LAYER-WISE SEMANTIC EVIDENCE REWRITING UNDER UNIFIED MASK-TOKEN TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending vision-language models to predict masks can also change their answers to semantic questions. What changes in the evidence behind those answers? We study layer-wise semantic evidence rewriting between reference and unified checkpoints. Our framework decomposes a decision margin into input and residual-block contributions. The decomposition accounts for normalization and is exact for a fixed forward pass. Comparing representations under checkpoint-sourced readouts separates representation-side changes from readout-side changes. A calibrated linear transport then maps block evidence between checkpoints. Across five model-benchmark comparisons spanning two backbone families, representation-side drift exceeds readout-side drift. Full transport reduces reconstruction error by 30.6–69.0% relative to uniform rescaling on the original evaluation splits. A retrospective group-disjoint reanalysis finds positive sample-level associations in four settings. Full transport also yields lower observed mean absolute errors than scalar, diagonal, and probe-mean baselines. Its advantage over diagonal transport persists when both include layer-wise intercepts. In HiMTok, selected late-layer groups are sensitive to residual-write rescaling, but no more sensitive than the last three blocks. These findings reveal a layer-wise structure in semantic change that aggregate accuracy alone cannot show.

1 INTRODUCTION

Large multimodal models increasingly combine language generation with pixel-level prediction through segmentation tokens, decoder-driving hidden states, or autoregressive mask codes (Lai et al., 2024; Rasheed et al., 2024; Ren et al., 2024; Zhang et al., 2024; Wang et al., 2025b;a). These interfaces bring dense visual tasks into models pretrained for semantic decisions. Grounding-oriented adaptation can also change general visual-question-answering behavior. F-LMM documents this trade-off and explores frozen-backbone grounding (Wu et al., 2025). RealUnify asks more broadly whether unifying understanding and generation yields functional benefits (Shi et al., 2026).

Accuracy tells us how often a model is right, not how the evidence behind its answers changes (Figure 1). We ask whether checkpoint differences amount to uniform rescaling or require a layer-dependent description. These changes can involve residual representations, the final normalization and output head, or their interaction. Distinguishing these sources requires looking beyond aggregate performance.

We study *layer-wise semantic evidence rewriting*: changes in decision-relevant residual-block contributions between a reference checkpoint and a unified checkpoint. An exact, normalization-aware decomposition accounts for input and residual-block contributions on a fixed forward pass. Checkpoint-sourced contrasts separate representation, readout, and interaction changes. A calibrated linear transport then maps paired block evidence between checkpoints (Figure 2). Different layer-wise maps can yield the same summed prediction. We therefore evaluate reconstruction, sample-level prediction, and real-forward sensitivity separately.

The empirical scope is checkpoint comparison rather than a mask-supervision-only training intervention. We analyze HiMTok and ALTo relative to InternVL2.5-8B and include GLaMM relative to LLaVA-v1.5-7B on MME. The released systems differ in training history and output interfaces. The comparisons describe checkpoint differences rather than isolated effects of mask supervision.

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.

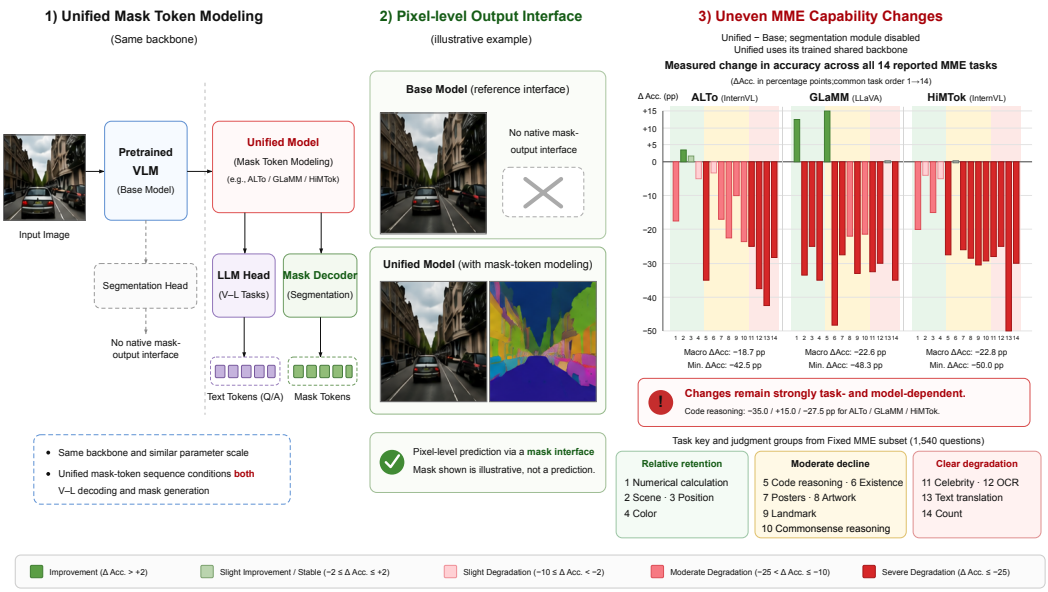


Figure 1: **Motivating checkpoint differences.** Pixel-level output interfaces accompany reported MME category-accuracy differences for ALTo, GLaMM, and HiMTok relative to their reference checkpoints. These motivating comparisons are distinct from the fixed-readout evidence measurements below. The bars show category-accuracy changes on a fixed 1,540-question MME subset covering 14 categories. Reported macro changes are unweighted averages across these categories.

Across all five comparisons, representation-side drift exceeds readout-side drift, and full transport reconstructs block evidence better than uniform rescaling.

A retrospective group-disjoint reanalysis in four HiMTok/ALTo settings retains positive sample-level associations. Full transport yields lower observed mean absolute error (MAE) than scalar, diagonal, and probe-mean references. Its advantage over diagonal transport persists with layer-wise intercepts, so mean shifts alone do not explain the result. Category-mean evidence ordering in the original audit is stronger for HiMTok than for ALTo. HiMTok interventions reveal sensitive late-layer groups, but no advantage over choosing the last three blocks.

Our contributions are threefold. **Evidence measurement:** exact, normalization-aware accounting with checkpoint-sourced representation/readout contrasts. **Rewriting characterization:** structural transport comparisons and group-disjoint sample-level audits, including an affine control for mean offsets. **Behavioral assessment:** real-forward tests that distinguish layer-group sensitivity from improved layer selection. Synthetic calibration, residual-spectrum analyses, and a three-seed training study are reported in the appendices.

2 Related Work

Pixel-grounded vision-language models. LISA, GLaMM, PixelLM, and OMG-LLaVA connect language-model representations to segmentation outputs (Lai et al., 2024; Rasheed et al., 2024; Ren et al., 2024; Zhang et al., 2024). Osprey instead emphasizes mask-conditioned region understanding (Yuan et al., 2024). HiMTok and ALTo learn discrete mask representations, while GETok develops tokens for spatial grounding (Wang et al., 2025b;a; Ren et al., 2026). Related work studies the segmentation-token interface, part-level grounding, and modular grounding architectures (Qian et al., 2025; Blume et al., 2025; Kang et al., 2026). Our analysis concerns the changes in internal semantic evidence across such checkpoints, rather than a new mask representation.

Representation diagnostics and adaptation. The tuned lens learns affine probes that expose intermediate vocabulary predictions (Belrose et al., 2023); centered kernel alignment (CKA) compares

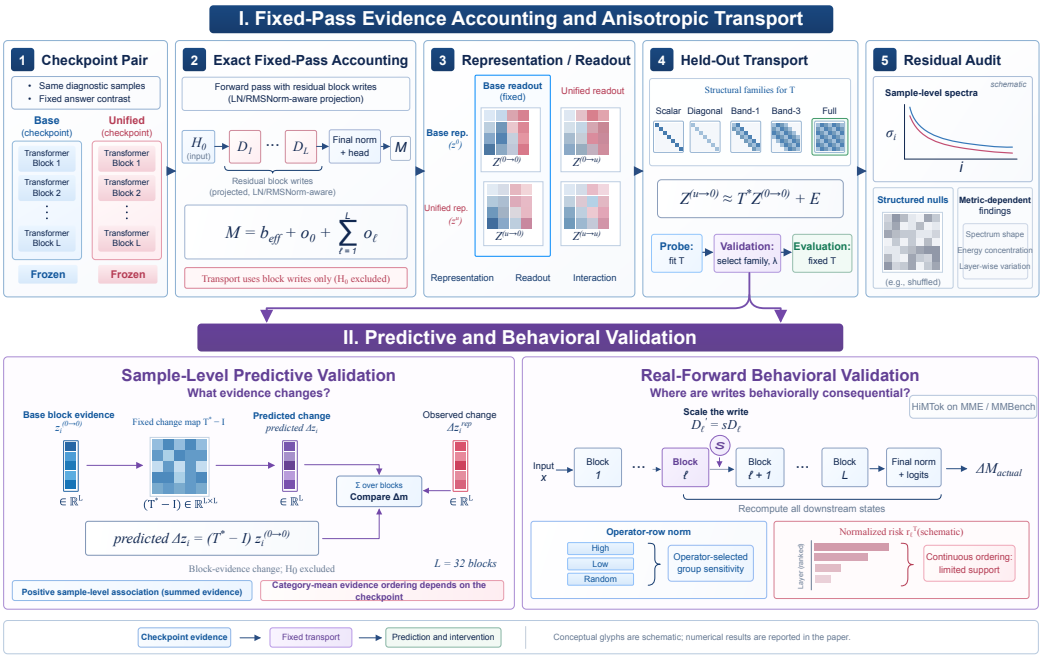


Figure 2: **Source-separated evidence diagnosis.** Fixed-pass accounting and representation/readout contrasts define the evidence matrices. A calibrated transport is evaluated by reconstruction and summed block-evidence prediction. Real-forward interventions assess operator-row-selected groups separately from normalized-risk ordering. The input contribution o_0 enters full-margin accounting but not the transport target.

representational similarity (Kornblith et al., 2019). Forgetting and multi-task methods instead regularize parameters or manipulate task gradients, including elastic weight consolidation (EWC), L2-SP, PCGrad, and CAGrad (Kirkpatrick et al., 2017; Li et al., 2018; Yu et al., 2020; Liu et al., 2021). Our object is a matrix of decision-evidence contributions indexed by block and sample. Its transport is a regularized linear regression, used to compare structural families under one sourced readout. The framework couples these checkpoint-sourced measurements to structural comparison and intervention-based assessment.

3 Evidence Accounting and Rewriting Geometry

3.1 Exact Fixed-Pass Decomposition

At the answer position, let the residual stream satisfy

$$H_\ell(x) = H_{\ell-1}(x) + D_\ell(x), \qquad \ell = 1, \dots, L, \quad D_0(x) = H_0(x). \tag{1}$$

For a correct decision token y^+ and a fixed distractor y^-, define

$$M(x; y^+, y^-) = \ell_{y^+}(x) - \ell_{y^-}(x), \qquad u = W_U[y^+] - W_U[y^-], \tag{2}$$

with optional output-bias difference b. Here $W_U[y]$ is a column-vector representation of the corresponding unembedding row.

For RMSNorm with gain g, let $r(h) = \sqrt{d^{-1}\,\|h\|_2^2 + \epsilon_{\text{rms}}}$ and $\tilde{u}_{\text{RMS}}(x) = (g \odot u)/r(H_L(x))$. For LayerNorm with gain g and bias c, let $P = I - d^{-1}\mathbf{1}\mathbf{1}^\top$, $s(h) = \sqrt{d^{-1}\,\|Ph\|_2^2 + \epsilon_{\text{LN}}}$, and $\tilde{u}_{\text{LN}}(x) = P(g \odot u)/s(H_L(x))$.

Proposition 1 (Exact fixed-pass decomposition)**.** *For a fixed realized forward trajectory,*

$$M(x) = b_{\text{eff}} + \sum_{\ell=0}^{L} o_\ell(x), \qquad o_\ell(x) = \tilde{u}(x)^\top D_\ell(x), \tag{3}$$

where $b_{\text{eff}} = b$ for RMSNorm and $b_{\text{eff}} = b + u^\top c$ for LayerNorm.

The identity follows by substituting the final normalization and unrolling the residual sum; the proof is in Section A.1. The effective readout depends on the final hidden state of the realized trajectory. Intervention effects are evaluated by recomputing that trajectory. The complete margin includes $o_0 = \tilde{u}^\top H_0$. Transport below uses only $\ell = 1, \ldots, L$, so its scalar target is a sum of block contributions rather than the complete margin.

