm	128 correct	128 parsed	512 correct	512 parsed
Anchor	35 (15.1)	100 (43.1)	71 (30.6)	194 (83.6)
GN	71 (30.6)	101 (43.5)	132 (56.9)	220 (94.8)
GS	104 (44.8)	158 (68.1)	143 (61.6)	221 (95.3)
AN	130 (56.0)	187 (80.6)	**156 (67.2)**	225 (97.0)
AS	**145 (62.5)**	218 (94.0)	154 (66.4)	230 (99.1)

All 1,160 formal outputs and both excluded pilots pass the specified release checks. The execution phase lasts 138.34 minutes, including per-call evidence checks rather than kernel-only time or billing. Including the earlier study, the documented extension now contains 4,440 formal calls and sixteen excluded pilots. Its freeze precedes this follow-up's outputs but follows earlier observations; it remains an internal freeze, not external preregistration.

K.3 Paired estimands and uncertainty

Let Y_{irab} and P_{irab} denote correctness and parse validity for base i, option-order index r, condition a, and output-token limit b. An unparsed successful output has $Y = P = 0$, and each condition retains denominator 232. We average the signed differences across the four option orders within each base to obtain d_i. For each persona p, S_p sums these base-level differences and n_p counts its bases. The observed and whole-persona bootstrap estimators are

$$\hat{\delta} = \frac{\sum_p S_p}{58}, \qquad \hat{\delta}^* = \frac{\sum_{p \in B^*} S_p}{\sum_{p \in B^*} n_p}. \tag{S2}$$

Each bootstrap draw samples ten personas with replacement. One shared 50,000-by-ten index matrix uses PCG64 seed 20260909 across conditions and contrasts. The ratio in Equation (S2) gives equal weight to bases rather than unequal-sized persona means. Its 2.5th and 97.5th percentiles form pointwise intervals. Exact two-sided tests enumerate all 1,024 persona signs, using an absolute-statistic tolerance of 1e−12. Sign symmetry is an assumption, not a consequence of randomized assignment.

The simple effects are $D_1 = AN - GN$, $D_2 = AS - GS$, $D_3 = GS - GN$, and $D_4 = AS - AN$. Family A contains the four 512-token effects, while family B contains the paired changes in these effects between output-token limits, $\Psi_k = D_k(512) - D_k(128)$. Each family has its own Holm correction; the original 128-token family remains separate. Within-condition changes and deletion ranges are descriptive. A pointwise interval does not override a corrected test, and nonsignificance does not establish equivalence. The first two changes are placement–budget interactions under neutral and semantic wording; the other two are wording–budget interactions at global and adjacent placement.

Table S26: Factorial effects and their changes between output-token limits (percentage points). Each cell gives estimate, pointwise 95% persona interval, and separately Holm-adjusted p. The 128, 512, and interaction families each contain four tests. Blue marks the two supported placement–budget interactions.

Contrast	128-token effect	512-token effect	Budget interaction
Neutral placement	+25.43 [16.67, 34.05] p = .0117	+10.34 [3.45, 16.95] p = .1328	**−15.09** **[−22.88, −7.63]** **p = .0234**

Contrast	128-token effect	512-token effect	Budget interaction
Semantic placement	+17.67 [13.36, 21.25] p = .0078	+4.74 [0.42, 9.48] p = .3984	**−12.93** **[−17.92, −8.47]** **p = .0078**
Global wording	+14.22 [6.36, 22.77] p = .0234	+4.74 [−1.67, 10.96] p = .4766	−9.48 [−16.82, −0.85] p = .0703
Adjacent wording	+6.47 [3.33, 9.65] p = .0234	−0.86 [−5.36, 3.81] p = .8594	−7.33 [−12.07, −2.54] p = .0625

Under the archived Qwen-only correction families, Table S26 supports attenuation of placement effects under both wording levels as the token limit increases. Neither 512-token placement contrast passes archived family A, although both estimates, pointwise intervals, and leave-one-persona-out ranges stay positive. Nonsignificance does not establish disappearance. Neither wording–budget interaction passes archived family B. These original tests remain for provenance; the main paper uses the expanded cross-model families in Section L.

K.4 Test details and within-arm changes

Table S27: Two-sided persona sign flips and leave-one-persona-out (LOPO) estimates. Holm4 is separate for family A (512 effects) and B (budget interactions). LOPO ranges are not confidence intervals.

Effect	Family	Raw p	Holm p	LOPO min	LOPO max
D1(512)	A	.0332	.1328	+8.96	+12.50
D2(512)	A	.1328	.3984	+3.37	+5.77
D3(512)	A	.2383	.4766	+2.88	+6.73
D4(512)	A	.8594	.8594	−2.40	+0.48
Ψ1	B	.0078	.0234	−17.79	−12.02
Ψ2	B	.0020	.0078	−13.94	−11.06
Ψ3	B	.0703	.0703	−12.98	−7.69
Ψ4	B	.0312	.0625	−8.65	−6.25

Table S28: Within-arm 512-minus-128 changes in percentage points. Intervals are pointwise whole-persona percentiles. These descriptions do not form an additional hypothesis-test family.

Arm	Accuracy change [95% CI]	Parse-coverage change [95% CI]
Anchor	+15.52 [11.40, 19.83]	+40.52 [32.63, 46.98]
GN	+26.29 [19.49, 33.04]	+51.29 [40.68, 60.27]
GS	+16.81 [10.59, 23.31]	+27.16 [18.42, 36.02]
AN	+11.21 [5.93, 16.82]	+16.38 [8.90, 23.73]
AS	+3.88 [0.00, 8.33]	+5.17 [1.72, 8.77]

Table S28 shows larger within-condition gains for the global mappings than for adjacent mappings. The four-outcome interactions in Table S26 directly test that difference. Within-condition increases alone do not: a significant effect at 128 tokens and a nonsignificant effect at 512 tokens are not sufficient evidence of an interaction. Table S27 retains the raw tests, declared corrections, and descriptive deletion ranges.

K.5 Extraction-rule sensitivity

To assess dependence on answer extraction, six post-freeze rules are applied uniformly to all 2,320 saved outputs: the frozen parser, a closed-field rule, first declaration, last declaration, correction priority, and conflict-as-ambiguous handling. No rule is selected for its score. Table S29 reports descriptive ranges across these rules; it neither replaces the frozen primary scorer nor supplies manually adjudicated labels.

Table S29: Six-rule extraction sensitivity (percentage points). Ranges cover all six uniformly applied rules.

Extraction	D1(512)	D2(512)	Ψ1	Ψ2
Frozen primary	10.34	4.74	−15.09	−12.93
Six-rule range	[9.91, 10.78]	[4.74, 5.60]	[−17.67, −15.09]	[−12.93, −12.07]

K.6 RETROSPECTIVE PREFIX ANALYSIS

We separately inspect supported declarations at successive prefixes of the same saved 512-token sequences. This retrospective prefix analysis is not a set of separately executed budget runs. This CPU diagnostic uses a closed-field, persistent, declaration-consensus rule rather than the primary response-level scorer. A complete declaration is available only if its recorded field persists through the endpoint. Available A–D declarations yield a label when they agree; incompatible declarations yield AMBIGUOUS, and absence yields NONE. Only A–D counts as parsed. Because persistence uses the saved endpoint, this is retrospective analysis, not an online stopping rule or an estimate of an optimal output-token limit.

Table S30: Saved-prefix effects versus the frozen primary scorer. Prefix values use the retrospective availability rule; primary values use the separately saved run at that endpoint. Intermediate prefixes have no independent primary run.

