Encoded but Not Reported: Label-Free Consistency Audits of Longitudinal Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models are asked how a disease changes between scans, but benchmarks score each question alone. We audit them without labels using a property of true change: change over an interval is the sum of changes over its parts. Answers obeying this *cocycle* rule form the gradient subspace of combinatorial Hodge theory, so a model's distance to it lower-bounds its error. Clipped at a predeclared threshold, this bound gives a distribution-free cohort certificate. We also probe frozen representations and repair answers by projection onto this subspace. Experiments on brain MRI and satellite imagery show that most open-weight models fail to beat a no-change predictor. On a glioma cohort with large changes, representations across model families encode change that answers do not report, and label-free probes recover most of it. Fine-tuning on projected answers corrects their direction but not their magnitude. Code, results and model outputs are available at `https://anonymous.4open.science/r/VLM-3j-ICLR27-725E/`.

1 Introduction

Vision-language models are increasingly asked to read serial imaging, where the quantity a radiologist reports from a follow-up scan is a *change* attached to an interval rather than a label attached to an image. Recent longitudinal benchmarks pose exactly such questions, for example how much a tumour grew between two visits (Al Ghallabi et al., 2026; Zhang et al., 2026). They score each question on its own, against annotations that are costly to collect, and summarise performance with composites that do not decompose: the headline composite of the most recent longitudinal MRI benchmark is nearly one of its own components (Table 27). None of them reports an image-ablated baseline, a chance floor, or a control in which the same scan is supplied twice.

Scoring questions one at a time ignores a structure that every estimate of change must have. Change composes across intervals, reverses when the interval is reversed, and vanishes when the two scans are the same: the change from visit 1 to visit 3 is the change from visit 1 to visit 2 plus the change from visit 2 to visit 3. We call an estimator that respects these three identities Δ-*consistent*, and the composition identity the *cocycle*. Measured ground truth satisfies them exactly, so a model whose answers violate them is wrong somewhere, and this can be detected without a single label. The same structure raises a second question that accuracy alone cannot answer. When a model's answers are poor, the change may be absent from what the model computes, or it may be present in its representation and lost only when the answer is produced. The two readings call for different remedies: the first for new architectures, the second for a better readout.

Recent work makes the failure concrete, but does not yet resolve it. The benchmark we build on (Al Ghallabi et al., 2026) reports reliable ordering of scans together with unreliable direction of change, and proposes architectural remedies. Methods that enforce agreement between related answers exist for language models (Mitchell et al., 2022; Kassner et al., 2021; Elazar et al., 2021), and the algebra of consistent pairwise comparisons is classical: an estimator that composes is the gradient part of the combinatorial Hodge decomposition (Jiang et al., 2011). These lines target different objects, namely learned relations between discrete answers and rankings from pairwise preferences. We ask a different question. Specifically, what does the composition identity of true change reveal

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.

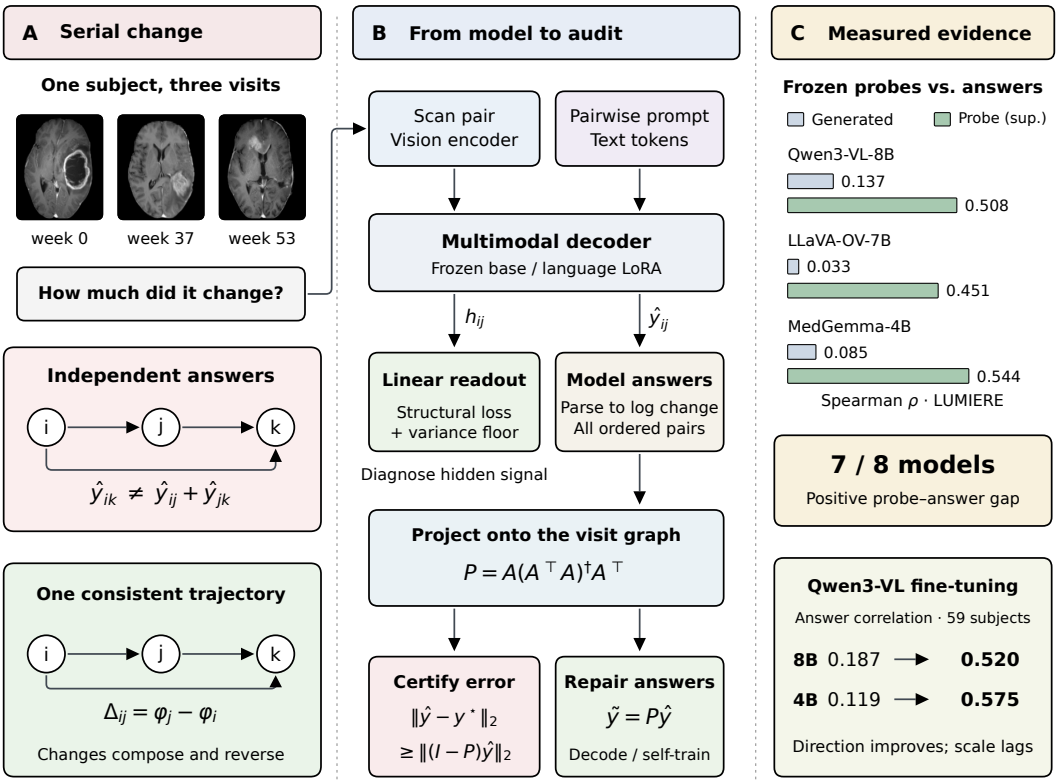

Figure 1: **From independent answers to a consistent trajectory.** **(A)** Real serial scans from one LUMIERE subject illustrate the setting. Independently generated log-change estimates can violate composition; consistent changes are differences of a scalar potential over visits. **(B)** The VLM exposes hidden states for linear probing and answers for projection onto the visit graph. The orthogonal residual lower-bounds error without labels; projected answers support decoding and self-training. **(C)** On matched held-out items, supervised linear probes outperform generated answers (three representative models shown); the gap's subject-level interval excludes zero for 7 of 8 models. Training on projected samples improves generated-answer correlation on 59 held-out subjects for Qwen3-VL-8B and 4B. Magnitude gains remain modest. Full results and uncertainty: Tables 22 and 5.

about a longitudinal model without labels, and is the information its answers lack absent from the model or present but unused?

Motivated by these questions, we propose a label-free consistency audit for longitudinal vision-language models (Figure 1). The audit asks a model about every ordered pair of a subject's visits and stacks the answers into one vector. Because true change lies in the linear subspace of consistent estimators, the distance of this vector to the subspace lower-bounds the model's error, and a threshold fixed in advance turns the distance into a distribution-free certificate over a cohort. A readout diagnosis then compares linear probes on the model's frozen representations with the model's own answers on the same items, including probes trained with no labels at all. Finally, a repair projects sampled answers onto the consistent subspace, at test time or as targets for fine-tuning.

We validate the audit both theoretically and empirically. The bound holds for any prediction and any consistent ground truth, and the certificate holds for any distribution of the residual. Experiments with 9 open-weight models on longitudinal brain MRI and satellite imagery show that most models do not beat a predictor that always answers no change, and that on a glioma cohort with large changes the representations of several model families encode change that the answers do not report. Fine-tuning on projected answers repairs the direction and coherence of the answers but not their magnitude.

Key contributions.

- We formulate longitudinal change estimation as a Δ-consistency problem and show that consistent answers are exactly the differences of one potential per visit, so a violation certifies that the answers are differences of no single quantity (Section 2).
- We prove that the distance to the consistent subspace is a label-free lower bound on error and, clipped at a pre-declared threshold, a distribution-free cohort certificate (Section 3.1).
- We show on identical held-out items that linear readouts of frozen representations recover measured change that the generated answers do not report, that input controls place this signal in the image pair, and that heads trained only to satisfy the constraint recover most of it (Section 4.2).
- We propose a repair by projection, at test time and in training, and show with matched controls that it corrects direction and coherence, mostly through antisymmetry, while magnitude remains largely uncorrected (Section 4.3).

Code, results and every model generation are released at `https://anonymous.4open.science/r/VLM-3j-ICLR27-725E/`.

2 Problem formulation

We first define the object of evaluation. Fix a subject observed at ordered timepoints $t_1 < \cdots < t_n$ with imaging I_i at t_i. A *change estimator* $\widehat{\Delta} : \{1, \ldots, n\}^2 \to \mathbb{R}$ returns an estimate $\widehat{\Delta}(i \to j)$ of how much a target quantity changed between visits t_i and t_j. We take a *relative* change in a positive quantity V (tumour or brain volume, building footprint area) in the logarithmic domain,

$$\Delta^\star(i \to j) \;=\; \log \frac{V(t_j) + \varepsilon}{V(t_i) + \varepsilon}, \tag{1}$$

with a fixed stabiliser $\varepsilon > 0$ that keeps complete responses finite; in the log domain relative change composes additively, so the constraints below are exact for every $\varepsilon > 0$.

Definition 1 (Δ-consistency). *A change estimator $\widehat{\Delta}$ is* Δ-consistent *if for all i, j, k:*

$$\textit{(identity)} \quad \widehat{\Delta}(i \to i) = 0, \tag{2}$$

$$\textit{(antisymmetry)} \quad \widehat{\Delta}(i \to j) = -\,\widehat{\Delta}(j \to i), \tag{3}$$

$$\textit{(cocycle)} \quad \widehat{\Delta}(i \to k) = \widehat{\Delta}(i \to j) + \widehat{\Delta}(j \to k). \tag{4}$$

Ground truth satisfies Definition 1 identically, since $\log \frac{V_k+\varepsilon}{V_i+\varepsilon} = \log \frac{V_j+\varepsilon}{V_i+\varepsilon} + \log \frac{V_k+\varepsilon}{V_j+\varepsilon}$ for any V_i, V_j, V_k. The constraints are therefore properties of change itself, not modelling assumptions about disease.

Proposition 2 (Consistency is equivalent to being a potential difference). *$\widehat{\Delta}$ satisfies the cocycle condition equation 4 for all i, j, k if and only if there exists $\varphi : \{1, \ldots, n\} \to \mathbb{R}$ with $\widehat{\Delta}(i \to j) = \varphi(j) - \varphi(i)$ for all i, j. Such a φ is unique up to an additive constant, and equation 4 implies equation 2 and equation 3.*

Proposition 2 is a statement about the reported outputs, not about the model that produced them. A consistent set of answers is representable as differences of one scalar potential, and a set that violates the cocycle is not, by any φ. It does not follow that a consistent model holds a disease state internally, nor that an inconsistent one lacks it: a model could represent burden accurately and emit inconsistent answers through a noisy readout, and $\widehat{\Delta}(i \to j) = t_j - t_i$ is consistent and estimates nothing about disease. Consistency is thus necessary for an estimator that reads a trajectory and not sufficient. Our goal is to audit a model's answers against this structure without labels, to locate where the change information is lost, and to repair the answers. Section 3 defines the audit that does so.

3 Label-free consistency audit

The audit has three parts, which mirror the three questions above. The first measures how far a model's answers lie from the consistent subspace and turns that distance into a lower bound and a cohort certificate. The second asks whether the model's representation carries what its answers lack. The third projects answers onto the subspace to repair them. Proofs are in Appendix C.

3.1 Label-free bound and certificate

Write a subject's predictions as a vector $\widehat{y} \in \mathbb{R}^E$ indexed by the observed ordered pairs, and let $A \in \mathbb{R}^{E\times n}$ be the signed incidence matrix of the timepoint graph, $(A\varphi)_{(i,j)} = \varphi_j - \varphi_i$. By Proposition 2 the consistent estimators are exactly range(A), which contains ground truth; let P be the orthogonal projector onto it.

Proposition 3 (Label-free lower bound)**.** *For any prediction $\widehat{y}$ and any consistent ground truth $y^\star \in$* range(A)*,*

$$\|\widehat{y} - y^\star\|_2 \;\geq\; \operatorname{dist}\big(\widehat{y}, \operatorname{range}(A)\big) \;=\; \big\|(I-P)\,\widehat{y}\big\|_2, \tag{5}$$

with equality if and only if the error $\widehat{y} - y^\star$ is orthogonal to range(A)*. The right-hand side is computable from model outputs alone.*

Splitting edge data into a gradient part in range(A) and a cyclic residual is the combinatorial Hodge decomposition, which Jiang et al. (2011) use to rank from pairwise comparisons and to measure their inconsistency (see also Hirani et al., 2010). Proposition 3 is its distance-to-a-subspace reading. What changes is what it audits: ground truth lies in the gradient subspace by construction, which turns the residual into a lower bound on error. Because the raw residual grows with the scale of the predictions, we compare models with the scale-free angle $\sin\theta = \|(I-P)\widehat{y}\|_2/\|\widehat{y}\|_2 \in [0,1]$.

Corollary 4 (Certification on unlabelled cohorts)**.** *Let $b_s = \|(I-P_s)\widehat{y}_s\|_2/\sqrt{m_s}$ be the per-subject bound on per-edge RMS error e_s, and fix a threshold $B > 0$ before any data are examined. For subjects drawn i.i.d. from a cohort,* $\min(b_s, B) \in [0, B]$ *and* $\mathbb{E}[\min(b, B)] \leq \mathbb{E}[b] \leq \mathbb{E}[e]$*, so any $(1-\delta)$ lower confidence bound on $\mathbb{E}[\min(b, B)]$ obtained from an inequality for bounded i.i.d. variables is a $(1-\delta)$ lower confidence bound on the model's expected error, whatever the distribution of b.*

Corollary 4 gives the practical payoff. A site holding unannotated serial imaging can certify a floor on a candidate model's expected error on its own distribution before commissioning annotation. We instantiate it with the Chernoff–Hoeffding bound in KL form (Hoeffding, 1963) and the empirical-Bernstein bound of Maurer & Pontil (2009) at $B = 1$ log-ratio unit, with a percentile bootstrap on the unclipped residual as an asymptotic reference (Appendix C.7). The threshold must be fixed independently of the sample, since a threshold taken from the data is not covered by the argument.

Two things lie outside the reach of the algebra. First, orientation.

Proposition 5 (Orientation is not identifiable)**.** *If $\widehat{y}$ satisfies the cocycle then so does $c\,\widehat{y}$ for every $c \in \mathbb{R}$, including $c < 0$, since* range(A) *is a linear subspace.*

A predictor trained only to be consistent is therefore determined at most up to a scalar, sign included, and fixing which end means growth costs one bit from outside the algebra; every label-free number below is scored sign-invariantly. Second, the measured pair. Post-chemoradiation enhancement may be pseudoprogression, which is why RANO requires later confirmation (Wen et al., 2010). A predictor reading only the measured pairwise log-ratio, discretised at 0.02, has a plug-in Bayes ceiling of 61.0% on 2001 LUMIERE pairs against a 49.4% majority baseline (Appendix C.6). This is a finite-sample diagnostic of that quantity, not an impossibility result for predictors of the raw scan pair.

3.2 Readout diagnosis

To ask whether the change is present in the model, we fit a linear head $p = w^\top h + b$ on the hidden state h at the final prompt token, with three-way splits over *subjects* and early stopping on validation (Appendix D), and we score the model's own generated answer on exactly the test items of every split. The supervised head regresses measured change. The label-free cocycle head uses no annotation and minimises

$$\mathcal{L}(w,b) = \operatorname*{mean}_{i<j<k}\big(p_{ik} - p_{ij} - p_{jk}\big)^2 + \lambda\,\operatorname{relu}\big(1 - \operatorname{Var}_{\text{train}}\,p\big), \qquad \lambda = 5, \tag{6}$$

over the forward triangles of the training subjects, where the second term excludes the perfectly consistent zero map. Replacing the first term by the mean of $(p_{ij} + p_{ji})^2$ over reverse pairs gives an antisymmetry head, and keeping both gives full Δ-consistency; by Proposition 5 these heads

are scored sign-invariantly. To locate the signal, the representation is re-extracted with the images removed, shuffled across subjects, or reduced to one scan supplied twice, and a tied head $w^\top(h_j - h_i)$ is fitted on one canonical representation per visit, which is a potential difference by construction. What a linear head recovers is accessible predictive information; the probes do not locate where it is lost.

3.3 Repair by projection

At test time we sample K predictions per ordered pair, across decoding samples and anatomical views, and treat each as a noisy observation of $\varphi(j) - \varphi(i)$. The weighted least-squares projection

$$\widehat{\varphi} \;=\; \arg\min_{\varphi\,:\,\varphi_1=0} \sum_{e=(i,j)} w_e\big(\varphi_j - \varphi_i - y_e\big)^2, \qquad \widehat{\Delta}_{\mathrm{TCR}}(i \to j) = \widehat{\varphi}_j - \widehat{\varphi}_i, \tag{7}$$

returns the nearest consistent estimator to the sampled evidence. It satisfies equation 2–equation 4 exactly, needs neither training nor model internals, and replaces the median of self-consistency with a structured aggregator (Appendix H). At training time the projected value, written back in the model's own answer format, is a target that needs no annotation; only the language pathway is adapted, with the vision tower frozen (Appendix M). The projection is orthogonal, so it preserves the model's own orientation.

4 Experiments

Longitudinal benchmarks score each question alone, so they cannot tell whether a model's answers are coherent, whether its errors are detectable without labels, or whether the model knows more than it reports. Our experiments test whether the audit answers these questions on real models, organised around three research questions (RQs):

(**RQ1**) Do vision-language models report longitudinal change reliably, and does the label-free certificate detect their failures?

(**RQ2**) Is the change information absent from the models, or present in their representations but unused by their answers?

(**RQ3**) Does projection onto the consistent subspace repair the answers, at test time and in training?

The benchmark we build on releases evaluation code but not its question-answer pairs, so we rebuild a substrate from public sources whose targets are measured, not annotated, and so satisfy the cocycle by construction: `LUMIERE` (Suter et al., 2022), post-treatment glioma; `OASIS-2` (Marcus et al., 2010), longitudinal ageing; `VS-MC-RC` (Kujawa et al., 2024), vestibular schwannoma; and `SpaceNet-7` (Van Etten et al., 2021), satellite mosaics as a cross-domain control (Table 1). Every probe asks how much the target changed between two timepoints, as a signed percentage mapped to the log domain, in both directions of every interval and per view, which gives antisymmetry directly and cocycle residuals post hoc. We evaluate nine open-weight models from four families, Qwen3-VL at four sizes (Bai et al., 2025), Qwen2.5-VL, InternVL3.5 at two sizes (Wang et al., 2025), LLaVA-OneVision and MedGemma (Sellergren et al., 2025), served with vLLM (Kwon et al., 2023), greedily for the main conditions and with K samples for projected decoding. We compare against the following conditions and baselines:

- **Trivial predictor** answers "no change"; it coincides with the best constant here and is Δ-consistent, so every error is read against it.
- **Blind** gives the question with no images.
- **Identity** supplies the same scan twice, so the correct answer is zero; **leading** adds a note reporting deterioration.
- **Deformation** rescales one lesion by a known factor, so the true change is exact.
- **Self-consistency** (Wang et al., 2023) takes the median of the same samples the projection consumes.

Appendix F reports every table in full, Appendix C.7 the coverage of the certificate, and Appendix M the fine-tuning recipe with its attribution controls.

Table 1: Substrates. Triples count $i < j < k$ visit combinations; $|\Delta|$ is the measured log-ratio change.

Domain	Substrate	Process	Subj.	Visits	Triples	Median $\lvert\Delta\rvert$	Ground truth
Neuro-oncology	LUMIERE	post-treatment glioma	91	599	8464	0.96	tumour volume
	VS-MC-RC	vestibular schwannoma	75	225	75	0.18	tumour volume
Neurodegeneration	OASIS-2	ageing and dementia	150	373	119	0.01	brain volume
Remote sensing	SpaceNet-7	urban development	30	120	120	0.04	building area

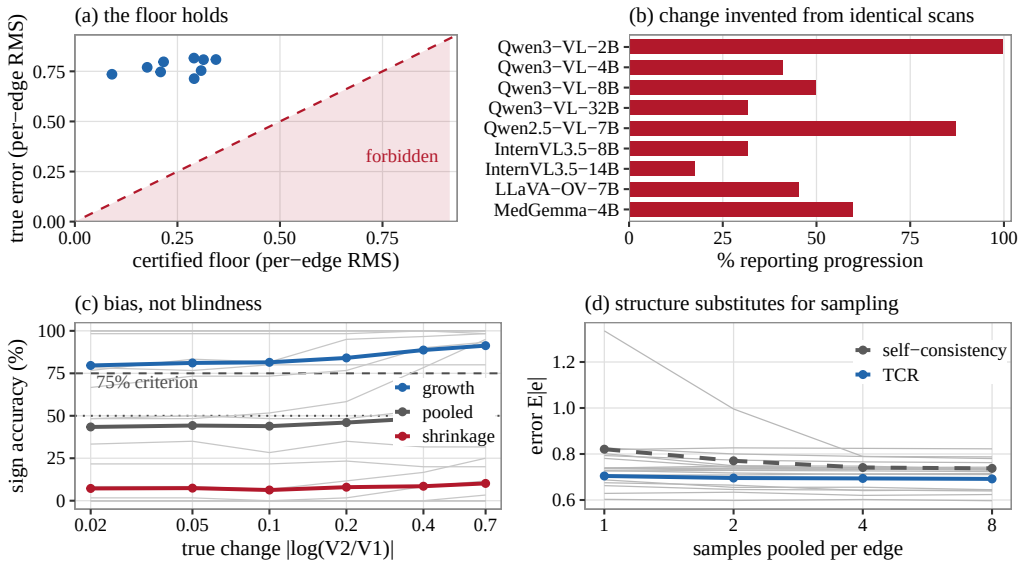

Figure 2: Audit results. (a) Realised error against the certified floor, one point per model; (b) change reported for identical scans; (c) sign accuracy under deformation; (d) error of the projection and of the median as samples grow.

4.1 RQ1: Reliability of reported change and of the certificate

We first answer RQ1 by comparing every model with the trivial predictor. Always answering "no change" achieves an error of 0.674; only 2 of 9 models fell below it, the best by 6.3% (Table 2), and predictions were range-collapsed ($\rho = 0.122$ with true change, spread 0.29 times the truth's; Table 10). Supplying the same scan twice makes the correct answer zero, yet Qwen3-VL-2B reported progression on 99.7% of such probes, panel average 51.5% (Figure 2b), and a leading note raised the rate for every model (Table 13). Blind, the models achieved 0.547 against 0.562 with images, which increased error for 5 of 9. Under controlled deformation no model reached 75% sign accuracy at any magnitude up to a 101% volume change, and 4 of 9 answered with one constant direction regardless of the truth (Figure 2c): a prior over the answer, not a limit on what they see (Appendix K).

The label-free side of the audit behaves as the theory predicts. Proposition 3 held in every cohort, no model fell below the diagonal (Figure 2a), and the fixed-threshold certificates lie below the realised error for every model at every confidence level tested (Tables 11 and 40), certifying floors of 0.058 to 0.279 log-ratio units at 95% from unlabelled data; 4 of 9 models are certified above 0.22, the smallest log-ratio implying a 20% volume error either way. On the same forward edges the bound captures 13.9% of the realised error against 12.0% for a matched triangle bound, and on the full directed graph it captures 32.1%, the difference coming from antisymmetry violations that no triangle sees (Appendix C.4). The certificate can reject a model and cannot clear one, since a collapsed model such as LLaVA-OV-7B (floor 0.058) is consistent, and it does not rank models that bunch at the trivial-predictor floor.

Table 2: Audit of the panel, in log-ratio units. 'Floor' is the 95% certified lower bound on per-edge RMS error. ✓/✗: error below the trivial predictor; floor below RMS error. †Medical model.

Family	Size	Against ground truth Error	Blind	Identity (%)	Label-free audit Cocycle	Floor	RMS error
Qwen3-VL	2B	0.701 ✗	0.584	99.7	0.145	0.170 ✓	0.770
	4B	0.725 ✗	0.643	40.8	0.545	0.251 ✓	0.809
	8B	0.719 ✗	0.595	49.8	0.288	0.279 ✓	0.808
	32B	0.646 ✓	0.608	31.7	0.439	0.259 ✓	0.754
Qwen2.5-VL	7B	0.743 ✗	0.219	87.2	0.164	0.144 ✓	0.797
InternVL3.5	8B	0.679 ✗	0.570	31.7	0.200	0.150 ✓	0.747
	14B	0.631 ✓	0.485	17.6	0.466	0.210 ✓	0.713
LLaVA-OV	7B	0.684 ✗	0.587	45.3	0.102	0.058 ✓	0.735
MedGemma†	4B	0.756 ✗	0.616	59.6	0.273	0.228 ✓	0.816
Trivial predictor		0.674	–	0.0	0.000	–	–

Table 3: Readout vs. answer: Spearman ρ with measured change on identical held-out items, LUMIERE unless stated. 'No labels': antisymmetry head, as $|\rho|$. ✓/✗: gap interval excludes zero.