3.2 Representation/readout separation

Let $a, b \in \{0, u\}$ denote the representation and readout checkpoints, respectively. With N samples and C benchmark categories, define

$$Z_{\ell i}^{a\to b} = (\tilde{u}_i^{a\to b})^\top D_\ell^a(x_i), \quad z_i^{a\to b} = Z_{:i}^{a\to b}, \quad G_{\ell\tau}^{a\to b} = \frac{1}{|I_\tau|}\sum_{i\in I_\tau} Z_{\ell i}^{a\to b}. \tag{4}$$

Thus $Z \in \mathbb{R}^{L\times N}$ and $G \in \mathbb{R}^{L\times C}$; $Z_{:i}$ denotes the entire i-th column. The same answer contrast is used for all four combinations. A sourced readout fixes its normalization parameters and decision-token head rows; normalization statistics are evaluated on the representation-source final state. For $X \in \{Z, G\}$, the three operational contrasts are

$$\Delta X_{\text{rep}} = X^{u\to 0} - X^{0\to 0}, \tag{5}$$

$$\Delta X_{\text{read}} = X^{0\to u} - X^{0\to 0}, \tag{6}$$

$$\Delta X_{\text{int}} = X^{u\to u} - X^{u\to 0} - X^{0\to u} + X^{0\to 0}. \tag{7}$$

Their matrix sum equals $X^{u\to u} - X^{0\to 0}$. We compare their Frobenius norms separately; these norms are not additive percentages of degradation. The transport analysis uses $Z^{0\to 0}$ and $Z^{u\to 0}$, both under the base-sourced readout.

3.3 Calibrated evidence transport

On the probe split, we fit

$$\widehat{T}_{\lambda,\mathcal{F}}^{(X)} = \arg\min_{T\in\mathcal{F}} \left\| X_p^{u\to 0} - T X_p^{0\to 0} \right\|_F^2 + \lambda \left\| T - I \right\|_F^2 . \tag{8}$$

The candidate families are scalar αI, diagonal, band-1, band-3, and full ridge. A band-k operator has $T_{ij} = 0$ for $|i - j| > k$; the diagonal family already permits independent per-layer rescaling. These five linear families omit an intercept. Each family tunes λ, and validation selects the family and regularization. As a control for layer-wise mean offsets, we additionally compare diagonal and full affine transport with probe-calibrated intercepts on the grouped splits (Section C.1 and Table 16). For the linear transport above, on evaluation records,

$$E_e^{(X)} = X_e^{u\to 0} - \widehat{T}_X^\star X_e^{0\to 0}, \qquad \mathrm{Err}_e^{(X)} = \frac{\left\| E_e^{(X)} \right\|_F}{\left\| X_e^{u\to 0} \right\|_F + \epsilon_{\text{err}}}. \tag{9}$$

The primary fit is sample-level ($X = Z$); category-refitted analyses are separate aggregation audits. We use "anisotropic transport" for the fitted block-dependent rescaling and cross-depth mixing. It describes checkpoint differences on the diagnostic distribution. Synthetic family recovery and residual spectra are detailed in Sections C.2 and C.3.

3.4 Sample-level evidence audit

With $\widehat{T}_Z^\star$ fixed after calibration, define

$$\widehat{\Delta z}_i = (\widehat{T}_Z^\star - I) z_i^{0\to 0}, \qquad \widehat{\Delta m}_i = \mathbf{1}^\top \widehat{\Delta z}_i, \tag{10}$$

and compare with $\Delta z_i^{\text{rep}} = z_i^{u\to 0} - z_i^{0\to 0}$ and $\Delta m_i^{\text{rep}} = \mathbf{1}^\top \Delta z_i^{\text{rep}}$. Lowercase m denotes summed block evidence; uppercase M denotes the complete decision margin. Category predictions average $\widehat{\Delta m}_i$ over I_τ.

Hypothesis 1 (Sample-level evidence association)**.** *Predicted and observed sample-level block-evidence changes are positively associated.*

A separate exploratory audit asks whether category means preserve the ordering of observed category-mean block-evidence changes. This aggregation endpoint is distinct from category accuracy.

Relation to scalar ridge prediction. Let $X_p = Z_p^{0\to 0}$ and $\Delta_p = Z_p^{u\to 0} - X_p$. For full transport at a fixed $\lambda > 0$, writing $A = T - I$ gives

$$\widehat{A}_\lambda = \Delta_p X_p^\top (X_p X_p^\top + \lambda I)^{-1}, \qquad \widehat{\Delta m}_i = \mathbf{1}^\top \widehat{A}_\lambda z_i^{0\to 0}. \tag{11}$$

The summed output equals direct, no-intercept ridge prediction with the same inputs and regularization. Moreover, $\Delta m_i^{\mathrm{rep}} - \widehat{\Delta m}_i = \mathbf{1}^\top E_{:i}$: reconstruction and scalar prediction summarize the same residual in different ways. The transport framework retains the full block-output matrix for source-separated structural comparison and behavioral assessment. Its scalar audit measures predictive association rather than introducing a separate estimator. The derivation and the rank-deficient, zero-regularization exception are given in Section A.4.

The layer-resolved map nevertheless contains information lost by summation. For any $K \in \mathbb{R}^{L\times L}$ with $\mathbf{1}^\top K = 0$, the operators T and $T + K$ give identical summed predictions on every input, although their blockwise predictions can differ. Scalar agreement alone therefore cannot determine the internal redistribution. We retain matrix reconstruction and real-forward interventions as separate assessments of the layer-resolved description.

4 Real-Forward Behavioral Validation

Intervention protocol. For a selected block ℓ and scale s, we replace its complete answer-position residual write by

$$D'_\ell(x) = sD_\ell(x), \qquad H'_\ell(x) = H_{\ell-1}(x) + D'_\ell(x), \tag{12}$$

and recompute all downstream blocks, final normalization, and logits. The effect is $\Delta M_\ell^{\mathrm{actual}} = M' - M$; $(s-1)o_\ell$ is a frozen-trajectory comparator. Local scales are $0.90, 0.95, 1.05, 1.10$. Magnitude sensitivity averages $|\Delta M_\ell^{\mathrm{actual}}|/(|s-1| + 10^{-12})$ over the four scales and evaluation records.

Layer selection. High- and low-score groups were selected using $s_\ell^{\mathrm{hist}} = \|(T_{\mathrm{hist}} - I)_{\ell:}\|_2$, where T_{hist} was fitted with the historical zero-centered penalty $\lambda \|T\|_F^2$. The final identity-centered fits yield the same high/low sets (Section D.1). Depth-matched random groups serve as controls. The selected groups are fixed before their evaluation intervention outcomes are used. A separate data-conditioned score r_ℓ^T and its continuous-ordering audit are defined in Section D.2; the group contrasts below assess sensitivity of the operator-selected groups.

5 Experiments

Checkpoint and evaluation scope. We compare InternVL2.5-8B (Chen et al., 2024) with HiMTok and ALTo on MME and MMBench (Fu et al., 2023; Liu et al., 2023), and GLaMM-FullScope with LLaVA-v1.5-7B (Liu et al., 2024) on MME. Each diagnostic manifest contains 1,000 records. The original transport reconstruction and predictive audits use the saved $600/200/200$ probe/validation/evaluation partitions. Intervention and continuous layer-risk analyses retain the original 200-record evaluation sets. The HiMTok/ALTo group-disjoint reanalysis uses separate $600/201/199$ partitions, shared between checkpoints within each benchmark. Its operators are fitted on grouped probe data and selected on grouped validation data; they do not replace the original operators (Sections A.6 and 5.3). Four-quadrant norms use the full-manifest G matrices. All measured architectures have $L = 32$ blocks. Table 1 separates cross-backbone reconstruction coverage from the four-setting predictive and HiMTok-only intervention analyses. The GLaMM/LLaVA analysis uses the same semantic queries and answer contrasts with checkpoint-specific native wrappers; exact initialization ancestry is unverified. These are compatible reference/unified comparisons, not retraining runs with identical data and budgets (Section B.1).

Table 1: **Comparison coverage.** Each row is a published-checkpoint comparison, not a matched-data retraining experiment. Accounting includes the sourced-readout contrasts; prediction refers to the scalar sum of block-evidence changes. A dash indicates an analysis not reported for that pair.

Unified	Reference	Accounting/transport	Prediction	Intervention
HiMTok	InternVL2.5-8B	MME, MMBench	MME, MMBench	MME, MMBench
ALTo	InternVL2.5-8B	MME, MMBench	MME, MMBench	—
GLaMM	LLaVA-v1.5-7B	MME	—	—

Statistical protocol. Predictive intervals use 2,000 paired-cluster bootstraps with the selected operator fixed. Intervention comparisons use 2,000 cluster bootstraps and 10,000 cluster sign flips; residual nulls refit and reselect operators within each replicate. Both audits are retrospective. In the original splits, 112 of 186 MME evaluation clusters and four of 200 MMBench clusters link to probe/validation records. The grouped reanalysis removes the identified cross-split links; its intervals still condition on the fitted and selected predictor. The intervention analyses retain the original evaluation membership (Section D.1).

5.1 Exact accounting and source separation

On fixed native-precision trajectories, FP32 direct-margin evaluation and write reconstruction have maximum absolute error below 10^{-5} for every reported checkpoint (Table 2). Representation-side drift exceeds readout-side drift in all five comparisons: the ratios are 2.675/2.747 for ALTo on MME/MMBench and 1.323 for GLaMM/MME. GLaMM's interaction and readout norms are 3.6592 and 4.4635, respectively. These are within-setting geometric magnitudes, not additive shares of accuracy loss. Native-BF16 logits and BF16 summation are separate precision checks (Section B.2); the identity audit concerns FP32 accounting, not exact additivity of native-precision execution.

5.2 Transport structure and calibration

Validation selects full ridge in all five settings, with lower evaluation error than scalar transport (Table 3). In the four grouped settings, full affine transport also achieves lower reconstruction error and summed block-evidence-change MAE than diagonal affine transport; both use unpenalized layer-wise intercepts estimated only from probe data (Table 16). These comparisons favor full transport within the tested families but do not uniquely identify a generating operator.

Synthetic calibration recovers the generating family in 93.46% of 40,000 main-grid trials. False-full selection is 1.59% among non-full generators and rises to 12.5–16.3% in the rank-1, zero-observation-noise strata. These experiments calibrate the selection procedure, including an unfavorable regime (Section C.2).

Residual spectral tests provide a secondary, model-dependent finding: support differs between leading-energy concentration and entropy ranks, and ALTo/MMBench has no significant primary variance-null metric after within-setting Holm correction. The full five-setting table and exploratory statistics are in Section C.3.

For HiMTok relative to Base, the diagnostic manifests underestimate full-set accuracy decline by 0.96 percentage points (pp) on MME and 1.40 pp on MMBench. Category changes have Spearman correlations of 0.821/0.829 and mean absolute errors of 2.49/6.69 pp. The largest MMBench category discrepancy is 21.66 pp. The audit concerns behavioral trend coverage; full-set and 1k records are nested (Figure 3 and Section B.4).

5.3 Fixed-operator predictive association

Group-disjoint retrospective audit. We reanalyze the existing caches with one frozen partition per benchmark, grouping records by their recorded image/question identifiers and sharing each partition between HiMTok and ALTo. The splits contain 600/201/199 probe/validation/evaluation records. After probe fitting and validation selection over the five families, all four settings select full transport. Both Pearson and Spearman correlations have 95% bootstrap intervals above zero (Table 4).

Table 2: Exact accounting and source-separated drift. Errors are absolute FP32 reconstruction errors on fixed native-precision caches. Rep./Read./Inter. are unnormalized Frobenius norms on category-mean G. Each drift triplet refers to a unified/reference pair; Base is InternVL2.5-8B and LLaVA is the GLaMM reference.

Task	Ckpt.	p95 err.	max err.	Rep.	Read.	Inter.
MME	Base	9.54×10^{-7}	1.91×10^{-6}	7.8437	2.4732	2.4096
MME	HiMTok	5.36×10^{-7}	1.43×10^{-6}			
MMBench	Base	3.81×10^{-6}	7.63×10^{-6}	6.9093	2.6224	2.1816
MMBench	HiMTok	1.43×10^{-6}	3.34×10^{-6}			
MME	ALTo	8.35×10^{-7}	1.91×10^{-6}	7.6623	2.8639	2.8059
MMBench	ALTo	3.34×10^{-6}	5.72×10^{-6}	6.8766	2.5034	2.1212
MME	LLaVA	1.19×10^{-6}	1.91×10^{-6}	5.9048	4.4635	3.6592
MME	GLaMM	1.43×10^{-6}	2.86×10^{-6}			

Table 3: Sample-level transport reconstruction errors on the original $600/200/200$ partitions. Regularization is selected separately for each family using validation data; all errors are evaluated on the 200-record evaluation splits. Full ridge is validation-selected in all five settings. Lower error is better. Selected regularization parameters are reported in Table 15.

Setting	Scalar	Diagonal	Band-1	Band-3	Full ridge
HiMTok/MME	0.9021	0.6501	0.4908	0.4565	0.3637
HiMTok/MMBench	0.9810	0.9209	0.8409	0.8144	0.6632
ALTo/MME	0.9316	0.7747	0.5755	0.5607	0.4603
ALTo/MMBench	0.9442	0.9218	0.8757	0.8206	0.6555
GLaMM/MME	0.7549	0.4600	0.3238	0.2861	0.2342

No probe-plus-validation refit is performed. The grouped operators are distinct from those used in the original-split analyses; selected parameters are reported in Section A.6.