Endpoint	Prefix D1	Prefix D2	Primary D1	Primary D2
64	36.21	20.26	–	–
128	27.59	17.67	25.43	17.67
256	11.21	8.19	–	–
384	9.48	6.47	–	–
512	9.91	4.74	10.34	4.74

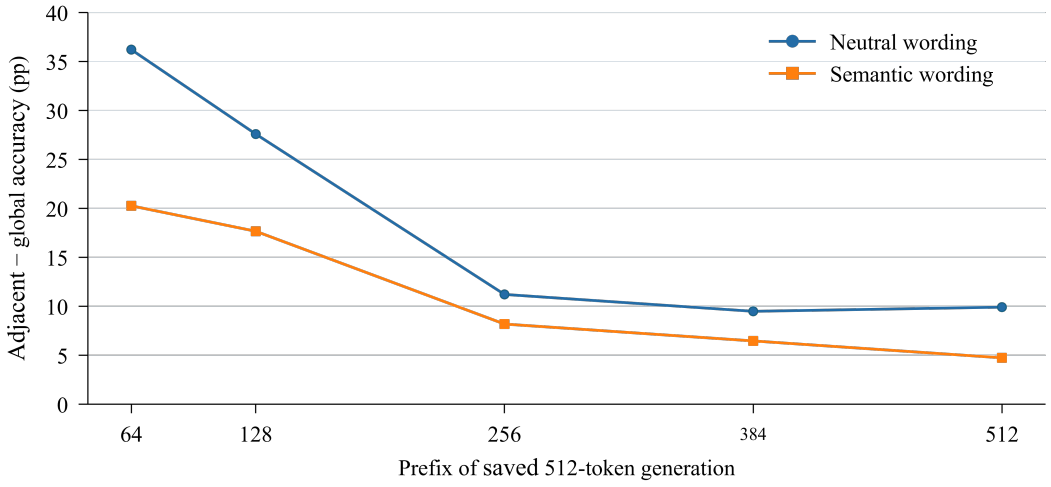

Figure S1: Retrospective declaration-availability curve from the same saved 512-token sequences. Points are dependent prefixes, not independent budget runs. The consensus rule differs from the frozen primary scorer, as Table S30 makes explicit. Lines summarize this diagnostic only; they do not certify budget sufficiency.

Table S30 makes the scoring-rule difference explicit. At 128 tokens, eight of 1,160 endpoint labels differ, all changing A–D to NONE. Five correct GN labels, one GS label, and one AS label disappear, increasing diagnostic D_1 by $5/232$, or 2.16 pp while leaving D_2 unchanged. At 512 tokens, ten labels differ: eight become AMBIGUOUS and two become NONE. Five change correctness; their net effect lowers D_1 by $1/232$, or 0.43 pp and leaves D_2 unchanged. These differences reconcile the diagnostic endpoints without changing primary scores.

All 1,160 comparable token prefixes match across the actual 128- and 512-token runs. A historical generation ending before token 128 also ends there in the longer-budget run. Equal text can nevertheless receive different labels under different extractors. Only the two separately executed full matrices enter the paired budget-interaction tests. The intermediate prefixes in Figure S1 are dependent observations from saved trajectories, not additional experimental replications.

K.7 REPORTING STATES AND INCOMPLETE OUTPUTS

To distinguish missing answers from wrong answers, each output is assigned U (unparsed), W (parsed but wrong), or C (correct). Tables S32–S33 retain every paired case, and their margins reproduce the condition totals. Comparisons restricted to jointly parsed cases condition on observed outcomes. They describe selection and reporting patterns but do not identify a causal formatting pathway.

Table S31: Generation-length and termination evidence. Median includes [Q1, Q3]. Cap means token length equals the limit. U/P cap distinguish unparsed/parsed outputs. Unavailable termination causes remain unknown.

Arm / limit	Median [Q1, Q3]	Range	Cap	U cap	P cap	Normal	Unknown
Anchor /128	128 [121.75, 128]	63–128	165	130	35	0	232
GN /128	128 [123.5, 128]	54–128	166	131	35	0	232
GS /128	128 [99.75, 128]	47–128	131	74	57	0	232
AN /128	120.5 [93, 128]	56–128	102	45	57	0	232
AS /128	110.5 [90.75, 128]	58–128	78	14	64	0	232
Anchor /512	160.5 [121.75, 242.25]	63–512	28	24	4	204	28
GN /512	173 [123.5, 237.5]	54–512	20	12	8	212	20
GS /512	138.5 [99.75, 223.5]	47–512	20	11	9	212	20
AN /512	120.5 [93, 187.25]	56–512	24	7	17	208	24
AS /512	110.5 [90.75, 142.25]	58–512	14	2	12	218	14

Table S32: Complete within-arm transitions from 128 to 512. Every row sums to 232 paired outputs; column order describes earlier → later state.

Arm	U→U	U→W	U→C	W→U	W→W	W→C	C→U	C→W	C→C
Anchor	38	57	37	0	65	0	0	1	34
GN	12	58	61	0	30	0	0	0	71
GS	11	24	39	0	54	0	0	0	104
AN	7	14	24	0	55	2	0	0	130
AS	2	3	9	0	72	1	0	1	144

At 512 tokens, seventy outputs remain unparsed. Fifty-six reach the limit with unknown termination causes; the other fourteen are below-cap Anchor outputs that explicitly choose E/F and end with EOS. Unparsed cap-hit counts are 24/12/11/7/2 for Anchor/GN/GS/AN/AS. The prespecified reporting threshold is at least 12/232 unparsed cap hits without confirmed normal termination in a condition. Anchor and GN reach this numerical threshold, leaving budget sufficiency uncertain. It does not automatically trigger a 1024-token experiment. An E/F response also does not identify the selected image or prove internal role confusion.

The tokenizer audit reconstructs all 2,320 raw decodes. Complete FINAL=<A-D> fields appear in 194/219/220/225/230 longer outputs and first complete after token 128 in 95/123/63/38/13. Two malformed GN/GS FINAL=<D fields are accepted by the frozen parser but excluded from this complete-field diagnostic. Seven longer outputs contain conflicting supported declarations. These checks characterize answer reporting, not correctness; additional tokens permit both computation and later declarations.

K.8 Placement transitions and option-order consistency

Table S33: Complete global-to-adjacent transitions. Each row retains 232 paired outputs; state definitions match Table S32.

Comparison	U→U	U→W	U→C	W→U	W→W	W→C	C→U	C→W	C→C
GN–AN /128	44	26	61	0	22	8	1	9	61
GS–AS /128	14	25	35	0	37	17	0	11	93
GN–AN /512	1	2	9	3	49	36	3	18	111
GS–AS /512	2	6	3	0	52	26	0	18	125

Table S34: Consistency over four option orders, counts out of 58 bases. Same original requires four parses of the same underlying candidate and can include consistently wrong selections.

Arm / limit	4 parsed	4 correct	Same original	Same letter
Anchor /128	11	0	0	1
GN /128	17	9	10	1
GS /128	35	15	16	4
AN /128	45	20	21	3
AS /128	51	25	26	6
Anchor /512	32	3	3	2
GN /512	49	19	23	3
GS /512	51	21	22	6
AN /512	55	27	29	3

Arm / limit	4 parsed	4 correct	Same original	Same letter
AS /512	56	27	31	5

Table S35: Distribution of correctly answered option orders per base. Each row sums to 58; an unparsed response is incorrect.

Arm / limit	0 correct	1	2	3	4
Anchor /128	32	19	5	2	0
GN /128	28	11	6	4	9
GS /128	15	17	6	5	15
AN /128	14	8	6	10	20
AS /128	7	13	7	6	25
Anchor /512	18	21	10	6	3
GN /512	10	12	7	10	19
GS /512	5	13	11	8	21
AN /512	8	7	7	9	27
AS /512	8	7	9	7	27

Tables S34–S35 examine whether gains persist across all four option orders of a base. Both adjacent conditions have 27/58 all-four-correct bases at 512 tokens. Same-original-choice counts are 29/58 for AN and 31/58 for AS, which differs from always selecting the same displayed letter. A broader post-freeze screen flags 41 unique outputs across budgets: ten incompatible complete declarations, ten unclosed FINAL forms, and 24 E/F declarations, with overlap. Its 512-token subset has 26 unique outputs, with counts of nine/two/seventeen under those flags. This screen uses broader definitions than the seven-conflict/two-malformed/fourteen-below-cap diagnostic. One flagged output belongs to the random core. The remaining risk list is not pooled into human agreement estimates.

K.9 Independent human answer extraction

The human audit fixes twenty outputs in each budget-by-condition cell, totaling 200. Selection ranks SHA-256 of protocol ID F58_BUDGET_FOLLOWUP_V1, seed 20260909, budget, condition, and case ID. It uses no response text, correctness, parse validity, or length, and is fixed before inspecting the 512-token outputs. Both completed annotation passes contain every selected row. The optional outcome-conditioned risk packet stays disabled; the separate 41-output screen does not enter the audit denominator.

Annotators receive separate offline packets with opaque IDs, complete raw outputs, the A–D answer space, and shared extraction instructions. Their task is to extract the stated answer, not solve the visual question. Labels are A/B/C/D/NONE/AMBIGUOUS, with a verbatim supporting span where applicable and an answer–explanation-conflict flag. Clear later corrections replace earlier declarations. Incompatible explicit choices without a clear correction are AMBIGUOUS; E/F or no definite answer is NONE. Response wording and length can reveal condition clues, so perfect blinding is not claimed.