Family	Size	Readout of h Supervised	No labels	Answer Generated	Gap 95% CI	Controls Text	Shuffled
Qwen3-VL	2B	**+0.456**	0.440	+0.031	[+0.22, +0.58] ✓	−0.003	+0.029
	4B	**+0.476**	0.460	+0.108	[+0.18, +0.54] ✓	+0.004	+0.035
	8B	**+0.508**	0.502	+0.137	[+0.16, +0.50] ✓	+0.002	+0.038
	32B	**+0.504**	0.478	+0.398	[−0.00, +0.24] ✗	+0.039	−0.008
Qwen2.5-VL	7B	**+0.293**	0.220	+0.082	[+0.03, +0.37] ✓	+0.002	−0.010
InternVL3.5	8B	**+0.582**	0.547	+0.206	[+0.23, +0.50] ✓	+0.001	−0.000
LLaVA-OV	7B	**+0.451**	0.382	+0.033	[+0.21, +0.60] ✓	+0.003	+0.004
MedGemma†	4B	**+0.544**	0.504	+0.085	[+0.30, +0.59] ✓	+0.000	+0.033
Qwen3-VL-8B	VS-MC-RC	**+0.098**	0.111	−0.022	[−0.19, +0.39] ✗	−0.007	−0.038
	SpaceNet-7	**+0.109**	0.192	+0.117	[−0.51, +0.44] ✗	+0.000	+0.016
	OASIS-2	**+0.052**	0.340	+0.023	[−0.19, +0.28] ✗	+0.005	−0.036

Takeaway 1

Most models do not beat a predictor that always answers no change, and they report change even between identical scans. The label-free certificate bounds this error from below: it can reject a model, but it cannot clear or rank one.

4.2 RQ2: Is the change encoded but not reported?

We next answer RQ2 with the readout diagnosis of Section 3.2 on Qwen3-VL-8B and LUMIERE. Over ten subject-level splits the supervised head reached $\rho = +0.508 \pm 0.046$ against $+0.137 \pm 0.051$ for the generated answer on the identical 544 test items per split (Table 19, Figure 3), a subject-level gap of $+0.368$ [$+0.163$, $+0.503$]. With sign and calibration fitted on the validation subjects only, the head gives 68.8% sign accuracy and MAE 1.011 against 53.6% and 1.156 for the answer under the same calibration (always-zero: 1.174).

The signal comes from the image pair (Table 20). With the images removed the head recovers nothing ($\rho = +0.002$), and likewise with the images shuffled across subjects ($+0.038$) or from the prompt's metadata ($+0.000$). One scan supplied twice recovers about six tenths of the pair ($+0.307$, $+0.312$), and the tied head on canonical per-visit representations gives $\rho = 0.499 \pm 0.063$ (Table 21), so a potential read off each visit and differenced recovers change about as well as the paired head. The

Table 4: Projected decoding against self-consistency on the same $K = 8$ samples, in log-ratio units. ✓/✗: projection error below self-consistency.

Family	Size	Error One sample	Error Self-cons.	Error **Projection**	Reduction vs. SC (%)	Cocycle Self-cons.	Cocycle **Projection**
Qwen3-VL	2B	1.206	0.780	**0.739**	**+5.2** ✓	0.160	**0.000**
	4B	1.202	0.722	**0.660**	**+8.5** ✓	0.333	**0.000**
	8B	1.229	0.772	**0.723**	**+6.4** ✓	0.215	**0.000**
	32B	1.062	0.644	**0.621**	**+3.7** ✓	0.216	**0.000**
Qwen2.5-VL	7B	1.184	0.740	**0.732**	**+1.1** ✓	0.090	**0.000**
InternVL3.5	8B	1.177	0.742	**0.722**	**+2.7** ✓	0.151	**0.000**
	14B	1.077	0.635	**0.619**	**+2.4** ✓	0.206	**0.000**
LLaVA-OV	7B	2.171	0.787	**0.743**	**+5.7** ✓	0.151	**0.000**
MedGemma†	4B	1.228	0.824	**0.733**	**+11.0** ✓	0.234	**0.000**
Mean					**+5.2**		

paired head is not itself a potential difference: its predictions on a pair and its reverse correlate at -0.622, not -1.

The same protocol on 8 models gives the same picture on `LUMIERE` (Table 3; per-model controls in Table 22). Supervised heads reach $\rho = +0.29$ to $+0.58$ against $+0.03$ to $+0.40$ for the answers, the text and shuffled controls are at chance for every model, and the gap's subject-level interval excludes zero for 7 of 8, across Qwen3-VL, Qwen2.5-VL, InternVL3.5, LLaVA-OneVision and MedGemma. The exception is Qwen3-VL-32B, whose answer already uses most of the signal (gap $[-0.002, +0.238]$). No overall gap is statistically resolved on `VS-MC-RC`, `SpaceNet-7` and `OASIS-2`, whose median $|\Delta|$ is 0.18, 0.04 and 0.01 against 0.96. On `OASIS-2` an antisymmetry head reaches $|\rho| = 0.340$ by reading visit order, not change: its correlation with change after partialling out order is $+0.016$, against $+0.494$ on `LUMIERE` (Table 23). Within a cohort the gap grows with the size of change, from $+0.11$ at $|\Delta| \leq 0.18$ to $+0.41$ above 0.69 on `LUMIERE` (Table 24); an intervention on magnitude is inconclusive (Appendix F), so this is an association, not a cause.

The same features admit heads trained with no annotation whatsoever. The cocycle, antisymmetry and full heads of equation 6 reach $|\rho| = 0.449$, 0.502 and 0.495, that is 88%, 99% and 98% of the supervised head on the same items, whereas the variance term alone reaches 0.081, the first principal component 0.087, and the best of ten components chosen on validation labels 0.286. The structure does the extracting, and antisymmetry alone extracts at least as much as the cocycle; whether the two select the same direction in feature space is untested.

Takeaway 2

On a cohort with large changes, the representations of models across families encode change that their answers do not report, and heads trained without labels recover most of it. On cohorts with small changes, no gap is resolved.

4.3 RQ3: Repair by projection

To answer RQ3 we first compare the projection with self-consistency, which consumes the same $K = 8$ samples and differs only in using an unstructured median. The projection reduced error by 5.2% on average and 11.0% at best (Table 4); the gains are modest, largest at $K = 1$, and shrink as self-consistency is given more draws (Figure 2d), so structure and sampling partly substitute.

We then fine-tuned Qwen3-VL-8B on its own projected samples, with no annotation anywhere, and evaluated on 1716 intervals from 59 unseen subjects (Table 5; subject-level bootstraps and paired comparisons in Appendix M.2). Cocycle violation in the model's own generations fell from 0.366 to 0.026, which the objective optimised directly, and correlation with measured change rose from $\rho = +0.187$ to $+0.520 \pm 0.008$ over three seeds (paired gain $[+0.253, +0.397]$), which it did not. Four controls differ from the projection only in the target string (Table 45). Distilling the greedy

Table 5: Consistency fine-tuning on Qwen3-VL-8B and its controls, pooled held-out set, mean over three seeds.

Targets	Run	Direction ρ	Direction Sign (%)	Magnitude Error	Magnitude sd(pred)	Magnitude Zero (%)	Coherence $\lvert\mathcal{R}\rvert$	Coherence $\lvert\mathcal{A}\rvert$
Reference	Base model	$+0.187$	55.4	1.153	0.279	0.0	0.366	0.733
	Trivial predictor	0.000	–	1.143	0.000	100.0	0.000	0.000
Label-free	Format only	$+0.204$	54.1	1.154	0.267	0.3	0.297	0.731
	Shrinkage only	$+0.229$	54.7	1.133	0.077	0.1	0.089	0.211
	Self-consistency median	$+0.226$	50.9	1.176	0.129	0.0	0.285	0.762
	Antisymmetry only	$+0.493$	43.8	1.128	0.045	41.0	0.021	0.018
	Cocycle projection	**+0.520**	**64.5**	**1.120**	**0.057**	**8.4**	**0.026**	**0.017**
	Scale-matched projection	$+0.537$	69.4	1.070	0.195	0.1	0.069	0.059
Labelled	Supervised reference	$+0.546$	71.0	1.227	1.695	0.0	1.195	0.946

answer as a bare answer line changes nothing ($\rho = +0.204$), shrinking it toward zero reaches $+0.229$, and distilling the unprojected median of the same samples reaches $+0.226$, because the samples carry the model's directional bias (92% increases). Antisymmetry alone cancels that bias and accounts for 92% of the gain in correlation; paired against it, the full projection adds $\Delta\rho = +0.027$ [$+0.002$, $+0.052$] and $+20.7$ points of sign accuracy [$+18.0$, $+23.2$], a modest cocycle-specific contribution. The same recipe on Qwen3-VL-4B raises ρ from $+0.119$ to $+0.575$, with antisymmetry accounting for 96% of the gain (Table 46).

Magnitude does not follow. Error moved from 1.153 to 1.120 against a trivial predictor at 1.143, and the response range contracted from 0.279 to 0.057, because averaging compresses the projected targets 3.3× below the model's own answers. Rescaling the targets to the base model's answer spread, still without labels, lowers error to 1.070, 6.4% below the trivial predictor, the best magnitude result here. A supervised reference trained with the same recipe reaches $\rho = +0.546$ but overshoots (sd 1.695) and does not beat the trivial predictor (1.227), so labels alone do not fix magnitude under this recipe.

Takeaway 3

Projection onto consistent answers lowers error at decoding time for every model, though modestly. Fine-tuning on projected answers improves their direction and coherence, mostly through antisymmetry, but not their magnitude.

5 Conclusion

Longitudinal questions ask for change, and change has an algebra that true answers obey. Checking a model's answers against this algebra bounds its error without labels and yields a distribution-free certificate, although a collapsed model is consistent and passes. On a glioma cohort with large changes, the weakness of current models is not for want of information, which linear heads read far above the generated answers across model families; on cohorts with much smaller changes no gap is resolved. Projection repairs direction and coherence, mostly through antisymmetry, while magnitude error stays close to that of the trivial predictor. Estimating the size of change, rather than its direction, remains future work.

Reproducibility Statement

Every substrate is rebuilt from public sources with scripted downloads: `LUMIERE` (Suter et al., 2022) from Figshare, `OASIS-2` (Marcus et al., 2010) from its open mirror, Vestibular-Schwannoma-MC-RC (Kujawa et al., 2024) through the TCIA public API together with its released segmentations, and `SpaceNet-7` (Van Etten et al., 2021) from its public S3 bucket. All models are open-weight and served with vLLM (Kwon et al., 2023) under greedy decoding for the main conditions, with a fixed seed for the sampled conditions. Ground truth is measured from released segmentations and released

morphometry rather than annotated by us, so it is reproducible bit-for-bit. Preprocessing, probe construction, inference, analysis, figures and tables are separate scripts run end-to-end by a single driver. The code, the summary results behind every table and figure, and all per-item model generations are released at `https://anonymous.4open.science/r/VLM-3j-ICLR27-725E/`; from the released results alone, one script rebuilds every number, table and figure of the paper in under a minute on a CPU. The four datasets are not redistributed, and the scripts download them from their providers; the theory in Sections 2 and 3 is self-contained and its proofs are in Appendix C.

Use of Large Language Models

During the preparation of this paper, we used a large language model to draft and revise text and to improve its grammar, word choice, and conciseness. We also used it to help identify relevant literature. The authors independently read all suggested papers and verified each citation against its original source before inclusion.

We used Claude Code to assist with designing and running the experiments and analyzing the results. The authors reviewed the code and verified all reported numerical results against the saved outputs. All claims are the authors' own, and we take full responsibility for the content of this paper.

References

Wafa Al Ghallabi, Ritesh Thawkar, Sara Ghaboura, Omkar Thawakar, Numan Saeed, Dana Al Nuaimi, Ajnas Alkatheeri, Salman Khan, and Fahad Shahbaz Khan. How good are foundation models in longitudinal MRI disease progression reasoning? In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, Lecture Notes in Computer Science, pp. 312–322. Springer, 2026. doi: 10.1007/978-3-032-38172-9_30.

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

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027, 2023. doi: 10.1109/CVPR52729.2023.01442.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Cosmin I. Bercea, Jun Li, Philipp Raffler, Evamaria O. Riedel, Lena Schmitzer, Angela Kurz, Felix Bitzer, Paula Roßmüller, Julian Canisius, Mirjam L. Beyrle, et al. NOVA: A benchmark for rare anomaly localization and clinical reasoning in brain MRI. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 115600–115633, 2025. doi: 10.52202/085713-3488.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 94327–94427, 2024. doi: 10.52202/079017-2992.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Xingxin He, Aurora Rofena, Ruimin Feng, Haozhe Liao, Zhaoye Zhou, Albert Jang, and Fang Liu. OmniMRI: A unified vision–language foundation model for generalist MRI interpretation. *arXiv preprint arXiv:2508.17524*, 2025.

Anil N. Hirani, Kaushik Kalyanaraman, and Seth Watts. Least squares ranking on graphs. *arXiv preprint arXiv:1011.1716*, 2010.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 4156–4165, 2023. doi: 10.1145/3580305.3599819. Dataset: Medical-Diff-VQA, PhysioNet.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMed-VQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, 2024. doi: 10.1109/CVPR52733.2024.02093.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1266–1279, 2022. doi: 10.18653/v1/2022.emnlp-main.82.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8849–8861, 2021. doi: 10.18653/v1/2021.emnlp-main.697.

Aaron Kujawa, Reuben Dorent, Steve Connor, Suki Thomson, Marina Ivory, Ali Vahedi, Emily Guilhem, Navodini Wijethilake, Robert Bradford, Neil Kitchen, et al. Deep learning for automatic segmentation of vestibular schwannoma: A retrospective study from multi-center routine MRI. *Frontiers in Computational Neuroscience*, 18:1365727, 2024. doi: 10.3389/fncom.2024.1365727. Data: Vestibular-Schwannoma-MC-RC, The Cancer Imaging Archive.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open Access Series of Imaging Studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. doi: 10.1162/jocn.2009.21407.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D. Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1754–1768, 2022. doi: 10.18653/v1/2022.emnlp-main.115.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6642–6651, 2019. doi: 10.1109/CVPR.2019.00681.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 596–608, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Yannick Suter, Urspeter Knecht, Waldo Valenzuela, Michelle Notter, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. The LUMIERE dataset: Longitudinal glioblastoma MRI with expert RANO evaluation. *Scientific Data*, 9(1):768, 2022. doi: 10.1038/s41597-022-01881-7.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6394–6403, 2021. doi: 10.1109/CVPR46437.2021.00633.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Patrick Y. Wen, David R. Macdonald, David A. Reardon, Timothy F. Cloughesy, A. Gregory Sorensen, Evanthia Galanis, John DeGroot, Wolfgang Wick, Mark R. Gilbert, Andrew B. Lassman, et al. Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology Working Group. *Journal of Clinical Oncology*, 28(11):1963–1972, 2010. doi: 10.1200/JCO.2009.26.3541.

Junyi Zhang, Jia-Chen Gu, Wenbo Hu, Yu Zhou, Robinson Piramuthu, and Nanyun Peng. TemMed-Bench: Evaluating temporal medical image reasoning in vision-language models. In *Conference on Language Modeling (COLM)*, 2026. URL https://openreview.net/forum?id=kyByX4OnSJ.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. DrVD-Bench: Do vision-language models reason like human doctors in medical image diagnosis? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 14020–14064, 2025. doi: 10.52202/085713-0420.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

A Related work

A first line of related work evaluates vision-language models on medical images. Medical VQA suites (Hu et al., 2024; Chen et al., 2024; Zhang et al., 2023) and brain-MRI benchmarks (Bercea et al., 2025; Zhou et al., 2025; He et al., 2025) process each image independently, and work with a prior image (Bannur et al., 2023; 2024; Hu et al., 2023; Zhang et al., 2026) poses one comparison at a time. Al Ghallabi et al. (2026) assemble the first multi-view, multi-sequence longitudinal MRI benchmark and report reliable ordering with unreliable direction; we corroborate this on a rebuilt substrate with measured ground truth. Our contribution is complementary: rather than a new set of questions, we check the answers against an identity that true change satisfies, which needs no annotation, and we test whether the missing information is accessible in the models' representations.

A second line of work studies consistent pairwise comparisons. Jiang et al. (2011) decompose pairwise comparisons into a gradient flow and cyclic residuals and read the residual as inconsistency, and Hirani et al. (2010) study the corresponding least-squares ranking problem on graphs. The machinery of Section 3 is theirs: the consistent estimators are the gradient subspace, and the projection of Section 3.3 is HodgeRank on a temporal graph. What changes is the use. On a temporal graph whose ground truth lies in the gradient subspace by construction, the residual becomes a lower bound on error, and clipping it at a pre-declared threshold yields a certificate on an unlabelled cohort (Hoeffding, 1963; Maurer & Pontil, 2009).

A third line of work enforces consistency in model outputs. *Self-consistency* (Wang et al., 2023) aggregates sampled reasoning paths by majority vote, while consistency regularisation (Sohn et al., 2020) and cycle constraints (Zhu et al., 2017; Shah et al., 2019) enforce agreement under transformations of the *input*. Closer are methods that jointly correct answers across related inputs through learned or logical relations (Mitchell et al., 2022; Kassner et al., 2021; Jung et al., 2022; Elazar et al., 2021). Our setting differs in that continuous answers are coupled through an exact linear identity, so the correction is a closed-form projection and the residual is a bound rather than a heuristic score.

B Notation

Table 6 lists every symbol the paper uses, in order of first appearance, with the section that introduces it.

Table 6: Symbols used throughout, in order of first appearance.

Symbol	Meaning
V_i	measured quantity at timepoint i (lesion volume, brain volume, footprint area)
ε	additive stabiliser in the log-ratio, $1000\,\mathrm{mm}^3$ unless stated
r_{ij}	log-ratio target $\log\big((V_j+\varepsilon)/(V_i+\varepsilon)\big)$
$\widehat{\Delta}(i\to j)$	the model's estimate of change from timepoint i to timepoint j
φ	latent trajectory; $\varphi(i)$ is the state at timepoint i
n	number of timepoints for one subject
E, m	observed ordered pairs, and their number
A	signed incidence matrix of the timepoint graph, $A \in \mathbb{R}^{m\times n}$
range(A)	the consistent set: exactly the estimates satisfying the cocycle
P	orthogonal projector onto range(A)
$\widehat{y}$, $y^\star$	stacked model predictions, and stacked ground truth
e	error vector $\widehat{y}-y^\star$
$\|(I-P)\widehat{y}\|_2$	the certified floor of Proposition 3; needs no labels
$\sin\theta$	scale-free inconsistency angle, $\|(I-P)\widehat{y}\|_2/\|\widehat{y}\|_2$
b_s	per-subject bound on per-edge RMS error, $\|(I-P_s)\widehat{y}_s\|_2/\sqrt{m_s}$
δ	certificate risk level; bounds hold with probability $1-\delta$
B, $\tilde{b}_s$	pre-declared clipping threshold, and the clipped residual $\min(b_s, B) \in [0, B]$
W, w_e	diagonal reliability weights used by the projection, $w_e = (1+s_e^2)^{-1}$
s_e	standard deviation of the samples on edge e
K	samples drawn per edge before aggregation
C	observable evidence conditioned on in the pairwise ceiling
L, $\widehat{L}$	true and predicted change category

C Deferred Proofs

C.1 Proposition 2

Proposition. $\widehat{\Delta}$ satisfies the cocycle condition for all i, j, k if and only if there exists φ with $\widehat{\Delta}(i\to j) = \varphi(j) - \varphi(i)$; such a φ is unique up to an additive constant, and the cocycle implies identity and antisymmetry.

Proof. ($\Leftarrow$) Immediate: $(\varphi(k)-\varphi(i)) = (\varphi(j)-\varphi(i)) + (\varphi(k)-\varphi(j))$.

($\Rightarrow$) Setting $i = j = k$ gives $\widehat{\Delta}(i \to i) = 2\widehat{\Delta}(i \to i)$, hence $\widehat{\Delta}(i \to i) = 0$, which is identity. Setting $k = i$ then gives $0 = \widehat{\Delta}(i \to j) + \widehat{\Delta}(j \to i)$, which is antisymmetry. Fix a reference index r and define $\varphi(i) := \widehat{\Delta}(r \to i)$. Applying the cocycle to (r, i, j) gives $\widehat{\Delta}(r \to j) = \widehat{\Delta}(r \to i) + \widehat{\Delta}(i \to j)$, i.e. $\widehat{\Delta}(i \to j) = \varphi(j) - \varphi(i)$. If φ, φ' both represent $\widehat{\Delta}$ then $\varphi - \varphi'$ is constant, since $\varphi(j) - \varphi'(j) = \varphi(i) - \varphi'(i)$ for all i, j. □

C.2 Proposition 3

Proof. $y^\star \in \text{range}(A)$, so $\|\widehat{y} - y^\star\|_2 \geq \inf_{z \in \text{range}(A)} \|\widehat{y} - z\|_2$, and the infimum of the distance to a closed subspace is attained at the orthogonal projection, giving $\|(I - P)\widehat{y}\|_2$. Equality holds exactly when $y^\star$ is that projection, i.e. when the error is orthogonal to the subspace. □

C.3 Rank of the incidence matrix, and what the bound uses

For a graph on n timepoints with c connected components, $\text{rank}(A) = n - c$, so $\text{range}(A)$ has dimension $n - c$ inside $\mathbb{R}^E$ and the orthogonal complement, the space the residual lives in, has dimension $E - n + c$. This is the cycle space of the graph. On a complete graph over $n = 4$ timepoints with both directions observed, $E = 12$ and the cycle space has dimension 9; a triangle statistic sees only its three-cycle generators, and only in one direction.

C.4 Triangles against the whole cycle space, compared on one quantity

The first version of this paper compared the projection bound with the mean absolute triangle residual divided by three, $\frac{1}{3}\,\mathbb{E}_t|\mathcal{R}_t|$, and reported the projection as $3\times$ tighter. The comparison was not valid. The triple statistic is a lower bound on the mean absolute error over the edges of the forward triangles it is computed on, whereas the projection bound was normalised over all observed directed edges; the two neither bound the same quantity nor live on the same edge set. A concrete failure: three timepoints, both directions observed, truth identically zero, forward answers $(+1, +1, -1)$ and reverse answers 0. The triple statistic returns 1.000, while the per-edge RMS error over the six directed edges is 0.707, which the projection bound returns exactly. On the panel the triple statistic exceeded the all-edge RMS error on 8 of 4,989 subject-view systems, so it was not a lower bound on the quantity it was compared against.

A matched comparator exists. Restrict to the forward edge set E_f (all observed pairs $i < j$), on which both bounds certify the same per-edge RMS error $\|e_f\|_2/\sqrt{|E_f|}$. For each triangle t with residual $\mathcal{R}_t = e_{ik} - e_{ij} - e_{jk} = u^\top e_t$, $\|u\|_2 = \sqrt{3}$ gives $\mathcal{R}_t^2 \leq 3\|e_t\|_2^2$. Summing over triangles and writing c_e for the number of triangles that contain edge e,

$$\sum_t \mathcal{R}_t^2 \;\leq\; 3\sum_e c_e\, e_e^2 \;\leq\; 3\, c_{\max}\|e_f\|_2^2, \qquad \text{so} \qquad \frac{\|e_f\|_2}{\sqrt{|E_f|}} \;\geq\; \sqrt{\frac{\sum_t \mathcal{R}_t^2}{3\, c_{\max}\, |E_f|}}. \tag{8}$$

This is a valid lower bound on the same quantity as the projection bound on E_f, and on a single triangle the two coincide at $|\mathcal{R}|/3$. On the panel the projection captures 13.9% of the forward-set RMS error against 12.0% for equation 8, a ratio of 1.16, and it is at least as large on every one of 4,954 systems. The advantage of solving the least-squares problem over the whole cycle space, rather than reading triangles one at a time, is therefore real but small. What the projection over all directed edges adds, 32.1% of the all-edge RMS error, comes from the reverse edges: the distance to $\text{range}(A)$ on the full directed graph is also a distance to antisymmetry, and antisymmetry violations are the larger part of what these models get wrong (Section 4.3).

