When Preservation Misleads: Scale-Dependent Audits of Frozen Video-Language Interfaces

Anonymous authors
Paper under double-blind review

Abstract

Visual-token compression changes both the selected evidence and the values delivered to a frozen language model. Explaining a performance change requires distinguishing these effects. We introduce matched-packet audits: paired value interventions that hold source identities, token count, order, positions, and layout fixed. In native PruneVid, aggregation increases directional concentration at all three tested strengths, but the largest concentration change does not coincide with the largest reference-relative recall gap. At Medium, held-out alignment increases recall by 0.0237 under video-level scoring references while preserving delivery-stage directional diversity. The improvement therefore cannot be explained by recovery of that diversity; the query-level scoring estimate is inconclusive. Controlled experiments expose two further interpretation problems. Orthogonal interventions preserve norms, pairwise inner products, and linear CKA while reducing recall by 80.0–87.9% across four frozen consumers. Spatial–newline comparisons change sign between absolute and high relative perturbation amplitudes, while a common low-relative-dose window shows no detected recall separation. These audits distinguish representation change, behavioral loss, and response to a correction. They show how explicit source references and controlled interventions can rule out misleading explanations of compression outcomes.

1 Introduction

Visual-token sequences consume substantial computation and context in video-language models. Token reduction addresses these costs through temporal memory Song et al. (2024); He et al. (2024), temporal compression Ren et al. (2024); Shen et al. (2025), pruning Chen et al. (2024); Huang et al. (2025b), and token aggregation or hybrid selection–merging Ren et al. (2023); Yang et al. (2025); Shang et al. (2025). When performance falls after compression, a practical question is whether to retain different source evidence or change the values delivered to the model. A frozen consumer makes this distinction important: its pretrained interface can fail even when no source token is removed.

An end-to-end comparison can change token identities, count, values, and packing layout together. Its score combines the effects of those changes. Feature-similarity summaries describe the representation, but do not determine how a fixed downstream network uses it. A useful audit must connect three questions: what changed in the packet, what changed in behavior, and which explanations survive a controlled correction?

We answer these questions with matched-packet audits (Fig. 1). A selected source packet becomes the explicit reference, and paired interventions alter delivered values while fixing source identities, token count, order, positions, layout, prompt, decoder, and consumer. The comparison turns an undifferentiated compression loss into two observable reference-path quantities: a source-packet contrast and a fixed-packet value contrast.

The native PruneVid audit provides a concrete diagnostic example. Aggregation concentrates token directions, yet the concentration change and the recall gap order the operating

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.

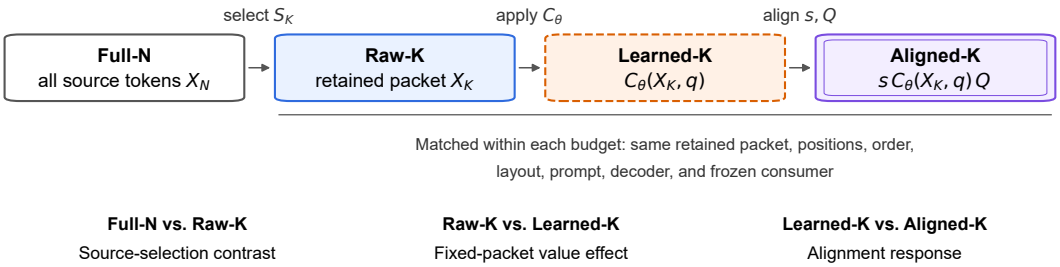

Figure 1: Matched-packet audit at a frozen visual interface. Full-N versus Raw-K changes the source packet. Within a budget, Raw-K provides the source reference; value-intervened packets retain its identities and consumer-side metadata. Learned-K and Aligned-K show one realization of this design. The paired behavioral contrast measures what changes after the packet values, rather than the selected sources, are changed.

points differently. A held-out alignment improves Medium recall under video-level scoring while mathematically preserving delivery-stage directional diversity. This rules out restoration of that diversity as the explanation for the gain. Reporting the source-reference and compressed means also reveals when a smaller relative gap reflects a changing reference.

Controlled experiments test the reasoning behind these diagnoses. Orthogonal interventions preserve selected geometric quantities while behavior degrades across four frozen consumers. Spatial–newline controls show that an apparent token-type ranking can reverse when perturbation amplitude is defined relative to source norms; a common low-dose window shows no detected recall separation. Here intervention scale means the magnitude of a value change, distinct from the fraction of tokens retained. The contribution is a source-referenced audit that separates measured changes from explanations of loss and recovery.

2 Related Work

Efficient video-language modeling. Video-language systems connect a pretrained vision backbone to a language model through a learned projection and token-packing interface Liu et al. (2023); Zhang et al. (2024); Bai et al. (2025b;a). Temporal-memory approaches Song et al. (2024); Ren et al. (2024); He et al. (2024) and token reduction through pruning or merging Chen et al. (2024); Rao et al. (2021); Bolya et al. (2023); Huang et al. (2025b) reduce these costs. Recent video compressors select representative subsets, allocate budgets across segments and tokens, organize high-ratio spatial–temporal pruning, or distill full-token semantics into compressed packets Cho et al. (2026); Ma et al. (2025); Ju et al. (2026); Guo et al. (2026). DeCo separates patch-level compression from semantic abstraction by using adaptive average pooling and leaving abstraction to the language model Yao et al. (2024). LangDC uses a lightweight language expert's hidden states as variable-length video representations, with semantic-density-aware supervision and downstream instruction tuning Wang et al. (2025). These methods address how to construct compact visual representations. Our question is how behavior changes along an explicit reference path: first with source selection, then with the values delivered after selection.

Representation similarity and functional equivalence. CKA and SVCCA quantify relationships between representations Kornblith et al. (2019); Raghu et al. (2017), while work on whitening, redundancy reduction, and dimensional collapse studies how feature geometry evolves during learning Ermolov et al. (2021); Zbontar et al. (2021); Bardes et al. (2022); Jing et al. (2022). Similarity diagnostics need not specify how a downstream network uses the available representation Hayne et al. (2024). Shesha measures geometric agreement between complementary feature subsets and is sensitive to orthogonal basis changes Raju (2026). The difference between geometric similarity and functional preservation motivates our controlled video-language interventions: holding specific geometric quantities constant makes the consumer's behavioral change directly observable.

Multimodal alignment and readout adaptation. Model re-basing and stitching make explicit that a permutation or basis change can preserve network function when downstream parameters are transformed with it Ainsworth et al. (2023); Bansal et al. (2021). Learned rotations can also be used deliberately to align multimodal representations with a language model's parameter space Yue et al. (2026). Implicit multimodal alignment studies perceptual and textual representations within frozen LLMs Shukor & Cord (2024). Modality Integration Rate quantifies cross-modal alignment quality during multimodal pretraining Huang et al. (2025a). Deep Pre-Alignment uses a perceiver VLM to align visual features before instruction tuning the target model Yu et al. (2026). The relevant distinction is whether the downstream readout adapts with the representation. Here it remains frozen, and a held-out source-referenced map measures the behavioral response to correcting the delivered features.

3 Methods

3.1 Matched Packets and Reference-Conditioned Contrasts

Let Full-N be the original visual packet and Raw-K an explicitly selected source packet. Learned-K applies a frozen value transform C_θ to Raw-K; Aligned-K applies a held-out direction-and-scale map to Learned-K. Within a budget, these three conditions share source identities, token count, order, positions, packing metadata, prompt, decoder, and frozen consumer. Writing $\bar{Y}$ for mean query score, the total score change follows the identity

$$\bar{Y}(\text{Learned-K}) - \bar{Y}(\text{Full-N}) = \underbrace{\bar{Y}(\text{Raw-K}) - \bar{Y}(\text{Full-N})}_{\text{source-packet contrast}} + \underbrace{\bar{Y}(\text{Learned-K}) - \bar{Y}(\text{Raw-K})}_{\text{fixed-packet value contrast}}. \quad (1)$$

This reference-path decomposition separates packet construction from value replacement. The value contrast measures the intervention's total endpoint effect, including changes to task-relevant information and its use by the frozen consumer.

For calibration packets C, R, the LLaVA control fits an uncentered Procrustes map in float64, using row vectors and right multiplication:

$$C^\top R = U\Sigma V^\top, \qquad Q = UV^\top, \qquad s = \max\left(0, \frac{\langle CQ, R\rangle_F}{\|CQ\|_F^2}\right). \quad (2)$$

The aligned packet is sCQ, using a global map fitted on 96 evaluation-disjoint calibration videos without answers. It changes neither token order nor consumer weights. The uncentered, unnormalized fit preserves the interface origin and includes source magnitudes in the fitting loss; $s \geq 0$ is a uniform gain. This restricts the correction to an orthogonal map and gain, without a translation. The native PruneVid fit uses the equal-video weighted, regularized variant specified in Appendix C.1.

3.2 Controlled Interventions

Orthogonal basis changes. For visual features $X \in \mathbb{R}^{n\times d}$ and $W^\top W = I$, $X' = XW$ preserves row norms and pairwise inner products, hence the Frobenius norm ratio and linear CKA equal one. The four-consumer test intervenes at each model's visual output boundary. A separate Qwen3-VL-4B dose experiment uses paired two-dimensional rotations at six angles from zero to $\pi/2$, with displacement $\|XW_\theta - X\|_F/\|X\|_F = 2\sin(\theta/2)$ and pairing seeds 17, 29, and 43. Appendix B.1 specifies the injection sites.

Spatial versus structural values. The LLaVA-Video-7B-Qwen2 packed boundary has 16 frames, each containing 169 spatial and 13 newline vectors ($N = 2{,}912$). Three existing source-constrained checkpoints (seeds 3401–3403, $\lambda_{\text{src}} = 10$) generate $Z = C_\theta(X, q)$ from the same full packet. Raw delivers X; Full delivers Z; Spatial-only substitutes Z only at spatial positions; Newline-only substitutes Z only at newline positions. All four retain the complete packet and metadata. The whole-group contrasts change different numbers of positions and perturbation energies (Appendix B.2).

A follow-up matches total perturbation energy between the two position groups. For $t \in \{\text{sp}, \text{nl}\}$, define $\Delta_t = M_t \odot (Z - X)$ and $E_t = \|\Delta_t\|_F^2$. Both conditions use $E_* = \min(E_{\text{sp}}, E_{\text{nl}})$ and deliver $X + \sqrt{E_*/E_t}\,\Delta_t$, retaining the queries and checkpoints (Appendix B.3).

A count-matched follow-up selects 208 spatial positions, one per frame–row block, using three fixed masks. We either substitute their transformed values (S208) or add the same newline displacement to each selected position (S208-R). Adding that displacement, $\delta_{\text{nl}} = z_{\text{nl}} - x_{\text{nl}}$, at the 208 newline positions defines N208. The repeated-displacement comparison matches count, direction, repetition, rank structure, and intended absolute energy. Appendix B.4 specifies mask selection and delivery checks.

Relative-amplitude controls instead deliver $x_i + \alpha\delta_{\text{nl}}$. Global relative matching (S208-RG) uses $\alpha_g = \text{mean}_i \, \|x_i\|_2/\|x_{\text{nl}}\|_2$; Local relative matching (S208-RL) uses $\alpha_i = \|x_i\|_2/\|x_{\text{nl}}\|_2$ at each position. Local scaling matches each relative displacement to the inherited N208 ratio $\|\delta_{\text{nl}}\|_2/\|x_{\text{nl}}\|_2 \approx 4.97$. These high-dose tests preserve displacement direction and rank-one structure but increase absolute energy (Appendix B.5).