Two distinct human annotators independently label all 200 outputs without access to gold, parser predictions, budgets, conditions, the analyst key, or each other's labels. Both files are locked before comparison. Six extraction-label disagreements and two conflict-flag disagreements occur on eight distinct outputs. A third human adjudicates exactly those eight under the same restrictions. Form trials and the earlier H57 packet remain excluded. The authors attest identity, independence, and applicable consent and ethics arrangements; software verifies records and arithmetic, not those personal facts.

Table S36: Completed human extraction audit. Proportion intervals are pointwise Wilson intervals. The unweighted κ interval uses 10,000 paired-output bootstrap draws, seed 20260910.

Quantity	Count / estimate	95% interval
Two-rater label agreement	194/200 = .970	[.9361, .9862]
Unweighted Cohen's κ	.961971	[.929761, .987367]
Conflict-flag agreement	198/200 = .990	[.9643, .9973]
Parser–adjudicated label agreement	**193/200 = .965**	[.9295, .9829]
Human-extracted task accuracy	96/200 = .480	[.4118, .5490]
Frozen-parser task accuracy	95/200 = .475	[.4069, .5440]

Quantity	Count / estimate	95% interval
Human–parser correctness agreement	**199/200 = .995**	[.9722, .9991]

Table S37: Extraction-label distributions. Adjudication replaces both labels for the eight outputs in the label-or-conflict disagreement union and retains the common label otherwise.

Label source	A	B	C	D	NONE	AMBIG.
Pass 1	54	30	50	27	38	1
Pass 2	52	30	52	27	37	2
Adjudicated	53	30	51	27	37	2

K.10 Agreement structure and sampled scope

Table S38: Complete extraction-label confusion matrix. Rows are pass 1 and columns are pass 2. The diagonal sums to 194.

Pass 1 / Pass 2	A	B	C	D	NONE	AMBIG.	Total
A	**52**	0	1	0	0	1	54
B	0	**30**	0	0	0	0	30
C	0	0	**49**	0	1	0	50
D	0	0	1	**26**	0	0	27
NONE	0	0	1	1	**36**	0	38
AMBIG.	0	0	0	0	0	**1**	1
Total	52	30	52	27	37	2	200

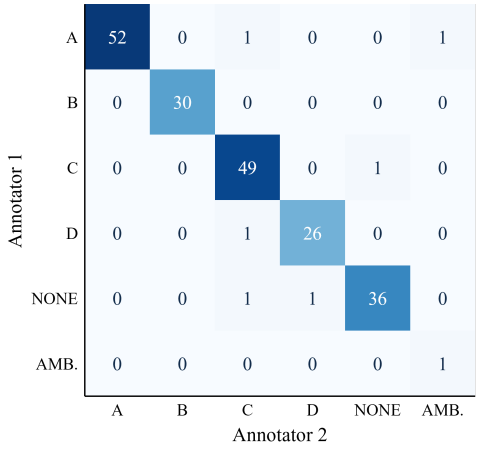

Figure S2: The same two-rater counts as Table S38. Color intensity shows count, not accuracy or a model confidence score. The six off-diagonal entries concern stated-answer extraction; the eight-output adjudication union additionally includes two distinct conflict-flag disagreements.

Table S39: Boolean answer/explanation-conflict matrix. Rows are pass 1; columns are pass 2. The two off-diagonal cases are disjoint from the six label disagreements.

Pass 1 / Pass 2	False	True	Total
False	196	2	198
True	0	2	2
Total	196	4	200

The confusion matrix yields 194/200 raw label agreement and $\kappa = .961971$. After adjudication, the parser matches 193 labels, and the seven label differences change correctness on only one output. Human extraction gives 96/200 correct, compared with 95/200 for the frozen parser. This supports extraction on the selected core, not every generation or the separate risk list. The audit evaluates reported answers rather than visual reasoning, and no human label replaces a frozen primary prediction or its declared test family.

K.11 Descriptive results within each sampled cell

Table S40 retains all ten sampled budget-by-condition cells, each with twenty outputs. Label agreement compares the frozen parser with adjudicated extraction. Correctness agreement compares their binary scores

against gold. The small-cell intervals are descriptive and pointwise; they are not new condition-level hypothesis tests or a re-estimate of the budget interaction.

Table S40: Human audit by budget and arm. Each count is accompanied by its Wilson 95% interval; no row is a separate confirmatory test.

Budget / arm	Label agreement	Human correct	Parser correct	Correctness agreement
128 Anchor	17/20 [.6396, .9476]	2/20 [.0279, .3010]	2/20 [.0279, .3010]	20/20 [.8389, 1.0000]
128 GN	19/20 [.7639, .9911]	4/20 [.0807, .4160]	4/20 [.0807, .4160]	20/20 [.8389, 1.0000]
128 GS	20/20 [.8389, 1.0000]	12/20 [.3866, .7812]	12/20 [.3866, .7812]	20/20 [.8389, 1.0000]
128 AN	20/20 [.8389, 1.0000]	7/20 [.1812, .5671]	7/20 [.1812, .5671]	20/20 [.8389, 1.0000]
128 AS	19/20 [.7639, .9911]	9/20 [.2582, .6579]	9/20 [.2582, .6579]	20/20 [.8389, 1.0000]
512 Anchor	18/20 [.6990, .9721]	7/20 [.1812, .5671]	6/20 [.1455, .5190]	19/20 [.7639, .9911]
512 GN	20/20 [.8389, 1.0000]	11/20 [.3421, .7418]	11/20 [.3421, .7418]	20/20 [.8389, 1.0000]
512 GS	20/20 [.8389, 1.0000]	15/20 [.5313, .8881]	15/20 [.5313, .8881]	20/20 [.8389, 1.0000]
512 AN	20/20 [.8389, 1.0000]	15/20 [.5313, .8881]	15/20 [.5313, .8881]	20/20 [.8389, 1.0000]
512 AS	20/20 [.8389, 1.0000]	14/20 [.4810, .8545]	14/20 [.4810, .8545]	20/20 [.8389, 1.0000]

The ten cells sum to 193 parser–human label agreements, 96 human correct answers, 95 parser correct answers, and 199 matching correctness indicators. The aggregate κ interval uses 10,000 paired-output bootstrap resamples with seed 20260910. These uncertainty summaries concern the reported sample, not simultaneous bounds over all metrics. In particular, 96.5% denotes extraction-label agreement, not task accuracy.

The sampling design is stratified across budgets and serializers, not conditioned on difficult outcomes. The broader post-freeze risk screen remains separate. This preserves the intended scope of the human validation. Automated annotation or format checks do not count as an independent human pass.

K.12 Artifact locations and post-lock replay

Paths in this section are relative to archived_studies/ in the anonymous companion, not the full author-side code tree. The registered budget files are data/budget_contrasts_v1.csv and data/budget_interactions_v1.csv; data/prefix_effects_v1.csv contains the retrospective diagnostic curve. The file data/endpoint_audit_summary_v1.json records 2,320 endpoint comparisons, eighteen differing labels, the 1,160-pair prefix check, and effect reconciliation without raw text. The original python verify_all.py command checks this historical package, not the new extension. The root README documents the combined replay.

The human-scoring linkage is joined only after both annotation passes and adjudication are locked. The file data/human_scoring_linkage_v1.jsonl contains opaque output IDs, text digests, paired labels, conflict flags, adjudication, budget, condition, pseudonymous base/case IDs, rotation (the option-order index), parser predictions, gold labels, and correctness. It excludes raw responses, supporting spans, free-form notes, identities, consent records, and private paths. The aggregate comparison is stored in data/human_scoring_aggregate_reference_v1.json.

In the external working copy, python -m src.recompute joins the 200 records by output_id and checks twenty rows per cell. It reconstructs 193/200 label agreement, 96/200 human correctness, 95/200 parser correctness, and 199/200 correctness agreement, writing results/recomputed_metrics.json in that working copy. Blinded collection addresses annotation bias; post-lock release enables numerical replay without disclosing personal information.

The author-side audit retains full generations, parser source, execution freezes, and token traces. Original human returns, administrative keys, and consent records remain private. The anonymous human_evidence/ directory separately retains approved verbatim support spans, labels, conflict fields, and adjudications. The thin 200-row linkage enables numerical comparisons; released spans support extraction decisions, not a complete raw-generation replay. Neither file-level checks nor statistical recomputation independently establish annotator identity or data collection.

The author-side code tree records the archived workflow under results/f58_budget_followup_v1/, earlier outputs under results/gpu_0909/, and the protocol at protocols/f58_budget_followup_v1.md. These names identify provenance; they are not all promised paths in the smaller anonymous companion. The public-code commitment covers our implementation and analysis, subject to third-party rights and private human records. Section L.6 distinguishes the derived-label replay from inference reproduction.