C.5 Why the log domain is not cosmetic

Relative change composes multiplicatively: $(1+p_{ij})(1+p_{jk}) = 1+p_{ik}$. Only after taking logarithms does composition become addition, which is what makes the consistent set a linear subspace and hence the projection well defined. Any $\varepsilon > 0$ in equation 1 preserves this exactly, since the identity $\log \frac{V_k+\varepsilon}{V_i+\varepsilon} = \log \frac{V_j+\varepsilon}{V_i+\varepsilon} + \log \frac{V_k+\varepsilon}{V_j+\varepsilon}$ holds for every V_i, V_j, V_k. The stabiliser therefore changes the target but never the algebra.

C.6 Pairwise unidentifiability and the ceiling it imposes

Both statements are deferred here from Sections 2 and 3, where only their consequence is used.

Proposition 6 (Pairwise unidentifiability)**.** *There exist subjects $A \neq B$ and timepoints such that the pairwise evidence agrees, $\Delta^\star_A(1 \to 2) = \Delta^\star_B(1 \to 2)$, while the confirmed trajectories diverge in sign, $\Delta^\star_A(2 \to 3) > 0 > \Delta^\star_B(2 \to 3)$. Consequently no function of the measured pairwise change $\Delta^\star(1 \to 2)$ alone can recover the confirmed progression label on such pairs. The statement does not extend to functions of the images: two pairs with equal measured change need not present equal image evidence, so nothing here bounds a predictor of (I_1, I_2).*

Appendix R lists the ten tightest witness pairs of 9295 found in `LUMIERE`. Partitioning the pairs into classes by their observable evidence C, a method restricted to C answers identically within a class, so:

Proposition 7 (Pairwise ceiling)**.** *Let $C(\cdot)$ map a pair to its observable evidence class and L be the confirmed direction. Any predictor of the form $\widehat{L} = f(C)$ satisfies* $\Pr[\widehat{L} = L] \leq \mathbb{E}_c\big[\max_\ell \Pr[L = \ell \mid C = c]\big]$.

Proof. Let $\widehat{L} = f(C)$. Conditioning on the class, $\Pr[\widehat{L} = L] = \mathbb{E}_c\big[\Pr[L = f(c) \mid C = c]\big] \leq \mathbb{E}_c\big[\max_\ell \Pr[L = \ell \mid C = c]\big]$, with equality when f selects the conditional mode. The bound is attained by the Bayes rule and therefore holds for every method restricted to C. □

Two caveats bound the reading. We take C to be the measured interval log-ratio, discretised at 0.02, so the estimate binds predictors reading that quantity rather than every function of the raw images. And it is a finite-sample plug-in on 2001 pairs across 337 classes, which rises monotonically with the resolution of C: 50.2% at bin width 1.0 against 61.0% at 0.02. The value quoted in the body is the loosest, and therefore the safest, point in that sweep (Table 16). The figure is therefore a restricted, finite-sample diagnostic of how much the measured pairwise change determines the confirmed direction, and not an impossibility result for image-based prediction.

C.7 The certificates in Corollary 4

Proposition 3 is a deterministic statement about one subject. A site that wants to certify a model before commissioning annotation needs a statement about its *cohort*, which requires a sampling assumption and a bounded variable. Let $b_1, \ldots, b_S$ be the per-subject bounds of Corollary 4, i.i.d. draws from an unknown distribution on $[0, \infty)$, let e_s be the subject's true per-edge RMS error, so that $b_s \leq e_s$ by Proposition 3, and let $\mu = \mathbb{E}[e]$.

Why the threshold must be fixed in advance. The distribution-free inequalities below hold for i.i.d. variables on a known interval $[0, B]$. The residual b_s has no such bound a priori, so we certify the clipped variable $\tilde{b}_s = \min(b_s, B)$, which lies in $[0, B]$ by construction and satisfies $\mathbb{E}[\tilde{b}] \leq \mathbb{E}[b] \leq \mu$. Any valid lower confidence bound on $\mathbb{E}[\tilde{b}]$ is therefore a valid lower confidence bound on μ, and clipping can only make the certificate more conservative. The argument requires B to be a constant: if B is read off the sample, the normalised variables $\tilde{b}_s/B$ are neither independent nor bounded by a fixed constant, and neither inequality applies. The first version of this paper made that mistake, setting B to $1.5\times$ the sample's own 99.9th percentile; the values it reported were therefore not covered by the proof, and Table 7 shows they differ from the corrected ones by up to 31%, in either direction. We now fix $B = 1$ log-ratio unit before any per-model computation. One unit is a per-edge RMS inconsistency of a factor e, the scale of the task itself, and Table 7 reports the certificate at $B \in \{0.5, 1, 2, 4\}$ so the dependence is visible: raising B clips fewer subjects but widens the concentration term, and the floor falls. At $B = 1$, 2 of 9 models have any clipped subject, at most 5.9% of the cohort.

Chernoff–Hoeffding in KL form. Writing $\bar{x} = \frac{1}{S}\sum_s \tilde{b}_s/B$, Hoeffding's inequality (Hoeffding, 1963) for $[0,1]$-valued i.i.d. variables gives, with probability at least $1 - \delta$, $\mathrm{kl}(\bar{x},\, \mathbb{E}[\tilde{b}]/B) \leq \log(1/\delta)/S$ on the event $\bar{x} > \mathbb{E}[\tilde{b}]/B$, so

$$\mu_{\mathrm{KL}} \;=\; B \cdot \min\Big\{ q \in [0, \bar{x}] : S\,\mathrm{kl}\big(\bar{x},\; q\big) \leq \log \tfrac{1}{\delta} \Big\}, \qquad \mathrm{kl}(p, q) = p \log \tfrac{p}{q} + (1-p) \log \tfrac{1-p}{1-q}, \quad (9)$$

is a $(1-\delta)$ lower confidence bound on $\mathbb{E}[\tilde{b}]$ and hence on μ. We solve it by bisection; $\mathrm{kl}(\bar{x}, \cdot)$ is strictly decreasing on $[0, \bar{x}]$, so the root is unique.

Empirical Bernstein. The KL bound pays for the full width B even when the residuals are concentrated far below it. The empirical-Bernstein inequality of Maurer & Pontil (2009) adapts to the sample variance V of $\tilde{b}_s/B$: with probability at least $1-\delta$,

$$\mu_{\mathrm{EB}} \;=\; B \cdot \left(\bar{x} - \sqrt{\frac{2V \log(2/\delta)}{S}} - \frac{7 \log(2/\delta)}{3(S-1)} \right)_+ \tag{10}$$

is a lower confidence bound on $\mathbb{E}[\tilde{b}]$. On this panel the two agree to within 0.011 at $S = 238$ (Table 11); at small S the constant term dominates, the empirical-Bernstein bound returns zero at $S = 30$ in the simulation below, and the KL bound is the better of the two.

Percentile bootstrap. Resample $\{b_s\}$ with replacement 10^4 times, take the mean of each resample, and report its δ quantile. This is the standard nonparametric interval on the unclipped residual and the estimate one would use in practice; its validity is asymptotic, and we report it only as a reference beside the two distribution-free bounds. It exceeds the KL bound by a factor of 1.28 on average across the panel.

Coverage check. A simulation cannot establish distribution-free validity; the proof does that, for a fixed B. What a simulation can show is the price of each construction and whether the asymptotic bootstrap and the superseded adaptive-threshold procedure behave as their guarantees, or lack of them, suggest. Table 8 draws $S \in \{30, 238\}$ residuals from 7 laws with known mean, including heavy-tailed, two-point and spike-plus-outlier laws with mass above B, 1000 times each. The two fixed-threshold bounds cover on 100.0% and 100.0% of draws at worst; the bootstrap falls below nominal in 3 of 14 cells, to 93.9% at worst. The adaptive-threshold procedure did not fail on any law we tried, which is not a guarantee and is the reason it was replaced.

Table 7: The cohort certificate at four pre-declared clipping thresholds B ($\delta = 0.05$). KL and EB are the Chernoff–Hoeffding and empirical-Bernstein bounds; 'legacy' uses a sample-dependent B.

			boot.	$B = 0.5$			$B = 1$			$B = 2$			$B = 4$			legacy	true
Model	S	max b_s	(asympt.)	clip %	KL	EB	clip %	KL	EB	clip %	KL	EB	clip %	KL	EB	(adaptive B)	error
Qwen3-VL-2B	238	0.707	0.214	8.8	0.188	0.180	0.0	0.170	0.165	0.0	0.144	0.129	0.0	0.113	0.056	0.169	0.770
Qwen3-VL-4B	238	0.887	0.297	30.7	0.249	0.236	0.0	0.251	0.244	0.0	0.217	0.208	0.0	0.177	0.135	0.239	0.809
Qwen3-VL-8B	238	0.853	0.329	26.1	0.281	0.271	0.0	0.279	0.277	0.0	0.242	0.241	0.0	0.200	0.168	0.267	0.808
Qwen3-VL-32B	238	1.745	0.320	26.1	0.238	0.223	5.9	0.259	0.243	0.0	0.248	0.221	0.0	0.204	0.148	0.232	0.754
Qwen2.5-VL-7B	238	0.568	0.191	0.8	0.165	0.165	0.0	0.144	0.147	0.0	0.120	0.111	0.0	0.093	0.038	0.150	0.797
InternVL3.5-8B	238	0.453	0.194	0.0	0.171	0.166	0.0	0.150	0.148	0.0	0.126	0.112	0.0	0.097	0.039	0.162	0.747
InternVL3.5-14B	238	1.716	0.255	13.9	0.212	0.201	1.7	0.210	0.199	0.0	0.184	0.164	0.0	0.148	0.092	0.174	0.713
LLaVA-OV-7B	238	0.219	0.093	0.0	0.069	0.071	0.0	0.058	0.053	0.0	0.044	0.017	0.0	0.029	0.000	0.075	0.735
MedGemma-4B†	238	0.894	0.279	10.1	0.245	0.242	0.0	0.228	0.231	0.0	0.196	0.194	0.0	0.158	0.122	0.215	0.816

Table 8: Coverage and tightness of the certificate under seven residual laws with known mean ($B = 1$, nominal 0.95).

		mass	coverage (%)				tightness		
Law	S	$> B$ (%)	boot.	KL	EB	legacy	boot.	KL	EB
gamma(2, 0.15)	30	1.1	96.5	100.0	100.0	100.0	0.81	0.43	0.00
gamma(2, 0.15)	238	1.0	94.2	100.0	100.0	100.0	0.93	0.77	0.75
lognormal(−1.6, 1.0)	30	5.5	97.7	100.0	100.0	100.0	0.69	0.39	0.00
lognormal(−1.6, 1.0)	238	5.5	96.6	100.0	100.0	100.0	0.87	0.68	0.64
lomax(2.5, 0.3) heavy tail	30	2.4	98.4	100.0	100.0	100.0	0.60	0.27	0.00
lomax(2.5, 0.3) heavy tail	238	2.6	97.7	100.0	100.0	100.0	0.82	0.62	0.52
uniform(0, 0.6)	30	0.0	95.2	100.0	100.0	100.0	0.83	0.43	0.00
uniform(0, 0.6)	238	0.0	93.9	100.0	100.0	100.0	0.94	0.77	0.78
two-point 0: .9, 2: .1	30	9.9	99.6	100.0	100.0	100.0	0.27	0.10	0.00
two-point 0: .9, 2: .1	238	10.0	94.7	100.0	100.0	100.0	0.69	0.30	0.06
spike + rare outlier 0.2: .98, 5: .02	30	2.0	97.9	100.0	100.0	100.0	0.69	0.25	0.00
spike + rare outlier 0.2: .98, 5: .02	238	2.0	97.0	100.0	100.0	100.0	0.79	0.53	0.54
beta(0.5, 3) on [0, 1]	30	0.0	96.6	100.0	100.0	100.0	0.68	0.25	0.00
beta(0.5, 3) on [0, 1]	238	0.0	96.8	100.0	100.0	100.0	0.88	0.65	0.54

C.8 A WORKED EXAMPLE ON THREE TIMEPOINTS

The smallest system that has any cycle space at all is one subject with three timepoints and all three forward intervals observed. Order the edges $\big((1,2),(2,3),(1,3)\big)$, so

$$A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \qquad \operatorname{rank}(A) = 2, \qquad \operatorname{range}(A)^{\perp} = \operatorname{span}\{u\}, \quad u = \tfrac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

The one-dimensional complement is exactly the cocycle direction: $u^{\top} y \propto y_{12} + y_{23} - y_{13}$, which is the cocycle residual. Hence $P = I - uu^{\top}$ and

$$\big\|(I - P)\widehat{y}\big\|_2 = |u^{\top}\widehat{y}| = \frac{\big|\widehat{y}_{12} + \widehat{y}_{23} - \widehat{y}_{13}\big|}{\sqrt{3}}. \tag{11}$$

At $n = 3$ the projection bound on the forward edges therefore equals $|\mathcal{R}|/3$, which is also the matched triangle bound equation 8, and the two separate only at $n \geq 4$, where the cycle space acquires dimensions no single triple sees. Concretely, take a model predicting $\widehat{y} = (+0.30, +0.30, +0.20)$ against a truth of $y^{\star} = (+0.10, +0.10, +0.20)$. The bound returns $|0.30{+}0.30{-}0.20|/\sqrt{3} = 0.231$, the true error is $\|(0.20, 0.20, 0)\|_2 = 0.283$, and the bound is valid and captures 82% of it. Replace the truth by $(+0.30, +0.30, +0.60)$ and the bound is unchanged at 0.231 while the true error rises to 0.400: the extra error went into range(A), where by construction nothing label-free can detect it. This is the mechanism behind Table 38 in miniature.

C.9 THE PROJECTION AT $\varepsilon \to 0$ AND AT $\varepsilon \to \infty$

The stabiliser ε in equation 1 changes the target the model is asked to predict, and Appendix C shows it never changes the algebra. It is worth noting the two limits, because they bound the design choice. As $\varepsilon \to 0$ the target tends to $\log(V_j/V_i)$, which is scale-invariant but diverges when a lesion is resected to zero volume, a real event in LUMIERE, where post-resection timepoints occur. As $\varepsilon \to \infty$ the target tends to $(V_j - V_i)/\varepsilon \to 0$, so every change is compressed toward zero and the task becomes trivial and uninformative. Our $\varepsilon = 1000\,\text{mm}^3$ sits nearer the first limit. It is one cubic centimetre: 7% of the median tracked volume in LUMIERE and $1/123$ of the largest, so it barely moves a typical interval, while the 16 of 599 timepoints whose tracked volume is exactly zero after resection, and the 58 that fall below ε, are regularised rather than sent to $\pm\infty$. Appendix F reports the full sensitivity sweep over this choice, and every qualitative conclusion is unchanged across two orders of magnitude in ε.

D IMPLEMENTATION DETAILS

Serving. All models are served with vLLM. Prompts are constructed as chat messages and rendered by each model's own processor rather than by a hand-written template. This is not cosmetic: several architectures expand a single image placeholder into a run of per-patch tokens, and bypassing that step misaligns image embeddings against text positions. The failure is silent, since the model emits fluent-looking token sequences rather than raising, so we verified every model's outputs parse and are on-topic before analysis. One checkpoint (the HF-converted InternVL3.5-8B) produced token garbage under vLLM and was replaced by the original release, which behaves normally. Under transformers with its own processor the converted checkpoint decodes normally, and its answers on the 2634 pair probes match the original release's exactly on 88% of items and in sign on 99% (Table 26), so the representation and the answers of Table 22 come from the same weights and agree in behaviour.

Answer extraction. Both response modes end with a fixed `ANSWER:` line, parsed with a strict pattern. We record how often a lenient fallback would have been needed; across the panel the strict pattern succeeds on essentially every response, so no result depends on lenient matching of words appearing in free-text reasoning. Generation length was capped for throughput, and two models lose a nontrivial fraction of *quantitative* answers to truncation before the answer line (Table 2); the categorical mode loses under 0.1% for every model, and it is the mode that carries the identity, suggestibility and detectability results.

Registration and orientation. LUMIERE is distributed atlas-registered. OASIS-2 and VS-MC-RC are handled with SimpleITK: 6-DOF rigid registration with Mattes mutual information over a three-level pyramid. Volumes are explicitly reoriented to RAS before slicing, because transposing a raw image buffer preserves the scanner's acquisition order rather than anatomical axes, which silently mislabels the three planes for sagittally-acquired MPRAGE.

Why VS-MC-RC is axial-only. 79% of that cohort's series are thin-slab acquisitions targeted at the internal auditory canal with fewer than 75 slices. Coronal and sagittal reformats would carry almost no through-plane information, and resampling follow-ups into a shared baseline frame discards large parts of studies whose field of view differs. We slice each timepoint in its own geometry at its own lesion centroid, state that in the prompt, and exclude the cohort from the view-invariance analysis rather than report reformats that are not real measurements.

Detectability sweep. The warped segmentation is measured by summing its linearly interpolated occupancy rather than by re-binarising it. Thresholding returns a bit-identical mask for every change below roughly 15%, because a 2% radial rescale displaces the lesion boundary by a fraction of a voxel, which would blank out precisely the low-magnitude region the experiment exists to probe.

Probe protocol. Representations are the hidden state at the final prompt token, before any generation. Splits are three-way over *subjects* (60/20/20 train/validation/test), never over pairs, so no timepoint seen in training appears at test time; features are standardised using training-split statistics only. Both heads are stopped on validation loss, each measured against its own objective. This is load-bearing rather than routine: the two heads converge at different rates, and at a fixed epoch budget the measured gap between them ranged from $5\times$ to $1.3\times$ purely as a function of how long we trained. Results are averaged over ten seeds.

Objectives, stated exactly. Every head is $p = w^\top h + b$ on the standardised representation, trained with AdamW (learning rate 10^{-3}, weight decay 10^{-2}, full batch), evaluated on the validation objective every ten epochs and stopped after thirty evaluations without improvement. The supervised head minimises the smooth-ℓ_1 loss to the measured log-ratio. The label-free heads minimise a structural term plus the variance term λ relu$(1 - \mathrm{Var}_{\text{train}}\, p)$ with $\lambda = 5$, which excludes the zero map and is inactive once the training variance exceeds one: the cocycle head uses $\mathrm{mean}_{i<j<k}(p_{ik} - p_{ij} - p_{jk})^2$ over the forward triangles of the training subjects, the antisymmetry head $\mathrm{mean}(p_{ij} + p_{ji})^2$ over reverse pairs, and the full head both. Table 21 reports these beside three heads that impose no constraint (the variance term alone, the first principal component, and the best of the top ten components chosen on validation labels) and two tied heads $w^\top(h_j - h_i)$ with no bias. The canonical one encodes each visit once, per view, under the pair prompt with the interval sentence removed and the scan supplied twice (`code/extract_visit_features.py`), so the difference features are antisymmetric and cocycle-consistent identically and the head is a potential difference by construction. The context-dependent one uses the single-scan representations of Table 20, extracted under each pair's own prompt, which states that pair's interval; its difference features are antisymmetric (residual 0.1% of the feature norm) but not cocycle-consistent (residual 54%), so a head on them is a tied antisymmetric control and not a potential difference. The first revision reported the second as if it were the first.

Matched scoring and controls. For every split the generated answer is parsed on exactly the test probes and scored beside the heads (Table 19). Two protocols are reported. The sign-invariant diagnostic fixes orientation and affine scale on the test labels and measures recovery up to those; the deployable protocol fixes both on the validation subjects and reads the sign from the prediction. The controls of Table 20 re-extract the representation on the identical probes with the images removed (the blind prompt for the same interval, replicated to each view), shuffled across subjects within a view, replaced by the first or the second scan supplied twice, or supplied in reversed order with the prompt unchanged; a further control concatenates the two single-scan representations. Metadata baselines fit a ridge regression on the same splits to the interval length the prompt states, and, as an oracle the model never sees, to the signed interval and visit indices. The reversed-order control is reported as the correlation between a head's predictions on a pair and on the same pair reversed: the prompt is direction-blind, so the reversed set is the reverse probes relabelled, and a correlation against truth on it would only restate the original number with its sign flipped.

E SUBSTRATES

LUMIERE. Longitudinal glioblastoma with expert RANO ratings and per-timepoint automated segmentations. We use the DeepBraTumIA atlas-space outputs, taking tumour volume as enhancing core plus necrotic core, and render axial, coronal and sagittal slabs at a patient-level tumour centroid held fixed across time so that image change reflects disease rather than slice selection.

OASIS-2. Longitudinal ageing and dementia. Ground truth is the released nWBV$\times$eTIV product, i.e. brain parenchymal volume, which is measured independently of anything in our pipeline. Median change is roughly two orders of magnitude smaller than in LUMIERE, which makes this cohort a near-null-signal control.

Vestibular-Schwannoma-MC-RC. Benign, slowly growing tumours with multi-year intervals and expert-reviewed masks released separately from the imaging. The release spreadsheet lists more rows than it ships, so the mapping is driven by the archive contents rather than the spreadsheet. Despite the disease, this cohort is not a natural-history series: 46.7% of consecutive intervals shrink and 21 of 75 subjects lose more than 40% of tumour volume between visits, which is treatment (Appendix H.4). We impose no monotonicity here.

SpaceNet-7. Monthly satellite mosaics of fixed locations with per-month building footprints. Included to test whether any of the machinery is medical; the tracked quantity is total footprint area.

F EXTENDED RESULTS

This section holds the numerical detail behind Section 4, in the order the body uses it. Table 9 breaks the single cocycle column of Table 2 into all five structural violations, and Table 10 adds the informativeness statistics, correlation, sign agreement and response range, that the body quotes as panel means. Table 11 reports the bound, the matched triangle comparator of Appendix C.4 and the realised error per model, with the certified floors of Corollary 4, and Table 12 splits error and violation by substrate. Table 13 gives the identity and suggestibility conditions in full, Table 14 the psychometric curve per model, and Table 15 the four decoders on the identical sample pool. Table 16 sweeps the resolution of the pairwise ceiling and Table 17 the consistency scaling law. Table 18 reports the frozen-feature readout, Tables 19 and 20 the matched comparison with the generated answer and the input controls, Table 27 the re-analysis of the source benchmark's composite promised in Section 1, and Tables 28 and 29 the sensitivity of our conclusions to the stabiliser ε and to the tolerance below which a change counts as unchanged. The last three tables, Tables 30, 31 and 32, are compact versions of the per-cohort, audit and detectability results.

Table 9: Structural violations per model with standard errors, in log-ratio units.

Model	Ident.	(sem)	Antisym.	(sem)	Cocycle	(sem)	ViewSD	Rate
Qwen3-VL-2B	0.095	0.002	0.307	0.005	0.145	0.004	0.029	0.094
Qwen3-VL-4B	0.177	0.004	0.560	0.011	0.545	0.016	0.341	0.278
Qwen3-VL-8B	0.102	0.003	0.547	0.009	0.288	0.007	0.125	0.150
Qwen3-VL-32B	0.017	0.001	0.435	0.012	0.439	0.017	0.266	0.183
Qwen2.5-VL-7B	0.081	0.001	0.406	0.004	0.164	0.004	0.061	0.126
InternVL3.5-8B	0.039	0.002	0.394	0.006	0.200	0.005	0.047	0.133
InternVL3.5-14B	0.026	0.002	0.444	0.012	0.466	0.016	0.243	0.228
LLaVA-OV-7B	0.104	0.001	0.171	0.002	0.102	0.002	0.012	0.049
MedGemma-4B	0.252	0.003	0.571	0.008	0.273	0.005	0.073	0.104

Size of change within one cohort. Across cohorts a gap is resolved only on `LUMIERE`, but the cohorts differ in anatomy, imaging, sample size and task as well as in the size of change. Table 24 holds the first four fixed: the heads of Table 22, fitted on the same ten splits, and the generated answers are scored on strata of the measured change $|\Delta|$. With Qwen3-VL-8B on `LUMIERE` the gap is $+0.11$ $[-0.25, +0.49]$ for $|\Delta| \leq 0.18$, $+0.14$ $[-0.11, +0.41]$ for $0.18 < |\Delta| \leq 0.69$ and $+0.41$ $[+0.17, +0.56]$ above. Across the 8 `LUMIERE` models the gap in the small stratum ranges from $+0.05$ to

Table 10: Informativeness of the generated answers. Range is sd(prediction)/sd(truth).