The low-dose follow-up fixes $r \in \{0.05, 0.10, 0.20, 0.40\}$ before inference and uses $u = \delta_{\text{nl}}/\|\delta_{\text{nl}}\|_2$ in both arms. It delivers $x_i + r\|x_i\|_2 u$ at 208 spatial or newline positions, with Raw as the shared $r = 0$ endpoint. This matches local relative amplitude, direction, count, and rank structure; source norms determine absolute energy (Appendix B.6).

Directional packet-concentration probe. To probe the shared high-dose spatial response, we reconstruct Raw, S208, S208-R, S208-RG, and S208-RL without new training or fitting. For the 208 treated rows H, let $U_i = H_i/\|H_i\|_2$ and $G = UU^\top$. Directional participation rank is $R_{\text{PR}} = \text{tr}(G)^2/\|G\|_F^2$. This is the eigenvalue participation ratio $(\sum_j \lambda_j)^2/\sum_j \lambda_j^2$ for eigenvalues λ_j of G. It varies continuously with spectral concentration and equals one for collinear rows. Exact rank can remain high with small residuals; the nuclear norm of this unit-row Gram matrix is always the row count. We use participation rank to describe directional concentration, measuring it at delivery and after the first decoder RMSNorm, alongside angular concentration and top-direction share (Appendix B.7).

Learned-transform control. The original LLaVA transform uses a tokenwise 256-dimensional residual bottleneck, trained on 6,127 external rows for 20 epochs with the cosine proxy $\mathcal{L}_i = 1 - \cos(K^{-1}\sum_k C_\theta(x_{ik}, q), a_i)$, where a_i averages frozen input embeddings of all training-answer token IDs, including special tokens. No consumer language-model loss is used. The source-constrained family adds normalized source MSE to this proxy. The 50% packet selects eight frames from one cached Full-N packet; the final checkpoint and calibration are fixed before evaluation. Appendix A.1 gives optimization and integrity checks.

3.3 Native Aggregation and Source References

We audit PruneVid's native `merge_frames_dynamic` aggregation with frozen PLLaVA-7B Xu et al. (2024); Huang et al. (2025b). The 16-frame, 12×12 pre-aggregation packet contains 2,304 tokens. With $\tau = 0.8$ and temporal-segment ratio 0.25, cluster ratios 0.75, 0.50, and 0.25 define Light, Medium, and Strong. Setting `selected_layer`$= -1$ disables downstream adaptive pruning. For each native output group, Representative-K selects the source vector nearest to its aggregate in squared Euclidean distance, breaking ties by source index. Compressed-K contains native aggregates; Repaired-K applies the Medium-fitted 96-video calibration, reused at all three strengths.

The native groups are computed from the original features and then held fixed across the reduced conditions. This measures the delivered-value effect conditional on the native algorithm's realized groups and layout. We define the recall gap $G = \bar{Y}(\text{Representative}) - \bar{Y}(\text{Compressed})$ and alignment gain $A = \bar{Y}(\text{Repaired}) - \bar{Y}(\text{Compressed})$. The original nearest-source reference is supplemented by earliest-index and Euclidean-medoid references within the same groups. These feature-reference checks differ from rescoring predictions against different text references.

We additionally measure directional participation rank on all K delivered rows and after the first decoder RMSNorm, pairing native groups across the three conditions (Appendix C.7).

Extensions test PruneVid-map transfer to MVBench and separately fitted maps for adapted ToMe and STTM (Appendix C.4). An independent MVBench map uses 90 videos pooled across 15 tasks (Appendix C.5).

3.4 Data, Endpoints, and Statistical Analysis

The main cohort contains 150 ordered queries from 138 Charades videos, using Charades-STA annotations Sigurdsson et al. (2016); Gao et al. (2017). Training, calibration, and evaluation sources are disjoint. The controlled LLaVA and native PruneVid evaluations use 16 uniformly sampled frames; model-specific orthogonal sampling is detailed in Appendix B.1. Prompts and greedy decoding remain fixed within each comparison. MVBench Li et al. (2024) evaluations use multiple-choice accuracy, with malformed outputs scored incorrect.

The historical controlled-experiment endpoint is substring keyword recall. Let $\mathcal{W}$ be the unique lowercased, whitespace-split reference tokens after removing tokens of length at most two and the fixed stopwords in Appendix A.2. Punctuation is retained, and

$$Y = \frac{|\{w \in \mathcal{W} : w \text{ is a substring of the lowercased prediction}\}|}{\max(|\mathcal{W}|, 1)}. \tag{3}$$

This historical rule permits substring matches and punctuation-related misses. We preserve it for the original analyses and report whole-word rescoring as a robustness endpoint for the token-type comparison. Native PruneVid already uses whole-word recall with `[a-z]+` tokenization and video-level descriptions; a post-hoc sensitivity analysis uses query-level action descriptions on the same predictions. These metrics measure lexical coverage. A post-hoc cross-consumer robustness analysis applies the controlled whole-word scorer to the cached 150-query identity and full-strength outputs. Its four paired differences form one fixed family with individual 95% and Bonferroni 98.75% intervals. Auxiliary teacher-forced NLL averages only reference-answer tokens within each query, using generation-matched conversation and visual formatting.

Behavioral point estimates weight queries equally. Paired percentile bootstrap draws resample videos, carrying all queries and conditions with each sampled cluster. The token-type ladder averages three checkpoints within query; spatial-mask conditions first average their three fixed masks. Its five whole-group primary recall contrasts receive individual 95% and Bonferroni 99% intervals. Count matching and high relative amplitude each define a separate two-comparison family against N208, with individual 95% and Bonferroni 97.5% intervals. These adjustments target nominal familywise 95% coverage within each stated family. The follow-ups form a sequential investigation: earlier results motivated later protocols, with focal contrasts fixed before their respective runs. This timing distinguishes within-run prespecification from independent confirmation. The single energy-matched recall contrast uses a two-sided 95% interval.

The low-dose primary endpoint is normalized trapezoidal substring-recall AUC(SP) minus AUC(NL) over $r \in \{0, 0.05, 0.10, 0.20, 0.40\}$. The concentration probe averages masks and checkpoints within video because queries from the same video share one visual packet. Its primary endpoint is first-RMSNorm R_{PR}(S208-RL) minus R_{PR}(S208-R). Both analyses resample 138 videos and report two-sided 95% intervals. Dose-specific and per-mask contrasts, non-primary concentration metrics, whole-word ladder rescoring, NLL, and generation-length measures are auxiliary. Their reported intervals are unadjusted. The native concentration audit has three primary delivery-stage Compressed-minus-Representative rank contrasts, with 95% and Bonferroni 98.333% intervals over 138 videos. Intervals condition on fitted checkpoints and maps; Appendix A.3 collects resampling settings and comparison families.

4 Results

4.1 Geometry-Preserving Interventions Change Consumer Behavior

A representation can preserve the tested geometric quantities and still become unusable to its frozen consumer. Orthogonal interventions preserve token norms, pairwise inner prod-

ucts, and linear CKA by construction. In Qwen3-VL-4B, recall pooled across three pairing seeds fell from 0.1976 at identity to 0.0282 at the largest prescribed angle. Full-strength interventions produced 80.0–87.9% relative recall drops across four frozen interfaces (Table 1). The same separation appeared on a non-generative endpoint: Qwen2.5-VL-7B accuracy fell from 0.6587 to 0.3780 on 1,500 MVBench Li et al. (2024) questions. Exact whole-word rescoring of the same cached Charades outputs produced 85.9–93.6% relative drops; all four identity-minus-full-strength Bonferroni 98.75% intervals excluded zero. Four previously executed rotations per consumer likewise produced 69.6–92.7% substring-recall drops. Norms, pairwise geometry, and linear CKA can therefore describe a preserved representation without certifying that the frozen consumer can use it. Appendix B.1 reports the complete dose series, rescoring, and rotation-seed results.

Table 1: Cross-consumer full-strength support on the canonical 150-query, 138-video Charades cohort. Each row is a separate frozen consumer and model-specific visual injection boundary; identity and full-strength use the same prompt, greedy decoder, and keyword scorer within a row. Full-strength uses seed 999; the dose series in Appendix B.1 uses separate interventions.

Consumer	Identity	Full-strength	Relative drop
LLaVA-Video-7B	0.2047	0.0410	−80.0%
Qwen2.5-VL-7B	0.2043	0.0248	−87.9%
Qwen3-VL-4B	0.1976	0.0383	−80.6%
Qwen3-VL-8B	0.1710	0.0296	−82.7%

Learned-transform control. The proxy-trained LLaVA transform provides a separate realization of the matched-packet comparison. At 50% retention, recall was 0.1551 for Raw-K, 0.0000 after the value transform, and 0.1167 after held-out alignment (alignment difference +0.1167, 95% CI [0.0854,0.1493]). A video-swap test showed that alignment also increased the correct-video advantage by +0.0533 [+0.0243, +0.0825]. Appendix A.4 reports the full comparison.

4.2 Token-Type Rankings Depend on the Intervention Scale

Absolute-magnitude controls produced greater recall loss at newline positions. Replacing the 208 newline values reduced recall from 0.2047 to 0.1181, whereas replacing all 2,704 spatial values yielded 0.2082. Matching total squared energy preserved a +0.0928 spatial–newline difference (95% CI [+0.0543, +0.1324]). A stricter control changed 208 spatial positions and applied the same repeated newline displacement used at the 208 newline positions. This matched count, direction, repetition, rank structure, and absolute energy; spatial recall was 0.2126, giving +0.0946 relative to the newline condition (Bonferroni 97.5% CI [+0.0514, +0.1395]).

The source states, however, differed sharply in scale: the newline vector had norm 1.0033, while the treated spatial vectors had mean norm about 54.2. Under exact local relative-amplitude matching at the inherited high dose ($r \approx 4.97$), the ordering reversed. Spatial recall fell to 0.0131 and the spatial–newline difference became −0.1050 (Bonferroni 97.5% CI [−0.1431, −0.0697]). A global mean-norm scale produced an intermediate point estimate of −0.0272 with an interval containing zero (Fig. 2b).

We then compared the two position groups over a prespecified common low-relative-dose window, $r \in \{0.05, 0.10, 0.20, 0.40\}$. Both curves remained close to the shared Raw baseline. Their normalized recall AUCs were 0.2057 for spatial and 0.2077 for newline positions, a difference of −0.0020 (95% video-cluster CI [−0.0106, +0.0068]). No dose-specific recall interval excluded zero. Whole-word recall gave the same result (AUC difference −0.0047 [−0.0136, +0.0042]). Unadjusted exploratory intervals favored the spatial arm on answer-token NLL (−0.0161 [−0.0182, −0.0140]) and showed output length 0.60 words higher [+0.04, +1.18]; Appendix C.6 shows representative truncation and expansion failures.