Together, these materials support the specified estimates of serializer effects, budget sensitivity, sampled extraction agreement, and changes in retrieval comparisons. They do not establish universal prompt optimality, an internal role-confusion state, a sufficient 512-token budget, or failure prevalence across systems. Archived controls, non-replications, unresolved provenance, and the distinction between primary scoring and retrospective diagnostics remain part of the evidence.

L Expanded Experiments And Analysis

L.1 Scope And Reuse Of Earlier Results

This extension examines cross-model budget sensitivity and a separate external cohort. Sections A–K retain the earlier experiments and their original protocols. The extension followed feedback on that manuscript; it was not externally preregistered. A broader plan listed 47,008 generation slots. Runtime and native-template checks, completed before the new formal results, motivated the narrower batch described here.

The final extension comprises 13,920 F58 generations and 1,440 external generations. Qwen3.5-9B, Gemma 3-4B, and InternVL3.5-8B-HF each use all 58 questions, four cyclic option orders, five arms, and four independently executed output-token limits: 64, 128, 256, and 512. The Qwen matrix is a same-hardware RTX 4090 rerun; the archived RTX 5090 128/512-token endpoints from Section K are retained for provenance and the human audit but are not mixed into the cross-model curves or tests. These totals exclude technical pilots and reused outputs; the other proposed experiments were not run.

All 15,360 scheduled new formal generations completed: 4,640 for each of the three F58 factorial models and 360 for each of the four external-cohort models. Formal journals contain no failed or duplicate generated calls. Technical gate calls are excluded from these totals. The complete factorial analysis therefore uses 13,920 saved outputs across twelve model–budget matrices. The Qwen producer loop completed all 4,640 calls in 350.2 minutes; this is recorded execution time rather than billed rental time.

L.2 Inputs And Execution Conditions

The F58 manifest fixes the 58 base questions, image assets, and four label-balanced orders. Within each model, the budget comparison preserves the checkpoint, source requests, preprocessing, greedy decoding, answer cleaner, and parser. The runs record checkpoint-content hashes, prepared input IDs and tensor hashes, generation arguments, and output-token journals. Gemma and InternVL use the previously frozen counterbalanced schedule; the Qwen rerun enumerates all four budgets for every case–arm combination in one frozen 4,640-key schedule. Wrong or unparsed successful generations are retained rather than rerun. All Qwen statistics in Section L use the contemporaneous same-hardware records; reconstructed settings from the archived endpoint study are not mixed into this analysis.

The final F58 four-budget matrices were executed on hosts reporting NVIDIA GeForce RTX 4090; the archived Qwen RTX 5090 endpoints in Section K are not combined with them. We report observed device identities rather than requested rental categories. Budget effects are within-model comparisons; cross-model differences also reflect processors, software stacks, and potentially host-specific numerical execution. The runs use PyTorch 2.12.1+cu130 and CUDA runtime 13.0.

Qwen, Gemma, and InternVL use Transformers 5.15.1; Molmo uses Transformers 4.57.1 with its pinned local processor. Models run in evaluation mode with greedy decoding and seed 20260909. Deterministic algorithms were not enabled, so these settings do not guarantee bitwise equivalence across hardware or software versions.

Native-template checks cover all 1,160 source requests per model. Gemma and InternVL preserve the intended mapping placement. Molmo2 places images before the mapping labels. Its input streams remain distinct, but the intended image adjacency is not preserved. Its factorial placement estimand is therefore NA, not a measured zero effect. Technical pilots check selected prepared inputs; they do not establish exhaustive processor equivalence. For each of the three valid factorial models, all 4,640 calls completed without technical failures or retries. Within each model, the 3,480 comparisons pairing the 128-token run with the 64-, 256-, and 512-token runs matched on actual input IDs, tensor hashes, and non-budget settings. Separate CPU tokenizer checks recovered the saved raw text for all 13,920 outputs. Journals record a distinct generation call for every scheduled budget. These paired checks do not create new independent questions.

The new Gemma 128-token run changed 47 of 1,160 answer labels relative to the archived run (Table S17). Eighteen incorrect answers became correct, thirteen correct answers became incorrect, and sixteen changed between incorrect labels. Recorded checkpoint files and actual input tensors match, but the executions used different GPUs and did not record every runtime detail. We do not attribute this difference to hardware. Gemma budget effects use only the new within-run pair; archived counts are neither replaced nor mixed into that pair.

L.3 Estimation And Statistical Tests

We average the four cyclic orders within each base question and weight the 58 bases equally. Intervals resample the ten whole personas, preserving all bases within a sampled persona. Each bootstrap estimate divides sampled effect sums by sampled base counts. We use 50,000 PCG64 resamples, seed 20260909, and the 2.5th and 97.5th percentiles. These are unadjusted pointwise intervals, not simultaneous bands or inverted sign-flip confidence sets. A positive interval therefore need not imply even an unadjusted sign-flip p below .05. Leave-one-persona-out estimates assess sensitivity to individual clusters.

The four endpoint effects are neutral placement (D1), semantic placement (D2), global wording (D3), and adjacent wording (D4), as defined in main Equations (2)–(3). For each effect, the budget interaction is its paired 512-minus-128 change. Two-sided tests enumerate all 1,024 persona-sign assignments, including zero-sum clusters. They assume symmetric persona-level differences under the null and are not randomized-treatment tests. A significant effect at one budget and a nonsignificant effect at another do not establish an interaction.

The fixed Holm family sizes are 16 for the 128-token effects, 16 for the 512-token effects, eight for placement–budget interactions, and eight for wording–budget interactions. Unsupported Molmo slots remain publicly NA; internal p=1 values only retain the planned family sizes. All Qwen endpoint and interaction p-values in Section L are recalculated jointly with Gemma and InternVL in these cross-model families. The archived Qwen-only corrections in Sections A–K remain separate. The actual 64/256-token runs are descriptive, independently executed budgets rather than retrospective prefixes or additional independent questions.

L.4 Counts And Extraction Sensitivity

Table S41 reports correct and parsed counts separately as C;P, each with denominator 232. Unparsed successful outputs count as incorrect. Technical failures are not relabelled as answers; a complete matrix requires every scheduled call. Tables S42–S44 give all endpoint effects and paired budget interactions with their declared tests. Gemma and InternVL retain the same primary answer label for every case and arm across all four budgets; their observed 512-minus-128 interactions are zero, which does not establish population equivalence. In the same-hardware Qwen rerun, the neutral and semantic placement interactions are −20.26 and −11.21 pp, with Holm8 p=.03125 and .109375, respectively. Thus, only the neutral interaction passes the declared .05 threshold. Qwen's 512-token placement effects are +5.60 and +8.62 pp, and neither passes the 16-slot endpoint family. InternVL's global-wording effect remains −13.79 pp at both endpoints (Holm16 p=.03125), providing a counterexample to universal benefit from explicit role words.

Table S41: All completed natural-task budget cells. Each C;P cell has N=232; models and budgets are repeated measurements of 58 bases. All 64/128/256/512 points are separately generated full-budget runs rather than truncated 512-token sequences. The four Qwen points come from one same-hardware RTX 4090 rerun; archived RTX 5090 endpoints are not mixed.

Model	Budget	Anchor C;P	GN C;P	GS C;P	AN C;P	AS C;P
Qwen	64	29;76	40;68	92;153	125;180	**144;218**
Qwen	128	39;100	69;101	99;160	129;184	**145;219**
Qwen	256	71;174	124;196	127;205	145;210	**150;228**
Qwen	512	84;198	138;219	132;222	151;222	**152;230**
Gemma	64	65;232	64;232	63;232	**80;232**	76;232
Gemma	128	65;232	64;232	63;232	**80;232**	76;232
Gemma	256	65;232	64;232	63;232	**80;232**	76;232
Gemma	512	65;232	64;232	63;232	**80;232**	76;232
InternVL	64	53;224	102;224	70;230	**116;232**	111;232
InternVL	128	53;224	102;224	70;230	**116;232**	111;232
InternVL	256	53;224	102;224	70;230	**116;232**	111;232
InternVL	512	53;224	102;224	70;230	**116;232**	111;232

Table S42: 128-token simple effects. D1=AN-GN, D2=AS-GS, D3=GS-GN, D4=AS-AN. Ten-persona bootstrap intervals are pointwise; Holm retains 16 planned slots including unsupported Molmo slots (public NA; p=1 is used internally only to retain the planned family size).