Same-target comparisons. Scalar, diagonal, and full transport are evaluated against the same $\Delta m_i^{\mathrm{rep}}$ targets. The probe-mean/Base-only reference is

$$\widehat{\Delta m_i}^{\mathrm{mean}} = \bar{m}_p^{u\to 0} - \mathbf{1}^\top z_i^{0\to 0}, \qquad \bar{m}_p^{u\to 0} = \frac{1}{N_p}\sum_{j\in p} \mathbf{1}^\top z_j^{u\to 0}. \tag{13}$$

It predicts a constant Unified block-evidence sum estimated from probe records, while retaining the sample-specific Base subtraction.

Comparisons use 2,000 shared paired-group bootstrap draws with predictors fixed. Full transport has lower point-estimate MAE than scalar, diagonal, and probe-mean/Base-only prediction in all 12 comparisons; 11 unadjusted 95% paired intervals exclude zero (Table 5). The exception is ALTo/MMBench versus scalar (-0.260, $[-0.531, 0.003]$). Full also has lower sign agreement there (142/199 versus 149/199), where 70.85% of targets are negative. These results support comparison-specific MAE improvements, not superiority on every endpoint. Section A.6 gives the protocol and target sign proportions; changed evaluation membership prevents attributing old–new differences solely to removing cross-split links.

Original-split association and exploratory aggregation. The original 200-record fixed-operator audits remain in Section A.5 and Table 7. Their sample-level Spearman correlations range from 0.407 to 0.815, with the reported conditional intervals above zero. These are original-split estimates, not the grouped estimates above. Category-mean evidence ordering in that original audit is less consistent: HiMTok reaches Spearman correlations of 0.6703 and 0.8271, whereas ALTo gives 0.1165 and 0.0812. ALTo predicts the wrong category-mean direction for code reasoning, color, and existence on MME, and for spatial relationship and image topic on MMBench. Thus, the sample-level association in Hypothesis 1 does not imply reliable category ordering across checkpoints. Category accuracy decline was not a prediction endpoint.

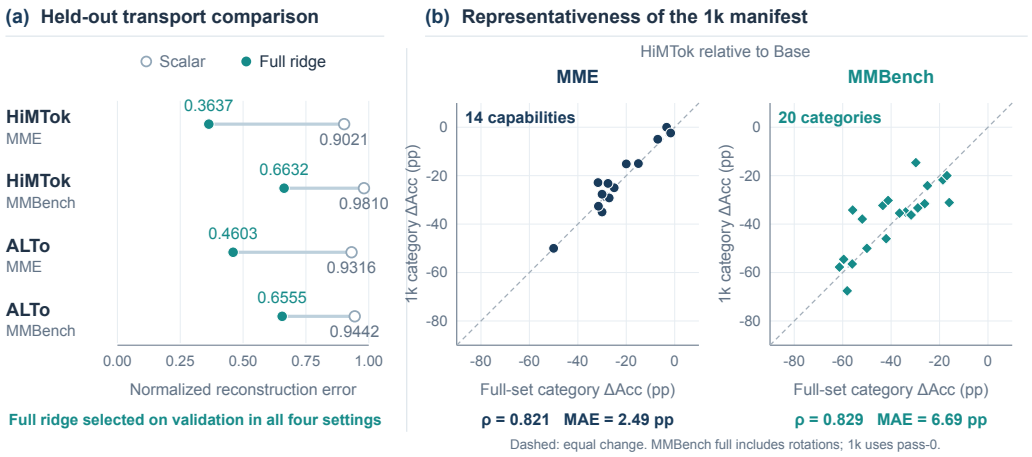


Figure 3: **Transport comparison and diagnostic-manifest coverage.** (a) Scalar/full-ridge reconstruction errors for the four InternVL-based settings. The additional GLaMM result is reported in Table 3. (b) Category-accuracy changes for the full and 1k HiMTok/Base evaluations. The sets are nested; MMBench full includes option rotations while the 1k manifest uses pass-0. This panel audits behavior, not the full-set distribution of internal evidence.

Table 4: Full-transport predictive association on the group-disjoint retrospective evaluation splits. Each setting contains 199 records, comprising 127 groups on MME and 195 on MMBench. Bracketed values are 95% percentile intervals from 2,000 paired-group bootstrap resamples with the fitted operator held fixed. MAE is measured in block-evidence units. Sign reports the number and percentage of predictions with the same sign as the observed change.

Setting	Pearson	Spearman	MAE	Sign
HiMTok/MME	0.7021 [0.6330, 0.7653]	0.6928 [0.6045, 0.7690]	1.2315	175/199 (87.94%)
HiMTok/MMBench	0.7967 [0.7340, 0.8494]	0.7664 [0.6834, 0.8342]	1.3772	187/199 (93.97%)
ALTo/MME	0.5866 [0.4926, 0.6714]	0.5611 [0.4431, 0.6591]	2.0200	136/199 (68.34%)
ALTo/MMBench	0.3967 [0.2755, 0.5048]	0.3900 [0.2653, 0.5068]	2.8089	142/199 (71.36%)

5.4 Real-forward behavioral localization

Operator-row-selected high-score layers in HiMTok have mean sensitivity 1.247 versus $0.342/0.417$ for low/random controls on MME, and 0.594 versus $0.203/0.222$ on MMBench. The four high-minus-control differences remain positive after Holm correction, with paired standardized effects $d_z = 1.21$–1.74 (Table 6). The groups and retrospective evaluation protocol are documented in Section D.1.

Sensitivity is absolute complete-margin change per unit write-rescaling magnitude, not accuracy percentage points. Random controls match high-score layers only within four blocks; low-score layers are not depth-matched. Additional controls show that the high groups are more sensitive than probe Base evidence-RMS top-three groups on both benchmarks, but less sensitive than the last three blocks on MME. On MMBench, the high and last-three groups are identical. Thus, the observed localization does not establish an advantage over last-three-block selection (Table 20). The reported d_z values standardize paired sample differences; intervals and sign-flip tests use question/image clusters.

The separate normalized-risk score gives inconsistent continuous ordering: Spearman correlation is -0.100 on MME (11 layers; interval $[-0.291, 0.091]$) and 0.333 on MMBench (nine layers;

Table 5: Full-minus-reference MAE differences on the grouped evaluation splits, with 95% paired-group bootstrap intervals. Negative values favor full transport. Units are original block evidence, not percentage points. Intervals are not multiplicity-adjusted. Probe mean predicts the probe-estimated unified mean minus each sample's Base evidence (Equation (13)).

Setting	Scalar	Diagonal	Probe mean
HiMTok/MME	-0.446 $[-0.611, -0.292]$	-0.547 $[-0.739, -0.360]$	-0.765 $[-0.949, -0.573]$
HiMTok/MMBench	-0.134 $[-0.260, -0.011]$	-0.176 $[-0.305, -0.047]$	-0.326 $[-0.488, -0.174]$
ALTo/MME	-0.471 $[-0.738, -0.201]$	-0.435 $[-0.705, -0.143]$	-0.583 $[-0.832, -0.335]$
ALTo/MMBench	-0.260 $[-0.531, +0.003]$	-0.324 $[-0.601, -0.064]$	-0.916 $[-1.305, -0.553]$

Table 6: HiMTok localization under real-forward write rescaling. High and low groups are selected using the operator-row score, with depth-matched random controls. Differences are high minus control in magnitude-sensitivity units, not accuracy percentage points. Intervals use 2,000 paired-cluster bootstrap resamples; p values use 10,000 cluster sign flips with Holm correction across the four comparisons.

Task	Comparison	Difference	95% CI	d_z	Holm p
MME	High–low	+0.9052	[0.8303, 0.9768]	1.7406	0.0004
MME	High–random	+0.8307	[0.7559, 0.9081]	1.5003	0.0004
MMBench	High–low	+0.3908	[0.3467, 0.4387]	1.2063	0.0004
MMBench	High–random	+0.3716	[0.3294, 0.4142]	1.2164	0.0004

$[0.167, 0.550]$). These intervals condition on the tested layers and fixed scores. Within the tested layers, the pre-specified operator-row group contrasts are better supported than the continuous normalized-risk ordering (Section D.2). Base–Unified difference-in-differences likewise do not show a consistent training-induced increase in sensitivity.

6 Discussion and Limitations

Full transport captures layer-dependent evidence changes that uniform rescaling misses. Its advantage over diagonal transport persists with layer-wise intercepts, so mean shifts alone do not explain the result. This is a checkpoint-calibrated diagnostic, not a forecast of adaptation or a new scalar estimator.

The grouped audit remains retrospective, with intervals conditional on fitted operators. Different training histories and interfaces prevent attribution to mask supervision alone. Category ordering is checkpoint-dependent. HiMTok interventions reveal local write sensitivity but no advantage over selecting the last three blocks; normalized-risk ordering is inconsistent. Online Evidence-Constrained Mask Training (OECMT) likewise shows no consistent retention advantage across three seeds (Appendix E).

7 Conclusion

We introduce a layer-wise diagnostic of semantic evidence rewriting based on exact fixed-pass accounting and checkpoint-calibrated transport. Across five comparisons, representation-side drift exceeds readout-side drift, and full transport reconstructs evidence better than uniform rescaling. Group-disjoint audits retain sample-level association in four settings. The framework separates structural reconstruction, evidence prediction, and behavioral sensitivity in the analysis of checkpoint differences.

Reproducibility Statement

The appendices specify the evidence construction, precision checks, transport fitting and selection, conditional bootstrap procedures, residual nulls, intervention groups, and training controls. The original diagnostic splits are retrospective and contain cross-split group links, as documented in Section A.5. Checkpoint provenance and run-specific implementation details are documented in Sections B.1 and E. The completed cache-only group-disjoint retrospective reanalysis is reported separately from the original predictive audit in Section A.6.

AI Use Statement

Generative AI tools assisted with manuscript organization, language editing, experimental-protocol suggestions, and LaTeX preparation. They also assisted in formulating mathematical claims and drafting and checking derivations and proofs. The authors are responsible for verifying the mathematical arguments, experimental results, references, and the final manuscript.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Ansel Blume, Jeonghwan Kim, Hyeonjeong Ha, Elen Chatikyan, Xiaomeng Jin, Khanh Duy Nguyen, Nanyun Peng, Kai-Wei Chang, Derek Hoiem, and Heng Ji. PARTONOMY: Large multimodal models with part-level visual understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38986–39015, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. VGent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41160–41170, 2026.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2825–2834, 2018.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how <SEG> token works. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24722–24731, 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41171–41181, 2026.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. RealUnify: Do unified models truly benefit from unification? a comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22488–22497, 2026.

Lingfeng Wang, Hualing Lin, Senda Chen, Tao Wang, Changxu Cheng, Yangyang Zhong, Dong Zheng, and Wuyue Zhao. ALTo: Adaptive-length tokenizer for autoregressive mask generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 18090–18114, 2025a.

Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. HiMTok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23267–23278, 2025b.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24710–24721, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28202–28211, 2024.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71737–71767, 2024.

A Formal Derivations and Predictive Protocol

A.1 Proof of Proposition 1

For RMSNorm, substitution gives

$$u^\top \mathrm{RMSNorm}(H_L) + b = \frac{(g \odot u)^\top H_L}{r(H_L)} + b = \tilde{u}_{\mathrm{RMS}}^\top \sum_{\ell=0}^{L} D_\ell + b.$$

For LayerNorm, using $P^\top = P$,

$$u^\top \mathrm{LN}(H_L) + b = \frac{(g \odot u)^\top P H_L}{s(H_L)} + u^\top c + b = \tilde{u}_{\mathrm{LN}}^\top \sum_{\ell=0}^{L} D_\ell + u^\top c + b.$$

The normalization statistics are those of the realized final state. The identities hold in exact arithmetic; Section B.2 distinguishes the corresponding FP32 accounting check from native BF16 kernel outputs.

For a fixed answer contrast and base-sourced readout, let $m_i^{a\to 0} = \mathbf{1}^\top z_i^{a\to 0}$. Then

$$M_i^{u\to 0} - M_i^{0\to 0} = \Delta m_i^{\mathrm{rep}} + o_{0,i}^{u\to 0} - o_{0,i}^{0\to 0}. \tag{14}$$

The readout bias cancels, but the input contribution need not: even a shared H_0 can be projected by different effective readouts because the final states differ. This is why the reported predictive target is the block-evidence sum.