Packet concentration at high relative amplitude. The high-dose spatial packets underwent a second, directly measurable change: their 208 treated rows collapsed toward the

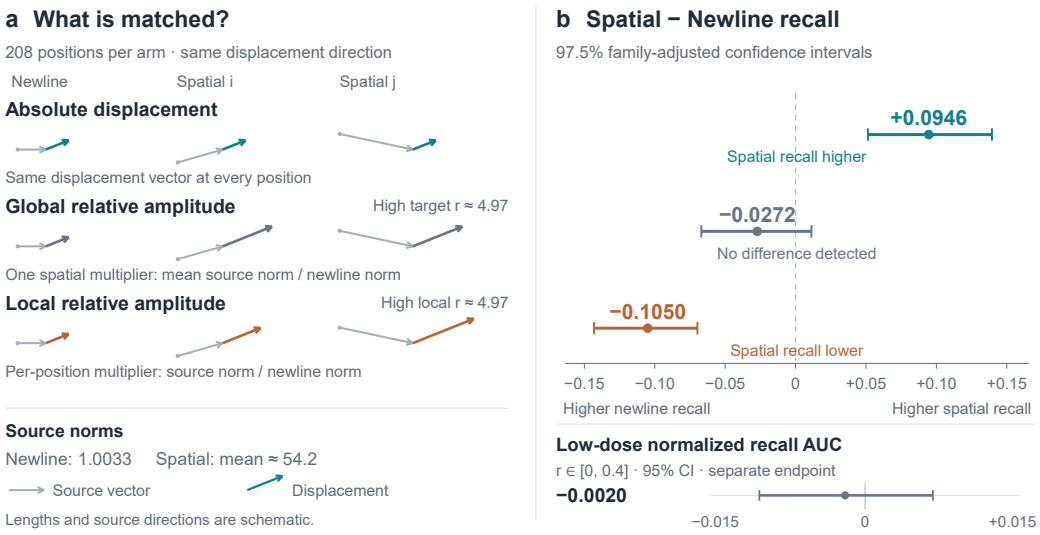

Figure 2: Intervention scale changes the spatial–newline comparison. **(a)** Gray arrows denote source vectors; colored arrows denote displacements with a common direction at 208 positions per arm. Spatial displacements are unchanged, scaled by the mean source norm, or scaled by each position's source norm relative to the newline norm. Lengths and source orientations are schematic. Relative matching uses the inherited high target $r \approx 4.97$: global matching uses the mean source norm; local matching matches r at every treated position. **(b)** Paired recall differences against N208. Bars are Bonferroni 97.5% video-cluster intervals from the original two-comparison families (S208/S208-R versus N208; S208-RG/S208-RL versus N208). The separate lower inset uses its own axis and reports normalized recall AUC over $r \in [0, 0.4]$: -0.0020, 95% CI $[-0.0106, +0.0068]$. Analyses use 150 queries from 138 videos, averaging three checkpoints and three spatial masks within query. Complete condition means appear in Table 15.

repeated displacement direction. After the first consumer RMSNorm, directional participation rank was 3.9955 under absolute matching (S208-R), 1.0502 under global relative matching (S208-RG), and 1.0375 under local relative matching (S208-RL). The prespecified S208-RL-minus-S208-R difference was -2.9580 (95% video-bootstrap CI $[-3.0247, -2.8925]$). Angular concentration rose from 0.6821 to 0.9908, and top-direction share rose from 0.4894 to 0.9818. Nearly the same concentration was already present in the delivered packet, where rank fell from 4.0276 to 1.0440: concentration begins at delivery and persists through normalization. Despite similar concentration, S208-RG and S208-RL had recalls of 0.0909 and 0.0131. Raw newline was already rank one, whereas Raw spatial rank was 3.9247 after RMSNorm. Concentration therefore needs a token-role-specific source reference (Appendix B.7).

Absolute matching perturbs the small-norm newline state more in relative terms; local relative matching delivers more absolute energy to spatial states. Appendices B.2–B.6 give the complete curves and energy audits.

4.3 Native PruneVid: Separating Change, Loss, and Alignment Response

Under video-level scoring references, feature displacement and reference-relative behavioral loss ordered the Medium and Strong PruneVid settings differently. Strong was farther from its Representative-K packet than Medium by cosine, norm ratio, and relative RMS error. Yet the Representative-K-minus-Compressed-K recall gap was -0.0075 at Strong and $+0.0204$ at Medium (Medium minus Strong $+0.0279$, 95% CI $[+0.0037, +0.0538]$; Fig. 3). Centered linear CKA was nearly unchanged at 0.9192 and 0.9198 (difference CI $[-0.0047, +0.0054]$). The absolute means explain the reversal: from Medium to Strong, Representative-K recall fell from 0.1618 to 0.1377 while Compressed-K changed from 0.1413 to 0.1452. The smaller gap thus reflected changes in both paths and must be read alongside their absolute scores.

At Medium, held-out alignment increased recall by +0.0237 [+0.0069, +0.0439], while Light and Strong showed no detected increase. Table 2 reports every operating point and Appendix C.1 reports the geometry and NLL results.

Table 2: Complete native PruneVid free-generation recall results. Entries in the last two columns are paired query-mean contrasts with 95% video-cluster bootstrap intervals (10,000 draws, seed 20260914). Full-N recall is 0.1736 at every point.

Point (K/N)	Representative	Compressed	Repaired	Rep.− Comp.	Repair− Comp.
Light (0.530)	0.1599	0.1565	0.1562	+0.0033 [−0.0107, +0.0187]	−0.0003 [−0.0139, +0.0137]
Medium (0.355)	0.1618	0.1413	0.1650	+0.0204 [+0.0043, +0.0383]	+0.0237 [+0.0069, +0.0439]
Strong (0.177)	0.1377	0.1452	0.1432	−0.0075 [−0.0251, +0.0109]	−0.0020 [−0.0129, +0.0103]

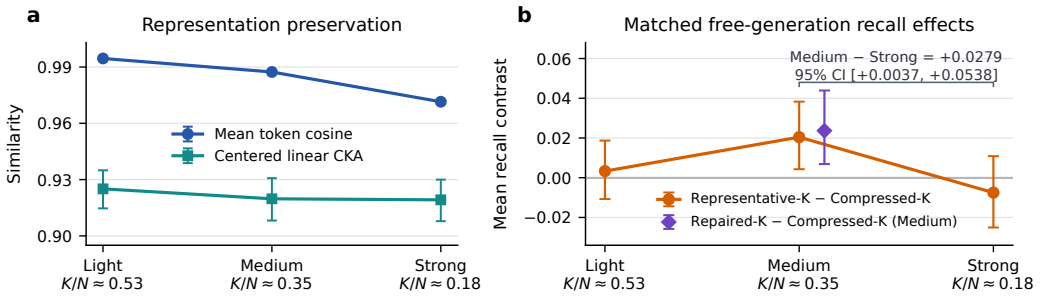

Figure 3: Native PruneVid audit on PLLaVA-7B/Charades. **(a)** Similarity of Compressed-K to Representative-K over 138 matched video pairs, on a magnified similarity axis. **(b)** Representative-K minus Compressed-K whole-word recall over 150 queries, using video-level scoring references; purple marks the Medium alignment response. Bars are 95% video-bootstrap or video-cluster intervals. The Medium–Strong gap was not established under query-level references (Appendix C.2).

Scoring-reference granularity. Rescoring tests how the native comparison changes with the text describing success. Query-level action descriptions differed from the video-level descriptions on 94 of 150 queries. Under query-level scoring, the Medium-minus-Strong gap was +0.0051 (95% CI [−0.0320, +0.0417]) and Medium's alignment response was +0.0124 [−0.0078, +0.0330]. The direct query-minus-video change in ordering was −0.0228 [−0.0580, +0.0102]. Thus the point-estimate ordering retained its sign, but the data did not establish that it persists across scoring references or that the references produce different effects. Appendix C.2 reports the complete three-strength, dual-reference family.

Feature source-reference sensitivity. At Medium, the earliest-index and medoid references gave recall gaps of +0.0178 and +0.0299, compared with +0.0204 for nearest-source selection. All six direct alternative-minus-original contrasts across Light, Medium, and Strong had 99.1667% intervals containing zero. Appendix C.3 reports the full selector comparison.

Native directional concentration. Aggregation increased directional concentration relative to matched source packets at all three strengths (Table 3). All three rank reductions survived the three-comparison adjustment. Strong had the largest reduction, but Medium had the largest recall gap. At Medium, alignment raised recall by +0.0237 while delivery-stage rank remained essentially unchanged at 3.2672. This invariance follows from the shared orthogonal map and uniform scaling, ruling out restoration of delivery-stage directional diversity as the explanation for the gain. Appendix C.7 reports the paired measurements and first-RMSNorm results.

Table 3: Native concentration and behavior. Rank contrasts are Compressed minus Representative at delivery, with three-comparison Bonferroni 98.333% intervals over 138 videos. G is Representative-minus-Compressed recall; A is Repaired-minus-Compressed recall. Historical behavior intervals remain in Table 2.

Point	Rank change [98.333% CI]	Recall gap G	Alignment gain A
Light	-0.0842 $[-0.0923, -0.0765]$	$+0.0033$	-0.0003
Medium	-0.2097 $[-0.2206, -0.1992]$	$+0.0204$	$+0.0237$
Strong	-0.5335 $[-0.5564, -0.5112]$	-0.0075	-0.0020

Task and method extensions. We detected no alignment benefit in the tested task and method extensions. Differences were -0.0047 on 1,500 MVBench questions, $+0.00056$ for adapted ToMe, and -0.00389 for adapted STTM; all three 95% intervals included zero. A newly fitted pooled MVBench map likewise changed accuracy by -0.0027 $[-0.0080, +0.0027]$. Appendix C.4 gives the configurations and complete intervals.

5 Discussion

Representation change, behavioral loss, and alignment response are distinct measurements. In native PruneVid, concentration increases at all three strengths, yet neither the recall gaps nor the alignment responses follow that change.

Alignment tests a specific explanation: its orthogonal map and uniform gain preserve delivery-stage directional diversity while Medium recall improves under video-level scoring. Diversity restoration cannot explain this gain. The smaller Strong gap also partly reflects a lower-scoring source reference, showing why a relative gap needs its absolute condition means.

Orthogonal interventions preserve selected geometric summaries yet disrupt behavior; token-type rankings vary with perturbation scale. At high doses, similar concentration also coexists with different recalls, so concentration describes input structure, not failure severity.

Using the audit to make a diagnosis. Reconstruct source values within the compressor's realized groups and compare Full-N with that reference (selection) and the reference with the native packet (value change). Test a correction when the latter degrades behavior; report absolute means and check whether the correction changes the proposed diagnostic.

Experiments use 150 Charades queries from 138 videos and the stated MVBench extensions. Free-generation lexical coverage is supplemented by whole-word recall and answer-token NLL.

6 Conclusion

Matched-packet audits distinguish representation change, behavioral loss, and response to correction. Native PruneVid concentrates token directions without a matching ordering of recall gaps; Medium alignment improves video-reference recall while preserving delivery-stage directional diversity. These results show why geometric invariants and token-type rankings alone cannot diagnose compression outcomes. The audit makes source references and value interventions explicit so proposed explanations can be tested against consumer behavior.

Reproducibility Statement. The appendices specify execution, scoring, calibration, and video-cluster inference. The anonymous supplement provides fixed manifests, cached scores, analysis scripts, and figure data for the reported audits.

AI Use Statement. Generative AI assisted with organization, language editing, LaTeX, experiment orchestration, and analysis and audit code. An exploratory model-assisted visual screen was conducted during the research and excluded from the reported evidence. The authors take responsibility for the analyses and final content.