Model	Effect	Estimate (pp)	95% CI	Raw p	Holm16 p
Qwen	D1	+25.86	[17.67, 34.58]	.002	**.031**
Qwen	D2	+19.83	[15.09, 24.14]	.002	**.031**
Qwen	D3	+12.93	[5.08, 21.43]	.018	.158
Qwen	D4	+6.90	[4.31, 9.55]	.004	.051
Gemma	D1	+6.90	[4.39, 9.48]	.004	.051
Gemma	D2	+5.60	[3.02, 8.47]	.008	.078
Gemma	D3	−0.43	[−2.23, 1.32]	1.000	1.000
Gemma	D4	−1.72	[−6.25, 2.27]	.625	1.000
InternVL	D1	+6.03	[0.44, 11.86]	.105	.844
InternVL	D2	+17.67	[12.05, 22.84]	.004	.051
InternVL	D3	−13.79	[−19.83, −8.77]	.002	**.031**
InternVL	D4	−2.16	[−8.19, 3.39]	.611	1.000

Table S43: 512-token simple effects. D1=AN-GN, D2=AS-GS, D3=GS-GN, D4=AS-AN. Ten-persona bootstrap intervals are pointwise; Holm retains 16 planned slots including unsupported Molmo slots (public NA; p=1 is used internally only to retain the planned family size).

Model	Effect	Estimate (pp)	95% CI	Raw p	Holm16 p
Qwen	D1	+5.60	[−0.45, 11.44]	.141	1.000
Qwen	D2	+8.62	[3.75, 13.39]	.023	.281
Qwen	D3	−2.59	[−8.80, 2.97]	.523	1.000
Qwen	D4	+0.43	[−3.39, 3.95]	1.000	1.000
Gemma	D1	+6.90	[4.39, 9.48]	.004	.059
Gemma	D2	+5.60	[3.02, 8.47]	.008	.102
Gemma	D3	−0.43	[−2.23, 1.32]	1.000	1.000
Gemma	D4	−1.72	[−6.25, 2.27]	.625	1.000
InternVL	D1	+6.03	[0.44, 11.86]	.105	1.000
InternVL	D2	+17.67	[12.05, 22.84]	.004	.059
InternVL	D3	−13.79	[−19.83, −8.77]	.002	**.031**
InternVL	D4	−2.16	[−8.19, 3.39]	.611	1.000

Table S44: Paired 512-minus-128 changes in the four effects. Placement changes (Ψ1/Ψ2) and wording changes (Ψ3/Ψ4) use separate Holm8 families. A change in significance between endpoints is not itself an interaction.

Model	Change	Estimate (pp)	95% CI	Raw p	Holm8 p
Qwen	Ψ1	−20.26	[−28.51, −12.28]	.004	**.031**
Qwen	Ψ2	−11.21	[−17.37, −5.36]	.016	.109
Qwen	Ψ3	−15.52	[−25.00, −6.90]	.012	.094
Qwen	Ψ4	−6.47	[−10.17, −3.02]	.016	.109
Gemma	Ψ1	+0.00	[0.00, 0.00]	1.000	1.000
Gemma	Ψ2	+0.00	[0.00, 0.00]	1.000	1.000
Gemma	Ψ3	+0.00	[0.00, 0.00]	1.000	1.000
Gemma	Ψ4	+0.00	[0.00, 0.00]	1.000	1.000
InternVL	Ψ1	+0.00	[0.00, 0.00]	1.000	1.000
InternVL	Ψ2	+0.00	[0.00, 0.00]	1.000	1.000
InternVL	Ψ3	+0.00	[0.00, 0.00]	1.000	1.000
InternVL	Ψ4	+0.00	[0.00, 0.00]	1.000	1.000

The frozen primary parser remains the main analysis. We apply all six extraction rules uniformly to all 13,920 saved F58 outputs, including strict-closed, first-answer, last-answer, correction-priority, and conflict-as-ambiguous alternatives. Out-of-range answers remain a separate category. Table S45 gives the full min/max effect ranges; these are sensitivity ranges rather than confidence intervals. Gemma's labels agree across all six rules. For InternVL, the strict-closed rule differs from the frozen primary rule on 314 outputs per budget, changing correctness for 125; these cases use plain FINAL=A-style declarations without angle brackets. InternVL's global-wording effect remains negative, ranging from −20.26 to −13.79 pp. Qwen's Ψ1 and Ψ2 ranges remain negative, [−21.12, −20.26] and [−11.64, −11.21]. Gemma and InternVL retain zero budget interactions under every rule. The extension risk inventory contains 74 output IDs: 58 same-hardware Qwen outputs and 16 InternVL outputs. Among the Qwen rows, 17 contain conflicting supported declarations, 35 contain out-of-range declarations, and 13 contain unclosed FINAL forms, with overlap. The InternVL rows are four out-of-range E responses repeated at four budgets. The archived 41-output Qwen screen in Section K remains separate. Neither inventory is a random sample or an estimate of population error rates.

Table S45: Full min/max ranges across all six frozen extraction rules, in percentage points. These are sensitivity ranges, not confidence intervals or independent replications. The frozen primary parser remains primary; no favorable rule is selected. Qwen rows cover all four independently generated budgets; the 512-minus-128 rows are paired interactions.

Model / budget	D1 or Ψ1 range	D2 or Ψ2 range	D3 or Ψ3 range	D4 or Ψ4 range
Qwen / 64	[36.64, 37.07]	[22.41, 22.84]	[22.41, 22.41]	[8.19, 8.19]
Qwen / 128	[25.86, 27.16]	[19.83, 20.26]	[12.93, 13.79]	[6.90, 6.90]
Qwen / 256	[9.05, 9.48]	[9.91, 9.91]	[1.29, 1.29]	[1.72, 2.16]
Qwen / 512	[5.60, 6.47]	[8.62, 9.05]	[−2.59, −2.59]	[−0.43, 0.43]
Qwen / 512−128	[−21.12, −20.26]	[−11.64, −11.21]	[−16.38, −15.52]	[−7.33, −6.47]
Gemma / 128	[6.90, 6.90]	[5.60, 5.60]	[−0.43, −0.43]	[−1.72, −1.72]
Gemma / 512	[6.90, 6.90]	[5.60, 5.60]	[−0.43, −0.43]	[−1.72, −1.72]
Gemma / 512−128	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
InternVL / 128	[2.59, 6.03]	[17.67, 18.53]	[−20.26, −13.79]	[−4.31, −2.16]
InternVL / 512	[2.59, 6.03]	[17.67, 18.53]	[−20.26, −13.79]	[−4.31, −2.16]
InternVL / 512−128	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]

The historical seven-output conflict screen and parser edge cases remain separate from the stratified 200-output human sample. Saved-length diagnostics appear in Table S50. At 64 tokens, Qwen, Gemma, and InternVL have 1,151, 533, and 78 cap-length responses, respectively. Qwen's unparsed-at-cap total falls from 464 at 64 tokens to 394 at 128, 139 at 256, and 60 at 512. Cap length alone does not identify truncation. The Qwen 64/128/256/512 points are distinct generation calls on the same recorded hardware, not offline slices of a 512-token output. Existing Gemma and

InternVL token-prefix diagnostics remain descriptive; separate call journals establish their execution schedules.

L.5 New External Cohort

The new MuirBench cohort contains 120 questions, with 24 from each of five tasks. Selection uses deterministic, outcome-independent ID hashing and task round-robin ordering. It excludes the archived 40 questions and their identified counterparts, exact decoded-RGB overlaps, and repeated members of transitive image-duplicate groups. These rules do not guarantee universal source-ID, near-duplicate, or pretraining exclusion. The anonymous companion's external_cohort/cohort_manifest_public.json links the numerical case aliases to public MuirBench row IDs, tasks, gold labels, counterpart IDs, target-image positions and digests, and two reference donor records. Its export_receipt.json distinguishes the original frozen manifest hash from the derived public projection's hash. The projection includes the 80 excluded archived or counterpart IDs and source-shard digests, but no question text, option text, model responses, or image bytes. It permits record linkage against the pinned upstream release; full selection replay also requires the source shards and frozen selection algorithm. Exact RGB checks do not rule out near duplicates, common latent sources, or training-data overlap.

All four models share the frozen cohort and a 512-token limit. Each technical gate repeats the first frozen question; these separate calls are excluded from the formal statistics, while that question's scheduled formal calls remain included. No-ref retains typed task images without injected references. Flat includes reference and task images without role-label blocks; RPS adds these blocks. For each question, the injected conditions use the same two fixed other-task reference images and image order. The original answer extractor remains fixed, including legal labels up to A–E. Molmo's comparison concerns its native transformed serialization, not preserved image adjacency. The archived 40-question study used 32 tokens and remains separate in Table S13.