A.2 Aggregation of a structured sample residual

Suppose the sample-level residual admits the factorization

$$Z_{:i}^{u\to 0} = T Z_{:i}^{0\to 0} + B\xi_i + e_i, \qquad B \in \mathbb{R}^{L\times r}, \quad \xi_i \in \mathbb{R}^r.$$

Using the same finite category sets I_τ as in Equation (4), define $\bar{\xi}_\tau = |I_\tau|^{-1} \sum_{i\in I_\tau} \xi_i$ and $\bar{e}_\tau = |I_\tau|^{-1} \sum_{i\in I_\tau} e_i$. Averaging the sample relation then gives

$$G^{u\to 0} = T G^{0\to 0} + B Q^\top + \bar{E}, \qquad \mathrm{rank}(BQ^\top) \le r,$$

where $Q_{\tau:} = \bar{\xi}_\tau^\top$ and $\bar{E}_{:\tau} = \bar{e}_\tau$. The equality follows by finite-sample averaging of the specified factorization. Whether empirical residuals exhibit concentration is tested separately in Section C.3.

A.3 Local relation to parameter updates

Let $\mathcal{Z}(\theta)$ vectorize block evidence for fixed inputs, answer contrasts, and base-sourced readout parameters. Assume this map has a locally Lipschitz Jacobian on the segment between θ^0 and $\theta^u = \theta^0 + \Delta\theta$, with Lipschitz constant K. Then

$$\mathcal{Z}(\theta^u) - \mathcal{Z}(\theta^0) = J_{\mathcal{Z}}(\theta^0)\Delta\theta + R, \qquad \|R\|_2 \le \tfrac{K}{2} \|\Delta\theta\|_2^2. \tag{15}$$

The bound follows by integrating the Jacobian along the segment and applying its Lipschitz bound. It gives a local functional interpretation of parameter change. Multiple parameter updates may produce the same observed evidence change; the fitted transport does not identify a unique parameter update. No small-update approximation is used to compute the reported checkpoint measurements.

A.4 Full transport and direct ridge prediction

Let $X = Z_p^{0\to 0} \in \mathbb{R}^{L\times N_p}$, $Y = Z_p^{u\to 0}$, and $\Delta = Y - X$. With $A = T - I$, the full-family objective in Equation (8) becomes

$$\widehat{A} = \arg\min_A \|\Delta - AX\|_F^2 + \lambda \|A\|_F^2 = \Delta X^\top (XX^\top + \lambda I)^{-1}, \qquad \lambda > 0. \tag{16}$$

Let $d = \Delta^\top \mathbf{1} \in \mathbb{R}^{N_p}$. The summed prediction obeys

$$\mathbf{1}^\top \widehat{A} x_i = \widehat{w}^\top x_i, \qquad \widehat{w} = (XX^\top + \lambda I)^{-1} X d. \tag{17}$$

Thus it equals no-intercept scalar ridge prediction with the same inputs, loss normalization, and regularization. At $\lambda = 0$, the same identity holds when X has full row rank. For a rank-deficient design, choosing minimum-norm T is not equivalent to choosing minimum-norm $A = T - I$; off-span predictions then depend on the tie-breaking convention. Selecting λ by a different scalar validation objective can produce a different selected predictor. The equality makes explicit that the prediction audit evaluates the functional output of transport rather than a separate estimator.

For an evaluation residual $e_i = z_i^{u\to 0} - \widehat{T}_Z^{\star} z_i^{0\to 0}$,

$$\Delta m_i^{\mathrm{rep}} - \widehat{\Delta m}_i = \mathbf{1}^\top e_i. \tag{18}$$

Reconstruction error and scalar prediction error therefore probe different summaries of the same residual. Positive correlation alone does not establish improvement over a simple reference that also subtracts the sample's base evidence.

A.5 FIXED-OPERATOR PREDICTIVE AUDIT

The audits cover HiMTok–InternVL and ALTo–InternVL on MME and MMBench. Each setting uses 200 evaluation records, comprising 186 clusters on MME and 200 on MMBench. Each model/benchmark pair uses its own previously selected operator; no family, parameter, threshold, or sign convention is fitted to these evaluation endpoints.

Pearson and Spearman correlations use 2,000 paired-cluster bootstrap resamples. Predicted and observed values are resampled together, while the selected operator remains fixed. The reported percentile intervals quantify evaluation-sample uncertainty conditional on that operator. Sign agreement and MAE are reported descriptively. Category results average predictions and observations within 14 MME capabilities or 20 MMBench categories and compute their Spearman correlation. Category-level $\Delta\mathrm{Acc}$ prediction is not evaluated.

These analyses reuse previously inspected evaluation records. Of 186 MME evaluation clusters, 112 link to records in probe or validation; the corresponding count is four of 200 for MMBench. Resampling evaluation clusters does not remove the cross-partition dependence, and the intervals do not incorporate fitting or selection uncertainty. The original record-level results are preserved as such.

Table 7: Fixed-operator predictive association on the original evaluation splits. Bracketed values are 95% paired-cluster bootstrap intervals for Pearson and Spearman correlations, based on 2,000 resamples with the selected transport held fixed. Sign denotes sample-level sign agreement, and MAE uses the original block-evidence units. Category-mean $\rho_{\Delta m}$ compares category-averaged predicted and observed block-evidence changes.

Setting	Pearson	Spearman	Sign	MAE	Category $\rho_{\Delta m}$
HiMTok/MME	0.7216 [0.6482, 0.7820]	0.6873 [0.6048, 0.7539]	85.0%	1.3261	0.6703
HiMTok/MMBench	0.8451 [0.7971, 0.8830]	0.8146 [0.7477, 0.8661]	94.0%	1.3451	0.8271
ALTo/MME	0.5259 [0.4010, 0.6344]	0.5001 [0.3686, 0.6181]	69.5%	2.1871	0.1165
ALTo/MMBench	0.4240 [0.3037, 0.5262]	0.4067 [0.2783, 0.5165]	76.5%	2.7402	0.0812

A.6 CACHE-ONLY GROUP-DISJOINT ROBUSTNESS

Grouped protocol. We retrospectively reanalyzed the existing 32-block evidence caches, excluding H_0. Groups were formed by transitive linkage of available exact image and question identities. For MME, this joint linkage yields 661 groups over the same 1,000 records, compared with 671 groups under the original question-ID grouping. One partition per benchmark was frozen before fitting, with no identified groups shared across splits. All 1,000 records were retained, yielding 600/201/199 probe/validation/evaluation records. The corresponding group counts were 401/133/127 for MME and 583/187/195 for MMBench. HiMTok and ALTo shared each benchmark's partition. We refitted

the five existing transport families on probe and selected regularization and family on validation; all four settings selected full transport.

The grouped operators are distinct from those used in the original-split analyses. For HiMTok/MME, HiMTok/MMBench, ALTo/MME, and ALTo/MMBench, respectively, the selected full-transport (μ, λ) pairs are $(0, 0)$, $(1.7783 \times 10^{-6}, 2.4239 \times 10^{-4})$, $(5.6234 \times 10^{-6}, 2.5274 \times 10^{-3})$, and $(3.1623 \times 10^{-6}, 4.3104 \times 10^{-4})$, where $\lambda = \mu \|Z_p^{0\to 0}\|_F^2/32$. Each scalar and diagonal comparator selects $\mu = \lambda = 0$. Regularization is selected separately for each family on grouped validation data, without subsequent probe-plus-validation refitting. The probe-mean/Base-only reference has no transport operator or regularization parameter.

Same-target comparisons. Scalar, diagonal, full transport, and the probe-mean/Base-only reference in Equation (13) are evaluated against the same $\Delta m_i^{\mathrm{rep}}$ targets. It predicts a constant unified block-evidence sum estimated only from probe, retaining the sample-varying Base subtraction. We used 2,000 shared paired-group bootstrap draws across predictors and both models within each benchmark, holding fitted operators and probe means fixed. Full-transport Pearson and Spearman intervals remain above zero in all four settings. Full has lower point-estimate MAE in all 12 comparisons; 11 descriptive paired intervals lie entirely below zero (Table 5). The ALTo/MMBench comparison against scalar remains uncertain.

Table 8 reports absolute MAE for all four predictors; paired differences and their intervals are reported in Table 5.

Table 8: Absolute MAE on the group-disjoint evaluation splits. All predictors use the same 199 evaluation records within each setting and the same representation-side block-evidence change target. Values are in block-evidence units. Probe mean denotes the reference in Equation (13).

Setting	Scalar	Diagonal	Full	Probe mean
HiMTok/MME	1.6778	1.7781	1.2315	1.9961
HiMTok/MMBench	1.5109	1.5533	1.3772	1.7033
ALTo/MME	2.4914	2.4549	2.0200	2.6030
ALTo/MMBench	3.0691	3.1333	2.8089	3.7249

Negative targets account for 78.89%, 90.45%, 54.77%, and 70.85% of evaluation records in table order; no targets are exactly zero. On ALTo/MMBench, full sign agreement is 142/199, below scalar's 149/199 and close to the negative-target fraction. Thus, lower MAE does not imply superiority on every metric.

Checks and scope. For grouped HiMTok/MME, the validation-selected full transport has $\mu = \lambda = 0$. The executed float64 implementation computes $T = Y_p X_p^\top (X_p X_p^\top)^+$, directly solving for T rather than the minimum-norm displacement $A = T - I$. The Gram pseudoinverse uses a relative cutoff of 10^{-12}, corresponding to a design-matrix relative cutoff of 10^{-6}. The absolute cutoffs are 1.10947×10^{-8} and 1.05331×10^{-4}, respectively; the effective design rank is $32/32$.

The specified solver exactly reproduces the saved operator and all 199 evaluation predictions. Direct no-intercept scalar least squares for $\mathbf{1}^\top (Y_p - X_p)$, using the same Gram cutoff, gives evaluation predictions differing from the transport predictions by at most 4.19902×10^{-10}. This exceeds the prescribed $\texttt{atol} = \texttt{rtol} = 10^{-12}$ criterion: the full-rank solutions are equivalent in exact arithmetic, but not numerically identical at this tolerance. The earlier check at $\mu = 1$ does not verify numerical equivalence of the selected zero-regularization predictor.

The analysis remains retrospective, and bootstrap intervals exclude fitting and selection uncertainty. Different evaluation membership prevents attributing old–new metric differences solely to cross-split dependence. These results support group-disjoint sample-level association and comparison-specific MAE improvements; no new category-ranking or accuracy-prediction claim is inferred.

B Evidence Construction and Numerical Audits

B.1 Checkpoint provenance and evaluation interfaces

Table 1 records the evaluated checkpoint pairs and diagnostic coverage. Checkpoint provenance was checked against available repository metadata and file-specific hashes. Configuration, tokenizer, and weight-index hashes are distinguished from weight-payload hashes. Model cards identify InternVL2.5-8B as the base for HiMTok and ALTo, but do not establish the exact ancestor revision. Repository revisions are incompletely recorded, and precise initialization ancestry remains unverified. Reference compatibility permits source swaps without establishing identical training data or budgets.

Recorded InternVL tokenizations use leading-space yes/no IDs 9970/1038 for MME and option-label IDs 493/556/487/553 for A/B/C/D on MMBench. MME contrasts are oriented toward the correct answer; MMBench uses the correct option against the fixed Base-selected highest-logit incorrect option among the available choices. LLaVA/GLaMM use bare yes/no IDs 4874/694. Swaps use these decision-specific readout rows rather than entire vocabulary heads. Evidence is read at the last active causal-prefix position after image integration, with actual input IDs and positions retained in the interface records.

The GLaMM/LLaVA comparison uses the same images, questions, and answer contrasts but checkpoint-specific native wrappers; neither full input sequences nor concrete decision-position indices are identical. Both roles use the recorded RGB CLIP preprocessing with short-edge resize and center crop to 336 pixels. The evaluation disables region inputs and concerns MME language decisions, not the full grounding pipeline. Together with the multi-task FullScope training history and unverified initialization revision, these differences limit the comparison to checkpoint-level diagnostic replication rather than a strictly identical-input, mask-supervision-only causal contrast.

All evaluated roles cover their 1,000-record frozen diagnostic manifests without runtime omissions. Manifest construction and subsampling are distinguished from runtime exclusions.

B.2 Exact reconstruction audit

The base and unified checkpoints use the same 1,000-record manifest. The numerical audit caches $H_0, D_1, \ldots, D_L$ at the decision position and reads the normalization gain, bias, epsilon, and output-head rows directly from the implementation.

The architecture and hook audits check the decision position, residual-write locations, actual norm parameters, head bias/tied-weight status, and paired sample/token alignment. The evaluated InternVL-based checkpoints use RMSNorm with $\epsilon = 10^{-5}$ and no norm or output-head bias; the CPU unit tests additionally cover LayerNorm centering and bias. On MMBench, the strongest incorrect option selected by Base is held fixed across checkpoints. HiMTok's own highest-logit incorrect option differs from the fixed Base-selected distractor on 532 of the 1,000 MMBench diagnostic samples. The audit retains the Base-selected distractor rather than substituting HiMTok's alternative.