Model	ρ	r	sign (%)	sd(pred)	sd(gt)	range	n
Qwen3-VL-2B	0.025	0.037	50.1	0.148	1.128	0.13	5816
Qwen3-VL-4B	0.100	0.134	52.2	0.523	1.128	0.46	5816
Qwen3-VL-8B	0.123	0.149	51.3	0.317	1.128	0.28	5816
Qwen3-VL-32B	0.352	0.376	41.0	0.663	1.128	0.59	5816
Qwen2.5-VL-7B	0.036	0.051	49.3	0.229	1.134	0.20	5746
InternVL3.5-8B	0.096	0.156	46.6	0.199	1.124	0.18	5784
InternVL3.5-14B	0.323	0.368	49.5	0.553	1.128	0.49	5813
LLaVA-OV-7B	0.012	0.010	50.0	0.066	1.128	0.06	5816
MedGemma-4B	0.033	0.082	49.9	0.229	1.148	0.20	5616

Table 11: The label-free bound in per-edge RMS units, the matched triangle comparator on forward edges, and the cohort certificate.

		all directed edges			forward edges (matched)						certificate		
Model	n	proj.	true	tight	proj.	triangle	true	tight_p	tight_t	$\sin\theta$	boot	KL	EB
Qwen3-VL-2B	555	0.176	0.770	22.9	0.058	0.052	0.772	7.5	6.7	0.988	0.214	0.170	0.165
Qwen3-VL-4B	555	0.344	0.809	42.5	0.192	0.163	0.779	24.6	20.9	0.905	0.297	0.251	0.244
Qwen3-VL-8B	555	0.314	0.808	38.8	0.108	0.094	0.808	13.4	11.7	0.935	0.329	0.279	0.277
Qwen3-VL-32B	555	0.308	0.754	40.8	0.163	0.140	0.735	22.1	19.0	0.714	0.320	0.259	0.243
Qwen2.5-VL-7B	555	0.216	0.797	27.1	0.064	0.056	0.802	8.0	7.0	0.965	0.191	0.144	0.147
InternVL3.5-8B	555	0.209	0.747	28.0	0.069	0.059	0.752	9.1	7.8	0.971	0.194	0.150	0.148
InternVL3.5-14B	555	0.291	0.713	40.8	0.167	0.143	0.698	24.0	20.5	0.846	0.255	0.210	0.199
LLaVA-OV-7B	555	0.090	0.735	12.3	0.036	0.032	0.737	4.9	4.3	0.989	0.093	0.058	0.053
MedGemma-4B	549	0.290	0.816	35.6	0.099	0.086	0.871	11.4	9.9	0.974	0.279	0.228	0.231

Table 12: Error, cocycle violation, identity fabrication and the bound check per model and substrate.

Model	Cohort	n	Error	Cocycle	Identity	Bound
Qwen3-VL-2B	LUMIERE	3058	1.199	0.164	0.122	yes
Qwen3-VL-2B	OASIS2	1948	0.023	0.020	0.018	yes
Qwen3-VL-2B	SPACENET	360	0.261	0.152	0.141	yes
Qwen3-VL-2B	VS	450	0.605	0.418	0.333	yes
Qwen3-VL-4B	LUMIERE	3058	1.247	0.774	0.351	yes
Qwen3-VL-4B	OASIS2	1948	0.030	0.024	0.025	yes
Qwen3-VL-4B	SPACENET	360	0.286	0.212	0.134	yes
Qwen3-VL-4B	VS	450	0.534	0.323	0.208	yes
Qwen3-VL-8B	LUMIERE	3058	1.197	0.366	0.224	yes
Qwen3-VL-8B	OASIS2	1948	0.062	0.072	0.000	yes
Qwen3-VL-8B	SPACENET	360	0.245	0.184	0.000	yes
Qwen3-VL-8B	VS	450	0.693	0.368	0.144	yes
Qwen3-VL-32B	LUMIERE	3058	1.078	0.626	0.026	yes
Qwen3-VL-32B	OASIS2	1948	0.017	0.009	0.000	yes
Qwen3-VL-32B	SPACENET	360	0.171	0.036	0.000	yes
Qwen3-VL-32B	VS	450	0.816	0.490	0.073	yes
Qwen2.5-VL-7B	LUMIERE	3058	1.197	0.178	0.085	yes
Qwen2.5-VL-7B	OASIS2	1945	0.175	0.139	0.091	yes
Qwen2.5-VL-7B	SPACENET	293	0.183	0.076	0.008	yes
Qwen2.5-VL-7B	VS	450	0.479	0.175	0.055	yes
InternVL3.5-8B	LUMIERE	3033	1.185	0.277	0.069	yes
InternVL3.5-8B	OASIS2	1943	0.018	0.007	0.002	yes
InternVL3.5-8B	SPACENET	360	0.190	0.076	0.034	yes
InternVL3.5-8B	VS	448	0.513	0.234	0.096	yes
InternVL3.5-14B	LUMIERE	3058	1.080	0.670	0.064	yes
InternVL3.5-14B	OASIS2	1945	0.022	0.014	0.000	yes
InternVL3.5-14B	SPACENET	360	0.178	0.043	0.000	yes
InternVL3.5-14B	VS	450	0.573	0.425	0.001	yes
LLaVA-OV-7B	LUMIERE	3058	1.195	0.125	0.115	yes
LLaVA-OV-7B	OASIS2	1948	0.023	0.027	0.095	yes
LLaVA-OV-7B	SPACENET	360	0.227	0.104	0.114	yes
LLaVA-OV-7B	VS	450	0.439	0.118	0.096	yes
MedGemma-4B	LUMIERE	3058	1.227	0.306	0.328	yes
MedGemma-4B	OASIS2	1748	0.070	0.107	0.062	yes
MedGemma-4B	SPACENET	360	0.367	0.293	0.225	yes
MedGemma-4B	VS	450	0.537	0.331	0.302	yes

Table 13: Identity, leading-note and blind conditions in full.

Model	$\lvert\widehat{\Delta}(i \to i)\rvert$	+lead	prog. (%)	+lead (%)	blind	imaged	benefit
Qwen3-VL-2B	0.095	0.121	99.7	100.0	0.584	0.629	−0.045
Qwen3-VL-4B	0.177	0.294	40.8	64.2	0.643	0.634	0.009
Qwen3-VL-8B	0.102	0.214	49.8	64.0	0.595	0.652	−0.058
Qwen3-VL-32B	0.017	0.058	31.7	48.1	0.608	0.600	0.009
Qwen2.5-VL-7B	0.081	0.203	87.2	66.0	0.219	0.241	−0.022
InternVL3.5-8B	0.039	0.124	31.7	59.1	0.571	0.583	−0.012
InternVL3.5-14B	0.026	0.098	17.6	40.0	0.485	0.460	0.025
LLaVA-OV-7B	0.104	0.205	45.3	64.2	0.587	0.586	0.001
MedGemma-4B	0.252	0.293	59.6	78.7	0.632	0.676	−0.044

Table 14: Sign accuracy (%) at each deformation magnitude (absolute log volume ratio); chance is 50%.

Model	0.02	0.05	0.10	0.20	0.40	0.70
Qwen3-VL-2B	50.0	50.0	50.0	50.0	50.0	50.0
Qwen3-VL-4B	40.8	43.3	38.3	41.7	42.5	44.2
Qwen3-VL-8B	51.7	50.8	50.8	50.8	50.0	49.2
Qwen3-VL-32B	48.3	50.0	46.7	51.7	57.5	55.0
Qwen2.5-VL-7B	50.0	50.0	50.0	50.0	50.0	50.0
InternVL3.5-8B	44.2	49.2	45.0	47.9	49.2	51.7
InternVL3.5-14B	41.7	40.0	40.0	41.7	53.3	62.5
LLaVA-OV-7B	50.0	49.2	50.8	51.7	50.0	50.0
MedGemma-4B	50.0	50.0	50.0	50.0	50.0	50.0

Table 15: Projected decoding in full, all decoders on the same $K = 8$ samples.

Model	n	1-samp	self-cons	TCR	TCR-iso	vs SC (%)	vs 1-s (%)	$\lvert\mathcal{R}\rvert_{\mathrm{SC}}$	$\lvert\mathcal{R}\rvert_{\mathrm{TCR}}$
Qwen3-VL-2B	444	1.206	0.780	0.739	0.735	5.2	38.7	0.160	0.000
Qwen3-VL-4B	444	1.202	0.722	0.660	0.654	8.5	45.1	0.333	0.000
Qwen3-VL-8B	444	1.229	0.772	0.723	0.714	6.4	41.2	0.215	0.000
Qwen3-VL-32B	444	1.062	0.644	0.621	0.619	3.7	41.6	0.216	0.000
Qwen2.5-VL-7B	444	1.184	0.740	0.732	0.730	1.1	38.2	0.090	0.000
InternVL3.5-8B	444	1.177	0.742	0.722	0.719	2.7	38.6	0.151	0.000
InternVL3.5-14B	444	1.077	0.635	0.619	0.617	2.4	42.5	0.206	0.000
LLaVA-OV-7B	444	2.171	0.787	0.743	0.741	5.7	65.8	0.151	0.000
MedGemma-4B	444	1.228	0.824	0.733	0.728	11.0	40.3	0.234	0.000

Table 16: Plug-in pairwise ceiling as a function of the evidence bin width.

Bin width	Classes	Ceiling (%)	Majority (%)	Singletons (%)
1.00	10	50.2	49.4	0.0
0.50	19	51.3	49.4	0.0
0.25	36	52.5	49.4	0.1
0.10	82	53.8	49.4	0.2
0.05	148	56.3	49.4	0.4
0.02	337	61.0	49.4	3.5

Table 17: Scale-invariant inconsistency angle by model family and size.

Model	Family	Params	$\sin\theta$
Qwen3-VL-2B	Qwen3-VL	2B	0.988
MedGemma-4B	MedGemma	4B	0.974
Qwen3-VL-4B	Qwen3-VL	4B	0.905
LLaVA-OV-7B	LLaVA-OV	7B	0.989
Qwen2.5-VL-7B	Qwen2.5-VL	7B	0.965
InternVL3.5-8B	InternVL3.5	8B	0.971
Qwen3-VL-8B	Qwen3-VL	8B	0.935
InternVL3.5-14B	InternVL3.5	14B	0.846
Qwen3-VL-32B	Qwen3-VL	32B	0.714

Table 18: Supervised and cocycle-only readout heads on frozen representations.

Model / head	objective	ρ	(sd)	sign (%)	MAE_{aff}	MAE_0	sd(pred)	collapse (%)	seeds
Qwen3-VL-8B, linear	supervised	0.508	0.046	68.9	1.002	1.174	0.879	0.0	10
Qwen3-VL-8B, linear	cocycle-only	0.449	0.072	66.4	1.048	1.174	0.976	0.0	10
Qwen3-VL-8B, MLP-256	supervised	0.477	0.056	66.9	1.016	1.174	0.914	0.0	10
Qwen3-VL-8B, MLP-256	cocycle-only	0.253	0.093	58.2	1.123	1.174	1.003	0.0	10
Qwen3-VL-32B, linear	supervised	0.504	0.042	68.5	1.013	1.174	0.849	0.0	10
Qwen3-VL-32B, linear	cocycle-only	0.448	0.073	65.7	1.049	1.174	0.963	0.0	10

Table 19: Readout heads against the generated answer on identical held-out items (Qwen3-VL-8B, LUMIERE). Upper block: orientation and scale fitted on test labels; lower block: deployable.

Estimator	ρ	$\lvert\rho\rvert$	sign (%)	MAE	MAE_0	n
sign-invariant diagnostic (orientation and scale fitted on the test labels)						
supervised probe	$+0.508 \pm 0.046$	0.508 ± 0.046	68.9 ± 2.5	1.002 ± 0.094	1.174 ± 0.127	544 ± 32
cocycle-only head	–	0.449 ± 0.072	66.4 ± 3.5	1.048 ± 0.122	1.174 ± 0.127	544 ± 32
generated answer	$+0.137 \pm 0.051$	0.137 ± 0.051	49.4 ± 3.1	1.151 ± 0.127	1.174 ± 0.127	544 ± 32
deployable (orientation and calibration from validation subjects; sign read from the prediction)						
supervised probe	$+0.508 \pm 0.046$	–	68.8 ± 2.8	1.011 ± 0.093	1.174 ± 0.127	544 ± 32
cocycle-only head	$+0.449 \pm 0.072$	–	66.1 ± 3.7	1.058 ± 0.115	1.174 ± 0.127	544 ± 32
generated answer (raw)	$+0.137 \pm 0.051$	–	55.1 ± 1.5	1.194 ± 0.123	1.174 ± 0.127	544 ± 32
generated answer, calibrated on validation	$+0.137 \pm 0.051$	–	53.6 ± 1.2	1.156 ± 0.127	1.174 ± 0.127	544 ± 32

Table 20: Readout heads on representations with the images removed, shuffled or reduced to one scan (Qwen3-VL-8B, LUMIERE).

Representation / input	supervised ρ	cocycle-only $\lvert\rho\rvert$
image pair, as asked	$+0.508 \pm 0.046$	0.449 ± 0.072
text only (no images; interval length visible)	$+0.002 \pm 0.036$	0.011 ± 0.011
images shuffled across subjects, same text	$+0.038 \pm 0.041$	0.020 ± 0.015
first scan supplied twice	$+0.307 \pm 0.041$	0.134 ± 0.058
second scan supplied twice	$+0.312 \pm 0.039$	0.155 ± 0.081
two single-scan representations, concatenated	$+0.548 \pm 0.046$	0.396 ± 0.134
metadata only: $\lvert\Delta t\rvert$, n_{tp} (what the prompt states)	$+0.000 \pm 0.000$	–
metadata oracle: signed Δt and visit indices	$+0.023 \pm 0.109$	–
signed Δt alone ($\widehat{\Delta}_{ij} \propto t_j - t_i$)	-0.012 ± 0.145	–
corr. of predictions on a pair and on the same pair reversed	-0.622 ± 0.055	-0.699 ± 0.104

Table 21: Supervised and label-free heads by training objective (Qwen3-VL-8B, LUMIERE), scored sign-invariantly.

Head	$\lvert\rho\rvert$	% of sup.	antisym.	labels
paired representation $h(I_i, I_j)$, linear head $w^\top h + b$				
supervised (labels)	0.508 ± 0.046	100	-0.61 ± 0.05	–
variance term only, λ relu$(1 - \text{Var})$	0.081 ± 0.055	16	$+0.83 \pm 0.09$	no
antisymmetry only + variance	0.502 ± 0.069	99	-0.87 ± 0.04	no
cocycle only + variance (Section 4.2)	0.449 ± 0.072	88	-0.69 ± 0.10	no
cocycle + antisymmetry + variance	0.495 ± 0.074	98	-0.88 ± 0.04	no
no constraint, no training				
first principal component	0.087 ± 0.023	17	$+0.90 \pm 0.02$	no
best of top ten PCs, chosen on validation labels	0.286 ± 0.075	56	$+0.17 \pm 0.16$	selection
tied head $w^\top (h_j - h_i)$, no bias, on one canonical representation per visit (fixed prompt)				
supervised	0.499 ± 0.063	98	-1.00 ± 0.00	–
cocycle + variance (cocycle residual identically zero)	0.318 ± 0.101	63	-1.00 ± 0.00	no
tied head on the interval-conditioned single-scan representations (context-dependent)				
supervised	0.533 ± 0.038	105	-1.00 ± 0.00	–
cocycle + variance	0.379 ± 0.129	75	-1.00 ± 0.00	no

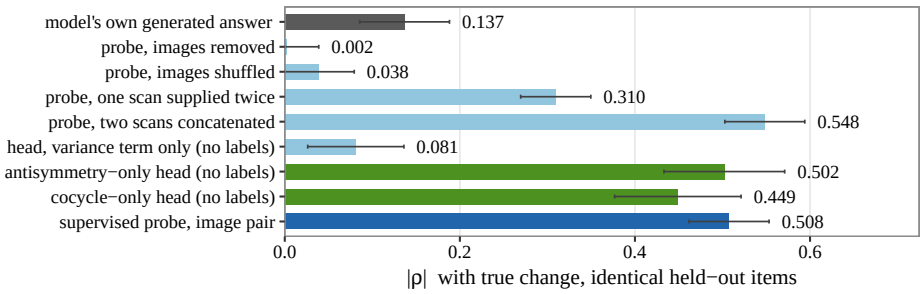

Figure 3: Correlation with measured change on identical held-out items (Qwen3-VL-8B, LUMIERE) for the generated answer, heads under input controls, the label-free head and the supervised head.

Table 22: The readout diagnosis across models and cohorts. 'gap' is the subject-level interval for the supervised head minus the generated answer.

Model / cohort	n	sup. ρ	gen. ρ	gap [95% CI]	text	shuffled	one scan	cocycle $\lvert\rho\rvert$	antisym. $\lvert\rho\rvert$	var. only
Qwen3-VL-2B / LUMIERE	544	+0.456	+0.031	[+0.22, +0.58]	−0.003	+0.029	+0.322	0.357	0.440	0.091
Qwen3-VL-4B / LUMIERE	544	+0.476	+0.108	[+0.18, +0.54]	+0.004	+0.035	+0.323	0.423	0.460	0.034
Qwen3-VL-8B / LUMIERE	544	+0.508	+0.137	[+0.16, +0.50]	+0.002	+0.038	+0.310	0.449	0.502	0.081
Qwen3-VL-32B / LUMIERE	544	+0.504	+0.398	[−0.00, +0.24]	+0.039	−0.008	+0.297	0.448	0.478	0.055
Qwen2.5-VL-7B / LUMIERE	544	+0.293	+0.082	[+0.03, +0.37]	+0.002	−0.010	+0.240	0.245	0.220	0.066
InternVL3.5-8B / LUMIERE	543	+0.582	+0.206	[+0.23, +0.50]	+0.001	−0.000	+0.335	0.509	0.547	0.084
LLaVA-OV-7B / LUMIERE	544	+0.451	+0.033	[+0.21, +0.60]	+0.003	+0.004	+0.335	0.154	0.382	0.028
MedGemma-4B† / LUMIERE	544	+0.544	+0.085	[+0.30, +0.59]	+0.000	+0.033	+0.359	0.480	0.504	0.049
Qwen3-VL-8B / OASIS-2	336	+0.052	+0.023	[−0.19, +0.28]	+0.005	−0.036	+0.049	0.044	0.340	0.022
Qwen3-VL-8B / VS-MC-RC	90	+0.098	−0.022	[−0.19, +0.39]	−0.007	−0.038	+0.161	0.085	0.111	0.037
Qwen3-VL-8B / SpaceNet-7	72	+0.109	+0.117	[−0.51, +0.44]	+0.000	+0.016	+0.070	0.128	0.192	0.090

+0.29 (median +0.16) and in the large stratum from +0.10 to +0.54 (median +0.42); under the matched-analysis summary the large stratum resolves a gap for 7 of them and the small stratum for 0 (under the pooled-row summary, 8 and 3). The two summaries agree on the ordering of the strata and differ in width, because a stratum holds only a few subjects per split; the small-stratum estimates are positive for every model, so a smaller gap at small changes within the glioma cohort is not excluded, and the size of change accounts for part of the difference between cohorts, not necessarily all of it. The mirror test on VS-MC-RC, whose changes are mostly small, gives a positive estimate only in its largest-change stratum (+0.31, 22 subjects of 182 pooled test rows); the pooled-row interval resolves it ([+0.06, +0.49]) and the matched-analysis interval does not ([−0.63, +1.12]), so it is a lead rather than a result. SpaceNet-7 has too few large changes to test. Stratifying on the label restricts the range inside each stratum and lowers every correlation, the head's and the answer's alike; the comparison is between the two on the same rows. This supports, without ruling out every alternative, that the size of change is a large part of what separates the cohorts.

An intervention on magnitude, and why it is inconclusive. Stratifying real pairs by measured change varies subjects and intervals along with the magnitude. The radial-warp probes of Appendix K do not: each of 30 subjects contributes one scan and twelve copies with the lesion rescaled by a known factor, so between probes of one subject only the magnitude varies. Table 25 applies the heads of Table 19, fitted on the real pairs of each split, to the warp probes of that split's held-out subjects, and also trains a head on the warp probes themselves under a subject-level split. The design isolates magnitude; the result does not settle the question. The head trained on real change transfers weakly to the synthetic change (ρ with the imposed log-ratio +0.16 for Qwen3-VL-8B; 0 of 8 models resolve a gap against the answer), a head trained on the warps reads them only in part (+0.18;

Table 23: Change or visit order: the antisymmetry head on OASIS-2 and LUMIERE; $o = \text{sign}(t_j - t_i)$.

	order	antisymmetry head, ρ with					supervised
Cohort	oracle	change	order o	change, fwd	change, rev	change | o	change | o
LUMIERE (glioma, not monotone)	−0.042 ± 0.125	+0.502 ± 0.069	−0.191 ± 0.134	+0.492 ± 0.077	+0.503 ± 0.064	+0.494 ± 0.066	+0.502 ± 0.044
OASIS-2 (atrophy, monotone)	−0.609 ± 0.133	+0.340 ± 0.146	−0.563 ± 0.111	+0.030 ± 0.094	−0.014 ± 0.089	+0.016 ± 0.087	+0.038 ± 0.053

Table 24: The readout gap by size of measured change within each cohort; 'split' and 'pooled' are two subject-level bootstrap summaries.