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. 3

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025a. 2

Shuai Bai et al. Qwen2.5-VL Technical Report, 2025b. 2

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021. 3

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022. 2

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023. 2, 23

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, volume 15139 of *Lecture Notes in Computer Science*, pp. 19–35, 2024. doi: 10.1007/978-3-031-73004-7_2. 1, 2

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. FLoC: Facility location-based efficient visual token compression for long video understanding. In *International Conference on Learning Representations*, 2026. 2

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021. 2

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.563. 5

Xiaoyang Guo, Guoping Luo, Jusheng Zhang, Keze Wang, and Wenhao Wang. Token-budget distillation: Transferring full-token semantics to compressed video vision-language models. In *ACM International Conference on Multimedia*, 2026. 2

Lucas Hayne, Heejung Jung, and R. McKell Carter. Does representation similarity capture function similarity? *Transactions on Machine Learning Research*, 2024. 2

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.01282. 1, 2

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models via modality integration rate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 218–227, 2025a. 3

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025b. doi: 10.18653/v1/2025.findings-acl.1024. 1, 2, 4

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-granular spatio-temporal token merging for training-free acceleration of video LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23990–24000, 2025. 23

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *ICLR*, 2022. 2

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, HuaiXi Wang, Yiyi Zhou, and Rongrong Ji. Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. *arXiv preprint arXiv:2603.22911*, 2026. 2

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019. 2

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024. doi: 10.1109/cvpr52733.2024.02095. 5, 6

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-1516. 2

Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang. MMG-Vid: Maximizing marginal gains at segment and token levels for efficient video LLMs. *arXiv preprint arXiv:2508.21044*, 2025. 2

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *NeurIPS*, 2017. 2

Prashant C. Raju. Geometric stability: The missing axis of representations. *arXiv preprint arXiv:2601.09173v5*, 2026. 2

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 2

Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. Testa: Temporal-spatial token aggregation for long-form video-language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.66. 1

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.01357. 1, 2

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/iccv51701.2025.02122. 1

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025. URL `https://proceedings.mlr.press/v267/shen25j.html`. 1

Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen LLMs to multimodal inputs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130848–130886, 2024. doi: 10.52202/079017-4159. 3

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pp. 510–526, 2016. doi: 10.1007/978-3-319-46448-0_31. 5

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.01725. 1, 2

Xiangchen Wang, Jinrui Zhang, Teng Wang, Haigang Zhang, and Feng Zheng. Seeing more, saying more: Lightweight language experts are dynamic video token compressors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 541–558, 2025. doi: 10.18653/v1/2025.emnlp-main.28. 2

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. PLLaVA: Parameter-free LLaVA extension from images to videos for video dense captioning, 2024. 4

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/cvpr52734.2025.01843. 1

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. 2

Tianyu Yu, Kechen Fang, Zihao Wan, Kaidong Zhang, Yicheng Zhang, Jun Song, Bo Zheng, and Yuan Yao. Deep pre-alignment for VLMs. *arXiv preprint arXiv:2605.15300*, 2026. Accepted by ICML 2026. 3

Tongtian Yue, Xuange Gao, Longteng Guo, Zijia Zhao, Zikang Liu, Jie Jiang, Hua Huang, and Jing Liu. Rose: Rotate your large language model to see. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19265–19275, 2026. 3

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 2

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning with Synthetic Data, 2024. 2

A Execution and Reproducibility

A.1 Packed Transform and Calibration

The LLaVA-Video-7B-Qwen2 control captures one post-pooling, post-newline packet with $N = 2{,}912$ and hidden dimension 3,584. Its eight-frame selector uses indices $[0, 2, 4, 6, 9, 11, 13, 15]$ from the 16-frame source packet, giving $K = 1{,}456$. All reduced conditions reuse these immutable source indices, positions, packing metadata, question, and decoding settings. No second frame extraction is used to form Raw-K.

The tokenwise residual bottleneck has width 256, RMS normalization ($\epsilon = 10^{-6}$), SiLU, a fixed-question bias, and a scale-one residual connection, with no token mixing or dropout. It runs in float32, with Xavier-uniform initialization and zero bias. AdamW uses learning rate 2×10^{-4}, weight decay 0.01, gradient clipping at 1, and cosine decay to zero. The original control uses seed 3401, 6,127 external training rows, 20 epochs, and 122,540 updates; only the final checkpoint is retained. Each update uses one training row (batch size 1), with no gradient accumulation. Training, calibration, development, and evaluation partitions are disjoint in video and query identifiers. Calibration contains 96 unique external videos; duplicate annotations use the lexicographically first query ID.

The target is the first nonempty semicolon-delimited segment of the official training-video description, whitespace-stripped and tokenized directly with special tokens, without a conversation wrapper or answer mask. If $t_{i\ell}$ are the returned token IDs, $a_i = L_i^{-1} \sum_\ell E_{\text{in}}(t_{i\ell})$. The cosine objective normalizes the mean transformed visual vector and a_i separately. The source-constrained checkpoints add $\lambda_{\text{src}} \|C_\theta(X, q) - X\|_F^2 / \|X\|_F^2$. The token-type follow-up uses all three final $\lambda_{\text{src}} = 10$ checkpoints from the earlier four-penalty, three-seed family,

without new training. Their training uses the same 50% packed boundary; their intervention evaluation uses full retention and no alignment map.

Alignment requires at least 20 calibration videos, $10D$ token pairs, full centered rank of both packets, full uncentered cross-moment rank, and finite nonzero norms and scale. Identity-repack checks precede evaluation. External-development gates require at least 20 videos and queries, finite outputs, centered-energy ratio at least 0.05, norm ratio in $[0.1, 10]$, Full-N and Raw-K mean recall at least 0.05, Raw-K/Full-N at least 0.6, at least 25% nonzero Raw-K scores, and zero generation or scoring errors.

A.2 Scoring and Answer-Token NLL

The substring scorer removes tokens of length at most two and the stopwords *the, a, an, is, are, was, were, in, on, at, to, of, and, or, person, does*. Reference tokens are unique, lowercased, and whitespace-split; punctuation is retained. Token-type whole-word rescoring lowercases, extracts `[a-z0-9]+` tokens, and uses unique reference words after the length and stopword filters. Native PruneVid instead uses `[a-z]+` and retains repeated reference words when averaging coverage. Both test exact word membership in the prediction. The native stopwords are *a, an, the, is, are, to, of, and, in, on, at, then, they, person, someone*. Token-type whole-word rescoring changes both reference tokenization and matching; it retains the controlled-experiment stopword list above. Scorer and scoring-reference identity stay fixed within each reported paired contrast.

For reference-answer token indices $\mathcal{A}_i$, NLL is $-|\mathcal{A}_i|^{-1} \sum_{t \in \mathcal{A}_i} \log p(y_{it} \mid y_{i,<t}, X_i)$. Only answer tokens contribute; prompt, visual, and padding positions are masked. Formatting and visual placement match the corresponding generation condition. NLL is averaged within query before cohort aggregation. It measures likelihood of the specified reference, while whole-word and substring recall measure lexical coverage.

A.3 Paired Inference and Integrity

All queries from a sampled video remain together, with multiplicity when a cluster is sampled repeatedly. Each replicate divides by its sampled query count, preserving query weighting. Shared draws are used for paired conditions and contrasts. Checkpoint or intervention-seed means are formed within query, not treated as additional independent videos. The orthogonal dose uses 100,000 draws (seed 20260820); the original packed comparison uses the same count and seed. Token-type analysis uses 100,000 draws (20260915); the energy-, count-, and high-relative-amplitude follow-ups use the same count and seed. The low-relative-dose curve uses 100,000 draws (20260921). Whole-word rotation rescoring and the packet-concentration probe each use 100,000 draws (20260922). The three-strength native recall intervals in Table 2 use 10,000 draws (20260914); method/task extensions use 10,000 (20260906). The earlier Medium-only dual-reference check uses 10,000 (20260907). Native geometric and behavioral ordering uses 50,000 (20260914). Dual-reference rescoring uses 100,000 (20260917); source-reference sensitivity uses 50,000 (20260915); independent MVBench calibration uses 20,000 (20260915). The follow-up families have five, two, two, three, six, and two primary contrasts for token type, count matching, high relative amplitude, dual reference, source reference, and independent MVBench, respectively, with Bonferroni intervals at 99%, 97.5%, 97.5%, 98.333%, 99.1667%, and 97.5%. The low-relative-dose analysis has one primary AUC contrast and reports an individual 95% interval. Whole-word rotation rescoring has four primary consumer contrasts, with individual 95% and Bonferroni 98.75% intervals. The concentration probe has one primary first-RMSNorm contrast and reports a two-sided 95% interval. The native concentration follow-up uses 100,000 shared video draws (20260922) and three delivery-stage Compressed-minus-Representative rank contrasts, with individual 95% and Bonferroni 98.333% intervals (Appendix C.7). Other intervals are individual exploratory 95% intervals. Percentile coverage is approximate and conditional on the fitted checkpoints and maps.

The packed implementation validates source indices, shapes, positions, order, metadata, and a single selector invocation. Shards require complete coverage, no overlap, identical query ordering, and zero recorded errors. A synthetic exact-inverse test verifies the alignment al-

gebra at float64 precision. Independent rescoring and cluster-multiplicity calculations verify reported statistics. The whole-group token-type, energy-matched, count-matched, high-relative-amplitude, low-relative-dose, source-reference, and task-matched follow-ups contain 1,800, 900, 2,700, 2,700, 7,200, 900, and 3,750 records, respectively, with complete condition coverage and no exclusions or imputation. The concentration probe contains 14,076 metric rows, or 4,692 per checkpoint, and validates packet construction, masks, casting, relative amplitude, and finiteness. Historical-output parity and independent statistical recomputation are checked alongside these construction tests.

A.4 Learned-Transform Control and Input Dependence

At the original 50% endpoint, Full-N, Raw-K, Learned-K, and Aligned-K recall are 0.2047, 0.1551, 0.0000, and 0.1167. Full-N minus Raw-K is +0.0496 [0.0176,0.0831]; Learned-K minus Raw-K is -0.1551 $[-0.1959, -0.1168]$; Aligned-K minus Learned-K is +0.1167 [0.0854,0.1493]. These are individual 95% video-cluster intervals.

A fixed correct-versus-mismatched-video control forms 69 pairs from 138 videos and swaps visual packets while retaining each query and reference. Correct-minus- mismatched recall advantages are 0.0903, -0.0022, and 0.0511 for Raw-K, Learned-K, and Aligned-K. Alignment changes that advantage by $+0.0533$ $[+0.0243, +0.0825]$. Its NLL advantage changes by $+0.2405$ $[+0.1085, +0.3673]$, where positive means lower NLL for the correct video. The 10,000 paired draws resample the 69 video pairs (seed 20260905). This control measures input dependence before and after alignment of the proxy-trained intervention.

B Core Intervention Protocols and Complete Outcomes

B.1 Orthogonal Interventions

The four-consumer comparison uses the same 150-query, 138-video Charades pool, question, greedy decoder, and historical keyword scorer. The seed-999 rotation acts at the LLaVA-Video-7B `mm_projector` output, the Qwen2.5-VL-7B visual-merger output, or the Qwen3-VL-4B/8B visual output (both `pooler_output` and all DeepStack features). These are model-specific interfaces, with identity and rotated conditions matched within each model. The separate Qwen3-VL-4B dose characterization uses pairing seeds 17, 29, 43 and angles $0, \pi/16, \pi/8, \pi/4, 3\pi/8, \pi/2$. Thus its largest-dose value and the seed-999 full-strength value describe different interventions. The Qwen2.5-VL-7B MVBench test uses 1,500 inputs, 100 per task in 15 categories (sampling seed 42), 16 frames at 1 fps, a multiple-choice prompt, and the same visual-merger rotation. Unrecognized answers count as incorrect.