RPS-minus-Flat is the primary paired comparison for each model. We use 50,000 whole-task ratio bootstrap resamples with PCG64 seed 20260909 and enumerate all 32 task-sign assignments. The four primary tests use Holm4. With five clusters, the minimum nonzero two-sided exact p is .0625; no primary test can reach .05, even if a percentile interval excludes zero. No-ref comparisons are descriptive. Repeating the tasks across models does not create additional independent task clusters.

Table S46 reproduces main Table 6 for this cohort; Table S47 gives all four primary tests and Table S48 all task strata. RPS-minus-Flat estimates are +27.50 pp for Qwen, +10.00 for Gemma, +6.67 for Molmo, and +16.67 for InternVL. All four primary Holm4 p values are .25. Their positive percentile intervals do not override the exact-test resolution limit. Qwen parsing rises from 92/120 under Flat to 119/120 under RPS. InternVL parses only 29/120 in each injected condition; the parsed sets overlap on 15 questions, with 14 distinct questions in each. Thus equal aggregate parsing rates do not remove extraction availability as an explanation. Across its three conditions, 265/360 outputs are unparsed and reach the 512-token cap without observed final EOS. This is a saved-token diagnostic, not a separately recorded internal stopping cause. RPS remains below No-ref for Qwen, Gemma, and Molmo; No-ref comparisons are descriptive. All 1,440 formal external outputs pass source/journal checks and CPU tokenizer re-decoding.

Table S46: New external-cohort C;P counts and RPS-minus-Flat paired differences, repeated from main Table 6 for a self-contained protocol record. N=120 in each cell.

Model	No-ref C;P	Flat C;P	RPS C;P	RPS - Flat, pp [95% CI]
Qwen3.5-9B	53;115	18;92	51;119	+27.50 [11.67, 43.33]
Gemma 3-4B	51;120	29;120	41;120	+10.00 [0.83, 20.83]
Molmo2-8B	26;120	11;119	19;120	+6.67 [1.67, 14.17]
InternVL3.5-8B-HF	23;37	7;29	27;29	+16.67 [9.17, 25.83]

Table S47: Four planned external RPS-minus-Flat tests. The task-cluster sign-flip distribution has 32 assignments and minimum two-sided p=.0625; no .05 significance claim is possible. No-ref contrasts are descriptive.

Model	Wins / harms	Raw sign-flip p	Holm4 p
Qwen3.5-9B	39 / 6	0.0625	0.250
Gemma 3-4B	23 / 11	0.1875	0.250
Molmo2-8B	8 / 0	0.1250	0.250
InternVL3.5-8B-HF	23 / 3	0.0625	0.250

Table S48: All five external task strata for all four models. Each cell has N=24. No task is excluded on the basis of performance; models do not multiply the number of independent task clusters.

Model	Task	No-ref C;P	Flat C;P	RPS C;P	RPS-Flat (pp)
Qwen3.5-9B	Attribute Similarity	0;23	2;21	5;24	+12.50
Qwen3.5-9B	Diagram Understanding	21;24	10;18	21;24	+45.83
Qwen3.5-9B	Difference Spotting	11;21	3;20	4;23	+4.17
Qwen3.5-9B	Geographic Understanding	8;23	1;19	7;24	+25.00
Qwen3.5-9B	Visual Retrieval	13;24	2;14	14;24	+50.00
Gemma 3-4B	Attribute Similarity	8;24	6;24	5;24	-4.17
Gemma 3-4B	Diagram Understanding	11;24	9;24	11;24	+8.33
Gemma 3-4B	Difference Spotting	8;24	4;24	5;24	+4.17
Gemma 3-4B	Geographic Understanding	9;24	7;24	10;24	+12.50
Gemma 3-4B	Visual Retrieval	15;24	3;24	10;24	+29.17
Molmo2-8B	Attribute Similarity	3;24	0;24	0;24	+0.00
Molmo2-8B	Diagram Understanding	9;24	6;24	7;24	+4.17
Molmo2-8B	Difference Spotting	9;24	4;23	9;24	+20.83
Molmo2-8B	Geographic Understanding	1;24	0;24	1;24	+4.17
Molmo2-8B	Visual Retrieval	4;24	1;24	2;24	+4.17
InternVL3.5-8B-HF	Attribute Similarity	4;13	0;8	2;2	+8.33
InternVL3.5-8B-HF	Diagram Understanding	9;9	3;5	8;8	+20.83
InternVL3.5-8B-HF	Difference Spotting	0;0	0;0	3;4	+12.50
InternVL3.5-8B-HF	Geographic Understanding	3;8	1;6	3;4	+8.33
InternVL3.5-8B-HF	Visual Retrieval	7;7	3;10	11;11	+33.33

L.6 Retrieval Reanalysis And Reproducibility Scope

Table S49 reports all 18 serializer-by-method interactions from the archived 128-token Gemma and InternVL runs. It covers All (696 questions), Image (58), and Text (638); Figure 4 shows All and Image. The Full–Naive, Retrieval-only–Naive, and Full–Retrieval-only contrasts are algebraically related. For both All and Image, conservative corrections retain eight planned slots for Full–Naive and 16 for the other two contrasts; Text is descriptive. This CPU reanalysis is not the proposed new four-model 512-token retrieval experiment. The historical Gemma checkpoint revision remains unknown.

Table S49: All 18 CPU-reanalyzed historical 128-token retrieval interactions. Conservative p retains original planned family sizes (8 for Full-Naive, 16 for other pairs where specified); Text results remain descriptive. Figure 4 shows All and Image. The three method differences are algebraically related, not independent replications.

Model / subset	Method contrast	Shift (pp)	95% CI	Raw p	Conservative p
Gemma / all	Full-Naive	+1.58	[0.44, 2.74]	0.062	0.375
Gemma / all	Retrieval-only-Naive	+0.00	[-0.74, 0.86]	1.000	1.000
Gemma / all	Full-Retrieval-only	+1.58	[0.60, 2.58]	0.031	0.438
Gemma / image_choice	Full-Naive	+18.97	[6.90, 31.67]	0.062	0.375
Gemma / image_choice	Retrieval-only-Naive	+0.00	[-7.02, 6.78]	1.000	1.000
Gemma / image_choice	Full-Retrieval-only	+18.97	[6.90, 30.51]	0.039	0.508
Gemma / text_choice	Full-Naive	+0.00	[-0.47, 0.45]	1.000	descriptive

Model / subset	Method contrast	Shift (pp)	95% CI	Raw p	Conservative p
Gemma / text_choice	Retrieval-only-Naive	+0.00	[-0.60, 0.65]	1.000	descriptive
Gemma / text_choice	Full-Retrieval-only	+0.00	[-0.65, 0.59]	1.000	descriptive
InternVL / all	Full-Naive	+3.30	[1.37, 5.51]	0.016	0.109
InternVL / all	Retrieval-only-Naive	+0.86	[-0.59, 2.51]	0.438	1.000
InternVL / all	Full-Retrieval-only	+2.44	[1.43, 3.50]	0.008	0.125
InternVL / image_choice	Full-Naive	+31.03	[17.24, 45.61]	0.008	0.062
InternVL / image_choice	Retrieval-only-Naive	+8.62	[1.72, 16.07]	0.125	1.000
InternVL / image_choice	Full-Retrieval-only	+22.41	[10.71, 35.00]	0.016	0.234
InternVL / text_choice	Full-Naive	+0.78	[-0.33, 1.97]	0.375	descriptive
InternVL / text_choice	Retrieval-only-Naive	+0.16	[-1.16, 1.63]	1.000	descriptive
InternVL / text_choice	Full-Retrieval-only	+0.63	[-0.50, 1.56]	0.438	descriptive

Human validation remains limited to the original 200 archived RTX 5090 Qwen outputs. The two independent annotation passes, eight third-person adjudications, and scientific linkage are unchanged. They do not validate the same-hardware Qwen rerun, Gemma, InternVL, or external-cohort outputs. Approved anonymous human evidence retains labels, verbatim support spans, and conflict fields; identities and free notes are excluded.

The portable numerical package recomputes statistics from anonymous derived-label records. It does not rerun the models, re-extract answers from raw text, or constitute an independent human reproduction. The author-side backup and the public numerical package have different scopes. The final anonymous ZIP is member-hash checked and extracted into a fresh CPU directory; its root verifier and the included Qwen four-budget tests record their checked quantities and environment versions in the verification receipt. The 200-row human linkage retains its original hash. All 1,440 external label rows join the public cohort projection by case, task, and gold. This metadata join does not repeat upstream candidate selection or image decoding. The root README documents the command and records the narrower replay scope.