Model / cohort	$\lvert\Delta\rvert$	n	subj.	mean $\lvert\Delta\rvert$	sup. ρ	gen. ρ	gap	95% CI, split	95% CI, pooled	cocycle $\lvert\rho\rvert$	zero (%)
Qwen3-VL-2B / LUMIERE	≤ 0.18	606	28	0.10	+0.318	+0.027	+0.291	[−0.26, +0.83]	[+0.02, +0.42]	0.280	0
	0.18–0.69	1596	55	0.42	+0.190	+0.034	+0.156	[−0.12, +0.51]	[+0.04, +0.27]	0.129	0
	> 0.69	3234	67	1.74	+0.543	+0.036	+0.507	[+0.29, +0.66]	[+0.42, +0.57]	0.429	0
Qwen3-VL-4B / LUMIERE	≤ 0.18	606	28	0.10	+0.258	+0.075	+0.183	[−0.16, +0.53]	[+0.01, +0.31]	0.360	0
	0.18–0.69	1596	55	0.42	+0.164	+0.020	+0.144	[−0.18, +0.62]	[+0.02, +0.30]	0.158	0
	> 0.69	3234	67	1.74	+0.558	+0.149	+0.409	[+0.17, +0.61]	[+0.29, +0.50]	0.509	0
Qwen3-VL-8B / LUMIERE	≤ 0.18	606	28	0.10	+0.293	+0.185	+0.108	[−0.25, +0.49]	[−0.06, +0.29]	0.387	0
	0.18–0.69	1596	55	0.42	+0.203	+0.063	+0.140	[−0.11, +0.41]	[+0.04, +0.24]	0.169	0
	> 0.69	3234	67	1.74	+0.577	+0.172	+0.405	[+0.17, +0.56]	[+0.34, +0.48]	0.541	0
Qwen3-VL-32B / LUMIERE	≤ 0.18	606	28	0.10	+0.230	+0.180	+0.050	[−0.34, +0.69]	[−0.11, +0.17]	0.290	5
	0.18–0.69	1596	55	0.42	+0.202	+0.153	+0.049	[−0.17, +0.30]	[−0.03, +0.11]	0.179	3
	> 0.69	3234	67	1.74	+0.578	+0.479	+0.100	[−0.01, +0.23]	[+0.04, +0.15]	0.529	1
Qwen2.5-VL-7B / LUMIERE	≤ 0.18	606	28	0.10	+0.260	+0.042	+0.218	[−0.22, +0.58]	[+0.02, +0.37]	0.272	0
	0.18–0.69	1596	55	0.42	+0.080	+0.039	+0.041	[−0.25, +0.35]	[−0.07, +0.16]	0.134	0
	> 0.69	3234	67	1.74	+0.359	+0.105	+0.254	[+0.03, +0.45]	[+0.18, +0.34]	0.304	0
InternVL3.5-8B / LUMIERE	≤ 0.18	606	28	0.10	+0.259	+0.136	+0.124	[−0.22, +0.55]	[−0.03, +0.27]	0.265	1
	0.18–0.69	1596	55	0.42	+0.213	+0.049	+0.164	[−0.09, +0.41]	[+0.07, +0.26]	0.174	2
	> 0.69	3229	67	1.74	+0.674	+0.256	+0.418	[+0.28, +0.54]	[+0.37, +0.47]	0.607	1
LLaVA-OV-7B / LUMIERE	≤ 0.18	606	28	0.10	+0.303	+0.143	+0.161	[−0.26, +0.56]	[−0.04, +0.30]	0.132	0
	0.18–0.69	1596	55	0.42	+0.147	−0.003	+0.150	[−0.13, +0.45]	[+0.06, +0.25]	0.040	0
	> 0.69	3234	67	1.74	+0.542	+0.042	+0.500	[+0.24, +0.70]	[+0.38, +0.60]	0.187	0
MedGemma-4B† / LUMIERE	≤ 0.18	606	28	0.10	+0.154	+0.097	+0.057	[−0.52, +0.62]	[−0.19, +0.30]	0.279	0
	0.18–0.69	1596	55	0.42	+0.152	+0.004	+0.148	[−0.18, +0.42]	[+0.02, +0.26]	0.166	0
	> 0.69	3234	67	1.74	+0.655	+0.119	+0.536	[+0.38, +0.69]	[+0.47, +0.59]	0.598	0
Qwen3-VL-8B / OASIS-2	≤ 0.18	3360	137	0.02	+0.052	+0.023	+0.029	[−0.19, +0.28]	[−0.03, +0.08]	0.044	3
Qwen3-VL-8B / VS-MC-RC	≤ 0.18	462	51	0.07	+0.009	+0.080	−0.071	[−0.52, +0.38]	[−0.24, +0.11]	0.113	1
	0.18–0.69	256	36	0.37	+0.147	−0.046	+0.193	[−0.30, +0.66]	[−0.04, +0.32]	0.189	0
	> 0.69	182	22	1.43	+0.222	−0.086	+0.309	[−0.63, +1.12]	[+0.06, +0.49]	0.268	0
Qwen3-VL-8B / SpaceNet-7	≤ 0.18	612	25	0.05	+0.072	+0.033	+0.039	[−0.34, +0.45]	[−0.06, +0.17]	0.098	33
	0.18–0.69	70	9	0.26	+0.294	+0.031	+0.263	[−1.17, +1.49]	[−0.14, +0.70]	0.487	37
	> 0.69	38	3	1.82	+0.012	+0.753	−0.741	–	–	0.699	24

1 of 8 resolved), and the generated answers sit at chance at every level for most models, including a doubling of the lesion (sign accuracy 47% at $|\Delta| = 0.7$ for Qwen3-VL-8B, against 59% for the transferred head and 66% for the warp-trained head). Two readings are consistent with this: the representation encodes a synthetically rescaled lesion differently from real progression, so the transfer does not test the same information; and 30 subjects give little power. Either way the intervention does not establish that magnitude causes the gap, and we do not claim it: the dependence on the size of change within cohorts stands as an association.

Table 25: The readout gap under an imposed magnitude, for heads transferred from real pairs and heads trained on the warps.

	pooled over levels, ρ with imposed Δ					sign accuracy (%) at imposed $\lvert\Delta\rvert$: transfer head / warp head / answer					
Model	transfer	warp	answer	gap, transfer	gap, warp	0.02	0.05	0.1	0.2	0.4	0.7
Qwen3-VL-2B	+0.033	+0.082	+0.029	[−0.41, +0.47]	[−0.51, +0.40]	52 / 49 / 50	48 / 47 / 50	55 / 52 / 50	47 / 52 / 50	53 / 50 / 50	54 / 56 / 50
Qwen3-VL-4B	+0.031	+0.105	+0.053	[−0.60, +0.34]	[−0.25, +0.46]	53 / 45 / 52	47 / 52 / 55	48 / 52 / 47	48 / 47 / 51	47 / 53 / 54	55 / 53 / 51
Qwen3-VL-8B	+0.162	+0.183	−0.039	[−0.12, +0.52]	[−0.27, +0.57]	51 / 48 / 49	50 / 51 / 49	48 / 51 / 49	51 / 52 / 48	55 / 58 / 50	59 / 66 / 47
Qwen3-VL-32B	+0.085	+0.255	+0.201	[−0.51, +0.29]	[−0.39, +0.42]	46 / 54 / 15	50 / 54 / 20	48 / 53 / 22	49 / 52 / 28	54 / 61 / 45	56 / 76 / 48
Qwen2.5-VL-7B	+0.062	+0.118	+0.019	[−0.20, +0.48]	[−0.37, +0.43]	48 / 54 / 50	46 / 53 / 50	50 / 47 / 50	49 / 48 / 50	49 / 54 / 50	54 / 63 / 50
InternVL3.5-8B	+0.178	+0.322	+0.068	[−0.15, +0.41]	[+0.01, +0.58]	52 / 52 / 32	54 / 54 / 40	49 / 53 / 40	52 / 61 / 46	59 / 64 / 49	67 / 79 / 52
LLaVA-OV-7B	+0.145	+0.135	+0.046	[−0.19, +0.42]	[−0.20, +0.42]	54 / 54 / 50	61 / 52 / 50	48 / 49 / 50	49 / 52 / 50	53 / 58 / 50	61 / 57 / 50
MedGemma-4B†	+0.044	+0.106	+0.068	[−0.36, +0.27]	[−0.32, +0.37]	48 / 52 / 50	48 / 52 / 50	47 / 52 / 50	50 / 52 / 50	52 / 55 / 50	57 / 62 / 50

Table 26: InternVL3.5-8B answers from the original release and from the converted checkpoint that produced its representation.

Answers from	parse (%)	zero (%)	distinct	ρ truth	sign (%)	MAE	sup. ρ	gen. ρ	gap [95% CI]
original release, vLLM	99.9	1.0	21	+0.250	50.5	1.187	+0.582	+0.206	[+0.23, +0.50]
converted checkpoint, transformers	99.9	0.8	24	+0.249	50.7	1.186	+0.582	+0.212	[+0.23, +0.51]

Table 27: Pairwise rank correlations among the quantities reported by the source benchmark.

Pair	Spearman ρ	p
SignAcc vs TAC	0.971	4.72e-10
RS vs TAC	0.937	9.25e-08
RS vs SignAcc	0.929	1.96e-07
SignAcc vs TEDS	0.926	2.56e-07
TAC vs TEDS	0.906	1.36e-06
TAC vs TrendF1	0.900	2.05e-06
RS vs TEDS	0.898	2.29e-06
RS vs TrendF1	0.848	3.29e-05
SignAcc vs TrendF1	0.812	1.34e-04
TEDS vs TrendF1	0.729	1.34e-03
FinalAcc vs TEDS	0.706	2.25e-03
FinalAcc vs RS	0.686	3.33e-03
FinalAcc vs TAC	0.679	3.79e-03
FinalAcc vs TrendF1	0.629	8.99e-03
FinalAcc vs SignAcc	0.606	1.29e-02
Coverage vs SignAcc	0.509	4.41e-02
Chronology vs FinalAcc	0.431	9.58e-02
Coverage vs TAC	0.412	1.13e-01
Chronology vs Coverage	-0.378	1.49e-01
Chronology vs TEDS	0.289	2.78e-01
Coverage vs RS	0.274	3.05e-01
Coverage vs TEDS	0.265	3.22e-01
Coverage vs TrendF1	0.215	4.25e-01
Chronology vs RS	0.173	5.22e-01
Chronology vs TrendF1	-0.133	6.24e-01
Coverage vs FinalAcc	-0.085	7.53e-01
Chronology vs SignAcc	0.018	9.48e-01
Chronology vs TAC	0.003	9.91e-01

Table 28: Sensitivity to the log-ratio stabiliser ε on LUMIERE.

Model	ε (mm^3)	n	error	predict-zero	cocycle	beats zero
Qwen3-VL-32B	10	3058	1.433	1.559	0.626	yes
Qwen3-VL-32B	100	3058	1.315	1.439	0.626	yes
Qwen3-VL-32B	1000	3058	1.078	1.191	0.626	yes
Qwen3-VL-32B	5000	3058	0.799	0.863	0.626	yes
Qwen3-VL-8B	10	3058	1.562	1.559	0.366	no
Qwen3-VL-8B	100	3058	1.443	1.439	0.366	no
Qwen3-VL-8B	1000	3058	1.197	1.191	0.366	no
Qwen3-VL-8B	5000	3058	0.885	0.863	0.366	no

Table 29: Sensitivity of sign agreement to the no-change tolerance.

Model	tolerance	sign agreement (%)
Qwen3-VL-32B	0.05	65.7
Qwen3-VL-32B	0.10	66.0
Qwen3-VL-32B	0.20	66.5
Qwen3-VL-32B	0.40	67.8
Qwen3-VL-8B	0.05	55.5
Qwen3-VL-8B	0.10	55.6
Qwen3-VL-8B	0.20	55.6
Qwen3-VL-8B	0.40	56.1

Table 30: Error and cocycle violation per substrate.

	LUMIERE		OASIS2		SPACENET		VS	
Model	err	coc	err	coc	err	coc	err	coc
Qwen3-VL-2B	1.199	0.164	0.023	0.020	0.261	0.152	0.605	0.418
Qwen3-VL-4B	1.247	0.774	0.030	0.024	0.286	0.212	0.534	0.323
Qwen3-VL-8B	1.197	0.366	0.062	0.072	0.245	0.184	0.693	0.368
Qwen3-VL-32B	1.078	0.626	0.017	0.009	0.171	0.036	0.816	0.490
Qwen2.5-VL-7B	1.197	0.178	0.175	0.139	0.183	0.076	0.479	0.175
InternVL3.5-8B	1.185	0.277	0.018	0.007	0.190	0.076	0.513	0.234
InternVL3.5-14B	1.080	0.670	0.022	0.014	0.178	0.043	0.573	0.425
LLaVA-OV-7B	1.195	0.125	0.023	0.027	0.227	0.104	0.439	0.118
MedGemma-4B†	1.227	0.306	0.070	0.107	0.367	0.293	0.537	0.331

Table 31: Fabricated change and progression rate on identity probes, with and without a leading note.

	fabricated magnitude		progression rate (%)	
Model	identity	+leading	identity	+leading
Qwen3-VL-2B	0.095	0.121	99.7	100.0
Qwen3-VL-4B	0.177	0.294	40.8	64.2
Qwen3-VL-8B	0.102	0.214	49.8	64.0
Qwen3-VL-32B	0.017	0.058	31.7	48.1
Qwen2.5-VL-7B	0.081	0.203	87.2	66.0
InternVL3.5-8B	0.039	0.124	31.7	59.1
InternVL3.5-14B	0.026	0.098	17.6	40.0
LLaVA-OV-7B	0.104	0.205	45.3	64.2
MedGemma-4B†	0.252	0.293	59.6	78.7

Table 32: Peak sign accuracy and dominant answer on the synthetic deformation sweep.

Model	peak sign acc. (%)	dominant answer (%)	JND
Qwen3-VL-2B	50.0	100.0	not estimable (constant)
Qwen3-VL-4B	44.2	50.1	not estimable
Qwen3-VL-8B	51.7	98.1	not estimable (constant)
Qwen3-VL-32B	57.5	75.4	not estimable
Qwen2.5-VL-7B	50.0	100.0	not estimable (constant)
InternVL3.5-8B	51.7	85.5	not estimable
InternVL3.5-14B	62.5	57.6	not estimable
LLaVA-OV-7B	51.7	78.9	not estimable
MedGemma-4B†	50.0	100.0	not estimable (constant)

G FIGURES

Six figures that the body refers to but has no room for. Figure 8 plots the label-free violation against ground-truth error with the certified floor drawn on it, Figure 9 gives the pooled psychometric curves, Figure 10 compares the three decoders, and Figure 11 shows all five structural violations per model. Figure 12 is the composite re-analysis and Figure 13 the identity and suggestibility behaviour. Three further figures resolve the panel means: Figure 5 gives one psychometric curve per model so no failure is hidden by averaging, Figure 6 splits error and violation by substrate, and Figure 7 plots the consistency scaling law by family.

Inconsistency falls with capacity. (Deferred from Section 4.) Being scale-invariant, the angle is comparable across models, and it fell with capacity in both families we can compare, from 0.988 at 2B to 0.714 at 32B for Qwen3-VL and from 0.971 at 8B to 0.846 at 14B for InternVL3.5, correlating with log-parameters at $r = -0.809$ ($p = 0.0082$, $n = 9$; Fig. 7). The largest model in the panel is still well above the floor, so the gain with capacity is small relative to the gap that remains.

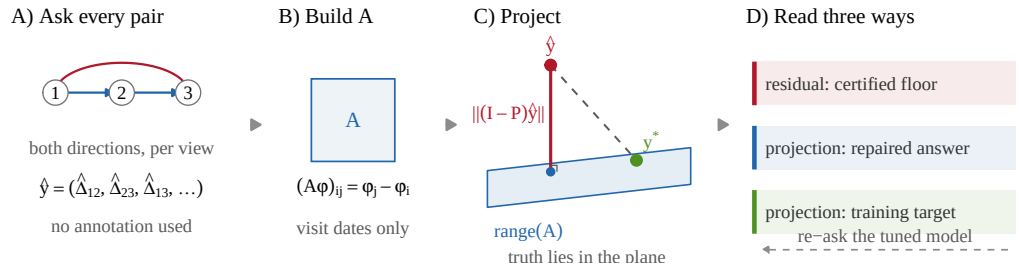

Figure 4: Geometry of the audit. (A) Stacked answers; (B) the incidence matrix and its column space; (C) the projection and its residual, which lower-bounds the error; (D) three uses of the projection.

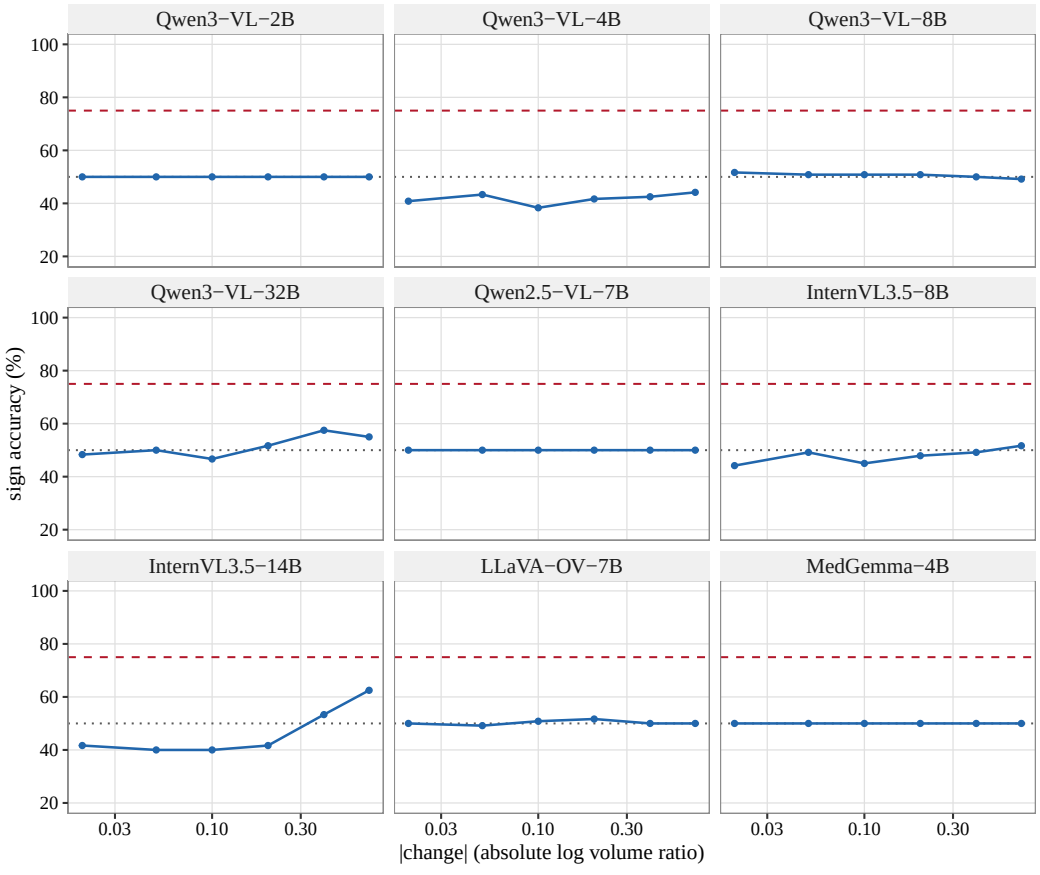

Figure 5: Psychometric curve per model; the dashed line marks the 75% criterion and the dotted line chance.

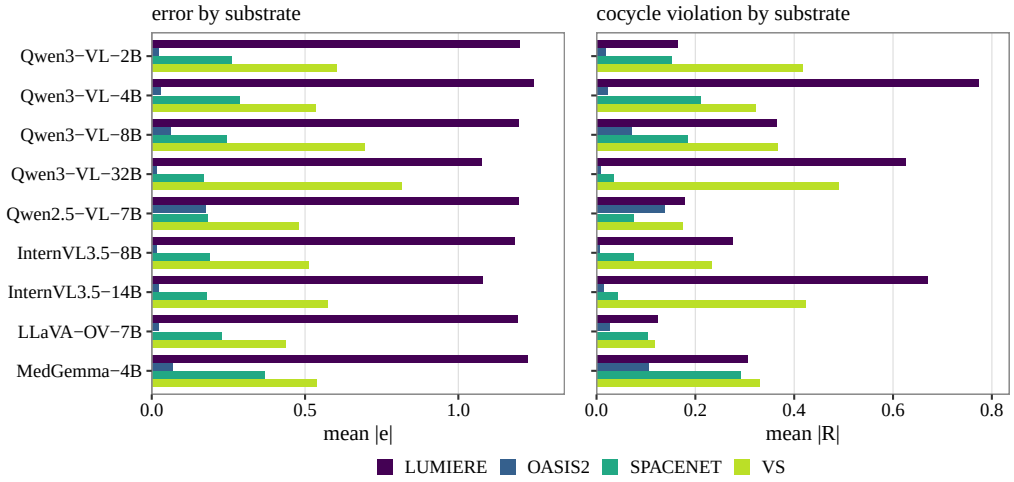

Figure 6: Error and cocycle violation by substrate.

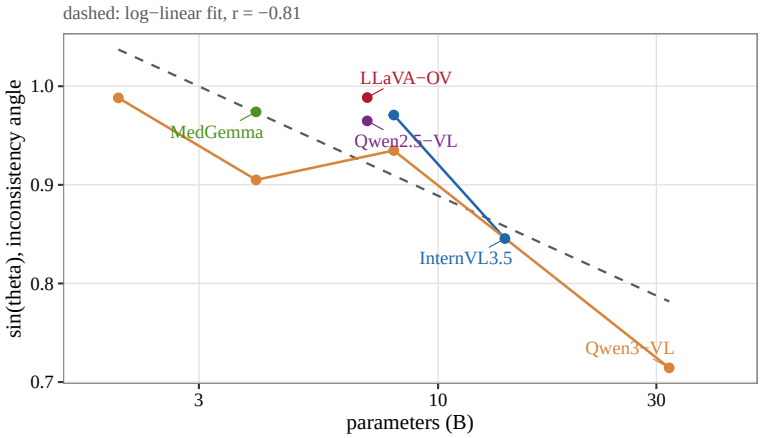

Figure 7: Inconsistency angle against model size, by family.

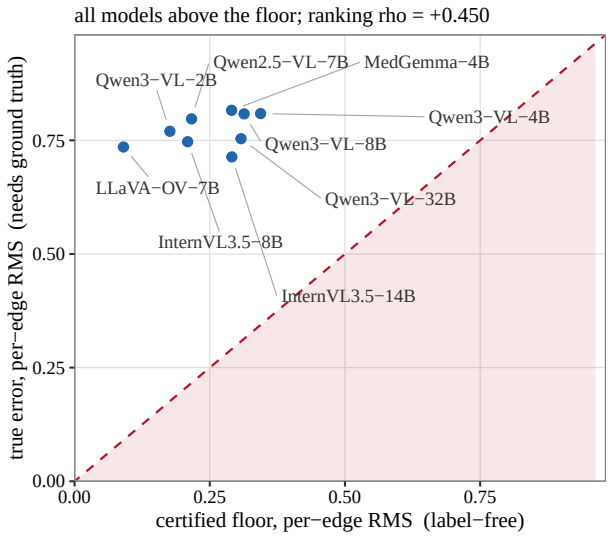

Figure 8: The label-free floor against ground-truth error per model; the region below the diagonal is unreachable.

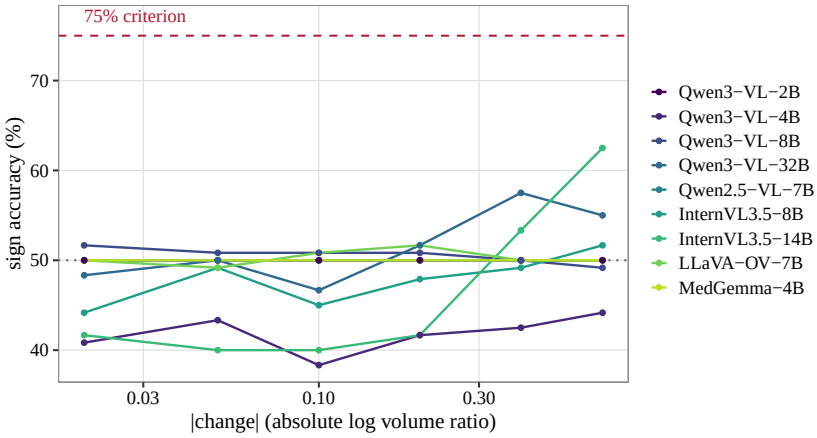

Figure 9: Sign accuracy against the magnitude of controlled deformation.

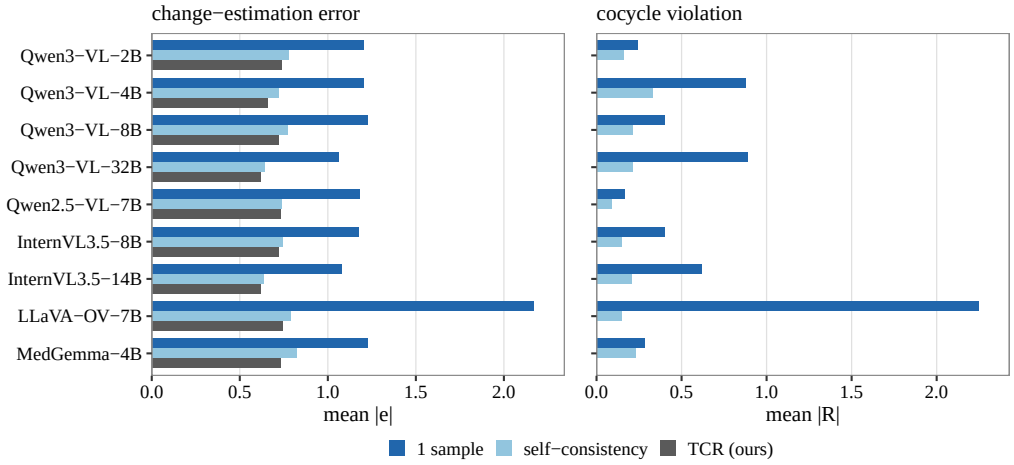

Figure 10: Projected decoding against single-sample and self-consistency decoding.