Precision paths and relative error. The model forward itself is native BF16. We distinguish three comparisons on the same captured trajectory: (A) FP32 direct evaluation from H_L and the actual norm/head versus FP32 write reconstruction; (B) native-BF16 raw logits versus the direct FP32 formula; and (C) BF16 direct accounting versus BF16 write reconstruction. For each path,

$$e_{\mathrm{abs}} = |M_{\mathrm{test}} - M_{\mathrm{ref}}|, \qquad e_{\mathrm{rel}} = \frac{e_{\mathrm{abs}}}{\max(|M_{\mathrm{ref}}|, 10^{-8})}. \tag{19}$$

The reference is the FP32 direct margin for A, the native raw margin for B, and the BF16 direct margin for C. Thus a zero raw margin need not be a zero denominator for A. The 10^{-5} numerical-precision check applies to Path A, not to B or C. This is FP32 accounting of a fixed native-precision trajectory, not a fully-FP32 model-forward experiment.

As shown in Table 9, the Path-A absolute p95 and maximum are below 10^{-5} in all four settings. On MME, both checkpoints have median absolute error 2.3842×10^{-7}; their overall stabilized relative p95 values are 9.5209×10^{-7} and 9.7406×10^{-7}.

BF16 robustness audit. Table 10 separates the native-logit precision gap (B) from BF16 reconstruction arithmetic (C). Their larger errors do not enter the Path-A identity test. The direct FP32

Table 9: Path-A FP32 fixed-pass error on 1,000 samples per benchmark and checkpoint. Both absolute and stabilized relative errors are reported as median, p95, and maximum.

Absolute error

Task	Ckpt.	Median	p95	Maximum
MME	Base	2.3842×10^{-7}	9.5367×10^{-7}	1.9073×10^{-6}
MME	HiMTok	2.3842×10^{-7}	5.3644×10^{-7}	1.4305×10^{-6}
MMBench	Base	1.9074×10^{-6}	3.8147×10^{-6}	7.6294×10^{-6}
MMBench	HiMTok	4.7684×10^{-7}	1.4305×10^{-6}	3.3379×10^{-6}

Stabilized relative error

Task	Ckpt.	Median	p95	Maximum
MME	Base	1.1754×10^{-7}	9.5209×10^{-7}	7.7119×10^{-5}
MME	HiMTok	9.6245×10^{-8}	9.7406×10^{-7}	7.9127×10^{-5}
MMBench	Base	3.0068×10^{-7}	4.2903×10^{-6}	7.1185×10^{-5}
MMBench	HiMTok	5.0629×10^{-7}	9.2468×10^{-6}	1.0341×10^{-3}

formula and native raw margin have 100% sign agreement on non-zero raw margins: 991/991 and 995/995 on MME, and 993/993 for each checkpoint on MMBench. Raw-zero cases are reported separately rather than counted as non-zero sign agreements.

Table 10: Precision-path robustness on the same 1,000-sample manifests. B: FP32 formula versus native-BF16 raw margin. C: BF16 direct accounting versus BF16 reconstruction. All entries are absolute errors, separate from the Path-A identity check.

		Path B: native-logit gap			Path C: BF16 arithmetic		
Task	Ckpt.	Median	p95	Max	Median	p95	Max
MME	Base	0.032388	0.082811	0.119394	0.001465	0.031250	0.062500
MME	HiMTok	0.010982	0.033894	0.052109	0.000000	0.015625	0.031250
MMBench	Base	0.034888	0.091554	0.120668	0.000000	0.125000	0.250000
MMBench	HiMTok	0.005133	0.018081	0.034859	0.015625	0.042969	0.062500

Near-zero-margin bins. We stratify each checkpoint by its own absolute native raw margin for the same fixed decision contrast. Let $B_1 = [0, 10^{-3})$, $B_2 = [10^{-3}, 10^{-2})$, $B_3 = [10^{-2}, 10^{-1})$, and $B_4 = [10^{-1}, \infty)$. Raw-zero samples are a subset of B_1, not an additional disjoint bin. Empty bins have $n = 0$ and NA statistics. These are within-checkpoint descriptions, not paired differences between potentially different bin members.

Table 11: Raw-margin-bin coverage and near-zero Path-A absolute error. Each row totals 1,000 samples. Raw-zero counts are already included in B_1.

Task	Ckpt.	B_1	B_2	B_3	B_4	Raw zeros	B_1 abs. p95
MME	Base	9	0	0	991	9	2.8759×10^{-7}
MME	HiMTok	5	0	18	977	5	2.9318×10^{-7}
MMBench	Base	7	0	0	993	7	1.9074×10^{-6}
MMBench	HiMTok	7	0	65	928	7	9.5367×10^{-7}

On MME, the near-zero bins contain nine Base and five HiMTok samples (Table 12). Their Path-A maximum absolute errors are 2.9057×10^{-7} and 3.2503×10^{-7}, respectively, while stabilized relative p95 values rise to 5.9481×10^{-5} and 6.5264×10^{-5}. We therefore use absolute error as the primary numerical check in this small-margin regime.

In MMBench's first bin, the Path-A maximum absolute errors remain 1.9074×10^{-6} and 9.5367×10^{-7} (Table 13). Each estimate describes only seven samples. By contrast, all seven Path-B reference margins per checkpoint are zero; the stabilized relative maxima reach 7.6824×10^{6} and 1.5863×10^{6}.

Table 12: Complete MME Path-A absolute and relative error by native raw-margin bin. NA denotes an empty bin, not zero error. All nonempty saved summaries have finite-error counts equal to the bin sample count.

Absolute error					
Ckpt.	Bin	n	Median	p95	Maximum
Base	B_1	9	2.2352×10^{-7}	2.8759×10^{-7}	2.9057×10^{-7}
Base	B_2	0	NA	NA	NA
Base	B_3	0	NA	NA	NA
Base	B_4	991	2.3842×10^{-7}	9.5367×10^{-7}	1.9073×10^{-6}
HiMTok	B_1	5	1.0245×10^{-7}	2.9318×10^{-7}	3.2503×10^{-7}
HiMTok	B_2	0	NA	NA	NA
HiMTok	B_3	18	1.7323×10^{-7}	4.0699×10^{-7}	5.8115×10^{-7}
HiMTok	B_4	977	2.3842×10^{-7}	5.3644×10^{-7}	1.4305×10^{-6}

Stabilized relative error					
Ckpt.	Bin	n	Median	p95	Maximum
Base	B_1	9	3.6651×10^{-6}	5.9481×10^{-5}	7.7119×10^{-5}
Base	B_2	0	NA	NA	NA
Base	B_3	0	NA	NA	NA
Base	B_4	991	1.1731×10^{-7}	8.8414×10^{-7}	7.3369×10^{-6}
HiMTok	B_1	5	7.9360×10^{-6}	6.5264×10^{-5}	7.9127×10^{-5}
HiMTok	B_2	0	NA	NA	NA
HiMTok	B_3	18	2.9400×10^{-6}	1.1102×10^{-5}	1.8084×10^{-5}
HiMTok	B_4	977	9.4547×10^{-8}	7.0139×10^{-7}	3.2195×10^{-6}

Table 13: Complete MMBench Path-A absolute and relative error by native raw-margin bin. NA denotes an empty bin, not zero error. All 1,000 FP32 references per checkpoint are non-zero, and no non-finite margin or error is recorded.

Absolute error					
Ckpt.	Bin	n	Median	p95	Maximum
Base	B_1	7	1.9074×10^{-6}	1.9074×10^{-6}	1.9074×10^{-6}
Base	B_2	0	NA	NA	NA
Base	B_3	0	NA	NA	NA
Base	B_4	993	1.9074×10^{-6}	3.8147×10^{-6}	7.6294×10^{-6}
HiMTok	B_1	7	7.1526×10^{-7}	9.5367×10^{-7}	9.5367×10^{-7}
HiMTok	B_2	0	NA	NA	NA
HiMTok	B_3	65	4.7684×10^{-7}	1.4305×10^{-6}	1.9074×10^{-6}
HiMTok	B_4	928	4.7684×10^{-7}	1.4305×10^{-6}	3.3379×10^{-6}

Stabilized relative error					
Ckpt.	Bin	n	Median	p95	Maximum
Base	B_1	7	2.4827×10^{-5}	5.9518×10^{-5}	7.1185×10^{-5}
Base	B_2	0	NA	NA	NA
Base	B_3	0	NA	NA	NA
Base	B_4	993	3.0040×10^{-7}	4.0019×10^{-6}	4.3241×10^{-5}
HiMTok	B_1	7	5.2098×10^{-5}	7.5762×10^{-4}	1.0341×10^{-3}
HiMTok	B_2	0	NA	NA	NA
HiMTok	B_3	65	9.4904×10^{-6}	5.9437×10^{-5}	2.1526×10^{-4}
HiMTok	B_4	928	4.7843×10^{-7}	3.6522×10^{-6}	1.2867×10^{-5}

Path C has two Base and six HiMTok zero-reference cases within that bin (four and ten over the full manifests). We retain these denominator counts and interpret near-zero cases primarily through absolute error.

For MME, binwise identity statistics refer to Path A, whereas Paths B/C are reported as overall robustness summaries. The separately saved raw-margin-versus-FP32-reconstruction comparison is a composite precision check and is not relabeled as Path B. No separate pre-stabilization zero-denominator counts are available in the MME summaries; raw-zero counts alone are not used to infer the reference-zero counts for Paths A/C.

B.3 Four-quadrant source swaps

The reported drift norms use category-mean G with 32 block-write rows; H_0 is excluded. Each benchmark uses its fixed 1,000-record manifest. MME matrices have shape 32×14 and use the correct yes/no answer versus its alternative. MMBench matrices have shape 32×20 and use the correct option versus the Base checkpoint's highest-logit incorrect option, held fixed across quadrants. The swap combines representation-source states with readout-source final-normalization parameters and compatible label-token rows. Normalization statistics are recomputed on the representation-source final state. The Frobenius norms are within-benchmark geometry summaries, not cross-benchmark standardized effects.

ALTo/MMBench and HiMTok/MMBench use the same Base evidence, with identical sample order, categories, input IDs, answer positions, candidate IDs, and fixed distractors. Their $G^{0\to0}$ matrices are identical. All four ALTo/MMBench quadrants are available at the sample level. For ALTo/MME, the four-quadrant comparison is reported at the category-mean G level, whereas the predictive audit uses paired sample-level representation-side evidence. ALTo/MMBench Path-A p95 and maximum errors are 3.3379×10^{-6} and 5.7220×10^{-6}, respectively, on the native-BF16 cache.

B.4 Diagnostic-manifest representativeness

The full-set and 1k evaluations use the same native-logit candidate-decision rule. MME full contains 2,374 questions in 1,187 groups and 14 categories; its 1k manifest contains 671 groups under the original question-ID grouping and is sampled through an earlier category-capped 1,540-question set. The joint image/question linkage used for the grouped reanalysis yields 661 groups on this same manifest (Section A.6). MMBench full contains 4,329 rows in 1,164 groups, including option rotations; the 1k manifest consists of 1,000 canonical pass-0 rows. The nested sets support a representativeness audit, not an independent replication of the same accuracy effect.

Table 14: Full-set versus 1k-manifest degradation.

Benchmark	Full ΔAcc.	1k ΔAcc.	Abs. gap	Category ρ	Category MAE
MME	-23.76	-22.80	0.96	0.8207	2.491 pp
MMBench	-39.20	-37.80	1.40	0.8286	6.692 pp

The category Spearman correlations are $0.8207/0.8286$ and category MAEs are $2.491/6.692$ pp on MME/MMBench. The maximum category discrepancy is 8.81 pp for MME count and 21.66 pp for MMBench nature relation. These behavioral comparisons do not establish full-benchmark representativeness of internal evidence geometry.

C Transport Identification and Residual Diagnostics

C.1 Fitting and model selection

Every family uses the identity-centered objective in Equation (8). The regularization grid is $\lambda = \mu \|X_p\|_F^2 /32$, with $\mu \in \{0\} \cup \mathrm{logspace}(-12, 4, 65)$. Validation-error ties within 10^{-10} prefer the simpler family; multiplier ties within 10^{-15} prefer smaller μ. At zero regularization, full transport is computed as $T = Y_p X_p^\top (X_p X_p^\top)^+$, with `rcond=1e-12` applied to the Gram matrix, where $Y_p = X_p^{u\to0}$ and $X_p = X_p^{0\to0}$. This directly solves for minimum-norm T, rather than minimum-norm

$A = T - I$. Structured least squares uses `rcond=None`. For the selected grouped HiMTok/MME model, the effective design rank is $32/32$; the threshold details and finite-precision prediction check are reported in Section A.6. The saved evaluation target norms are nonzero, so the reported errors use $\epsilon_{\text{err}} = 0$. The structural families form nested spaces, but their held-out errors, not parameter counts, determine predictive selection.