Table S50 supplies the saved-length diagnostics introduced in Section L.4.

Table S50: Saved output-length diagnostics for every factorial cell (N=232). At-cap counts describe sequence length, not a verified internal stopping cause. Do not equate a parsed declaration with a correct answer.

Model	Budget	Arm	Median tokens	At cap	Unparsed at cap
Qwen	64	Anchor	64	229	155
Qwen	64	GN	64	230	164
Qwen	64	GS	64	231	79
Qwen	64	AN	64	231	52
Qwen	64	AS	64	230	14
Qwen	128	Anchor	128	155	130
Qwen	128	GN	128	168	131
Qwen	128	GS	128	131	72
Qwen	128	AN	120	102	48
Qwen	128	AS	109	76	13
Qwen	256	Anchor	165	56	50
Qwen	256	GN	165	52	36
Qwen	256	GS	140	48	27
Qwen	256	AN	120	41	22
Qwen	256	AS	109	28	4
Qwen	512	Anchor	165	28	25

Model	Budget	Arm	Median tokens	At cap	Unparsed at cap
Qwen	512	GN	165	19	13
Qwen	512	GS	140	23	10
Qwen	512	AN	120	28	10
Qwen	512	AS	109	17	2
Gemma	64	Anchor	56.00	80	0
Gemma	64	GN	61.00	106	0
Gemma	64	GS	59.00	96	0
Gemma	64	AN	64.00	129	0
Gemma	64	AS	64.00	122	0
Gemma	128	Anchor	56.00	6	0
Gemma	128	GN	61.00	7	0
Gemma	128	GS	59.00	5	0
Gemma	128	AN	66.00	5	0
Gemma	128	AS	66.00	4	0
Gemma	256	Anchor	56.00	0	0
Gemma	256	GN	61.00	0	0
Gemma	256	GS	59.00	1	0
Gemma	256	AN	66.00	0	0
Gemma	256	AS	66.00	0	0
Gemma	512	Anchor	56.00	0	0
Gemma	512	GN	61.00	0	0
Gemma	512	GS	59.00	0	0
Gemma	512	AN	66.00	0	0
Gemma	512	AS	66.00	0	0
InternVL	64	Anchor	47.00	29	0
InternVL	64	GN	41.00	9	2
InternVL	64	GS	40.00	5	0
InternVL	64	AN	47.00	20	0
InternVL	64	AS	43.00	15	0
InternVL	128	Anchor	47.00	0	0
InternVL	128	GN	41.00	0	0
InternVL	128	GS	40.00	0	0
InternVL	128	AN	47.00	0	0
InternVL	128	AS	43.00	0	0
InternVL	256	Anchor	47.00	0	0
InternVL	256	GN	41.00	0	0
InternVL	256	GS	40.00	0	0
InternVL	256	AN	47.00	0	0
InternVL	256	AS	43.00	0	0
InternVL	512	Anchor	47.00	0	0
InternVL	512	GN	41.00	0	0
InternVL	512	GS	40.00	0	0
InternVL	512	AN	47.00	0	0
InternVL	512	AS	43.00	0	0

M Legacy Protocol Coverage Completion

This section completes the missing InternVL counterbalanced binding and Qwen/Molmo fixed-retrieval cells under the legacy 128-token Official/RPS protocol. The new jobs contain 8,816 formal generations: 464 for InternVL binding and 8,352 for six Qwen/Molmo retrieval cells. Seven separate four-call technical pilots contribute 28 generations and are excluded from every result. All seven formal and seven pilot jobs have successful producer-exit receipts and passed

CPU closure audits. Archived Qwen/Gemma/Molmo binding and Gemma/InternVL retrieval outputs are reused, not rerun. Sections A–L remain historical records; Table S49 is the earlier two-model retrieval analysis, while S51–S52 and the main paper report the completed four-model comparison.

The historical references to Figure 4 in Section L and Table S49 correspond to the Gemma/InternVL subset of the current figure. Tables S51–S52 provide the complete four-model counts and contrasts.

M.1 Fixed Inputs And Execution Provenance

The completion uses the released question prompt, Official and RPS request builders, native model adapter, response cleaner, and parser. Greedy generation has a requested and verified effective maximum of 128 new tokens; this cap is not an observed response length. Both serializers retain each question, retrieved context, image content and order, checkpoint, and scoring rule. Retrieval uses all 696 questions in each of the three byte-pinned Naive, Retrieval-only, and Full manifests, including 58 image-choice and 638 text-choice questions. Input-size pilots and technical continuation gates never condition inclusion on correctness or parsing. Molmo is a native end-to-end serializer comparison here, not evidence for the excluded adjacency-preserving factorial.

Before the new Molmo retrieval jobs produced any output, their original Naive pilot failed with a TypeError in the generation-trace wrapper: Transformers 4.57.1 passed an additional positional argument to _prepare_generation_config. A separately frozen compatibility revision accepts and forwards positional arguments in that wrapper; these two insertions are the only runner-byte changes. InternVL and Qwen retain the original runner (SHA-256 prefix c7ac2f47ea98), while all six successful Molmo formal/pilot jobs use the compatibility runner (aec5b768202b) and distinct compat_v1 job/output identifiers. The validator bytes, conditions, 128-token budget, seed, native adapter, parser, and scorer remain unchanged. The original failed job and its chained journal are retained: one started attempt, one technical failure, zero generated events, and zero successful outputs. This failed attempt is separate from the 28 successful pilot outputs excluded from analysis; it contributes no research observation and causes no outcome-dependent reroll.

The completed formal metadata inventory contains 810 distinct source assets and 3 model checkpoints. Each asset has recorded source-file and decoded-RGB hashes; paired actual vision tensors were checked. The formal cells were run in BF16 on the GPU recorded below. Within-model checkpoint and decoding comparisons are matched, but archived and new model stacks differ; their cross-model differences do not isolate architecture. The historical Gemma revision remains unknown.

OpenGVLab/InternVL3_5-8B-HF: recorded revision 3974c115d44f0e13eae8b654f680b1e561c71ae9; checkpoint inventory SHA-256 98e17a5676e92b46e75933e172494b97ffe38e95951336026b252c5c24f931e6. Runtime: Python 3.12.3, PyTorch 2.12.1+cu130, Transformers 5.15.1, CUDA 13.0, NVIDIA GeForce RTX 4090 (reported device memory 47.37 GiB), torch.bfloat16.

Qwen/Qwen3.5-9B: recorded revision c202236235762e1c871ad0ccb60c8ee5ba337b9a; checkpoint inventory SHA-256 5a567c49d2a9aac8740b531700916b409e49b79081912b737cbd850576494b54. Runtime: Python 3.12.3, PyTorch 2.12.1+cu130, Transformers 5.15.1, CUDA 13.0, NVIDIA GeForce RTX 4090 (reported device memory 47.37 GiB), torch.bfloat16.

allenai/Molmo2-8B: recorded revision e28fa28597e5ec5e0cca2201dd8ab33d48bc4a1b; checkpoint inventory SHA-256 899edbd422483ff2c5d5eef6dd82a8c4a4d4a028d464269dd320039def264de9. Runtime: Python 3.12.3, PyTorch 2.12.1+cu130, Transformers 4.57.1, CUDA 13.0, NVIDIA GeForce RTX 4090 (reported device memory 47.37 GiB), torch.bfloat16.

M.2 Saved Evidence And Closure Checks

CPU closure verifies the complete selected-key matrix, immutable source and manifest hashes, paired saved image/tensor evidence, chained attempt journals, and exactly one generated and one finalized success record per planned call. Generated token IDs are freshly decoded with the

matching cached tokenizer, then checked against the stored raw text, frozen cleanup, and parser outcomes. Input token IDs are checked against their recorded tensor-byte hashes. Incorrect and unparsed outputs remain in fixed denominators. Resume does not regenerate a saved completed output or reroll generated-but-unfinalized evidence; a pre-generation failure needs a documented technical repair. The CPU audit loads no model weights and performs no GPU replication; it checks weight-inventory consistency but does not freshly hash weight shards or regenerate processor tensors.

M.3 Counterbalanced Binding Results

InternVL produces 65;200 correct;parsed outputs under Official and 124;219 under RPS, each out of 232. The RPS-minus-Official estimate is +25.43 pp, with a pointwise 95% persona-cluster interval [17.4, 33.2], raw sign-flip p=0.00391, and Holm4 p=0.01172. The 232 variants are four cyclic rotations of the same 58 base questions, not 232 independent questions. Choices, gold labels, and traced choice-image hashes are remapped together. Main Table 2 combines this result with the three byte-preserved historical model outputs.