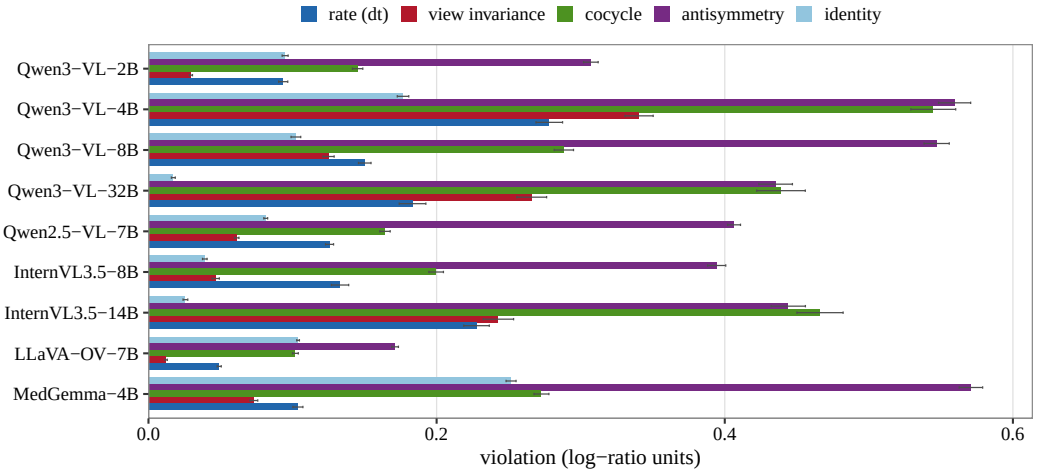

Figure 11: Structural violations per model, all computable without ground truth.

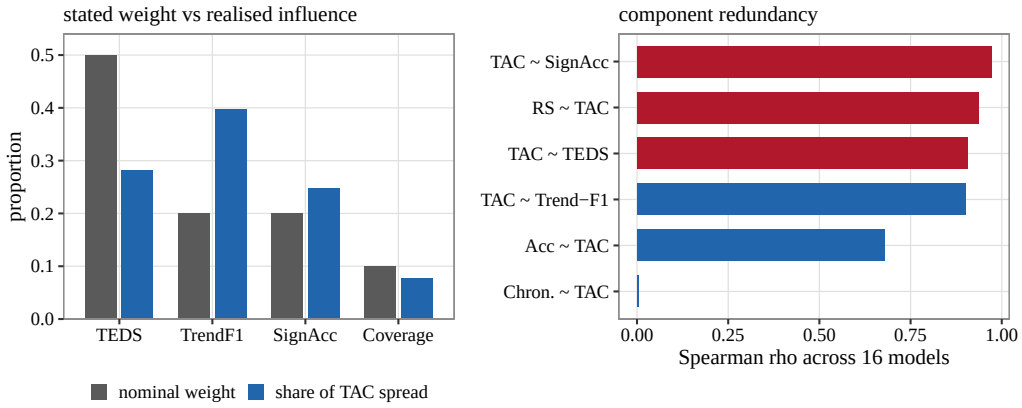

Figure 12: Audit of the source benchmark's composite: component weights against variance shares (left) and rank correlations (right).

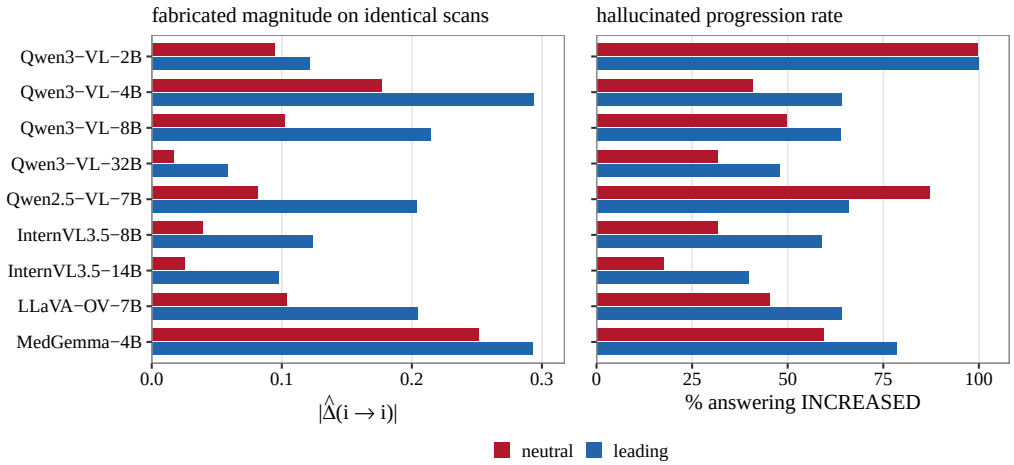

Figure 13: Answers on identity probes, with and without a leading note.

H Consistency-Projected Decoding in Full

H.1 The unconstrained projection

Let a subject have n timepoints and let E be the set of ordered pairs for which the model produced an answer, possibly several times and from several views. Write y_e for the aggregated observation on edge e and $w_e > 0$ for a weight. The decoder solves

$$\widehat{\varphi} \;=\; \arg\min_{\varphi:\varphi_1=0} \sum_{e=(i,j)\in E} w_e\big(\varphi_j - \varphi_i - y_e\big)^2, \qquad \widehat{\Delta}(i\to j) = \widehat{\varphi}_j - \widehat{\varphi}_i. \tag{12}$$

In matrix form this is weighted least squares against the incidence matrix, $\min_\varphi \|W^{1/2}(A\varphi - y)\|_2^2$, whose normal equations $A^\top W A\,\varphi = A^\top W y$ involve the weighted graph Laplacian $A^\top W A$. Fixing $\varphi_1 = 0$ removes the one-dimensional null space that Proposition 2 identifies (the additive constant); if the observed graph is disconnected, one gauge condition is needed per component. The solution is unique given that gauge, and by construction the output satisfies identity, antisymmetry and the cocycle exactly, so its residual is zero, which is why we evaluate TCR on error against ground truth rather than on the violation it removes by definition.

H.2 Weights

We set $w_e = (1 + s_e^2)^{-1}$ where s_e is the standard deviation of the samples on edge e. Sample spread is a per-edge reliability signal that costs nothing extra to collect, and down-weighting edges the model is unsure about is the natural use of it. Uniform weights change the numbers little; the ordering of decoders is unaffected.

H.3 The monotone variant

Where the underlying process is monotone (atrophy in ageing, construction in a growing city), the increments carry a sign constraint. Which cohorts qualify is a question about data, not about disease names, and we answer it by measurement in Appendix H.4. Parametrising by consecutive increments $d_t \geq 0$ and $\varphi_t = \varphi_1 + s\sum_{u<t} d_u$ with $s = \pm 1$ fixed by the direction of the process, every edge becomes a non-negative combination of increments, and the problem is a non-negative least squares in d. This is projection onto the intersection of range(A) with a convex cone, so it remains a projection and the guarantees are unchanged. We do not use it for post-treatment glioma, where the process is not monotone and the constraint would be false.

H.4 Monotonicity, measured rather than assumed

Applying a sign constraint asserts something about the disease, and an assertion of that kind should be checked against the ground truth it will be imposed on. Table 33 does so for all four substrates. Ageing atrophy holds up: 75.8% of consecutive OASIS-2 intervals lose brain volume, and no subject loses more than 40% between visits. Urban development holds up more strongly still, at 93.3%. Post-treatment glioma was never claimed to be monotone and is not.

The schwannoma prior did not survive, and we had assumed it. The VS-MC-RC release is described as vestibular schwannoma, which grows if left alone, and we initially set an increasing constraint on that basis. In the released data 46.7% of consecutive intervals *shrink*, and 21 of 75 subjects lose more than 40% of tumour volume between consecutive visits, one from $27{,}098$ to $1{,}186\,\mathrm{mm}^3$ in 230 days. Involution on that scale is radiosurgery, not natural history: this is a routine-clinical cohort, and a substantial fraction of it was treated between scans. We therefore impose no sign constraint on VS. The consequence for the reported numbers is nil, because the consistency-projected decoding experiment samples LUMIERE and SpaceNet-7 only, so the constraint was never exercised; the consequence for anyone reusing the code is not nil, and the flag is corrected in `code/probes.py` with the measurement that motivated it. We record it because the general lesson is the one this paper is about: a structural assumption that sounds obviously right is still an empirical claim, and this one was false.

Table 33: Monotonicity priors tested against the ground truth.

Cohort	assumed	shrinking (%)	median r	agree (%)	big drops	applied
LUMIERE	none	51.0	−0.0238	–	71/87	none
OASIS2	decreasing	75.8	−0.0079	75.8	0/150	decreasing
VS	increasing	46.7	+0.0576	53.3	21/75	none
SPACENET	increasing	6.7	+0.0261	93.3	1/30	increasing

H.5 Why this is not self-consistency

Self-consistency aggregates K samples of *one* quantity by majority or median. The quantities here are not independent: they are differences of a single latent trajectory, so the samples jointly over-determine $n - 1$ free parameters through $|E|$ observations. TCR exploits that; a per-edge median cannot. The comparison in Table 15 holds the sample pool fixed so the difference is attributable to the aggregator alone rather than to extra compute.

I Ablations of the Projection

Section H derives consistency-projected decoding and Table 4 reports one configuration of it. This section asks which parts of that configuration are doing the work. Every variant below re-aggregates the identical saved generations, so no number here required a GPU and every difference is attributable to the aggregator rather than to a different sample of model behaviour.

Structure substitutes for sampling. Table 34 truncates each edge's sample pool to its first K draws and re-runs both aggregators. The projection helps most where there is least to average. At $K = 1$ self-consistency degenerates to a single draw and has nothing to smooth, and the reduction TCR buys is at its largest for seven of nine models, reaching 41.5% for LLaVA-OV-7B, whose single draws are noisiest. As K grows the two converge (Fig. 14), because per-edge averaging eventually removes the same sampling noise the projection was removing structurally. This is the practical reading of the method: it is a way of spending the constraints you already have instead of the compute you would otherwise need. Two models refuse to converge: MedGemma-4B's advantage *widens* with K, from 9.9% to 11.3%, and Qwen3-VL-4B's holds above 12% throughout. Their residual gap at $K = 8$ is the part of the inconsistency that is not sampling noise at all but a systematic incoherence no amount of resampling will average away.

Everything around the projection is close to arbitrary. Table 35 varies the centre statistic, the reliability weights, the monotonicity constraint and the edge set. Across all nine models the spread between the best and worst weighting scheme is under 0.01 in log-ratio units, against a between-decoder gap an order of magnitude larger. The mean is marginally better than the median for six of nine models, which is the opposite of the usual robustness intuition and reflects that the parse failures have already been removed upstream, so the pool contains no wild values for a median to protect against. Isotonic projection helps by a hair wherever it is licensed. The only choice that matters is the edge set: dropping the long-range pairs and keeping only adjacent timepoints costs more, for all nine models, than the entire span of the weighting schemes, which is the expected sign, since the redundant long edges are precisely the over-determination the projection exists to exploit.

The three views are not three measurements. Table 36 runs the projection per view and on the pooled evidence, scored on the intersection of pairs that every configuration can answer so the columns are matched. Pooling the views is worth essentially nothing over axial alone: 0.9953 against 0.9954 for InternVL3.5-14B, and the pattern holds across the panel. The projection still improves on self-consistency *within* each view by roughly the margin it achieves overall. Anatomical views therefore behave like three correlated looks at one piece of evidence rather than three independent measurements of it, which is the same conclusion the multi-view experiment in Section 4 reaches from the opposite direction. It is worth stating what the unmatched version of this table looked like, because it was the first one we computed: pooled views appeared to reduce error from 1.00 to 0.62, an enormous gain, entirely because the pooled configuration can answer 804 pairs while a single

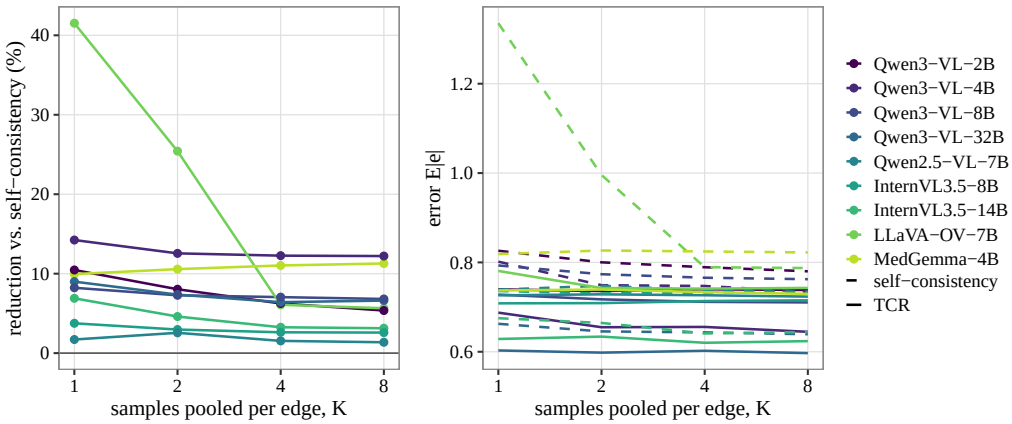

Figure 14: Error reduction of the projection over self-consistency (left) and absolute error (right) as K grows.

view can answer only 444, and the extra 360 are the easy ones. The gain was an artefact of the denominator.

Gauge invariance holds numerically. Proposition 2 fixes the latent trajectory only up to an additive constant, so re-anchoring φ at any timepoint must leave every predicted *difference* untouched. Table 37 re-solves each system under every possible anchor and reports the worst disagreement across all subjects and all nine models: 1.2×10^{-14} at worst, which is double-precision round-off on systems of this size. This is a test of the implementation rather than of the theory, but it is the kind of test that catches a sign error in an incidence matrix, and it did during development.

Table 34: Projection against self-consistency as the sample budget K grows.

	$K=1$			$K=2$			$K=4$			$K=8$		
Model	SC	TCR	%	SC	TCR	%	SC	TCR	%	SC	TCR	%
Qwen3-VL-2B	0.826	0.740	10.5	0.800	0.736	8.0	0.789	0.740	6.3	0.780	0.738	5.4
Qwen3-VL-4B	0.802	0.688	14.2	0.749	0.655	12.6	0.747	0.656	12.3	0.734	0.645	12.2
Qwen3-VL-8B	0.793	0.728	8.2	0.774	0.717	7.3	0.766	0.712	7.0	0.762	0.710	6.8
Qwen3-VL-32B	0.663	0.603	9.0	0.646	0.598	7.4	0.643	0.602	6.4	0.639	0.597	6.6
Qwen2.5-VL-7B	0.739	0.726	1.7	0.748	0.729	2.6	0.738	0.726	1.5	0.733	0.723	1.4
InternVL3.5-8B	0.736	0.708	3.8	0.731	0.709	3.0	0.733	0.714	2.6	0.734	0.715	2.6
InternVL3.5-14B	0.675	0.629	6.9	0.665	0.634	4.6	0.641	0.620	3.3	0.644	0.624	3.1
LLaVA-OV-7B	1.335	0.781	41.5	0.996	0.743	25.4	0.789	0.741	6.1	0.787	0.743	5.6
MedGemma-4B†	0.819	0.737	9.9	0.827	0.739	10.6	0.825	0.734	11.0	0.822	0.730	11.3

Table 35: Design choices inside the projection at $K = 8$.

	centre		weights				mono.	adj. only
Model	median	mean	unif.	inv-var	count	inv-span	isotonic	(no long edges)
Qwen3-VL-2B	0.738	0.737	0.738	0.738	0.738	0.738	0.734	0.740
Qwen3-VL-4B	0.645	0.650	0.650	0.645	0.650	0.651	0.639	0.671
Qwen3-VL-8B	0.710	0.706	0.713	0.710	0.713	0.714	0.702	0.718
Qwen3-VL-32B	0.597	0.599	0.597	0.597	0.597	0.598	0.595	0.614
Qwen2.5-VL-7B	0.723	0.722	0.724	0.723	0.724	0.725	0.721	0.727
InternVL3.5-8B	0.715	0.706	0.716	0.715	0.716	0.718	0.712	0.722
InternVL3.5-14B	0.624	0.619	0.627	0.624	0.627	0.630	0.622	0.653
LLaVA-OV-7B	0.743	0.742	0.743	0.743	0.743	0.743	0.741	0.747
MedGemma-4B†	0.730	0.731	0.729	0.730	0.729	0.729	0.724	0.734

Table 36: Projection by anatomical view on matched pairs.

	all views		axial		coronal		sagittal	
Model	SC	TCR	SC	TCR	SC	TCR	SC	TCR
Qwen3-VL-2B	1.211	1.199	1.211	1.199	1.211	1.202	1.209	1.201
Qwen3-VL-4B	1.115	1.038	1.115	1.038	1.143	1.091	1.177	1.145
Qwen3-VL-8B	1.193	1.146	1.193	1.146	1.211	1.168	1.206	1.172
Qwen3-VL-32B	1.023	0.952	1.023	0.952	1.080	1.035	1.084	1.048
Qwen2.5-VL-7B	1.185	1.175	1.184	1.175	1.201	1.196	1.210	1.204
InternVL3.5-8B	1.184	1.161	1.185	1.161	1.187	1.158	1.205	1.188
InternVL3.5-14B	1.028	0.995	1.028	0.995	1.058	1.011	1.047	1.024
LLaVA-OV-7B	1.212	1.207	1.212	1.207	1.208	1.203	1.212	1.208
MedGemma-4B†	1.225	1.180	1.225	1.180	1.242	1.178	1.233	1.195

Table 37: Gauge invariance: largest change in any prediction when the anchor timepoint moves.

Model	$\max \lvert \widehat{\Delta}_{\text{anchor}=0} - \widehat{\Delta}_{\text{anchor}=a} \rvert$
Qwen3-VL-2B	$2.9e-16$
Qwen3-VL-4B	$6.4e-15$
Qwen3-VL-8B	$1.1e-15$
Qwen3-VL-32B	$1.2e-14$
Qwen2.5-VL-7B	$9.2e-16$
InternVL3.5-8B	$1.1e-15$
InternVL3.5-14B	$3.4e-15$
LLaVA-OV-7B	$3.8e-16$
MedGemma-4B†	$6.8e-16$

J Where the Certificate Is Tight, and Where It Is Not

Proposition 3 is exact and assumption-free, but exactness is a statement about validity, not about usefulness. A valid bound that always returns zero would be useless. This section measures how much of the realised error the bound actually recovers, resolved along the two axes that plausibly matter: which substrate the subject comes from, and how many timepoints the subject has.

The bound is loosest exactly where the models are worst. Table 38 reports the captured fraction $\|(I-P)\hat{y}\|/\|\hat{y}-y^{\star}\|$ per model per substrate, and Fig. 15 plots it. It ranges from 0.10 to 0.98. On OASIS-2, where the true changes are small and models are close to the trivial predictor, the bound recovers up to 98% of the error: almost everything the model gets wrong there, it gets wrong incoherently. On LUMIERE, where errors are largest, the bound recovers as little as 10%. This is not a defect of the estimator but the structure of the problem. A model that is badly but *consistently* wrong has placed its error inside range(A), where no label-free method can reach it: the predictions compose perfectly and describe a coherent trajectory that simply is not the patient's. Proposition 3 certifies a floor, and the floor is informative in proportion to how much of the failure is incoherence. Reporting this is more useful than reporting the pooled mean, because it tells a practitioner when to trust the certificate.

Longer studies hide more of their error. A subject with n timepoints contributes $n-1$ free parameters against up to $n(n-1)/2$ measured edges, so the subspace the projection removes has codimension $(n-1)(n-2)/2$: one direction at $n=3$, three at $n=4$. More is removed, so more error might be expected to surface. The opposite happens. Table 39 and the right panel of Fig. 15 show the captured fraction falling for every one of the nine models when moving from three-timepoint to four-timepoint studies, from 0.645 to 0.312 for Qwen3-VL-8B, and by a similar factor throughout. The additional error incurred on longer studies is therefore predominantly *consistent* error. Models asked to describe a longer trajectory do not become more incoherent; they become more confidently wrong, which is worse, and which the label-free certificate is structurally unable to see.

The certificate survives at every confidence level we tested. Table 40 sweeps δ from 0.20 to 0.01 for both estimators. All thirty-six certificates lie below the realised error. The Bernoulli–KL bound is uniformly the more conservative, as it must be, since it assumes nothing about the shape of the

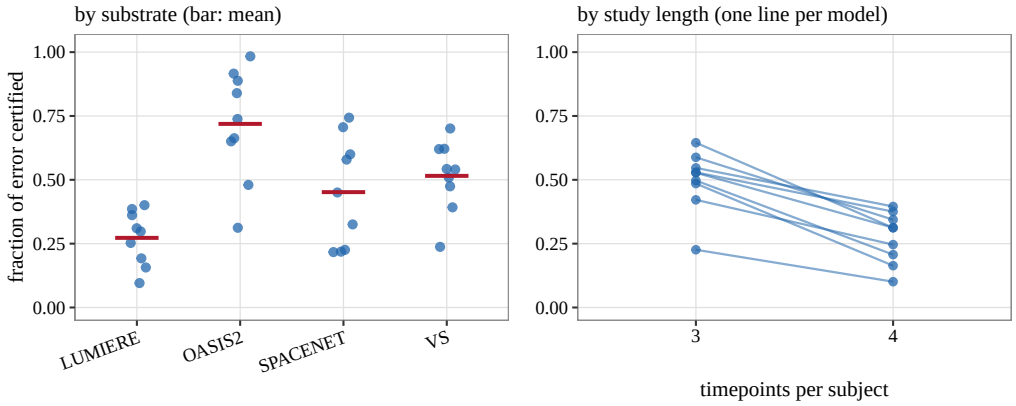

Figure 15: Share of the realised error recovered by the certificate, by substrate (left) and by number of timepoints (right).

per-subject residual distribution, only that the residuals are bounded and independent across subjects, and the price of that assumption-freedom is a factor of 1.3 against the percentile bootstrap, averaged over the panel. Both remain informative at $\delta = 0.01$. Nothing in this table consults ground truth; the final column is shown for the reader's reference and plays no part in producing any of the bounds to its left.

Table 38: The label-free bound, realised error and captured fraction by substrate.

	LUMIERE			OASIS2			SPACENET			VS		
Model	bound	error	tight	bound	error	tight	bound	error	tight	bound	error	tight
Qwen3-VL-2B	0.210	1.340	0.157	0.017	0.026	0.663	0.187	0.322	0.579	0.435	0.702	0.620
Qwen3-VL-4B	0.574	1.432	0.401	0.030	0.035	0.839	0.236	0.334	0.706	0.320	0.629	0.509
Qwen3-VL-8B	0.421	1.358	0.310	0.074	0.081	0.916	0.177	0.295	0.600	0.560	0.799	0.701
Qwen3-VL-32B	0.442	1.223	0.362	0.006	0.020	0.312	0.044	0.204	0.217	0.659	1.060	0.621
Qwen2.5-VL-7B	0.258	1.340	0.192	0.180	0.183	0.983	0.046	0.209	0.219	0.224	0.570	0.392
InternVL3.5-8B	0.338	1.337	0.253	0.010	0.020	0.480	0.073	0.225	0.325	0.282	0.593	0.475
InternVL3.5-14B	0.478	1.240	0.386	0.021	0.028	0.738	0.048	0.214	0.226	0.367	0.680	0.540
LLaVA-OV-7B	0.127	1.328	0.096	0.020	0.030	0.651	0.118	0.263	0.450	0.119	0.499	0.237
MedGemma-4B†	0.414	1.390	0.298	0.077	0.086	0.888	0.317	0.427	0.743	0.334	0.616	0.542

Table 39: The label-free bound by number of timepoints per subject.

	$n_{\text{tp}} = 3$			$n_{\text{tp}} = 4$		
Model	bound	tight	n	bound	tight	n
Qwen3-VL-2B	0.167	0.485	251	0.184	0.164	304
Qwen3-VL-4B	0.191	0.547	251	0.470	0.396	304
Qwen3-VL-8B	0.263	0.645	251	0.356	0.312	304
Qwen3-VL-32B	0.259	0.588	251	0.348	0.344	304
Qwen2.5-VL-7B	0.194	0.497	251	0.235	0.207	304
InternVL3.5-8B	0.133	0.422	252	0.272	0.246	303
InternVL3.5-14B	0.174	0.529	251	0.387	0.375	304
LLaVA-OV-7B	0.064	0.226	251	0.112	0.101	304
MedGemma-4B†	0.193	0.528	246	0.370	0.313	303

Table 40: Certified floors at four confidence levels δ.