Probe-calibrated affine control. On the frozen grouped $600/201/199$ partitions, we compare diagonal and full affine transport, $Y \approx TX + c\mathbf{1}^\top$, with an unpenalized layer-wise intercept. We estimate $\bar{x}_p$ and $\bar{y}_p$ on probe only, fit the centered probe matrices with the identity-centered penalty, and predict $\widehat{y}_i = T(x_i - \bar{x}_p) + \bar{y}_p$. Both families retain the existing regularization grid, the uncentered-probe scale $\lambda = \mu \|X_p\|_F^2 / 32$, and validation selection by reconstruction error. Evaluation reconstruction errors use the original target norm; scalar MAE compares $\mathbf{1}^\top(\widehat{y}_i - x_i)$ with $\mathbf{1}^\top(y_i - x_i)$. For HiMTok/MME, HiMTok/MMBench, ALTo/MME, and ALTo/MMBench, respectively, full affine transport selects (μ, λ) equal to $(0, 0)$, $(1.7783 \times 10^{-6},\ 2.4239 \times 10^{-4})$, $(5.6234 \times 10^{-6},\ 2.5274 \times 10^{-3})$, and $(3.1623 \times 10^{-6},\ 4.3104 \times 10^{-4})$. Diagonal affine transport selects $\mu = \lambda = 0$ in all four settings. The layer-wise intercepts remain unpenalized.

Validation selects full affine in all four settings. It also yields lower evaluation reconstruction error and scalar MAE than diagonal affine (Table 16). In table order, the full-minus-diagonal MAE differences and paired-group 95% intervals are -0.4376 $[-0.6025, -0.2749]$, -0.1912 $[-0.3202, -0.0658]$, -0.4185 $[-0.6218, -0.2045]$, and -0.3700 $[-0.6469, -0.1170]$. Intervals reuse the 2,000 saved group-bootstrap draws with fitted operators and intercepts held fixed; they are pointwise rather than simultaneous intervals. Thus, the observed advantage of full over diagonal transport persists when both allow layer-wise mean shifts. Adding an intercept does not itself uniformly improve the corresponding no-intercept results. The comparison remains retrospective and does not identify a causal mechanism.

C.2 Synthetic identifiability

For synthetic calibration, we generate $X \in \mathbb{R}^{32 \times N}$ from Gaussian draws with the empirical mean and layer covariance estimated only on each benchmark's 600-sample Base probe split, without whitening or layer rescaling. We construct $Y = T_{\text{true}}X + E_{\text{str}} + \Xi$, checking that each T_{true} belongs to its declared minimal family among scalar, diagonal, band-1, band-3 and full. For residual ranks $r \in \{1, 2, 4, 8\}$, E_{str} uses one shared orthonormal layer factor and independent Gaussian sample scores, with dataset-level RMS scaled to 0.10 times the clean $T_{\text{true}}X$ RMS; $r = 0$ sets $E_{\text{str}} = 0$. The independent Gaussian observation noise Ξ has standard deviation equal to the noise ratio times clean RMS. Thus noise ratio zero does not remove a nonzero-rank structured residual.

Each configuration uses the same 60/20/20 fitting/selection/evaluation protocol, identity-centered all-family regularization and numerical tie rules as the real transport analysis, with 200 repetitions. The $N = 1000$ main grid covers both benchmark covariance sources, five families, five residual ranks and noise ratios 0/.05/.10/.20, giving 200 configurations and 40,000 trials. The sample-size study uses $N = 250/500$ on the subset $r \in \{0, 4\}$ and noise ratios .05/.10. In total, 280 configurations represent 56,000 datasets, including 4,000 exact rank-0 bridge reuses. These are synthetic matrix trials, not additional model forwards.

Main-grid family-selection accuracy is 37,385/40,000 (93.46%; Wilson 95% interval [93.22%, 93.70%]). False-full counts a selected full family when the generating family is non-full: 510/32,000 (1.59%; Wilson 95% interval [1.46%, 1.74%]). Its denominator excludes full ground truth; it is a model-selection error rate, not a structured-null residual-test false-positive rate. Rank-0 false-full is 0/6,400, but rank-1/noise-0 rates reach 16.25% and 12.50% under MME/MMBench covariance. Tables 17 and 18 retain both overall recovery and the unfavorable strata. Main-grid mean relative operator error is 0.2790 and held-out prediction error is 0.1317; family selection, operator recovery and prediction are distinct endpoints.

Rank recovery separates the known structured component E_{str}, the oracle residual $Y - T_{\text{true}}X$, and the estimated residual $Y - \hat{T}X$. The recorded rank diagnostic uses threshold $\max(10^{-10}, 10^{-8}\sigma_1)$ when noise is zero. Otherwise, the threshold is $1.15\,\sigma_{\text{noise}}(\sqrt{32} + \sqrt{N_e})$. This is a numerical/spectral-edge diagnostic, not a structured-null test. The constructed structured component has its declared numerical rank in every trial. With oracle T, recovery is exact through noise .10; at .20, rank-4

Table 15: Validation-selected regularization λ for each transport family in Table 3. All entries correspond to the original $600/200/200$ partitions and the identity-centered objective. Nonzero values are rounded to five significant digits. These parameters do not refer to the grouped reanalysis or historical intervention-selection operators.

Setting	Scalar	Diagonal	Band-1	Band-3	Full
HiMTok/MME	0	7.8550×10^{-3}	7.8550×10^{-4}	1.3968×10^{-2}	4.4172×10^{-6}
HiMTok/MMBench	2.3903	7.5586×10^{-4}	4.2505×10^{-4}	4.2505×10^{-4}	1.3441×10^{-4}
ALTo/MME	0	2.4840×10^{-3}	2.4840×10^{-1}	2.4840×10^{-1}	2.4840×10^{-5}
ALTo/MMBench	13.441	1.3441×10^{-3}	1.3441×10^{-3}	2.3903×10^{-1}	7.5586×10^{-4}
GLaMM/MME	0	1.4268	2.5372×10^{-4}	4.5118×10^{-4}	1.4268×10^{-4}

Table 16: Affine transport on the grouped evaluation splits. Each setting contains 199 evaluation records. Errors are relative Frobenius reconstruction errors; MAE is measured in summed block-evidence-change units. Both families include probe-calibrated, unpenalized layer-wise intercepts.

	Reconstruction error		MAE	
Setting	Diagonal	Full	Diagonal	Full
HiMTok/MME	0.5934	0.3219	1.6540	1.2164
HiMTok/MMBench	0.9167	0.6427	1.5691	1.3780
ALTo/MME	0.7117	0.4371	2.4560	2.0375
ALTo/MMBench	0.8909	0.6154	3.1508	2.7808

recovery falls to approximately 64%/66% under MME/MMBench covariance and rank-8 recovery is 0%. With estimated T, noiseless residuals acquire small additional directions, so exact numerical-rank recovery is unreliable despite similar leading-energy profiles; at noise .05/.10 it agrees with the declared rank in all $N = 1000$ cells. These results support conditional recovery, not universal absence of full-selection bias or exact-rank recovery. Post-selection bootstrap coverage has not been computed.

For the separate rank-0 diagnostic at $N = 1000$, oracle residuals produce nonzero estimated rank in 0/2,000 noiseless trials, whereas estimated-T residuals do so in 35/2,000 (1.75%). Across noise ratios .05/.10/.20, both counts are 0/6,000. These denominators include all five generating families and differ from the non-full-only selection denominator. No nested structured-null false-positive calibration is inferred from these diagnostic counts.

C.3 Residual spectra and structured nulls

For a nonzero residual with singular values σ_j, define cumulative energy and stable rank by

$$\eta_k = \frac{\sum_{j\le k}\sigma_j^2}{\sum_j \sigma_j^2}, \qquad r_{\mathrm{stable}} = \frac{1}{\eta_1}. \tag{20}$$

Table 17: Synthetic recovery by generating family at $N = 1000$, pooled over both covariance sources, all five residual ranks and all four noise levels. Each family has 8,000 trials. Operator error is the saved mean relative Frobenius error of $\hat{T}$ against T_{true}, not prediction error. False-full is not applicable when the generating family is full; its pooled denominator is 32,000 rather than 40,000.

True family	Correct/trials	Correct (%)	Operator err.	False-full (%)
Scalar	6,727/8,000	84.09	0.0544	0.91
Diagonal	7,241/8,000	90.51	0.2734	1.81
Band-1	7,558/8,000	94.48	0.3604	1.89
Band-3	7,859/8,000	98.24	0.3795	1.76
Full	8,000/8,000	100.00	0.3271	—
All five	37,385/40,000	93.46	0.2790	1.59

Table 18: Synthetic sensitivity to residual rank, noise and sample count. (a) False-full percentages at $N = 1000$; every cell pools 1,600 non-full trials across both covariance sources. Noise zero still includes E_{str} for $r > 0$. (b) Matched sample-size subset: both covariance sources, all five families, $r \in \{0, 4\}$ and noise ratios .05/.10. Each N has 8,000 trials, including 6,400 non-full trials. This subset does not test the rank-1/noise-0 regime.

(a) False-full rate (%)

Residual rank	Noise 0	Noise .05	Noise .10	Noise .20
0	0.00	0.00	0.00	0.00
1	14.38	6.69	1.94	0.06
2	4.94	2.38	0.31	0.13
4	0.69	0.31	0.00	0.00
8	0.00	0.00	0.00	0.06

(b) Matched sample-size comparison

N	Correct (%)	Operator err.	False-full (%)
250	95.04	0.2980	0.39
500	95.46	0.2695	0.31
1000	96.28	0.2402	0.08

The two entropy ranks use different normalizations:

$$r_\sigma = \exp\left(-\sum_j p_j \log p_j\right), \qquad p_j = \frac{\sigma_j}{\sum_i \sigma_i}, \tag{21}$$

$$r_E = \exp\left(-\sum_j q_j \log q_j\right), \qquad q_j = \frac{\sigma_j^2}{\sum_i \sigma_i^2}. \tag{22}$$

Zero weights contribute zero to the entropy. Squaring the singular values emphasizes the leading modes; consequently an energy spectrum can be concentrated while many smaller singular values still raise r_σ. This explains why the statistics answer different questions, without assigning an untested cause to checkpoint-specific outcomes.

For each probe, validation, and evaluation split, the transformed inputs are $X = Z^{0\to 0}$ and $Y = Z^{u\to 0}$, containing the 32 block-write rows and excluding H_0; the nulls do not directly randomize the fitted residual.

Sample-wise layer permutation draws a separate layer permutation for each sample and applies the same permutation to its paired X/Y columns. The variance-preserving surrogate instead generates X and Y separately using a sequential RNG stream, without sharing their surrogate realizations. It alternates row and column mean-and-variance rescaling using population variance ($\mathrm{ddof} = 0$), with up to eight starts and 200 iterations per start. Initializations cycle through Gaussian, row-shuffled, and column-shuffled matrices.

Stopping checks the maximum relative row/column variance error against 10^{-6}, with denominator floor 10^{-12}. If the limit is reached, the candidate with the smallest recorded maximum variance error is returned. The saved maximum variance errors are below tolerance in all five settings. However, the stopping rule does not test mean errors, and final mean-error diagnostics were not retained. The recorded errors support numerical variance matching, not preservation of means, covariance, or singular-value spectra. Convergence is assessed from variance error, not merely from successful replicate completion.

In this implementation, each replicate recomputes the regularization scale from the transformed probe data, refits all five transport families, reselects regularization and family on null validation, and evaluates the corresponding null-evaluation residual. The reported tests retain 1,000 replicates per null. One-sided empirical tails use $(b+1)/1001$, counting null energy at least as high, or rank/reconstruction error at most as low, as observed. The lower attainable value is $1/1001$. Capability-label permutation is restricted to the separate capability-level audits.

Table 19: Holm-adjusted p values against the variance-preserving null. Adjustment is performed within each setting across the three reported metrics. The alternatives are higher Top-1 energy and lower entropy ranks r_σ and r_E. Values are rounded to three decimal places.

Setting	Top-1 energy	r_σ	r_E
HiMTok/MME	0.003	1.000	0.003
HiMTok/MMBench	0.003	1.000	0.003
ALTo/MME	1.000	1.000	0.003
ALTo/MMBench	1.000	1.000	1.000
GLaMM/MME	0.003	1.000	0.003

For the paper-level summary, Top-1 energy, r_σ, and r_E form one within-setting Holm family. The adjusted values in Table 19 are rounded to three decimals; $3/1001$ rounds to 0.003. The family does not combine all settings or all exploratory endpoints. We make no claim of external preregistration.

HiMTok/MME and HiMTok/MMBench retain Top-1 and r_E concentration under the variance-preserving null, while r_σ is not supported. ALTo/MME retains only r_E; ALTo/MMBench retains none of the three. GLaMM/MME retains Top-1 and r_E but not r_σ. The layer-permutation null separates the reported statistics more strongly. Because r_σ is unsupported under the variance-preserving null in every setting, a joint criterion requiring both Top-1 and r_σ cannot pass across both nulls.