The cached 150-query Charades outputs are also rescored by exact whole-word membership using 100,000 paired video-cluster draws (seed 20260922). Four previously executed full-strength rotations provide a descriptive direction check for every consumer; rotation seeds specify interventions, not independent data or model replicates.

Table 4: Complete pooled dose-characterization summary for the canonical 150-query, 138-video Charades cohort using the frozen Qwen3-VL-4B consumer. Rows are prescribed orthogonal displacements pooled over pairing seeds 17, 29, and 43, which define intervention specifications rather than independent datasets.

Displacement	Mean recall	95% video-cluster CI
0.0000	0.1976	[0.1563, 0.2407]
0.1960	0.1815	[0.1443, 0.2202]
0.3902	0.1813	[0.1458, 0.2184]
0.7654	0.1704	[0.1385, 0.2034]
1.1111	0.1190	[0.0966, 0.1429]
1.4142	0.0282	[0.0188, 0.0385]

Table 5: Post-hoc exact whole-word rescoring of the cached 150-query Charades identity and full-strength outputs. Differences are Identity minus Full. Bonferroni 98.75% video-cluster intervals cover the fixed family of four consumer comparisons; all exclude zero.

Consumer	Identity	Full	Difference [98.75% CI]	Relative drop
LLaVA-Video-7B	0.2795	0.0395	+0.2400 [+0.1818, +0.3001]	85.9%
Qwen2.5-VL-7B	0.3043	0.0193	+0.2850 [+0.2290, +0.3436]	93.6%
Qwen3-VL-4B	0.2495	0.0283	+0.2212 [+0.1608, +0.2843]	88.6%
Qwen3-VL-8B	0.2131	0.0172	+0.1959 [+0.1398, +0.2551]	91.9%

Table 6: Descriptive sensitivity to the full-strength rotation draw, separate from the canonical four-consumer comparison in Table 1. LLaVA and Qwen2.5 use their fixed 50-query control cohorts; Qwen3 uses the canonical 150-query cohort. Each consumer has four rotation seeds. Seeds define interventions and are not independent data or model replicates.

Consumer	Queries	Identity recall	Rotation recall range (relative drop)
LLaVA-Video-7B	50	0.2137	0.0157–0.0543 (74.6–92.7%)
Qwen2.5-VL-7B	50	0.1877	0.0157–0.0233 (87.6–91.7%)
Qwen3-VL-4B	150	0.1976	0.0191–0.0601 (69.6–90.3%)
Qwen3-VL-8B	150	0.1710	0.0213–0.0341 (80.1–87.5%)

B.2 Spatial and Newline Intervention Protocol

The three source-constrained checkpoints in Appendix A.1 are evaluated at full retention. Each packet has 2,704 spatial and 208 newline positions. For $Z = C_\theta(X, q)$, Spatial-only selects Z at spatial positions and X elsewhere; Newline-only uses the complementary assignment. Every untreated position equals its source, and every treated position equals the full-transform value. Newline source values are checked against the model's structural vector. All four conditions use the same complete transform input, including the historical contiguous tensor layout, while preserving count, order, shape, dtype, and consumer metadata.

Inference uses 4-bit LLaVA-Video-7B-Qwen2, SDPA, average spatial pooling at stride 2, 16 sampled frames, the fixed action-description question, `qwen_1_5`, greedy decoding, and a 128-token output limit. Raw and Full reproduce historical predictions and NLL for every query and checkpoint. Tables 7 and 8 report the complete whole-group means and five primary recall contrasts; interaction means Full minus Spatial minus Newline plus Raw. The remaining auxiliary family is exploratory, with individual 95% intervals. All outputs are nonempty.

For seeds 3401, 3402, and 3403, Newline minus Raw recall differences are -0.0840, -0.0880, and -0.0880; Spatial minus Newline differences are $+0.0875$, $+0.0915$, and $+0.0915$. These are descriptive checkpoint-specific means; inference uses checkpoint averages within query.

Table 7: Full-retention token-type outcomes, averaged within query over three checkpoints. NLL is in nats per answer token; words use whitespace splitting. All outputs are nonempty.

Condition	Substring recall	Whole-word recall	NLL	Words
Raw	0.2047	0.2795	4.2760	33.47
Full transform	0.1661	0.2292	4.2879	37.06
Spatial-only	0.2082	0.2891	4.2414	32.88
Newline-only	0.1181	0.1960	4.6547	40.52

Table 8: Five primary substring-recall contrasts. Individual 99% intervals account for the five-comparison family. Interaction is Full–Spatial–Newline+Raw.

Contrast	Estimate	95% CI	99% CI
Full–Raw	−0.0386	[−0.0742, −0.0031]	[−0.0854, +0.0080]
Spatial–Raw	+0.0036	[−0.0170, +0.0237]	[−0.0238, +0.0301]
Newline–Raw	−0.0866	[−0.1262, −0.0482]	[−0.1391, −0.0364]
Spatial–Newline	+0.0902	[+0.0516, +0.1296]	[+0.0395, +0.1422]
Interaction	+0.0445	[+0.0065, +0.0851]	[−0.0052, +0.0989]

Table 9: Exploratory token-type contrasts with individual 95% intervals. Signs are literal differences; lower NLL is better. All nonempty-fraction contrasts equal zero in this cohort.

Contrast	Whole-word recall	NLL	Output words
Full–Raw	−0.0503 [−0.0866, −0.0140]	+0.0119 [−0.0898, +0.1214]	+3.5889 [−1.1134, +8.6225]
Spatial–Raw	+0.0097 [−0.0116, +0.0309]	−0.0346 [−0.0398, −0.0295]	−0.5978 [−2.6082, +1.3651]
Newline–Raw	−0.0834 [−0.1256, −0.0424]	+0.3787 [+0.2242, +0.5372]	+7.0511 [+1.5890, +12.5599]
Spatial–Newline	+0.0931 [+0.0512, +0.1367]	−0.4133 [−0.5714, −0.2598]	−7.6489 [−13.0972, −2.2538]
Interaction	+0.0235 [−0.0197, +0.0693]	−0.3323 [−0.4429, −0.2227]	−2.8644 [−9.9234, +4.1778]

B.3 Energy-Matched Spatial and Newline Interventions

Following the whole-group results, we evaluated an energy-matched intervention on the same 150 queries, 138 videos, and three source-constrained checkpoints. For each query and checkpoint, the spatial and newline masks define $\Delta_t = M_t \odot (Z - X)$ and total squared energy $E_t = \|\Delta_t\|_F^2$. The shared target is $E_* = \min(E_{\text{sp}}, E_{\text{nl}})$, and $\alpha_t = \sqrt{E_*/E_t}$ gives Spatial-matched and Newline-matched packets $X + \alpha_t \Delta_t$. If either energy is zero, both packets equal X; no such pairs occurred. Untreated positions retain X exactly. When $\alpha_t = 1$, treated positions receive Z directly to avoid floating-point subtraction-and-addition error. The decoder and preprocessing match the whole-group experiment.

Across the 414 video–checkpoint combinations, spatial energy ranged from 63,158.0 to 74,974.2. Newline energies were 5,163.8418, 5,160.3467, and 5,181.3184 for checkpoints 3401, 3402, and 3403, respectively. Thus $\alpha_{\text{nl}} = 1$ throughout, while α_{sp} ranged from 0.26244 to 0.28642. After fp16 delivery, energy-to-target ratios ranged from 0.99964361 to 1.00056962. The maximum deviation was 0.057%, below the 0.543% tolerance set at ten times the maximum residual in a forward-only precision check.

Total energy is matched while the position counts remain 2,704 spatial and 208 newline tokens. Newline perturbations repeat one vector across positions within each checkpoint; perturbation directions, rank, and per-token amplitudes therefore retain their group-specific structure.

Checkpoint scores are averaged within query, then 138 video clusters are resampled jointly across conditions in 100,000 percentile bootstrap draws (seed 20260915). The single primary contrast is Spatial-matched minus Newline-matched substring recall, with a two-sided 95% interval. Other endpoints use exploratory pointwise 95% intervals. Analysis uses Python 3.12.9 and NumPy 2.2.3.

Table 10: Energy-matched outcomes. Means average three checkpoints within each of 150 queries. Differences are Spatial-matched minus Newline-matched, with 95% video-cluster intervals. Recall is primary; remaining endpoints are exploratory.

Endpoint	Spatial-matched	Newline-matched	Difference [95% CI]
Recall	0.2108	0.1181	+0.0928 [+0.0543, +0.1324]
Whole-word recall	0.2888	0.1960	+0.0928 [+0.0508, +0.1359]
NLL	4.2679	4.6547	−0.3869 [−0.5452, −0.2327]
Output words	33.01	40.52	−7.52 [−13.08, −1.96]
Nonempty fraction	1.0000	1.0000	0.0000 [0.0000, 0.0000]

Checkpoint-specific recall differences were +0.0908, +0.0937, and +0.0937. All 450 Newline-matched predictions and NLL values matched their corresponding whole-group Newline-only outputs exactly.

B.4 Count-Matched and Repeated-Displacement Controls

The count-matched experiment uses the same 150 queries, 138 videos, three source-constrained checkpoints, transform inputs, and decoder as the preceding token-type analyses. Three masks, fixed by seeds 5501, 5502, and 5503 before evaluation, select one of the 13 spatial columns in every frame–row block. Each mask therefore contains $16 \times 13 = 208$ spatial positions, has no overlap with the newline positions, and is reused for every query and checkpoint.

For mask M, S208 delivers $X + M \odot (Z - X)$. S208-R delivers $X + M \odot \delta_{\mathrm{nl}}$, where $\delta_{\mathrm{nl}} = z_{\mathrm{nl}} - x_{\mathrm{nl}}$. The implementation verifies that all 208 newline displacements within a packet are identical and that S208-R and S208 use the same mask. N208 is the original newline-only condition. S208-R and N208 thus apply the same vector 208 times and have the same intended squared perturbation energy; their receiving slots differ. Treated S208 values equal Z exactly, untreated values equal X exactly, and S208-R values equal the cast delivery of $X + \delta_{\mathrm{nl}}$.

S208 delivered 1.0074 times the N208 energy on average (range 0.9380–1.1208), with mean per-position energy 25.0326 versus 24.8479 for N208. The S208-R delivered-energy ratio was 0.999972 (range 0.999893–1.000027). Scores are first averaged over the three masks, then over the three checkpoints within query. The two primary substring-recall contrasts use 100,000 paired video-cluster draws over 138 videos (seed 20260915), with individual 95% and Bonferroni 97.5% intervals. Whole-word recall, NLL, output length, and nonempty fraction are auxiliary endpoints.

Table 11: Complete count-matched outcomes. Condition means average three masks and three checkpoints within each of 150 queries. Contrast intervals are individual exploratory 95% intervals except that the two recall contrasts also have the family-adjusted 97.5% intervals reported in the final column.

Endpoint	S208	S208-R	N208	Differences from N208 [95% CI]
Recall	0.2116	0.2126	0.1181	$+0.0935$ $[+0.0556, +0.1324]$; $+0.0946$ $[+0.0566, +0.1335]$
Whole-word recall	0.2904	0.2920	0.1960	$+0.0943$ $[+0.0537, +0.1361]$; $+0.0960$ $[+0.0557, +0.1379]$
NLL	4.2730	4.2730	4.6547	-0.3817 $[-0.5400, -0.2274]$; -0.3817 $[-0.5400, -0.2256]$
Output words	33.60	33.60	40.52	-6.92 $[-12.43, -1.43]$; -6.93 $[-12.49, -1.49]$
Nonempty fraction	1.0000	1.0000	1.0000	0.0000; 0.0000

The two Bonferroni 97.5% recall intervals are $[+0.0502, +0.1380]$ for S208 minus N208 and $[+0.0514, +0.1395]$ for S208-R minus N208.