	$\delta = 0.2$			$\delta = 0.1$			$\delta = 0.05$			$\delta = 0.01$			true
Model	boot	KL	EB	boot	KL	EB	boot	KL	EB	boot	KL	EB	error
Qwen3-VL-2B	0.223	0.186	0.186	0.218	0.177	0.175	0.214	0.170	0.165	0.207	0.157	0.143	0.770
Qwen3-VL-4B	0.309	0.270	0.267	0.303	0.260	0.255	0.297	0.251	0.244	0.287	0.236	0.220	0.809
Qwen3-VL-8B	0.340	0.298	0.299	0.334	0.288	0.288	0.329	0.279	0.277	0.320	0.263	0.254	0.808
Qwen3-VL-32B	0.338	0.277	0.267	0.328	0.267	0.255	0.320	0.259	0.243	0.305	0.243	0.217	0.754
Qwen2.5-VL-7B	0.197	0.159	0.165	0.194	0.151	0.156	0.191	0.144	0.147	0.186	0.132	0.127	0.797
InternVL3.5-8B	0.202	0.165	0.167	0.198	0.157	0.157	0.194	0.150	0.148	0.187	0.137	0.127	0.747
InternVL3.5-14B	0.268	0.227	0.221	0.261	0.218	0.210	0.255	0.210	0.199	0.245	0.195	0.175	0.713
LLaVA-OV-7B	0.095	0.067	0.069	0.094	0.062	0.061	0.093	0.058	0.053	0.090	0.050	0.036	0.735
MedGemma-4B†	0.287	0.245	0.251	0.282	0.236	0.241	0.279	0.228	0.231	0.272	0.212	0.209	0.816

K Sensitivity or Bias? Re-reading the Detectability Sweep

Table 32 reports that no model reaches 75% sign accuracy at any tested magnitude, and concludes that no just-noticeable difference is estimable. That is the correct summary of the pooled curve, and it is also incomplete in a way that changes what should be done about it. The synthetic sweep applies a smooth radial diffeomorphism to a single scan, so the two images differ only by a controlled rescaling of the lesion and the realised volume ratio is measured exactly. The stimulus set is balanced by construction: 360 growth probes and 360 shrinkage probes, 60 at each of twelve magnitudes from $|\log(V_2/V_1)| = 0.02$ to 0.7. Under that design the marginal rate at which a model answers INCREASED is directly interpretable, and it should be near one half.

Four of nine models answer INCREASED to every single probe. Table 41 gives the marginals. MedGemma-4B, Qwen2.5-VL-7B and Qwen3-VL-2B answer INCREASED on 100.0% of 720 balanced probes; Qwen3-VL-8B does so on 98.1%. Their sign accuracy is consequently 100% on growth and 0% on shrinkage, and averages to chance. The flat psychometric curve is not a sensitivity floor. It is a prior over the answer that the image evidence never overrides.

Why the distinction matters. A sensitivity floor and a response bias look identical in pooled accuracy and call for opposite remedies. A floor is an information problem: it would be relieved by higher-resolution inputs, more slices, or a stronger visual encoder. A bias is a decision problem: better inputs will not touch it, because the model is not using the inputs it already has. The two can be separated only by splitting on the direction of the true change, which is what Table 42 does. Figure 16 plots the same split per model with Wilson intervals. Read a growth row and its shrinkage row together: a model with a genuine floor improves on *both* as the change grows. A purely biased model sits near 100 on one and near 0 on the other at every magnitude, including at $|\log(V_2/V_1)| = 0.7$, a doubling of lesion volume that is unmissable to a human reader. The panel splits three ways. InternVL3.5-14B and Qwen3-VL-32B improve on both directions ($48.3 \to 95.0$ and $8.3 \to 25.0$; $66.7 \to 93.3$ and $0.0 \to 11.7$): a real sensitivity gradient sitting underneath a strong prior, and the only two models for which more visual evidence demonstrably changes the answer. The three constant responders and Qwen3-VL-8B do not move at all in either direction across a thirty-five-fold range of stimulus magnitude. Qwen3-VL-4B and LLaVA-OV-7B are the least biased in their marginals and the flattest in their curves, which is the least useful combination of the two: balanced guessing.

Consequences for the rest of the paper. This does not disturb the main results, and it sharpens two of them. The consistency floor of Section 4 is if anything understated: a model that always answers INCREASED satisfies antisymmetry badly and the cocycle in a specific degenerate way, and its low measured violation reflects the collapse of its response range rather than any grasp of the algebra, which is exactly the confound the scale-free angle $\sin\theta = \|(I - P)\hat{y}\|/\|\hat{y}\|$ was introduced to remove. And the readout result of Section 4 becomes more pointed: a representation that supports a linear decode of change at $|\rho| = 0.51$ is sitting underneath a decision rule that ignores the image entirely on this task. The information is present and the head is not using it.

Table 41: Answer rates and sign accuracy by direction on the balanced deformation sweep.

		answer given (%)			sign accuracy (%)		
Model	n	INCR.	DECR.	UNCH.	growth	shrink	gap
Qwen3-VL-2B	720	100.0	0.0	0.0	100.0	0.0	100.0
Qwen3-VL-4B	720	50.1	31.8	18.1	50.8	32.5	18.3
Qwen3-VL-8B	720	98.1	0.1	1.8	98.6	0.0	98.6
Qwen3-VL-32B	720	75.3	2.1	22.6	78.9	3.6	75.3
Qwen2.5-VL-7B	720	100.0	0.0	0.0	100.0	0.0	100.0
InternVL3.5-8B	720	87.2	0.8	11.9	88.6	1.1	87.5
InternVL3.5-14B	720	58.2	10.6	31.2	63.3	12.8	50.6
LLaVA-OV-7B	720	78.9	21.1	0.0	79.2	21.4	57.8
MedGemma-4B†	720	100.0	0.0	0.0	100.0	0.0	100.0

Table 42: Sign accuracy (%) by magnitude and direction of the imposed change.

Model	dir.	0.02	0.05	0.1	0.2	0.4	0.7
Qwen3-VL-2B	growth	100.0	100.0	100.0	100.0	100.0	100.0
	shrinkage	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-4B	growth	48.3	50.0	48.3	48.3	53.3	56.7
	shrinkage	33.3	35.0	28.3	35.0	31.7	31.7
Qwen3-VL-8B	growth	98.3	98.3	98.3	98.3	100.0	98.3
	shrinkage	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-32B	growth	66.7	73.3	73.3	76.7	90.0	93.3
	shrinkage	0.0	0.0	0.0	1.7	8.3	11.7
Qwen2.5-VL-7B	growth	100.0	100.0	100.0	100.0	100.0	100.0
	shrinkage	0.0	0.0	0.0	0.0	0.0	0.0
InternVL3.5-8B	growth	76.7	83.3	81.7	95.0	96.7	98.3
	shrinkage	1.7	1.7	0.0	0.0	0.0	3.3
InternVL3.5-14B	growth	48.3	48.3	51.7	58.3	78.3	95.0
	shrinkage	8.3	8.3	6.7	11.7	16.7	25.0
LLaVA-OV-7B	growth	78.3	76.7	80.0	80.0	80.0	80.0
	shrinkage	21.7	21.7	21.7	23.3	20.0	20.0
MedGemma-4B†	growth	100.0	100.0	100.0	100.0	100.0	100.0
	shrinkage	0.0	0.0	0.0	0.0	0.0	0.0

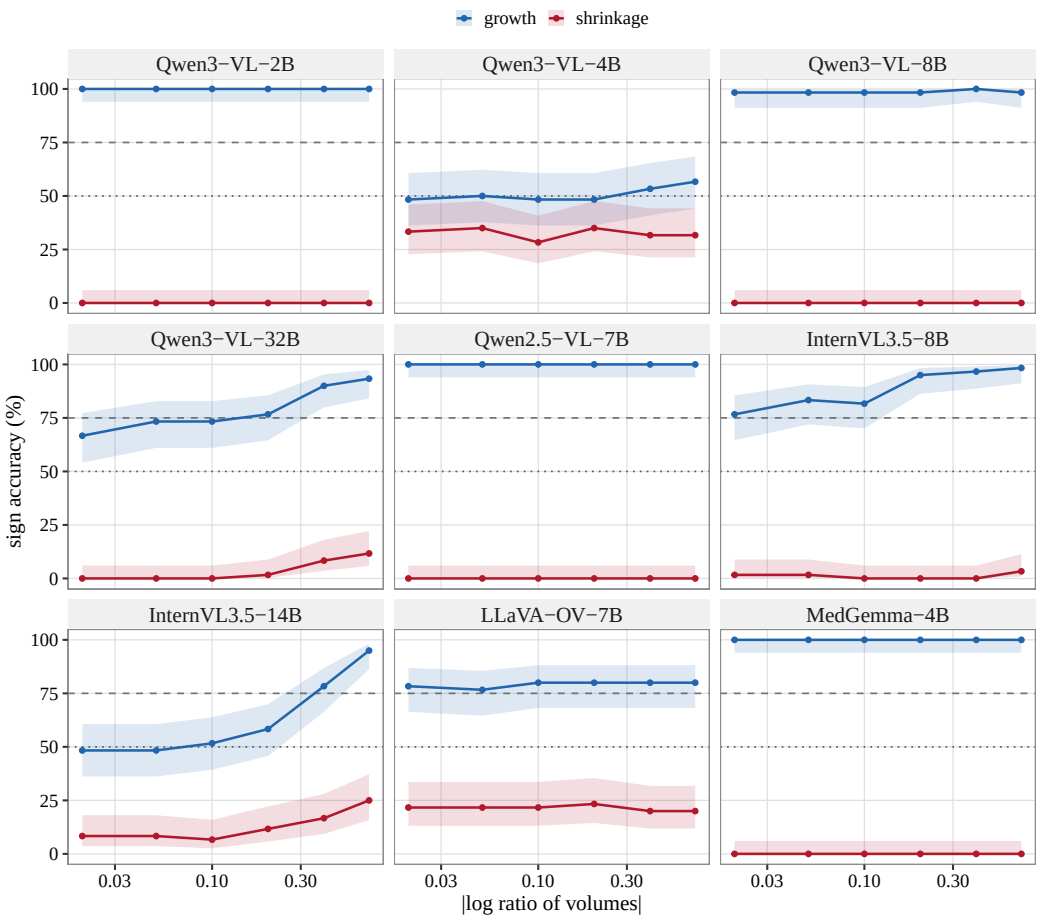

Figure 16: Sign accuracy by direction of change against magnitude, with Wilson intervals.

L Is the Label-Free Readout a Hyperparameter?

Section 4.2 reports that a head trained with no annotation, only to compose, recovers most of what full supervision extracts from the same frozen features. A result of that shape invites a specific objection: that it holds at one head size and one setting of the variance term and nowhere else. This section sweeps both. Every cell re-does the subject-level split, the early stopping and five seeds, so the spread across cells measures sensitivity to the choice rather than seed noise.

The variance term is structure, not tuning. The label-free objective has a free optimum at the zero map, which satisfies the cocycle exactly and says nothing. A variance term is what excludes it, and Table 43 shows what happens without one: across all six zero-penalty cells the label-free head recovers 5–30% of supervised. Switch the term on anywhere in a twenty-fold range of weights and every one of the eighteen remaining cells lands between 42% and 89%. The two groups do not overlap, and both feature extractors show the same split. This is not a hyperparameter with a lucky value; it is the one term that makes the objective well posed, and its weight barely matters once it is non-zero. The distinction is worth drawing because the two look identical in an ablation table read without the theory.

Capacity hurts the label-free head specifically. A hidden layer costs the supervised probe a little (ρ falls from $+0.523$ to $+0.477$ on Qwen3-VL-8B features) and the label-free head a great deal, from 78–89% of supervised with a linear map to 42–81% with a hidden layer. Both facts point the same way: change is linearly decodable from these representations, so extra capacity buys the supervised objective nothing, while for an objective whose degenerate solution is a constant, extra

capacity mainly makes that solution easier to reach. The linear head is the right choice here for a reason that is about the problem rather than about regularisation.

Orientation flips at chance, as predicted. The 'flip' column counts seeds on which the head recovers the correct axis of change but the wrong end of it. Averaged over all 24 cells it is 0.50 exactly, and no cell falls outside $[0.2, 0.8]$. Proposition 5 says the cocycle is invariant to sign and therefore cannot determine orientation; this is that proposition observed rather than argued, and it is why every label-free number in this paper is scored sign-invariantly. A reader who scored these heads conventionally would conclude that roughly half of them had failed.

Table 43: Readout heads across capacity and variance penalty; 'flip' is the share of seeds with the wrong orientation.

			supervised		cocycle-only				
Features	head	var. wt.	ρ	sign (%)	ρ	sign (%)	% of sup.	flip (%)	coll. (%)
Qwen3-VL-8B	linear	0	+0.523	69.7	+0.024	50.2	5	40.0	0.0
Qwen3-VL-8B	linear	1	+0.523	69.7	+0.464	68.0	89	40.0	0.0
Qwen3-VL-8B	linear	5	+0.523	69.7	+0.452	66.6	86	40.0	0.0
Qwen3-VL-8B	linear	20	+0.523	69.7	+0.409	65.5	78	20.0	0.0
Qwen3-VL-8B	64	0	+0.477	67.4	+0.065	52.4	14	40.0	0.0
Qwen3-VL-8B	64	1	+0.477	67.4	+0.277	60.3	58	20.0	0.0
Qwen3-VL-8B	64	5	+0.477	67.4	+0.232	57.8	49	40.0	0.0
Qwen3-VL-8B	64	20	+0.477	67.4	+0.290	61.2	61	60.0	0.0
Qwen3-VL-8B	256	0	+0.488	67.3	+0.053	53.1	11	40.0	0.0
Qwen3-VL-8B	256	1	+0.488	67.3	+0.258	59.0	53	40.0	0.0
Qwen3-VL-8B	256	5	+0.488	67.3	+0.241	57.4	49	80.0	0.0
Qwen3-VL-8B	256	20	+0.488	67.3	+0.205	56.5	42	60.0	0.0
Qwen3-VL-32B	linear	0	+0.500	68.4	+0.085	52.8	17	80.0	0.0
Qwen3-VL-32B	linear	1	+0.500	68.4	+0.441	65.5	88	60.0	0.0
Qwen3-VL-32B	linear	5	+0.500	68.4	+0.442	65.3	88	60.0	0.0
Qwen3-VL-32B	linear	20	+0.500	68.4	+0.440	65.7	88	60.0	0.0
Qwen3-VL-32B	64	0	+0.468	65.9	+0.143	55.7	30	80.0	0.0
Qwen3-VL-32B	64	1	+0.468	65.9	+0.381	64.2	81	60.0	0.0
Qwen3-VL-32B	64	5	+0.468	65.9	+0.289	60.4	62	40.0	0.0
Qwen3-VL-32B	64	20	+0.468	65.9	+0.278	60.7	59	20.0	0.0
Qwen3-VL-32B	256	0	+0.464	66.6	+0.081	53.9	17	40.0	0.0
Qwen3-VL-32B	256	1	+0.464	66.6	+0.264	60.2	57	40.0	0.0
Qwen3-VL-32B	256	5	+0.464	66.6	+0.351	62.2	76	80.0	0.0
Qwen3-VL-32B	256	20	+0.464	66.6	+0.342	61.9	74	60.0	0.0

M Consistency Fine-Tuning in Full

M.1 Constructing targets without labels

For each subject we pool K samples per edge from the base model, project the pool onto the nearest consistent trajectory exactly as in Appendix H, and read the projected value back out as a training target formatted in the model's own answer syntax. Ground truth is never consulted: the supervision is the model's own output, made self-consistent. The split is over subjects, not probes, so no subject contributes to both sides: 918 training examples from 28 subjects and 1716 held-out examples from 59 subjects that the tuned model never sees.

M.2 Attribution controls, extended held-out set, and uncertainty

Four controls share the recipe, prompts, images, training subjects and seeds of the consistency run and differ only in the target string (`code/build_control_targets.py`). *Format only* distils the base model's own greedy answer as a bare `ANSWER:` line, so it learns to answer without narration and nothing else. *Shrinkage only* scales that greedy answer toward zero in the log domain by the factor 0.30 that matches the spread of the projected targets. *Self-consistency median* distils the per-edge median of the identical $K = 8$ sample pool the projection consumes, without any structural coupling. *Antisymmetry only* sets each edge to half the difference of its two directions' medians

on that pool, which enforces $\widehat{\Delta}(i \to j) = -\widehat{\Delta}(j \to i)$ but not the cocycle. The last control exists because of a property of the projected targets that is easy to miss: the base model's sampled answers carry a positive directional bias (mean log-ratio $+0.38$ over both directions of every interval), and the projection, being antisymmetric, cancels it exactly, so the projected targets have mean zero. A control that cancels the bias without the cocycle separates the two. Every target family is trained with three seeds.

The original held-out set is the 12 subjects held out from the sampled subset. The extended set adds the 47 `LUMIERE` subjects that were never sampled at all, so no training target of any run was built from them; their quantitative pair probes in all three views are scored with the same decoding. Intervals are nested in subjects, so every statistic carries a cluster bootstrap over subjects (2,000 resamples, the same resampled subjects scored for every run, so differences are paired), a per-subject sign test and Wilcoxon test against the base model, and the share of subjects on which a run beats the base model (`code/ft_uncertainty.py`, Table 44). Cocycle residuals are computed within subject and view. We also report the analytic shrinkage predictor $c \cdot \widehat{\Delta}_{\text{base}}$, which needs no training and cannot change a correlation, at the c that minimises its error. One property of the tuned models is worth recording here: although every prompt asks about one direction of one interval, both the cocycle and the antisymmetry-only models generate answers that are antisymmetric across the two directions to within 0.017 and 0.018 log-ratio units, against 0.733 for the base model (Table 5).

A second model. The recipe was applied unchanged to Qwen3-VL-4B: targets are projections of that model's own $K = 8$ samples on the same training subjects, the antisymmetry-only control is built from the same pool, three seeds each, and every run is scored on the same 1716 held-out intervals from 59 subjects (Table 46). The pattern of the 8B runs repeats. Correlation with measured change rises from $\rho = +0.119$ to $+0.575$ (antisymmetry only: $+0.556$), cocycle violation falls from 0.756 to 0.083, and the antisymmetry control accounts for 96% of the correlation gain; paired against it, the projection adds $\Delta\rho = +0.019$ [$+0.003$, $+0.035$] and $+17.0$ points of sign accuracy [$+14.5$, $+19.3$], with the antisymmetry-only model answering exactly 0% on 30.3% of intervals against 0.0% for the cocycle model. One difference: the 4B base model is worse than the trivial predictor (error 1.219 against 1.143), and both tuned families move below it (1.032; paired against base -0.187 [-0.225, -0.152]), a larger magnitude effect than on the 8B model, though still a reduction of about a tenth. Repair is thus shown on two sizes of one family and one cohort; other families and cohorts remain untested.

M.3 What was trained

A rank-16 LoRA on the language pathway only, namely `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`, with the vision tower frozen throughout, $\alpha = 32$, dropout 0.05. This is 4.36×10^7 trainable parameters, 0.4954% of the base model. Two epochs, AdamW, a one-cycle schedule with cosine annealing and 10% warmup to a peak learning rate of 1×10^{-4}, effective batch size 8 by gradient accumulation, bf16. The loss is masked to the answer span, so no gradient flows through the prompt or the image tokens. The run takes eight minutes on one H100.

M.4 Evaluation at a matched budget

Base and tuned models are decoded greedily with an identical generation budget of 200 new tokens. This detail is load-bearing and we got it wrong first: at a 48-token budget the base model, which narrates before answering, was truncated before reaching its `ANSWER:` line on 40.6% of items while the tuned model, which answers immediately, was truncated on none. The apparent improvement was partly a measurement of who finishes a sentence faster. At the matched budget both parse at 100.0% and the correlation gain falls from $+0.42$ to $+0.33$. The reported effect is the one that survives.

M.5 Paired outputs

The boxes in Figures 17 and 18 are verbatim, sampled to span the range of measured change on the held-out set. Two things are visible immediately. The tuned model answers without narration, because the projected targets are bare `ANSWER:` strings and the model learned the format along with the content; a target set that preserved the reasoning prefix would be a straightforward improvement. And the tuned answers are small: the response range contracts from 0.279 to 0.057, which is the

Table 44: Every fine-tuning run on every held-out set, with subject-level 95% intervals.