ALTo/MMBench has observed Top-1/Top-2/Top-4 energy $0.4032/0.7313/0.9179$, $r_\sigma = 11.5584$, and $r_E = 4.6315$. The corresponding variance-null raw p values are 1.0000, 0.0420, 0.0280, 1.0000, and 0.5944. Top-2/Top-4 are exploratory, unadjusted findings outside the primary three-metric family. ALTo/MME has $r_\sigma/r_E = 14.9497/6.2856$. HiMTok/MME and HiMTok/MMBench have $r_E = 3.6951/4.2348$, stable rank $1.6295/2.2339$, and $r_\sigma = 12.0824/11.0520$.

C.4 Spectral stability and capability aggregation

The HiMTok spectral bootstrap uses 1,000 paired-sample resamples within each fixed split, refitting and reselecting transport; it is distinct from the fixed-operator cluster bootstrap used for predictive association. Mean absolute leading-left cosine similarity is 0.996 $[0.984, 0.999]$ on MME and 0.943 $[0.639, 0.999]$ on MMBench. Right-A compares directions on a fixed evaluation anchor after refitting: 0.974 $[0.934, 0.989]$ and 0.966 $[0.924, 0.985]$. Right-B resamples evaluation, obtains its leading left vector u_b, and projects it onto the common anchor using $(E_b^{\mathrm{anchor}})^\top u_b$: 0.973 $[0.933, 0.988]$ and 0.928 $[0.664, 0.982]$.

These are percentile stability intervals. Direction stability does not change a residual-null outcome. The analysis distinguishes a sample-selected residual, its capability average under the same transport, and a capability-refitted residual. Averaging and refitting are different operations; the capability audits do not substitute for sample-level structure. The wider MMBench direction intervals also limit interpretation of a single leading transferable vector.

D Intervention Analysis

D.1 Operator-row groups and inference

The historical high/low groups were selected using $s_\ell^{\mathrm{hist}} = \|(T_{\mathrm{hist}} - I)_{\ell:}\|_2$, where T_{hist} denotes the sample-level full transport fitted under the original $600/200/200$ probe/validation/evaluation partition. These historical operators used the zero-centered penalty $\lambda \|T\|_F^2$, rather than the identity-centered penalty of the final transport protocol. With $\lambda = \mu \left\| Z_p^{0\to 0} \right\|_F^2 / 32$, the historical parameters are $\mu = 10^{-8}$, $\lambda \approx 4.4172 \times 10^{-6}$ on MME and $\mu = 10^{-6}$, $\lambda \approx 1.3441 \times 10^{-4}$ on MMBench. The MMBench parameters are documented by a saved historical-protocol reproduction, not by a direct parameter field in the initial execution metadata.

The final-protocol check confirmed that the high/low sets were unchanged despite the change in operator: high $\{29, 30, 32\}$ and low $\{1, 2, 5\}$ on MME; high $\{30, 31, 32\}$ and low $\{1, 2, 5\}$ on

MMBench. The random controls, $\{25, 26, 28\}$ on MME and $\{26, 27, 29\}$ on MMBench, were frozen separately in the subsequent final protocol. They were sampled with seed 2027 from eligible unordered triples excluding high/low layers and admitting a one-to-one depth match within four blocks to the high group. The low group is not depth-matched. The reported comparisons use these frozen groups for both Base and Unified on the original 200-record evaluation subsets, not the grouped 199-record reanalysis.

Simple layer-selection controls. We compare the retained high groups with the last three blocks, $\{30, 31, 32\}$, and a Base probe evidence-RMS control. For the latter, we compute $q_\ell = \sqrt{N_p^{-1} \sum_{i \in p} (Z_{\ell i}^{0 \to 0})^2}$ on the original 600-record probe split and freeze the three highest-scoring blocks before examining their intervention outcomes. The RMS groups are $\{25, 27, 32\}$ on MME and $\{20, 29, 31\}$ on MMBench. The MMBench high and last-three groups are identical and do not constitute an independent comparison.

All comparisons use the original 200-record evaluation sets and the same four local intervention scales. Shared layers are retained in the paired group differences. In particular, the MME high-minus-last-three difference is $(S_{i,29} - S_{i,31})/3$ for each record. Inference uses 2,000 paired-cluster bootstrap resamples and 10,000 two-sided cluster sign flips with plus-one correction. Comparisons within each benchmark share resampling draws; the three non-identical comparisons form a separate Holm family.

The high groups are more sensitive than the probe-RMS controls on both benchmarks. However, the last three blocks are more sensitive than the high group on MME and identical to it on MMBench (Table 20). Thus, these comparisons do not establish an advantage over last-three-block selection. They describe sensitivity within the Unified checkpoint, not a causal effect of training.

Local scales are $0.90, 0.95, 1.05, 1.10$; large scales are $0.5, 1.5$. The group-localization analysis contains 28,800 records across checkpoints and benchmarks on the original 200-record evaluation subsets. Sensitivity first averages $|\Delta M^{\text{actual}}|/(|s-1| + 10^{-12})$ within sample/layer, then across group layers. Confidence intervals use 2,000 paired-cluster resamples. Two-sided cluster sign flips use 10,000 replicates with plus-one correction. Four Unified comparisons and four secondary Base–Unified contrasts form separate Holm families. The paired effect size d_z is computed at the sample level.

Group selection excludes evaluation intervention outcomes. The evaluation is nevertheless retrospective, and its cross-split group links are those in Section A.5. Within-evaluation cluster inference does not remove these links. The sign-flip procedure is not a randomized layer-assignment experiment.

Stability of historical layer selection. We use 200 paired-group bootstrap resamples of the original probe split per benchmark, holding validation fixed and repeating the historical fitting and selection procedure. This procedure selects among all five transport families by validation error, with zero-centered regularization tuned only for full transport. Layers are ranked by the historical operator-row score. Both benchmarks select full transport in all 200 resamples, with no failed replicates.

On MME, the selection frequencies of historical high layers 29, 30, and 32 are 79.0%, 54.5%, and 100%; layers 25 and 26 are selected in 13.0% and 53.5% of resamples. On MMBench, historical high layers 30, 31, and 32 are selected in 91.0%, 92.5%, and 94.5%; layers 22 and 29 are selected in 2.0% and 20.0%. All other layers have zero selection frequency. The complete historical top-three set is recovered in 84/200 resamples on MME and 156/200 on MMBench (42% and 78%, respectively). Thus, consistent selection of an individual layer does not imply stability of the entire set, particularly on MME.

The implementation follows the documented historical protocol and, on the original data, reproduces the saved operators within 10^{-10}, together with their selected parameters and top-three sets. The resulting frequencies describe probe-sampling sensitivity conditional on the fixed historical validation set and protocol; cross-split group dependence remains.

Table 20: Simple layer-selection controls in HiMTok. Differences are high minus control in magnitude-sensitivity units, not accuracy percentage points. Intervals are 95% paired-cluster percentile intervals; p values are Holm-adjusted across these three comparisons.

Task	Control	Difference	95% CI	Holm p
MME	Last three	-0.0513	$[-0.0672, -0.0351]$	0.0003
MME	Probe RMS	$+0.2809$	$[0.2588, 0.3016]$	0.0003
MMBench	Probe RMS	$+0.2573$	$[0.2276, 0.2896]$	0.0003

D.2 Continuous normalized-risk assessment

The separate data-conditioned score is

$$r_\ell^T = \frac{\left\| [(\widehat{T}_Z^\star - I) Z_p^{0\to 0}]_{\ell:} \right\|_2}{\left\| Z_{p,\ell:}^{0\to 0} \right\|_2 + 10^{-12}}. \tag{23}$$

Here, $\widehat{T}_Z^\star$ is the final identity-centered HiMTok transport used in the original reconstruction and predictive audits, not the historical selection operator T_{hist} or a grouped-reanalysis operator. The selected regularization is $\lambda \approx 4.4172 \times 10^{-6}$ on MME and $\lambda \approx 1.3441 \times 10^{-4}$ on MMBench. The score is computed on the original 600-record probe split and held fixed; intervention sensitivity is measured on the original 200-record evaluation subset. This normalized score is distinct from the operator-row norm used for historical group selection. The original continuous-risk analysis includes MME blocks $\{1, 2, 4, 5, 16, 25, 26, 28, 29, 30, 32\}$ and MMBench blocks $\{1, 2, 5, 26, 27, 29, 30, 31, 32\}$, each with complete four-scale coverage on the original 200-record evaluation set. These layer sets are unchanged by the subsequent simple-control interventions; the reported continuous correlations are not recomputed.

Table 21: Continuous normalized-risk association on the tested HiMTok layer subsets. Intervals resample evaluation clusters while keeping the tested layers, transport, and scores fixed.

Task	Tested layers	Spearman	Conditional 95% interval
MME	11	-0.100	$[-0.291, 0.091]$
MMBench	9	0.333	$[0.167, 0.550]$

The 2,000 paired-cluster resamples share evaluation indices across layers. The intervals describe sample-induced variation in association within fixed tested layer sets, not uncertainty from sampling layers from all 32 blocks. MMBench supports a conditional positive association; MME does not. These results do not imply a uniformly reliable continuous ordering, nor do they invalidate the separate operator-row group contrasts.

D.3 Base–Unified sensitivity contrasts

Let $\bar{S}_g^m$ denote mean sensitivity for checkpoint $m \in \{0, u\}$ and group $g \in \{H, L, R\}$. For $c \in \{L, R\}$,

$$\text{DiD}_{Hc} = (\bar{S}_H^u - \bar{S}_c^u) - (\bar{S}_H^0 - \bar{S}_c^0). \tag{24}$$

These compare changes in the high-minus-control magnitude gap, not signed margin effects or accuracy loss. Neither MME contrast is significant, while both MMBench contrasts are negative. Thus the within-Unified group localization does not establish a general training-induced increase in sensitivity. The $1/10001$ entries are finite-resampling bounds with plus-one correction.

Existing large-scale summaries use MME $N = 1000$, layers 4/16/28, and both checkpoints. Capability ranking uses same-checkpoint ΔAcc with ties retained. The output-level departure $C = \Delta M^{\text{actual}} - \Delta M^{\text{frozen}}$ and the downstream evidence sum $\sum_{j>\ell}(o'_j - o_j)$ are distinct summaries; the published inference in this paper is based on the local group and continuous analyses above.

Table 22: Base–Unified differences in high-minus-control sensitivity. Cluster-bootstrap intervals and Holm correction use a separate family of these four secondary comparisons.

Task	Contrast	DiD	95% interval	d_z	Raw p	Holm p
MME	High–low	-0.0621	$[-0.1809, 0.0489]$	-0.0749	0.299770	0.599540
MME	High–random	$+0.0080$	$[-0.1053, 0.1203]$	$+0.0104$	0.886211	0.886211
MMBench	High–low	-0.4927	$[-0.5814, -0.3991]$	-0.7195	$1/10001$	0.000400
MMBench	High–random	-0.4789	$[-0.5729, -0.3792]$	-0.7033	$1/10001$	0.000400

E Evidence-Constrained Training Study

E.1 Policy and controller specification

OECMT (Online Evidence-Constrained Mask Training) isolates the mask path in Stage A, opens a controlled shared subset in Stage B, and preserves but freezes that shared subset in Stage C. For shared layer group ℓ, the combined gradient is

$$g_{\ell,t} = g_{\ell,t}^{\text{lang}} + \gamma_{\ell,t} g_{\ell,t}^{\text{mask}}, \qquad 0 \le \gamma_{\ell,t} \le 1. \tag{25}$$

The gate scales only the shared mask-gradient component; private parameters receive their mask gradients without this attenuation.

Gradient pressure and compatibility. Language and mask gradients are accumulated separately in FP32, normalized by the global valid assistant-token count and mask-sample count, respectively, and summed across ranks. Before gating or clipping, we compute

$$p_{\ell,t} = \frac{\|g_{\ell,t}^{\text{mask}}\|_2/\sqrt{n_\ell}}{\|g_{\ell,t}^{\text{lang}}\|_2/\sqrt{n_\ell} + \epsilon_g}, \tag{26}$$

$$c_{\ell,t} = \frac{\langle g_{\ell,t}^{\text{lang}}, g_{\ell,t}^{\text{mask}} \rangle}{\|g_{\ell,t}^{\text{lang}}\|_2 \|g_{\ell,t}^{\text{mask}}\|_2 + \epsilon_c}, \tag{27}$$

where each vector concatenates all shared LoRA parameters in that layer and n_ℓ is their scalar parameter count. Unused task gradients contribute zeros. Thus, compatibility is computed over the whole layer group, not by averaging per-tensor cosines. We use $\epsilon_g = 10^{-8}$ and $\epsilon_c = 10^{-12}$.