All nine checkpoint–mask recall differences were positive for both controls. The absolute repeated-displacement comparison therefore separates two representation states whose source scales also differ: the newline source norm is 1.0033, whereas the mean spatial source norm is approximately 54.2. The following controls expose how that scale difference changes the ordering.

B.5 Relative-Amplitude Repeated-Displacement Controls

The relative-amplitude experiment was frozen before behavioral inference and reuses the three checkpoints, three spatial masks, 150 queries, 138 videos, decoder, historical Raw and N208 records, and repeated newline displacement from Appendix B.4. For every packet, define $r_{\mathrm{nl}} = \|\delta_{\mathrm{nl}}\|_2 / \|x_{\mathrm{nl}}\|_2$. S208-RG uses one scale $\alpha_g = \mathrm{mean}_i \, \|x_i\|_2 / \|x_{\mathrm{nl}}\|_2$ for all treated spatial positions. S208-RL uses $\alpha_i = \|x_i\|_2 / \|x_{\mathrm{nl}}\|_2$ and therefore satisfies $\|\alpha_i \delta_{\mathrm{nl}}\|_2 / \|x_i\|_2 = r_{\mathrm{nl}}$ at every treated position. Both retain the displacement direction and the repeated rank-one construction; absolute energy consequently differs across conditions.

Across conditions, the target relative ratio averaged 4.9684. S208-RL's delivered ratio averaged 4.9684; its scales ranged from 12.19 to 393.34. S208-RG used scales from 42.78 to 70.77

(mean 57.10). Its single global scale matches the ratio defined by the mean spatial source norm, not every local ratio. Relative scaling raised absolute delivered energy to 3,283 and 4,218 times the N208 energy on average for S208-RG and S208-RL, respectively. These large values are a consequence of matching a displacement that is about five times the newline source norm at spatial source vectors whose norms are about 54 times larger. S208-RL energy was also uneven across treated positions. After deduplicating repeated queries from the same video, the largest position contributed a median 13.3% of each packet's delivered energy across 1,242 video–checkpoint–mask audits, with a maximum of 25.8%. The three mask-specific medians were 15.0%, 2.6%, and 14.4%.

Scores are averaged over masks and checkpoints within query. The two primary recall contrasts use 100,000 shared paired video-cluster draws over 138 videos and report individual 95% and Bonferroni 97.5% intervals.

Table 12: Relative-amplitude outcomes. Contrasts use N208 as the reference. Adjusted intervals are Bonferroni 97.5% intervals for the two primary recall contrasts.

Condition	Recall	Difference	95% CI	97.5% CI
N208	0.1181	reference	—	—
S208-RG	0.0909	−0.0272	[−0.0616, +0.0063]	[−0.0669, +0.0113]
S208-RL	0.0131	−0.1050	[−0.1381, −0.0741]	[−0.1431, −0.0697]

The S208-RG intervals contained zero for all three masks. The S208-RL mask-specific differences were −0.1090, −0.1115, and −0.0945, with each 95% interval excluding zero. The normalization rule changes the observed contrast. Absolute matching yields a positive spatial-minus-newline difference, global relative scaling detects no difference, and high-dose local relative matching yields a negative difference. Because absolute energy and relative amplitude cannot be matched simultaneously when source norms differ, this ladder establishes that the observed token-type ordering is conditional on the scale definition.

B.6 Low-Relative-Dose Response Curves

The low-dose protocol was frozen before behavioral inference. It reuses the three checkpoints, three spatial masks, 150 queries, 138 videos, decoder, and historical Raw records from the preceding controls. For checkpoint-specific $u = \delta_{\mathrm{nl}}/\|\delta_{\mathrm{nl}}\|_2$ and $r \in \{0.05, 0.10, 0.20, 0.40\}$, the newline arm delivers $x_{\mathrm{nl}} + r\|x_{\mathrm{nl}}\|_2 u$ at all 208 newline positions. The spatial arm delivers $x_i + r\|x_i\|_2 u$ at the 208 positions in each frozen mask. Both arms therefore use the same local relative amplitude, direction, changed-position count, and repeated rank-one structure. Absolute energy is not matched: the spatial source vectors are much larger than the newline source vector.

Spatial scores are averaged over the three masks and then the three checkpoints within query; newline scores are averaged over checkpoints. The shared historical Raw value supplies $r = 0$. For each arm, normalized AUC is the trapezoidal area over $r \in \{0, 0.05, 0.10, 0.20, 0.40\}$ divided by 0.40. The single primary contrast is recall AUC(SP) minus AUC(NL), with 100,000 shared video-cluster draws (seed 20260921) and a two-sided 95% interval.

Table 13: Low-relative-dose recall curves. Spatial mask columns average three checkpoints; Spatial averages the three fixed masks within query. Newline averages the same checkpoints. Dose-specific intervals are auxiliary.

r	SP-5501	SP-5502	SP-5503	Spatial	Newline	SP–NL [95% CI]
0.00	shared Raw			0.2047	0.2047	reference
0.05	0.2095	0.2022	0.2055	0.2057	0.2097	−0.0039 [−0.0128, +0.0026]
0.10	0.2132	0.2027	0.2069	0.2076	0.2128	−0.0052 [−0.0126, +0.0015]
0.20	0.2151	0.2062	0.2095	0.2103	0.2060	+0.0043 [−0.0059, +0.0160]
0.40	0.2052	0.1969	0.1908	0.1976	0.2061	−0.0085 [−0.0304, +0.0126]

The normalized recall AUCs were 0.2057 for spatial and 0.2077 for newline positions, a difference of −0.0020 (95% CI [−0.0106, +0.0068]). The observed means remained close to

Raw throughout the tested window, and no dose-specific recall interval excluded zero. Table 14 reports the auxiliary AUC endpoints.

Table 14: Low-relative-dose AUC endpoints. Recall is primary; the remaining endpoints are exploratory and use unadjusted 95% intervals. Lower NLL is better.

Endpoint	Spatial AUC	Newline AUC	SP–NL [95% CI]
Recall	0.2057	0.2077	-0.0020 $[-0.0106, +0.0068]$
Whole-word recall	0.2837	0.2884	-0.0047 $[-0.0136, +0.0042]$
NLL	4.2554	4.2715	-0.0161 $[-0.0182, -0.0140]$
Output words	33.63	33.03	$+0.60$ $[+0.04, +1.18]$
Nonempty fraction	1.0000	1.0000	0.0000 [0.0000,0.0000]

The maximum delivered relative-dose error was 1.08×10^{-4}; the maximum high-dose construction probe error was 4.77×10^{-7}. Mean delivered energy relative to N208 ranged from 0.000101 to 0.00648 in the newline arm and from 0.427 to 27.32 in the spatial arm across the four doses, reflecting the source norm difference.

Table 15 summarizes the condition means and paired contrasts across the absolute, high-relative, and low-relative scales shown in Fig. 2.

Table 15: Scale-conditional spatial–newline comparisons at full token retention. The first three rows report mean recall and Bonferroni 97.5% intervals from their respective two-comparison families. The low-dose row reports normalized recall AUC over $r \in [0, 0.4]$ and its prespecified 95% video-cluster interval. Positive differences favor spatial positions.

Scale definition	Regime	Spatial	Newline	Difference [CI]
Absolute displacement	full	0.2126	0.1181	$+0.0946$ $[+0.0514, +0.1395]$
Global relative amplitude	high	0.0909	0.1181	-0.0272 $[-0.0669, +0.0113]$
Local relative amplitude	high ($r \approx 4.97$)	0.0131	0.1181	-0.1050 $[-0.1431, -0.0697]$
Local relative amplitude	low AUC ($r \leq 0.4$)	0.2057	0.2077	-0.0020 $[-0.0106, +0.0068]$

B.7 Directional Packet-Concentration Probe

This probe reuses the Exp55–Exp57 cohort, checkpoints, masks, and condition definitions without training or fitting. One visual forward pass per video and checkpoint reconstructs Raw, S208, S208-R, S208-RG, S208-RL, and N208 in the same scope. Metrics are recorded on the delivered rows and after the frozen first decoder input RMSNorm. The extraction wrapper returns the Raw packet unchanged; a two-video smoke reproduced every aligned historical Raw generation exactly.

For the 208 selected rows H, define normalized directions $U_i = H_i / \|H_i\|_2$ and $G = UU^\top$. The primary metric is directional participation rank,

$$R_{\mathrm{PR}}(U) = \frac{\mathrm{tr}(G)^2}{\|G\|_F^2}. \tag{4}$$

A value near one indicates that the rows occupy one dominant direction. Supporting metrics are angular concentration $\|208^{-1} \sum_i U_i\|_2$, top-direction share $\lambda_{\max}(G)/\,\mathrm{tr}(G)$, mean pairwise cosine, and alignment with the fixed newline displacement direction. Row normalization makes these directional statistics independent of row magnitude.

The video is the independent unit ($n = 138$) because queries from the same video share one visual packet. Spatial metrics are averaged over three masks and three checkpoints within video. The single primary comparison is first-RMSNorm R_{PR}(S208-RL) minus R_{PR}(S208-R), evaluated with 100,000 paired percentile-bootstrap draws over videos (seed 20260922). Its estimate was -2.9580 (95% CI $[-3.0247, -2.8925]$), a 74.0% decrease from S208-R.

Table 16: Directional concentration after the first consumer RMSNorm. Means average three masks and three checkpoints within each of 138 videos. Historical recall is included to connect the probe to the behavioral ladder; it was not recomputed by this feature-only extraction.

Condition	R_{PR}	Angular conc.	Top-dir. share	Recall
Raw	3.9247	0.6855	0.4946	0.2047
S208	3.9971	0.6820	0.4893	0.2116
S208-R	3.9955	0.6821	0.4894	0.2126
S208-RG	1.0502	0.9876	0.9758	0.0909
S208-RL	1.0375	0.9908	0.9818	0.0131
Raw newline	1.0000	1.0000	1.0000	—
N208	1.0000	1.0000	1.0000	0.1181

The concentration was already present at the delivered interface. Packed-input participation rank was 4.0276 for S208-R, 1.0583 for S208-RG, and 1.0440 for S208-RL. The first RMSNorm retained, rather than created, this directional collapse. The pattern was stable across masks: first-RMSNorm S208-RL rank ranged from 1.0372 to 1.0377, whereas S208-R ranged from 3.9459 to 4.0255. S208-RG and S208-RL were both nearly rank one but differed in recall by 0.0778. They therefore share a concentrated high-dose packet state, while their behavioral difference requires information beyond this packet summary. Raw newline and N208 are rank one by construction, confirming that participation rank must be interpreted against the source path and token role.

C Native Compression, References, and Boundaries

C.1 Native Audit and Calibration

Native PruneVid uses the PLLaVA-7B aggregation path specified in the main text, with all settings except cluster ratio fixed across the complete three-point grid. The original condition means and historical 95% intervals are in Table 2. Source-reference packets have one original vector per native output group, in the native group order. The audited grouping implementation must reproduce native aggregates before constructing references.