Uncertainty uses 50,000 PCG64 whole-persona bootstrap draws with seed 20260909 and the question-weighted ratio-of-sums estimator. All four rotations remain within their persona cluster. Exact two-sided sign flips use all 1,024 assignments across ten personas and assume symmetric persona effects under the null; they are not randomized causal inference. The expanded binding family contains four tests. Coverage-completion families were fixed before the new outcomes, after the historical results were known; this is not prospective preregistration of the original experiment.

M.4 Complete Retrieval Counts

Table S51 reports correct;parsed counts for every model, subset, method, and serializer. All equals Image plus Text for both counts and denominators. C denotes correct answers and P denotes outputs accepted by the frozen parser; a parsed answer need not be correct. No additional human labels are inferred from these automatic counts.

Table S51: Complete four-model retrieval C;P counts at 128 output tokens. O/R denote Official/RPS and RO denotes Retrieval-only. Gemma/InternVL outputs are historical; Qwen/Molmo are new. Pale red/bold marks the maximum C across three methods per model, subset, and serializer, including ties, not significance.

Model	Subset / N	Naive O	Naive R	RO O	RO R	Full O	Full R
Qwen	All / 696	276;655	293;671	289;659	298;666	266;664	285;682
Qwen	Image / 58	15;22	26;37	15;27	23;35	10;27	32;45
Qwen	Text / 638	261;633	267;634	274;632	275;631	256;637	253;637
Gemma	All / 696	178;696	177;696	193;696	192;696	178;696	188;696
Gemma	Image / 58	6;58	5;58	5;58	4;58	5;58	15;58
Gemma	Text / 638	172;638	172;638	188;638	188;638	173;638	173;638
Molmo	All / 696	215;696	215;696	234;696	227;696	200;695	199;696
Molmo	Image / 58	31;58	29;58	25;58	16;58	16;57	15;58
Molmo	Text / 638	184;638	186;638	209;638	211;638	184;638	184;638
InternVL	All / 696	186;675	186;673	197;686	203;674	167;669	190;676
InternVL	Image / 58	26;53	32;55	22;56	33;56	10;46	34;55
InternVL	Text / 638	160;622	154;618	175;630	170;618	157;623	156;621

M.5 Serializer By Method Contrasts

Table S52 reports all 36 contrasts: four models, three overlapping subsets, and three method pairs. Each estimate subtracts the Official method difference from the RPS difference, as in main Equation (4). Full-Naive equals (RO-Naive)+(Full-RO) for each question, so these contrasts are algebraically dependent. The All/Image Full-Naive family has eight tests; the two other All/Image pairs form a separate 16-test family. The 12 Text contrasts and their Holm values are explicitly exploratory. Pointwise intervals are not simultaneous intervals or substitutes for sign-flip tests. A positive interaction does not establish a method win. Historical estimates, intervals, and raw p values remain unchanged; expanded corrections are newly computed from complete families.

Table S52: All 36 serializer-by-method interactions at 128 output tokens. Estimates and pointwise 95% intervals are in percentage points. Raw p is the exact persona-level two-sided sign-flip result. Holm p uses

Primary8, Secondary16, or exploratory Text12 as stated. No All/Image contrast meets Holm-adjusted p<.05; Text is descriptive.

Model	Subset	Contrast	Change pp	95% interval	Raw p	Holm p	Family
Qwen	All	Full-Naive	+0.29	[-0.8, 1.4]	0.8125	1	Primary8
Qwen	All	RO-Naive	-1.15	[-2.6, 0.3]	0.24219	1	Secondary16
Qwen	All	Full-RO	+1.44	[0.8, 2.2]	0.00781	0.125	Secondary16
Qwen	Image	Full-Naive	+18.97	[3.3, 36.4]	0.10156	0.40625	Primary8
Qwen	Image	RO-Naive	-5.17	[-15.8, 5.2]	0.5625	1	Secondary16
Qwen	Image	Full-RO	+24.14	[11.7, 39.7]	0.01562	0.21875	Secondary16
Qwen	Text	Full-Naive	-1.41	[-3.0, 0.2]	0.1875	1	Text12 expl.
Qwen	Text	RO-Naive	-0.78	[-1.9, 0.6]	0.40625	1	Text12 expl.
Qwen	Text	Full-RO	-0.63	[-1.5, 0.3]	0.35938	1	Text12 expl.
Gemma	All	Full-Naive	+1.58	[0.4, 2.7]	0.0625	0.375	Primary8
Gemma	All	RO-Naive	+0.00	[-0.7, 0.9]	1	1	Secondary16
Gemma	All	Full-RO	+1.58	[0.6, 2.6]	0.03125	0.375	Secondary16
Gemma	Image	Full-Naive	+18.97	[6.9, 31.7]	0.0625	0.375	Primary8
Gemma	Image	RO-Naive	+0.00	[-7.0, 6.8]	1	1	Secondary16
Gemma	Image	Full-RO	+18.97	[6.9, 30.5]	0.03906	0.42969	Secondary16
Gemma	Text	Full-Naive	+0.00	[-0.5, 0.5]	1	1	Text12 expl.
Gemma	Text	RO-Naive	+0.00	[-0.6, 0.6]	1	1	Text12 expl.
Gemma	Text	Full-RO	+0.00	[-0.7, 0.6]	1	1	Text12 expl.
Molmo	All	Full-Naive	-0.14	[-0.9, 0.6]	1	1	Primary8
Molmo	All	RO-Naive	-1.01	[-1.9, -0.1]	0.12109	0.96875	Secondary16
Molmo	All	Full-RO	+0.86	[-0.1, 1.8]	0.21875	1	Secondary16
Molmo	Image	Full-Naive	+1.72	[-5.5, 8.6]	1	1	Primary8
Molmo	Image	RO-Naive	-12.07	[-20.0, -3.4]	0.0625	0.625	Secondary16
Molmo	Image	Full-RO	+13.79	[3.4, 23.7]	0.07812	0.70312	Secondary16
Molmo	Text	Full-Naive	-0.31	[-0.8, 0.0]	0.5	1	Text12 expl.
Molmo	Text	RO-Naive	+0.00	[-0.7, 0.6]	1	1	Text12 expl.
Molmo	Text	Full-RO	-0.31	[-0.7, 0.0]	0.5	1	Text12 expl.
InternVL	All	Full-Naive	+3.30	[1.4, 5.5]	0.01562	0.10938	Primary8
InternVL	All	RO-Naive	+0.86	[-0.6, 2.5]	0.4375	1	Secondary16
InternVL	All	Full-RO	+2.44	[1.4, 3.5]	0.00781	0.125	Secondary16
InternVL	Image	Full-Naive	+31.03	[17.2, 45.6]	0.00781	0.0625	Primary8
InternVL	Image	RO-Naive	+8.62	[1.7, 16.1]	0.125	0.96875	Secondary16
InternVL	Image	Full-RO	+22.41	[10.7, 35.0]	0.01562	0.21875	Secondary16
InternVL	Text	Full-Naive	+0.78	[-0.3, 2.0]	0.375	1	Text12 expl.
InternVL	Text	RO-Naive	+0.16	[-1.2, 1.6]	1	1	Text12 expl.
InternVL	Text	Full-RO	+0.63	[-0.5, 1.6]	0.4375	1	Text12 expl.

M.6 Human Evidence And Numerical Replay Scope

The independent human-extraction audit remains exactly the original 200 archived Qwen outputs; its annotation records, adjudications, and linkage are unchanged. It does not validate the Section L same-hardware Qwen rerun, the new InternVL binding outputs, or the Qwen/Molmo retrieval responses. The anonymous numerical extension replays statistics from derived-label records and source hashes without releasing raw responses or private runtime paths. This software replay does not regenerate model answers, re-extract raw text, or constitute independent human reproduction. Author-side token/journal closure and public derived-label replay have different scopes.

The anonymous package separately includes both exact inference runners: the original version used by InternVL/Qwen and the positional-argument compatibility version used by Molmo. Their manifest identifies model-specific source selection and hashes of the same required historical dependencies and frozen input manifests. Those dependencies, full retrieval contexts, images, and weights are not bundled with either runner; source inspection is not a self-contained inference rerun. The author-side archive retains the executed source closure, original failure, and raw records. The compatibility revision does not expand third-party redistribution permissions; hash verification and anonymization do not establish those permissions.

REFERENCES

Viet Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal visual memory from explicit and implicit evidence. arXiv:2605.28806, 2026.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A comprehensive benchmark for robust multi-image understanding. In International Conference on Learning Representations, 2025a.

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