Evidence monitoring. Let $\mathcal{C}$ index the controller-buffer groups I_τ, and let $[v]_+$ denote the element-wise positive part of v. Define

$$d_{\ell,t} = \frac{1}{|\mathcal{C}|} \sum_{\tau \in \mathcal{C}} \frac{\left\| \left[Z_{\ell,I_\tau}^{0\to 0} - Z_{\ell,I_\tau}^{\theta_t \to 0} \right]_+ \right\|_2}{\left\| Z_{\ell,I_\tau}^{0\to 0} \right\|_2 + \epsilon_Z}, \tag{28}$$

$$D_t^{\text{ctrl}} = \frac{1}{|\mathcal{C}|} \sum_{\tau \in \mathcal{C}} \frac{\left\| Z_{:,I_\tau}^{\theta_t \to 0} - Z_{:,I_\tau}^{0\to 0} \right\|_F}{\left\| Z_{:,I_\tau}^{0\to 0} \right\|_F + \epsilon_Z}. \tag{29}$$

The first measures evidence decreases and the second total drift, with $\epsilon_Z = 10^{-8}$. Groups receive equal weight, including zero-reference groups through the stabilized denominators. The executed pilot uses a single `visual_yes_no` group containing 915 records, rather than multiple balanced capability groups. Evidence contains the 32 block-write contributions, excluding H_0, and uses the fixed reference head and normalization gain with the RMS denominator recomputed from the current representation.

Executed gate and feedback. Let bars denote EMA control values. The EMA coefficient is 0.9, initialized from the first observation. The p/c EMAs update every Stage-B step, whereas the d/D EMAs update only when the full controller buffer is measured, at steps $0, 50, 100, 150$. Between refreshes, their latest values are reused. For steps $t = 50, \ldots, 199$, the gate is

$$\gamma_{\ell,t} = \sigma\big(b_0 + \tfrac{1}{2}\log(\bar{p}_{\ell,t} + 10^{-8}) + \tfrac{1}{2}\bar{c}_{\ell,t} - \bar{d}_{\ell,t} - \lambda_t \bar{d}_{\ell,t}\big), \tag{30}$$

where σ is the logistic sigmoid and $b_0 = -2.1972245773362196$. Steps $0, \ldots, 49$ instead use $\gamma_{\ell,t} = 0.1$ and disable feedback. All controller quantities are detached from gradient computation.

After combining shared and private gradients, we apply global norm clipping at 1.0 and perform one AdamW update. On each post-warm-up step, feedback then updates

$$R_t = \exp(-\bar{D}_t^{\mathrm{ctrl}}), \qquad \lambda_{t+1} = \max\{0, \lambda_t + 0.1(0.95 - R_t)\}, \tag{31}$$

with $\lambda_0 = 0$. The current gate uses λ_t, not the updated λ_{t+1}. Here, R_t is a geometric control proxy, not retained accuracy or a guaranteed constraint.

Uniform Gate replays the arithmetic mean of the 32 actually executed same-seed OECMT gates,

$$\gamma_t^{\mathrm{Uniform}} = \frac{1}{32} \sum_{\ell=1}^{32} \gamma_{\ell,t}^{\mathrm{OECMT}}, \tag{32}$$

including the warm-up steps. Every shared layer receives this same scalar. This matches mean gate attenuation, not parameter-count-weighted attenuation or actual AdamW parameter updates.

Trainable and frozen parameters. The study uses InternVL2.5-8B with rank-4 shared LoRA adapters (Hu et al., 2022) ($\alpha = 8$, dropout 0.05) on the `wqkv`, `wo`, `w1`, and `w3` projections, together with the restored rank-128 private branch. The saved runtime allowlists contain 256 shared tensors and 519 private tensors. The private list comprises 258 low-rank factor pairs across language, vision, and projector branches, plus the mask embedding, mask head, and boundary head. The codec/decoder and parameters outside the trainable allowlists remain frozen.

Stage B updates both lists. Stage C preserves and freezes the learned shared parameters and updates the same private list using a fresh optimizer. The general-language route retains the learned shared LoRA adapters while excluding private segmentation-specific parameters. Training uses BF16 parameters and native AdamW states, without FP32 master weights. Controller settings and parameter scopes are held fixed across seeds; the controller buffer does not represent the full range of MME/MMBench capabilities.

E.2 Phase-1 OECMT pilot details

All four policies start from the same Stage-A checkpoint. Stage B runs for 200 updates with matched language and mask batches; Stage C preserves and freezes the learned shared parameters and trains the private mask path for 100 updates. Joint SFT uses a shared learning rate of 10^{-5}, whereas Lower-LR uses 3×10^{-6}. Uniform Gate and OECMT both use a shared learning rate of 10^{-5}, with the policies specified above. The private learning rate is 3×10^{-5} for all four policies. Each Stage-B update uses 48 language and 48 mask examples, giving 9,600 exposures per task. Stage C uses 4,800 mask exposures from the same fixed training-manifest prefix. The language stream is derived from LLaVA-Instruct-150K, with images from COCO train2017. The mask stream is a locally converted mixture of COCO 2017 instance segmentation, COCO-Stuff 2017, and ADE20K (ADEChallengeData2016), using their training-source images and annotations. Stage B uses the first 9,600 records of each fixed training sequence.

Evaluation scope. The fixed evaluations comprise 4,250 segmentation validation records, 1,540 questions from MME `selected_100`, and 1,000 records from MMBench DEV_EN v1.0. The segmentation evaluation comprises 2,500 COCO instance, 1,250 COCO-Stuff, and 500 ADE20K records. It uses the project-defined seed-42 validation partition of the converted training-source pool, not the official validation splits of the source datasets. These are image–mask task records, not necessarily distinct images. Manifest identities and hashes are retained in the implementation audit. The MME training-study evaluation is distinct from the 1,000-record internal diagnostic manifest. Semantic evaluation uses the current checkpoint's general-language route with its learned shared parameters, not a Base-only fallback. Accuracy is computed over the legal answer candidates, rather than from the fixed-distractor evidence contrast.

Soft-on-hard cIoU. The reported segmentation endpoint uses probability-based mask decoding along a self-generated hard-code trajectory. In the inspected implementation, the model generates and feeds back 32 hard mask codes while retaining the code probabilities at each step. These probabilities

form codebook-weighted latents for mask decoding; soft embeddings are not fed back into the language model. The decoder output is channel-averaged and transformed by $\mathrm{sigmoid}(5(v - 0.5))$.

Evaluation operates on the 256×256 codec grid. Targets are right/bottom-padded to square and bilinearly resized; predictions are not restored to the original image resolution in this evaluation path. Predictions and targets are clamped to $[0, 1]$ and binarized using the strict threshold > 0.5. The reported cIoU is the sum of binary intersections divided by the sum of binary unions over the evaluation set, with value 1 when the total union is zero. Thus, "soft" describes decoding, not a continuous soft-IoU metric.

Implementation-version qualification. The evaluator for seeds 2028 and 2029 matches its recorded execution hash, but the historical versions of some decoding, preprocessing, and metric dependencies remain unverified. The definitions above describe the inspected implementation. The exact seed-2027 evaluator snapshot is unavailable.

Table 23: Nearest measured Stage-B comparisons for seed 2027. Differences are OECMT minus the comparator. An asterisk denotes a boundary-nearest match outside the comparator's observed segmentation range and should not be interpreted as strict matching.

OECMT / comparator	ΔcIoU (pp)	ΔMME (pp)	ΔMMBench (pp)
B50 / Joint B50	+0.1096	+0.1299	0.0000
B50 / Uniform B50	+0.0053	0.0000	-0.1000
B50 / Lower-LR B100*	+0.0011	+0.2597	-0.2000
B100 / Uniform B100*	+0.0418	+0.0649	0.0000
B200 / Uniform B150	+0.0455	-0.2597	-0.1000

These single-seed comparisons show local trade-offs rather than a consistent Pareto dominance claim.

Table 24: Single-seed Phase-1 pilot after Stage C (seed 2027). Semantic loss is in percentage points; negative values denote improvement. Stage-C gain is the cIoU increase from B200 to C100 in percentage points; semantic loss is measured relative to the common Stage-A/B0 reference, not relative to B200.

Method	C100 cIoU	Stage-C gain	MME loss	MMBench loss
OECMT	0.592903	+0.5565	+0.2597	0.0000
Joint SFT	0.591933	+0.4209	0.0000	+0.1000
Lower-LR	0.591612	+0.3313	+0.1299	+0.1000
Uniform Gate	0.591571	+0.5662	0.0000	-0.1000

The seed-2027 pilot shows competitive but non-dominating behavior: OECMT attains the highest final cIoU, while Uniform Gate has the largest Stage-C gain and different semantic-retention trade-offs.

E.3 Three-seed OECMT–Uniform comparison

We extend only OECMT and Uniform Gate to seeds 2028 and 2029, retaining the completed seed-2027 runs. All runs start from the same Stage-A checkpoint and use the same data order, trainable parameter sets, controller settings, and B200/C100 budgets. Seed variation affects shared LoRA initialization, dropout, and mask-image augmentation, without reshuffling the training records. Uniform Gate replays the actual per-step mean gate schedule from OECMT with the same seed, rather than reusing the seed-2027 schedule. Joint SFT and Lower-LR remain single-seed references.

We evaluate B50, B100, B150, B200, and C100 for each method and seed. Table 25 reports Stage-B segmentation and semantic-retention trajectories. Retention is the accuracy difference from the unique common Stage-A/B0 reference, so positive values indicate improvement. The shared B0 is not treated as three independent observations.

The Stage-B results vary with training step and benchmark. At B100, OECMT has higher mean segmentation and both semantic scores, but this ordering does not persist at B150 or B200. Same-step comparisons match the update budget, not necessarily segmentation quality; they do not establish an equal-quality retention advantage or consistent Pareto dominance.

Table 25: Three-seed Stage-B trajectories. Entries are mean ± sample standard deviation across seeds 2027, 2028, and 2029. cIoU is expressed in percent; retention is expressed in percentage points relative to the common B0 reference. Standard deviations are not confidence intervals.

Method	Step	cIoU (%)	MME retention	MMBench retention
OECMT	B50	58.710 ± 0.225	-0.108 ± 0.187	0.000 ± 0.100
Uniform	B50	58.722 ± 0.190	-0.087 ± 0.075	$+0.067 \pm 0.058$
OECMT	B100	59.069 ± 0.042	-0.065 ± 0.065	$+0.100 \pm 0.100$
Uniform	B100	58.975 ± 0.056	-0.130 ± 0.065	$+0.033 \pm 0.115$
OECMT	B150	58.692 ± 0.139	-0.152 ± 0.037	0.000 ± 0.000
Uniform	B150	58.590 ± 0.298	-0.087 ± 0.150	$+0.133 \pm 0.058$
OECMT	B200	58.752 ± 0.171	-0.065 ± 0.195	$+0.067 \pm 0.208$
Uniform	B200	58.686 ± 0.183	$+0.065 \pm 0.065$	$+0.067 \pm 0.058$

Table 26: Final Stage-C results over three seeds. All entries are percentages, reported as mean ± sample standard deviation. Each method's semantic accuracies are unchanged from its own B200 checkpoint; Stage C updates only the private mask path.

Method	C100 cIoU	MME accuracy	MMBench accuracy
OECMT	59.089 ± 0.211	84.351 ± 0.195	87.567 ± 0.208
Uniform Gate	58.997 ± 0.215	84.481 ± 0.065	87.567 ± 0.058

At C100, OECMT has slightly higher cIoU in all three seeds, with a mean paired difference of 0.093 percentage points. Its mean MME accuracy is 0.130 percentage points lower than Uniform Gate, while mean MMBench accuracy is equal and the per-seed differences have mixed signs. For seeds 2028 and 2029, comparisons across all eight ranks of each of the four runs found no changes in the examined general-route parameters, buffers, or inference configuration during Stage C. The seed-2027 checks cover only the narrower shared-LoRA freeze scope.

Semantic accuracies are identical between B200 and C100 within every run. This observed accuracy equality is distinct from the state-invariance checks and does not by itself establish parameter equivalence. The reported Stage-C segmentation gains therefore coexist with unchanged measured semantic accuracy; the cross-method semantic differences are already present at B200.

The three-seed comparison does not establish a consistent semantic-retention advantage over uniform attenuation. No significance or equivalence test is inferred from these descriptive means and standard deviations.

F Compute and Reproducibility

The intervention analysis reused 14,400 records and added 14,400 interventions plus 800 baselines: 15,200 new forwards, 0.9837 summed GPU-hours, and 18.75 GiB peak allocation per worker. This is incremental diagnostic cost, not OECMT training cost.

The above diagnostic cost excludes the subsequent last-three-block and probe-RMS controls. These controls reused 7,200 intervention records and added 1,600 intervention forwards and 400 baseline forwards. Their additional allocated cost was 0.0573 GPU-hours, including model loading, checks, and failed attempts.

The OECMT–Uniform multi-seed extension completed four additional B200/C100 training runs and 20 complete checkpoint evaluations. Its incremental allocated cost was 66.0203 GPU-hours: 36.8447 for training and associated loading, checks, and failure costs, and 29.1756 for evaluation. This excludes previously incurred Stage-A and seed-2027 costs and is not the total cost of all OECMT experiments.

Checkpoint provenance, manifest coverage, and available execution-version bindings are documented in Sections B.1 and E. The compute figures above describe incremental analyses rather than the total project cost.