For calibration video v, let $C_v, R_v \in \mathbb{R}^{K_v \times D}$ be native and representative packets. The fit forms $M = V^{-1} \sum_v K_v^{-1} C_v^\top R_v$ without centering. With $B = M/\|M\|_F + 10^{-6} I = U \Sigma W^\top$, the direction is $Q = UW^\top$ and

$$s = \max\left(0, \frac{\sum_v K_v^{-1} \langle C_v Q, R_v \rangle_F}{\sum_v K_v^{-1} \|C_v Q\|_F^2}\right),$$

requiring finite parameters and $s > 0$. Medium's map uses 96 external videos and is reused at Light and Strong. The independent MVBench fit uses the same routine with 90 new videos. No feature means, questions, or answer labels enter these calibration fits.

Preservation summaries compare native and representative packets using centered linear CKA, Frobenius norm ratio $\|C\|_F/\|R\|_F$, relative RMS error $\|C - R\|_F/\|R\|_F$, and mean rowwise cosine. They are averaged over 138 videos. This exploratory characterization followed inspection of the three-point behavioral pattern. All strengths and all four summaries are retained. The NLL gap is Compressed minus Representative, so positive values indicate worse compressed likelihood. Its Medium-minus-Strong contrast is -0.0001 $[-0.0200, +0.0208]$; the recall ordering is not established on this endpoint. At Medium, Compressed minus Repaired NLL is $+0.0008$ $[-0.0030, +0.0048]$.

Table 17: Compressed-K-to-Representative-K preservation summaries in the native PruneVid grid. Values are video means. The final row gives paired video-bootstrap differences with 95% intervals.

Point	Centered CKA	Norm ratio	Relative RMS	Mean cosine
Light	0.9251	0.9682	0.2036	0.9946
Medium	0.9198	0.9618	0.2309	0.9874
Strong	0.9192	0.9530	0.2581	0.9715
Medium−Strong	+0.0006	+0.0088	−0.0272	+0.0159
95% CI	[−0.0047, +0.0054]	[+0.0050, +0.0126]	[−0.0331, −0.0211]	[+0.0152, +0.0165]

Table 18: Auxiliary answer-token NLL in the native PruneVid grid. Lower is better. Full-N NLL is 4.6896 at every point.

Point	Representative	Compressed	Repaired	Comp.−Rep.
Light	4.7542	4.7614	4.7578	+0.0072
Medium	4.7211	4.7354	4.7346	+0.0143
Strong	4.7092	4.7236	4.7209	+0.0144

C.2 Scoring-Reference Granularity

The original endpoint scores video-level descriptions; query-level action descriptions differ on 94 of 150 queries. Both use the same whole-word scorer, including repeated reference words, with fixed outputs and alignment maps. The follow-up rescores 1,800 cached condition rows under both references, giving 3,600 scores. Define G = Representative − Compressed and A = Repaired − Compressed recall. Positive G is a gap; positive A is an alignment gain. All 27 contrasts appear below.

This post-hoc sensitivity analysis followed inspection of historical results. Before computing the follow-up, three focal comparisons were fixed: video-level $G_M - G_S$, query-level $G_M - G_S$, and their query-minus-video difference. The 100,000 shared draws resample 138 videos (seed 20260917), retain all sampled queries with multiplicity, and use linear-interpolation percentiles. The three 98.333% intervals address the stated follow-up family. They supplement the original video-level 95% analysis. Different draw counts and seeds yield slightly different percentile endpoints for the same point estimate; other intervals are exploratory 95%. All historical video-level scores reproduce exactly, and query expansion agrees with an independent cluster-multiplicity computation.

All three adjusted intervals include zero. The direct reference difference in ordering belongs to this three-comparison family; the Medium alignment-gain difference has an exploratory 95% interval. Displayed contrasts are calculated from unrounded scores; subtracting displayed means may differ in the last reported digit.

Table 19: Post-hoc three-strength rescoring of fixed PruneVid outputs. G is Representative-K minus Compressed-K recall; A is Repaired-K minus Compressed-K recall. Entries are query-weighted mean differences with exploratory individual 95% video-cluster intervals (150 queries, 138 videos; 100,000 shared draws, seed 20260917).

Reference	Strength	G [95% CI]	A [95% CI]
Video-level	Light	+0.0033 [−0.0107, +0.0185]	−0.0003 [−0.0142, +0.0134]
Video-level	Medium	+0.0204 [+0.0038, +0.0381]	+0.0237 [+0.0068, +0.0440]
Video-level	Strong	−0.0075 [−0.0254, +0.0106]	−0.0020 [−0.0129, +0.0100]
Query-level	Light	−0.0061 [−0.0300, +0.0151]	−0.0086 [−0.0279, +0.0096]
Query-level	Medium	+0.0002 [−0.0265, +0.0270]	+0.0124 [−0.0078, +0.0330]
Query-level	Strong	−0.0049 [−0.0284, +0.0187]	+0.0022 [−0.0108, +0.0162]

Table 20: Paired strength and scoring-reference contrasts from the same post-hoc analysis. $D_{MS} = G_M - G_S$, $D_{LM} = G_L - G_M$, and $D_{LS} = G_L - G_S$; Δ denotes query-level minus video-level scoring. The three D_{MS} rows form this follow-up's fixed comparison family. Their 98.333% intervals target nominal familywise 95% coverage by Bonferroni adjustment; all other 95% intervals are exploratory. A dash denotes a contrast outside this family.

Contrast	Estimate [95% CI]	98.333% CI
Video-level D_{MS}	$+0.0279$ $[+0.0035, +0.0536]$	$[-0.0020, +0.0596]$
Video-level D_{LM}	-0.0171 $[-0.0398, +0.0046]$	–
Video-level D_{LS}	$+0.0108$ $[-0.0129, +0.0358]$	–
Query-level D_{MS}	$+0.0051$ $[-0.0320, +0.0417]$	$[-0.0407, +0.0498]$
Query-level D_{LM}	-0.0063 $[-0.0436, +0.0270]$	–
Query-level D_{LS}	-0.0012 $[-0.0378, +0.0316]$	–
ΔG_L	-0.0094 $[-0.0265, +0.0056]$	–
ΔA_L	-0.0083 $[-0.0218, +0.0045]$	–
ΔG_M	-0.0202 $[-0.0451, +0.0033]$	–
ΔA_M	-0.0112 $[-0.0316, +0.0072]$	–
ΔG_S	$+0.0026$ $[-0.0159, +0.0218]$	–
ΔA_S	$+0.0042$ $[-0.0065, +0.0157]$	–
ΔD_{MS}	-0.0228 $[-0.0580, +0.0102]$	$[-0.0666, +0.0172]$
ΔD_{LM}	$+0.0108$ $[-0.0182, +0.0396]$	–
ΔD_{LS}	-0.0121 $[-0.0404, +0.0141]$	–

C.3 Feature Source-Reference Sensitivity

This protocol-fixed follow-up substitutes two source selectors within the same native groups at every strength, retaining aggregates, map, token counts, positions, masks, prompts, 16-frame sampling, greedy decoding, and the 128-token limit. Nearest minimizes squared Euclidean distance to the aggregate; First selects the smallest source index; Medoid minimizes the sum of unsquared Euclidean distances to unique source members without aggregation weights. Distances use float64 and ties select the smallest source index.

The six primary contrasts are alternative minus Nearest recall, two selectors at each of three strengths. They use video-level whole-word recall and 50,000 shared video-cluster draws (seed 20260915), with 95% and family-adjusted 99.1667% intervals. All six include zero at both levels. Tables below retain all six primary contrasts, all reference-relative recall and NLL gaps, and selector agreement. The latter two are descriptive/exploratory. Medoid changes only 1.04–2.94% of selected tokens; First changes 10.80–42.15%.

The 45 historical reproduction records match predictions, scores, source IDs, token counts, and NLL. Analysis uses Python 3.12.9 and NumPy 2.2.3.

Table 21: Six primary source-reference recall contrasts. Each is alternative minus Nearest; 99.1667% intervals adjust the six-comparison family.

Point	Alternative	Difference	95% CI	99.1667% CI
Light	First	-0.0012	$[-0.0173, +0.0140]$	$[-0.0231, +0.0191]$
Light	Medoid	-0.0048	$[-0.0197, +0.0089]$	$[-0.0254, +0.0134]$
Medium	First	-0.0026	$[-0.0215, +0.0157]$	$[-0.0282, +0.0221]$
Medium	Medoid	$+0.0095$	$[-0.0071, +0.0268]$	$[-0.0126, +0.0331]$
Strong	First	-0.0050	$[-0.0226, +0.0117]$	$[-0.0289, +0.0175]$
Strong	Medoid	$+0.0002$	$[-0.0150, +0.0146]$	$[-0.0207, +0.0197]$

Table 22: Reference minus native compression ($R - C$), with descriptive, unadjusted 95% intervals. Positive recall favors the reference; negative NLL favors the reference.

Point	Reference	Recall difference	NLL difference
Light	Nearest	$+0.0033$ $[-0.0107, +0.0185]$	-0.0072 $[-0.0153, +0.0013]$
Light	First	$+0.0021$ $[-0.0175, +0.0199]$	-0.0121 $[-0.0266, +0.0001]$
Light	Medoid	-0.0015 $[-0.0207, +0.0152]$	-0.0084 $[-0.0170, +0.0006]$
Medium	Nearest	$+0.0204$ $[+0.0041, +0.0380]$	-0.0143 $[-0.0277, -0.0015]$
Medium	First	$+0.0178$ $[+0.0036, +0.0331]$	-0.0060 $[-0.0242, +0.0122]$
Medium	Medoid	$+0.0299$ $[+0.0107, +0.0508]$	-0.0143 $[-0.0276, -0.0018]$
Strong	Nearest	-0.0075 $[-0.0254, +0.0108]$	-0.0144 $[-0.0344, +0.0055]$
Strong	First	-0.0125 $[-0.0313, +0.0058]$	-0.0050 $[-0.0285, +0.0178]$
Strong	Medoid	-0.0073 $[-0.0254, +0.0107]$	-0.0153 $[-0.0356, +0.0051]$

Table 23: Selected-token agreement, counting each of 138 videos once. Pooled agreement weights tokens; video mean weights videos equally. The last column counts videos with identical selections.

Point	Selectors	Equal / total tokens	Pooled (%)	Video mean (%)	Identical videos
Light	First / Nearest	150278 / 168467	89.20	86.61	0 / 138
Light	Medoid / Nearest	166716 / 168467	98.96	98.51	2 / 138
Light	First / Medoid	150235 / 168467	89.18	86.57	0 / 138
Medium	First / Nearest	88643 / 112581	78.74	76.48	0 / 138
Medium	Medoid / Nearest	110873 / 112581	98.48	97.97	0 / 138
Medium	First / Medoid	88623 / 112581	78.72	76.45	0 / 138
Strong	First / Nearest	32482 / 56149	57.85	56.08	0 / 138
Strong	Medoid / Nearest	54499 / 56149	97.06	96.44	0 / 138
Strong	First / Medoid	32459 / 56149	57.81	56.03	0 / 138

C.4 Task and Method Boundary Protocols

The transferred PruneVid map is evaluated on 1,500 MVBench questions, 100 in each of 15 tasks, with 16 frames and the official `check_ans` parser. All outputs have valid format. The historical intervals resample 1,422 relative-path video IDs. Canonicalizing shared Charades aliases gives 1,416 source videos; a subsequent source-cluster check leaves the primary interval unchanged at displayed precision. The historical interval is retained. Repair changes five incorrect answers to correct and twelve correct answers to incorrect; 677 are correct in both, and 806 incorrect in both.

The Charades ToMe and STTM cases each use frozen 4-bit LLaVA-Video-7B-Qwen2, SDPA, 150 queries from 138 videos, their own external 96-video calibration, and historical substring recall. ToMe Bolya et al. (2023) performs one balanced merging pass over 169 spatial tokens per frame with $r = 84$. Thirteen newline vectors retain their values and are excluded from matching and alignment, yielding 85 content outputs plus 13 newline vectors per frame ($K = 1{,}568$, 53.8% retention). Newline values are preserved, but their original Full-N grid positions are not: the 13 follow the 85 content outputs. Representative-K uses original destination-anchor values. This adaptation was fixed before canonical outcomes.

STTM Hyun et al. (2025) uses revision `336bf360118b1cde3caa8e0a826f05e15ab91a83`, decoder layer 2, root level 1, spatial threshold 0.85, temporal threshold 0.65, unweighted aggregation, the fast implementation, and no per-head similarity. Newline insertion is disabled in all conditions. Full-L2 has 2,704 tokens; reduced packets average 490.87 (18.15%, range 108–1,251). Representative-K uses the minimum source index per group. Native partition parity, position IDs, and masks are checked. The reported results quantify the fitted-map responses under these stated configurations. The structural-token adaptations were not separately ablated.

Both cases use 10,000 paired video-cluster draws (seed 20260906). For ToMe, native minus Representative recall is $+0.01856$ $[-0.00101, +0.03989]$, and its NLL difference is -0.00799 $[-0.01473, -0.00104]$. Repair changes NLL by $+0.00104$ $[-0.00135, +0.00326]$. STTM native minus Representative recall is -0.00200 $[-0.02887, +0.02471]$; neither its negative-gap criterion nor its positive-repair criterion is met.

Table 24: PruneVid MVBench results, distinct from the Qwen orthogonal-control MVBench experiment. Each condition contains 1,500 questions.

Condition	Correct	Accuracy	Malformed	Mean tokens
Full-N	706	0.4707	0	2304.0
Representative-K	696	0.4640	0	602.6
Compressed-K	689	0.4593	0	602.6
Repaired-K	682	0.4547	0	602.6

Table 25: Additional aggregation cases. Means are computed from the original query rows. Repaired-minus-native intervals come from the paired bootstrap.

Case	Full	Representative	Native	Repaired
ToMe	0.2047	0.1451	0.1637	0.1642
STTM	0.1843	0.1412	0.1392	0.1353
ToMe repair contrast: $+0.00056$ $[-0.00889, +0.01064]$				
STTM repair contrast: -0.00389 $[-0.01351, +0.00338]$				

C.5 Independent MVBench Calibration

The follow-up fixes its two primary comparisons before new evaluation, after the earlier transfer result. Seed 20260915 selects six calibration videos and 50 test questions per task: 90 unique calibration videos and 750 questions from 723 canonical test sources. Both splits exclude earlier PruneVid calibration/test sources and the 1,500-question MVBench evaluation; shared aliases are canonicalized first.

The native configuration is the 16-frame Medium setting in Section 3. TaskMatched-K fits one label-free map to the 90 videos pooled across tasks; Transferred-K uses the old 96-video Charades map. Reduced conditions share groups, counts, positions, masks, prompts, and greedy decoding (100-token limit). The fit sets differ in size and domain; task-specific fitting is untested.

Official-parser accuracy is question-weighted. The 20,000 shared cluster draws give 95% and 97.5% intervals for both primary contrasts; both include zero. Native minus Representative accuracy is $+0.0053$ $[-0.0027, +0.0134]$, so this cohort provides no clear aggregation deficit to repair.

Table 26: Independent MVBench follow-up: five paired conditions on 750 questions. Accuracy is higher-is-better. All outputs have valid answer format.

Condition	Correct answers (of 750)	Accuracy	Mean visual tokens
Full-N	354	0.4720	2304.0
Representative-K	355	0.4733	597.9
Compressed-K	359	0.4787	597.9
Transferred-K	355	0.4733	597.9
TaskMatched-K	357	0.4760	597.9

Table 27: Two primary paired accuracy contrasts. M denotes TaskMatched-K, C Compressed-K, and T Transferred-K. "Improved / harmed" counts questions changed from incorrect to correct / correct to incorrect.

Contrast	Difference	95% CI	97.5% CI	Improved / harmed
M − C	-0.0027	$[-0.0080, +0.0027]$	$[-0.0093, +0.0027]$	1 / 3
M − T	$+0.0027$	$[-0.0027, +0.0080]$	$[-0.0027, +0.0093]$	3 / 1

C.6 Concrete Output Examples

These cached illustrations use the first qualifying query in canonical order. For checkpoint 3401, the newline criteria are: (i) whole-word coverage drops by at least 0.5 with lower substring recall; (ii) output grows by at least ten words with both coverage scores lower. The MVBench criterion is a correct-to-incorrect alignment change. Outputs are quoted verbatim.

A partial answer after newline replacement. Charades video `ZF7YA` uses the question "Describe what the person does in the video." and query annotation "a person runs down a few stairs." **Raw (complete):** "The person in the video is seen descending a carpeted staircase, wearing a gray t-shirt and dark shorts. They stop midway down the stairs to pick up a black object from the floor, which appears to be a piece of clothing or a bag. The person then sits on the floor near the staircase, adjusting or examining the black object."

Spatial-only (opening sentence): "The person in the video is seen descending a carpeted staircase, wearing a gray t-shirt and dark shorts."

Spatial-matched (opening sentence): "The person in the video is seen descending a carpeted staircase, wearing a gray t-shirt and dark shorts."

Newline-only / Newline-matched (complete): "The person-"

The scored whole-word reference set is {runs, down, few, stairs}; "person" is a stopword. Raw matches "down" and "stairs", giving whole-word recall 0.50; Newline-only matches none, giving 0.00. Under the historical substring rule, punctuation is retained, so the final reference token is "stairs." and only "down" matches Raw; substring recall therefore decreases from 0.25 to 0.00. The newline output is nonempty but contains no action description.

A longer answer with lower reference coverage. Charades video `P8G8V` uses the same description prompt and the scoring reference "person they casually drink a cup of coffee." **Raw (complete):** "The person is standing in the kitchen, holding a smartphone in one hand and a cup in the other. They appear to be engaged with their phone while occasionally taking sips from the cup."

Newline-only (complete): "In the video, a woman is standing in a kitchen. She is wearing a gray tank top and dark gray pants. The woman is holding a phone in her right hand and a cup in her left hand. She appears to be looking at the phone while holding the cup. The kitchen has white cabinets and a black countertop. There is a microwave on the counter next to the woman. There are also some items on the counter, including a red object and a yellow object. The floor is wooden with a patterned rug."

Output length increases from 34 to 94 whitespace-delimited words. Both substring and whole-word recall decrease from 0.40 to 0.20. The matched reference words are "they" and "cup" for Raw, and "cup" for Newline-only. The longer answer adds scene details but omits the drinking action described in Raw.

Alignment changes a correct count. MVBench query `Action Count:00005` asks: "How many times did the person launch the object on the slanted plane?" The options are (A) 5, (B) 3, and (C) 4; the dataset answer is (B) 3. **Compressed-K (complete):** "(B) 3"

Repaired-K (complete): "(C) 4"

This is one of the twelve correct-to-incorrect changes in the transferred-map evaluation.

C.7 Native Directional-Concentration Audit

We apply the directional measures of Appendix B.7 to native PruneVid packets, pairing Representative-K, Compressed-K, and Repaired-K within each realized grouping at Light, Medium, and Strong. All K delivered rows enter the measurement. Mean token counts are

1,220.78, 815.80, and 406.88, respectively; conditions share groups, token counts, ordering, and layout within each strength. The 96-video Medium calibration map is reused unchanged. No map is refitted and no new outputs are generated.

We measure the delivered packet and its output from the first frozen decoder RMSNorm, evaluating row-normalized Gram statistics in float64 from the actual runtime tensors. The resulting 2,484 records cover 138 videos, three strengths, three conditions, and two stages. Videos, not records or token rows, are the independent sampling units. This follow-up was designed after the historical behavioral results. Its three primary contrasts were fixed before metric extraction: delivery-stage Compressed-minus-Representative directional rank, one at each strength. We use 100,000 shared paired video-bootstrap draws (seed 20260922), reporting individual 95% and Bonferroni 98.333% intervals. Repair, first-RMSNorm, and other geometric summaries are exploratory. Geometry means weight videos equally; the historical recall means retain their original query weighting and video-level whole-word scoring references.

Concentration and behavior. All three native aggregation contrasts reduce directional rank, with the largest reduction at Strong (Tables 3 and 29). Angular concentration increases at each strength as well (Table 28). The largest recall gap, however, occurs at Medium, where alignment improves recall by +0.0237 without restoring delivery-stage directional diversity. These are within-strength source-referenced comparisons: token counts and source groups differ across strengths.

Why alignment preserves delivery-stage rank. Let U contain the normalized rows of the compressed packet C. The repaired packet is sCQ, where $Q^\top Q = I$ and $s > 0$ is a shared scalar. Its normalized rows are UQ, and

$$(UQ)(UQ)^\top = UU^\top.$$

The full row-normalized Gram matrix, and hence directional participation rank and angular concentration, is unchanged in exact arithmetic. Before the runtime cast, the largest observed absolute rank difference was below 3×10^{-8}. Small delivery-stage differences in Table 29 arise after the bfloat16 cast. The learned coordinate weights of RMSNorm need not preserve this invariance; its repair contrasts are small but nonzero. Thus, neither a reduction in native rank nor its response to this alignment map supplies a behavioral diagnosis on its own.

Table 28: Native packet concentration means over 138 videos. R, C, and A denote Representative-K, Compressed-K, and Repaired-K. Rank and angular concentration use all K rows; a larger angular concentration indicates more similar directions.

		Directional rank		Angular concentration	
Point	Condition	Delivery	RMSNorm	Delivery	RMSNorm
Light	R	3.2830	1.2753	0.7274	0.9391
	C	3.1988	1.2687	0.7322	0.9403
	A	3.1988	1.2691	0.7322	0.9402
Medium	R	3.4769	1.2809	0.7158	0.9380
	C	3.2672	1.2660	0.7271	0.9408
	A	3.2672	1.2664	0.7271	0.9407
Strong	R	3.7802	1.3226	0.6957	0.9298
	C	3.2466	1.2859	0.7230	0.9365
	A	3.2466	1.2862	0.7230	0.9364

Table 29: Paired native rank changes. Delivery C−R is the primary three-comparison family; its 98.333% intervals appear in Table 3. Intervals below are individual 95% intervals. A−C entries are multiplied by 10^4 to display small delivery-stage and first-RMSNorm changes.

Point	Contrast	Delivery [95% CI]	RMSNorm [95% CI]
Light	C−R	-0.0842 $[-0.0907, -0.0779]$	-0.0066 $[-0.0072, -0.0059]$
	$(\mathrm{A}-\mathrm{C}) \times 10^4$	$+0.1648$ $[+0.1326, +0.1976]$	$+3.9717$ $[+3.6958, +4.2536]$
Medium	C−R	-0.2097 $[-0.2187, -0.2010]$	-0.0149 $[-0.0156, -0.0141]$
	$(\mathrm{A}-\mathrm{C}) \times 10^4$	$+0.0948$ $[+0.0562, +0.1336]$	$+3.4072$ $[+3.1146, +3.7085]$
Strong	C−R	-0.5335 $[-0.5521, -0.5152]$	-0.0367 $[-0.0380, -0.0354]$
	$(\mathrm{A}-\mathrm{C}) \times 10^4$	$+0.0267$ $[-0.0463, +0.1006]$	$+3.3495$ $[+2.9776, +3.7375]$