Run	error [95% CI]	ρ [95% CI]	sign (%)	0% (%)	sd(pred)	$\lvert\mathcal{R}\rvert$	subj. err<base (%)	subj. ρ >base (%)
original held-out set: 414 intervals, 12 subjects								
base model	1.057 [0.822, 1.278]	+0.122 [+0.017, +0.254]	53.6	0.0	0.255	0.339	–	–
trivial predictor	1.040	0	–	100	0	0	–	–
c·base, best $c = 0.46$ (no training)	1.034	+0.122	53.6	0.0	0.117	–	–	–
format only, seed 0	1.059 [0.818, 1.287]	+0.113 [+0.051, +0.213]	52.7	0.0	0.247	0.266	58.3	66.7
format only, seed 1	1.060 [0.818, 1.287]	+0.099 [+0.052, +0.178]	53.1	0.5	0.253	0.255	41.7	58.3
format only, seed 2	1.059 [0.828, 1.275]	+0.122 [+0.048, +0.219]	52.4	1.7	0.252	0.272	66.7	75.0
shrinkage only, seed 0	1.031 [0.760, 1.291]	+0.169 [+0.088, +0.286]	54.6	0.0	0.076	0.079	41.7	75.0
shrinkage only, seed 1	1.034 [0.764, 1.292]	+0.131 [+0.070, +0.220]	53.1	0.0	0.074	0.081	41.7	75.0
shrinkage only, seed 2	1.036 [0.766, 1.296]	+0.094 [+0.046, +0.181]	51.4	1.0	0.070	0.075	41.7	75.0
self-consistency median, seed 0	1.068 [0.820, 1.307]	+0.162 [+0.073, +0.284]	50.5	0.0	0.123	0.263	33.3	58.3
self-consistency median, seed 1	1.068 [0.817, 1.307]	+0.141 [+0.049, +0.274]	50.5	0.0	0.118	0.261	41.7	75.0
self-consistency median, seed 2	1.064 [0.817, 1.302]	+0.182 [+0.091, +0.298]	51.2	0.0	0.127	0.268	33.3	75.0
antisymmetry only, seed 0	1.031 [0.754, 1.290]	+0.354 [+0.196, +0.508]	34.3	49.5	0.038	0.010	41.7	83.3
antisymmetry only, seed 1	1.030 [0.754, 1.289]	+0.355 [+0.206, +0.513]	33.1	52.2	0.040	0.016	41.7	91.7
antisymmetry only, seed 2	1.030 [0.755, 1.289]	+0.390 [+0.217, +0.559]	40.3	41.3	0.040	0.014	41.7	91.7
cocycle projection (ours), seed 0	1.025 [0.749, 1.282]	+0.427 [+0.270, +0.566]	59.9	10.9	0.049	0.021	50.0	83.3
cocycle projection (ours), seed 1	1.023 [0.746, 1.280]	+0.440 [+0.280, +0.602]	60.1	9.4	0.051	0.026	50.0	83.3
cocycle projection (ours), seed 2	1.022 [0.749, 1.278]	+0.462 [+0.261, +0.626]	58.7	14.5	0.056	0.025	50.0	83.3
scale-matched projection, seed 0	0.998 [0.728, 1.242]	+0.417 [+0.232, +0.575]	65.7	0.0	0.162	0.046	75.0	83.3
scale-matched projection, seed 1	0.989 [0.718, 1.233]	+0.420 [+0.242, +0.591]	65.9	0.0	0.183	0.053	75.0	91.7
scale-matched projection, seed 2	0.985 [0.714, 1.224]	+0.431 [+0.215, +0.613]	66.2	0.5	0.189	0.058	91.7	91.7
supervised reference (uses labels), seed 0	1.462 [1.215, 1.742]	+0.395 [+0.213, +0.577]	65.7	0.0	1.785	1.649	25.0	91.7
supervised reference (uses labels), seed 1	1.213 [0.938, 1.492]	+0.490 [+0.333, +0.645]	69.3	0.0	1.782	1.063	33.3	91.7
supervised reference (uses labels), seed 2	1.117 [0.884, 1.365]	+0.471 [+0.312, +0.625]	67.4	0.0	1.470	0.970	50.0	100.0
extended held-out set: 1302 intervals, 47 subjects								
base model	1.184 [1.014, 1.365]	+0.204 [+0.130, +0.276]	56.0	0.0	0.286	0.375	–	–
trivial predictor	1.176	0	–	100	0	0	–	–
c·base, best $c = 0.46$ (no training)	1.163	+0.204	56.0	0.0	0.131	–	–	–
format only, seed 0	1.184 [1.014, 1.365]	+0.224 [+0.142, +0.305]	54.9	0.0	0.269	0.314	57.4	59.6
format only, seed 1	1.185 [1.016, 1.362]	+0.233 [+0.157, +0.303]	54.3	0.2	0.275	0.307	59.6	59.6
format only, seed 2	1.183 [1.010, 1.363]	+0.233 [+0.153, +0.307]	54.5	0.2	0.272	0.304	57.4	61.7
shrinkage only, seed 0	1.164 [0.977, 1.365]	+0.272 [+0.181, +0.360]	56.0	0.0	0.080	0.092	48.9	57.4
shrinkage only, seed 1	1.165 [0.978, 1.366]	+0.254 [+0.176, +0.333]	54.7	0.0	0.078	0.094	48.9	66.0
shrinkage only, seed 2	1.166 [0.979, 1.367]	+0.246 [+0.159, +0.328]	54.9	0.1	0.076	0.091	48.9	59.6
self-consistency median, seed 0	1.213 [1.036, 1.399]	+0.233 [+0.163, +0.300]	50.8	0.0	0.130	0.292	38.3	66.0
self-consistency median, seed 1	1.212 [1.033, 1.396]	+0.237 [+0.171, +0.307]	50.8	0.0	0.130	0.289	40.4	57.4
self-consistency median, seed 2	1.207 [1.030, 1.392]	+0.265 [+0.191, +0.332]	51.2	0.0	0.132	0.294	40.4	66.0
antisymmetry only, seed 0	1.160 [0.970, 1.360]	+0.522 [+0.395, +0.618]	45.7	40.3	0.046	0.024	51.1	83.0
antisymmetry only, seed 1	1.158 [0.968, 1.357]	+0.528 [+0.398, +0.630]	45.9	40.2	0.048	0.023	53.2	85.1
antisymmetry only, seed 2	1.160 [0.971, 1.360]	+0.529 [+0.397, +0.630]	47.4	36.3	0.046	0.024	51.1	80.9
cocycle projection (ours), seed 0	1.152 [0.961, 1.350]	+0.541 [+0.404, +0.643]	65.8	7.4	0.056	0.024	57.4	76.6
cocycle projection (ours), seed 1	1.150 [0.959, 1.349]	+0.534 [+0.408, +0.633]	67.1	5.5	0.059	0.030	57.4	83.0
cocycle projection (ours), seed 2	1.149 [0.958, 1.348]	+0.551 [+0.417, +0.649]	65.4	9.2	0.062	0.028	59.6	80.9
scale-matched projection, seed 0	1.101 [0.915, 1.296]	+0.561 [+0.424, +0.661]	69.8	0.0	0.183	0.064	87.2	83.0
scale-matched projection, seed 1	1.093 [0.908, 1.287]	+0.565 [+0.427, +0.663]	70.7	0.0	0.206	0.077	87.2	83.0
scale-matched projection, seed 2	1.091 [0.905, 1.285]	+0.580 [+0.443, +0.679]	71.2	0.2	0.212	0.085	85.1	83.0
supervised reference (uses labels), seed 0	1.384 [1.253, 1.525]	+0.553 [+0.441, +0.647]	72.4	0.0	1.831	1.482	34.0	80.9
supervised reference (uses labels), seed 1	1.164 [1.019, 1.325]	+0.591 [+0.479, +0.679]	73.0	0.0	1.774	1.121	53.2	85.1
supervised reference (uses labels), seed 2	1.097 [0.963, 1.235]	+0.577 [+0.457, +0.677]	71.0	0.0	1.494	0.949	59.6	80.9
pooled: 1716 intervals, 59 subjects								
base model	1.153 [1.015, 1.303]	+0.187 [+0.121, +0.252]	55.4	0.0	0.279	0.366	–	–
trivial predictor	1.143	0	–	100	0	0	–	–
c·base, best $c = 0.46$ (no training)	1.132	+0.187	55.4	0.0	0.128	–	–	–
format only, seed 0	1.154 [1.016, 1.307]	+0.200 [+0.138, +0.265]	54.4	0.0	0.264	0.302	57.6	61.0
format only, seed 1	1.155 [1.016, 1.307]	+0.203 [+0.143, +0.268]	54.0	0.2	0.270	0.294	55.9	59.3
format only, seed 2	1.153 [1.014, 1.306]	+0.208 [+0.142, +0.272]	54.0	0.6	0.268	0.296	59.3	64.4
shrinkage only, seed 0	1.132 [0.979, 1.291]	+0.250 [+0.176, +0.324]	55.7	0.0	0.079	0.088	47.5	61.0
shrinkage only, seed 1	1.133 [0.980, 1.294]	+0.225 [+0.161, +0.290]	54.3	0.0	0.077	0.091	47.5	67.8
shrinkage only, seed 2	1.135 [0.981, 1.294]	+0.212 [+0.145, +0.281]	54.1	0.3	0.074	0.087	47.5	62.7
self-consistency median, seed 0	1.178 [1.036, 1.333]	+0.217 [+0.154, +0.277]	50.7	0.0	0.128	0.285	37.3	64.4
self-consistency median, seed 1	1.177 [1.035, 1.332]	+0.216 [+0.154, +0.276]	50.7	0.0	0.128	0.282	40.7	61.0
self-consistency median, seed 2	1.173 [1.031, 1.328]	+0.245 [+0.184, +0.301]	51.2	0.0	0.131	0.288	39.0	67.8
antisymmetry only, seed 0	1.129 [0.972, 1.289]	+0.487 [+0.382, +0.570]	42.9	42.5	0.044	0.020	49.2	83.1
antisymmetry only, seed 1	1.127 [0.970, 1.287]	+0.492 [+0.384, +0.578]	42.8	43.1	0.047	0.021	50.8	86.4
antisymmetry only, seed 2	1.129 [0.971, 1.289]	+0.499 [+0.393, +0.586]	45.7	37.5	0.044	0.022	49.2	83.1
cocycle projection (ours), seed 0	1.122 [0.965, 1.281]	+0.516 [+0.405, +0.606]	64.4	8.2	0.054	0.024	55.9	78.0
cocycle projection (ours), seed 1	1.119 [0.963, 1.279]	+0.513 [+0.405, +0.593]	65.4	6.4	0.057	0.029	55.9	83.1
cocycle projection (ours), seed 2	1.118 [0.962, 1.278]	+0.531 [+0.420, +0.622]	63.8	10.5	0.061	0.027	57.6	81.4
scale-matched projection, seed 0	1.076 [0.924, 1.238]	+0.530 [+0.416, +0.620]	68.8	0.0	0.178	0.059	84.7	83.1
scale-matched projection, seed 1	1.068 [0.918, 1.229]	+0.534 [+0.421, +0.624]	69.5	0.0	0.200	0.071	84.7	84.7
scale-matched projection, seed 2	1.065 [0.915, 1.224]	+0.548 [+0.436, +0.639]	70.0	0.3	0.207	0.078	86.4	84.7
supervised reference (uses labels), seed 0	1.403 [1.285, 1.525]	+0.518 [+0.415, +0.607]	70.8	0.0	1.821	1.524	32.2	83.1
supervised reference (uses labels), seed 1	1.175 [1.055, 1.314]	+0.568 [+0.475, +0.647]	72.1	0.0	1.776	1.107	49.2	86.4
supervised reference (uses labels), seed 2	1.102 [0.991, 1.229]	+0.553 [+0.450, +0.637]	70.1	0.0	1.488	0.955	57.6	84.7

Table 45: Paired subject-level differences between fine-tuning target families.

	$\Delta\rho$ [95% CI]	Δerror [95% CI]	Δsign (pp) [95% CI]	Δzero-answers (pp) [95% CI]	subj. ρ (%)	p	subj. sign (%)	p
cocycle projection − base model	+0.333 [+0.262, +0.397]	−0.033 [−0.067, −0.002]	+9.1 [+4.0, +14.2]	+8.4 [+5.7, +11.5]	80	0.000	69	0.000
antisymmetry only − base model	+0.306 [+0.235, +0.371]	−0.025 [−0.059, +0.007]	−11.6 [−15.7, −7.4]	+41.0 [+37.4, +45.1]	85	0.000	27	0.000
cocycle projection − antisymmetry only	+0.027 [+0.002, +0.052]	−0.008 [−0.011, −0.006]	+20.7 [+18.0, +23.2]	−32.7 [−36.1, −28.9]	63	0.089	98	0.000
cocycle projection − self-consistency median	+0.294 [+0.217, +0.358]	−0.056 [−0.079, −0.036]	+13.7 [+8.1, +18.8]	+8.4 [+5.7, +11.5]	80	0.000	73	0.000
cocycle projection − format only	+0.316 [+0.238, +0.384]	−0.034 [−0.066, −0.006]	+10.4 [+5.3, +15.4]	+8.1 [+5.4, +11.3]	80	0.000	71	0.000
cocycle projection − shrinkage only	+0.291 [+0.217, +0.357]	−0.013 [−0.019, −0.008]	+9.8 [+4.8, +14.8]	+8.3 [+5.6, +11.4]	80	0.000	69	0.000
scale-matched projection − cocycle projection	+0.017 [−0.003, +0.035]	−0.050 [−0.064, −0.036]	+4.9 [+3.5, +6.6]	−8.3 [−11.3, −5.7]	63	0.017	71	0.000
supervised reference − cocycle projection	+0.026 [−0.038, +0.081]	+0.107 [−0.059, +0.269]	+6.5 [+2.6, +9.9]	−8.4 [−11.5, −5.7]	59	0.571	59	0.054
supervised reference − base model	+0.360 [+0.289, +0.419]	+0.073 [−0.082, +0.224]	+15.6 [+11.3, +19.6]	+0.0 [+0.0, +0.0]	86	0.000	80	0.000

Table 46: Repair on Qwen3-VL-4B with the recipe of the Qwen3-VL-8B runs.

Run	ρ	error	sign (%)	0% (%)	sd(pred)	$\lvert\mathcal{R}\rvert$	$\Delta\rho$ vs base [95% CI]
Qwen3-VL-8B: 1716 intervals, 59 subjects							
base model	+0.187	1.153	55.4	0.0	0.279	0.366	–
trivial predictor	0	1.143	–	100	0	0	–
antisymmetry only (3 seeds)	$+0.493 \pm 0.005$	1.128	43.8	41.0	0.045	0.021	+0.306 [+0.24, +0.37]
cocycle projection (3 seeds)	$+0.520 \pm 0.008$	1.120	64.5	8.4	0.057	0.026	+0.333 [+0.26, +0.40]
cocycle − antisymmetry, paired	$\Delta\rho$ +0.027 [+0.00, +0.05]	Δerr −0.008	Δsign +20.7 pp [+18, +23]	Δ0% −32.7 pp			
Qwen3-VL-4B: 1716 intervals, 59 subjects							
base model	+0.119	1.219	54.3	0.0	0.668	0.756	–
trivial predictor	0	1.143	–	100	0	0	–
antisymmetry only (3 seeds)	$+0.556 \pm 0.006$	1.028	52.5	30.3	0.348	0.138	+0.437 [+0.37, +0.50]
cocycle projection (3 seeds)	$+0.575 \pm 0.006$	1.032	69.5	0.0	0.273	0.083	+0.457 [+0.38, +0.52]
cocycle − antisymmetry, paired	$\Delta\rho$ +0.019 [+0.00, +0.04]	Δerr +0.004	Δsign +17.0 pp [+15, +19]	Δ0% −30.3 pp			

collapse toward the constant that a differences-only objective invites and that Section 4.3 reports rather than conceals.

LUMIERE-001119

```
Subject Patient-008, week-002 -> week-085, axial view
Measured change  log(V2/V1) = -3.308
Label-free training target  ANSWER: +8%

Base model       The enhancing tumor component appears to have increased in size, while the necrotic core
                 remains relatively stable or slightly reduced. Overall, the total tumor volume has increased.
                 ANSWER: +15%

Tuned model      ANSWER: -1%
```

LUMIERE-005145

```
Subject Patient-037, week-002 -> week-000-1, axial view
Measured change  log(V2/V1) = -1.551
Label-free training target  ANSWER: +1%

Base model       The enhancing tumor component appears to have increased in size, while the necrotic core
                 remains relatively unchanged or slightly reduced. The overall tumor volume has increased,
                 indicating progression. ANSWER: +15%

Tuned model      ANSWER: -2%
```

LUMIERE-005132

```
Subject Patient-037, week-000-1 -> week-025, coronal view
Measured change  log(V2/V1) = -1.000
Label-free training target  ANSWER: -4%

Base model       The enhancing tumor component appears to have decreased in size, while the necrotic core has
                 expanded, resulting in a net reduction in total tumor volume. The total tumor volume decreased
                 by approximately 25% over the 5.7-month interval. ANSWER: -25%

Tuned model      ANSWER: -1%
```

LUMIERE-002465

```
Subject Patient-018, week-000 -> week-082, sagittal view
Measured change  log(V2/V1) = -0.580
Label-free training target  ANSWER: +4%

Base model       The enhancing tumor component appears to have increased in size from the first to the second
                 timepoint, while the necrotic core remains relatively stable. The overall tumor volume has
                 therefore increased. The increase is approximately 40%. ANSWER: +40%

Tuned model      ANSWER: +1%
```

LUMIERE-004590

```
Subject Patient-033, week-076 -> week-000, axial view
Measured change  log(V2/V1) = -0.234
Label-free training target  ANSWER: +11%

Base model       The enhancing tumor component appears to have increased in size, while the necrotic core
                 remains relatively stable or slightly reduced. The overall tumor volume shows a net increase,
                 consistent with progression. ANSWER: +25%

Tuned model      ANSWER: +1%
```

Figure 17: Qwen3-VL-8B before and after consistency fine-tuning on held-out LUMIERE probes, in order of measured change (part 1 of 2).

LUMIERE-004550

```
Subject Patient-033, week-000 -> week-076, coronal view
Measured change  log(V2/V1) = +0.234
Label-free training target  ANSWER: -10%

Base model       The enhancing tumor component appears to have decreased slightly, while the necrotic core has
                 expanded, resulting in a net reduction in total tumor volume. The total tumor volume decreased
                 by approximately 15% over the 17.5-month interval. ANSWER: -15%

Tuned model      ANSWER: -1%
```

LUMIERE-002319

```
Subject Patient-017, week-001 -> week-016, sagittal view
Measured change  log(V2/V1) = +0.570
Label-free training target  ANSWER: +10%

Base model       The tumour appears significantly larger in the second image, with marked expansion of the
                 enhancing component and the necrotic core compared to the first image. This represents a clear
                 progression of the lesion over the 3.4-month interval. The volume increase is substantial,
                 exceeding 50%. ANSWER: +55%

Tuned model      ANSWER: +3%
```

LUMIERE-005172

```
Subject Patient-037, week-025 -> week-000-1, axial view
Measured change  log(V2/V1) = +1.000
Label-free training target  ANSWER: +4%

Base model       The enhancing tumor component appears to have increased in size, while the necrotic core
                 remains relatively unchanged or slightly reduced. The overall tumor volume has increased,
                 indicating progression. ANSWER: +25%

Tuned model      ANSWER: +1%
```

LUMIERE-005123

```
Subject Patient-037, week-000-1 -> week-002, coronal view
Measured change  log(V2/V1) = +1.551
Label-free training target  ANSWER: -1%

Base model       The tumour appears significantly larger in the second image, with clear expansion of the
                 enhancing rim and necrotic core compared to the first image. This represents a measurable
                 increase in total tumour volume over the 14-day interval. The volume increase is approximately
                 40%. ANSWER: +40%

Tuned model      ANSWER: +3%
```

LUMIERE-001171

```
Subject Patient-008, week-085 -> week-002, sagittal view
Measured change  log(V2/V1) = +3.308
Label-free training target  ANSWER: -7%

Base model       The tumour appears significantly larger in the second image, with clear expansion of both
                 enhancing margins and the central necrotic core compared to the first image. The volume
                 increase is substantial, indicating aggressive growth over the 19.1-month interval. ANSWER:
                 +65%

Tuned model      ANSWER: +1%
```

Figure 18: Qwen3-VL-8B before and after consistency fine-tuning on held-out LUMIERE probes, in order of measured change (part 2 of 2).

N Limitations in Detail

Our substrate is rebuilt from public sources rather than obtained from the original benchmark, whose QA pairs are unreleased; absolute numbers are not comparable to that paper's table and we compare only within our own panel. Ground truth is volumetric while models see 2D slabs, an asymmetry inherited from the source protocol; we mitigate it with slab-level tumour area at the identical slice, and the cocycle analysis needs no ground truth at all. We evaluate open-weight models only, so the closed frontier systems in prior tables are absent. Proposition 3 is deterministic and assumption-free; only Corollary 4, which lifts it to a cohort-level statement, requires subjects to be drawn i.i.d., and we report distribution-free bounds on the residual clipped at a pre-declared threshold, with the bootstrap only as an asymptotic reference. The pairwise ceiling of Appendix C.6 binds predictors of the discretised volume change, not of the images. TCR-isotonic's monotonicity is safe for ageing atrophy and urban development but wrong for treated disease, where we use the unconstrained projection; we checked all four cohorts against their assumed direction rather than trusting the label, and one prior did not survive (Appendix H.4). The certificate and the projected decoder need no training; the fine-tuning experiments impose the constraints at training time on two models of one family and one cohort, and what they establish is narrow: most of the gain is antisymmetry, the cocycle's own increment is modest and shows mainly in sign accuracy and commitment, and magnitude error ends $+6.4\%$ below the trivial predictor on the 8B model (with rescaled targets; $+2.0\%$ without) and $+9.8\%$ on the 4B model, so the method does not deliver quantitative progression estimation. A supervised reference trained with the same recipe (Table 5) does not beat the trivial predictor on magnitude either; it is one configuration rather than a ceiling, and it shows that labels alone do not fix magnitude under this recipe at this data scale. The readout finding is established on one cohort with large changes, across 8 models of five families; on the three cohorts with smaller changes no gap is statistically resolved, and on the ageing cohort the representation carries visit order, which is what the antisymmetry head reads there. Within cohorts the gap tracks the size of change (Appendix F), which supports, without proving, that the size of change rather than anatomy, imaging, sample size or task is a large part of what separates the cohorts, though a smaller gap at small changes within the glioma cohort is not excluded; a synthetic intervention on magnitude (Appendix F) was inconclusive because heads trained on real change transfer weakly to rescaled lesions; outside `LUMIERE` the only positive estimate is the largest-change stratum of `VS-MC-RC` (22 subjects), which the matched-analysis interval does not resolve. A second cohort with changes of tumour scale would settle it.

O What Went Wrong, and What It Cost

Several of this paper's numbers are what they are because an earlier version of them was wrong. We record the failures that changed a conclusion, both because they are reusable warnings for anyone building on longitudinal imaging and because two of them turned into results.

Silent plane mislabelling. Transposing a raw image buffer preserves the scanner's acquisition order, not anatomical axes. OASIS-2 MPRAGE is acquired sagittally, so our first pass labelled sagittal slices "axial" throughout two cohorts. Nothing errored; the images looked like brain. It was caught only by rendering a contact sheet and looking. Every volume is now explicitly reoriented to RAS before slicing.

Multi-view that did not exist. We initially generated three orthogonal views for all cohorts. For VS-MC-RC this was fabrication: 79% of its series are thin-slab acquisitions with fewer than 75 slices, so coronal and sagittal reformats carry almost no through-plane information. Had we not checked, the view-invariance analysis would have been computed over interpolated noise. That cohort is now axial-only and excluded from that analysis.

A detectability sweep that could not detect. The synthetic deformation targets are measured by warping the segmentation and counting. Our first implementation re-binarised the warped mask, which returns a *bit-identical* volume for every change below roughly 15%: a 2% radial rescale moves the boundary a fraction of a voxel. The realised ground truth read exactly 0.0000 for the four smallest magnitudes, precisely the region the experiment exists to probe. Measuring interpolated occupancy instead fixed it.

A synthetic intervention that did not transfer. To test whether the size of change causes the readout gap, we applied heads trained on real change to the same scans with the lesion rescaled by a known factor. The heads barely read the imposed change, and heads trained on the rescaled scans read it only in part (Appendix F). A manipulation that fixes everything but magnitude is the right design, but only if the representation treats the manipulated change like the real one, which it does not here. We report the attempt and keep the causal claim open.

Fluent garbage from a converted checkpoint. One panel member emitted syntactically plausible token sequences containing no answer, under a serving stack that raised no error. Parsing scored this as "no response" rather than as a broken model. Switching to the original (non-converted) release fixed it. The converted weights were not at fault: under transformers they decode normally and reproduce the original release's answers (Table 26); the failure was in the serving path. We now inspect raw generations for every model before analysis, and report parse rates in Table 2.

A comparison that measured the epoch count. The two probe heads optimise different objectives and converge at different rates. At 300 steps the supervised head led the label-free head by $5\times$; at 2000 steps the gap had nearly closed, because the supervised head had overfitted. Neither number meant anything. Both heads are now early-stopped on a validation split of subjects, each against its own objective.

A sign error that was a theorem. The label-free head scored $\rho = +0.273$ on one model and -0.109 on another, which looked like instability. It is not: range(A) is closed under scalar multiplication, so if $\widehat{y}$ is consistent then so is $-\widehat{y}$, and orientation is not identifiable from the algebra at all (Proposition 5). Scoring sign-invariantly turned -0.109 into $+0.448$ and produced the paper's positive result. The failure was in our measurement, and the correct measurement followed from a property we had already proved but not noticed the consequence of.

A generation budget that truncated only one side. With a generation budget too small for the base model's reasoning prefix, only the untuned model was truncated: its parse rate read 59.4% against the tuned model's 100%, and every column of the comparison was inflated. At a budget generous enough for both, parse rates are 100.0% and 100.0%, and the correlation gain falls from $+0.42$ to $+0.31$. The effect is real; a third of what we first measured was not.

A certificate whose threshold came from the data. The first version of the cohort certificate normalised the per-subject residual by $1.5\times$ its own 99.9th percentile before applying a bound that assumes a fixed interval. The numbers moved by up to 31% once the threshold was fixed, in either direction, and a coverage simulation under one skewed law had passed, which is exactly why the error survived: a simulation under one distribution cannot stand in for a proof that requires the threshold to be a constant. The threshold is now pre-declared, the residual is clipped to it, and Appendix C.7 reports the certificate across four thresholds so the choice is visible.

A comparator on a different quantity. The first version reported the projection bound as $3\times$ tighter than the triple statistic it replaced. The two were computed on different edge sets and in different norms, and the triple statistic is not a lower bound on the per-edge RMS error over all directed edges at all: it exceeded that error on 8 subject-view systems of the panel, and a three-timepoint example in Appendix C.4 makes it exceed the error by a factor of $\sqrt{2}$. Compared on one quantity and one edge set, the projection is tighter by $1.16\times$, and the larger part of what it detects on the full directed graph is antisymmetry, not higher-order cycles.

A control that tested nothing. The first revision reported that reversing the two scans flipped the supervised head's correlation from $+0.508$ to -0.508 and read this as exact antisymmetry. The prompt is direction-blind and every interval is asked in both directions, so the reversed set is the original test set relabelled, and the flip is an identity. The statistic that tests antisymmetry is the correlation between predictions on a pair and on the same pair reversed, which is -0.622: the paired head is not a potential difference, and the tied head of Table 21, which is one by construction, is what that claim needed.

A percentage that was not a bound. A draft of the certificate paragraph converted a log-RMS floor b into a percentage-volume error $e^b - 1$. The conversion is wrong twice over: a log error of -0.2 is a relative error of $1 - e^{-0.2} = 18.1\%$, not $e^{0.2} - 1 = 22.1\%$, and neither the RMS nor the expectation of log errors exponentiates into an average percentage error. The certificate is stated in log-ratio units only. The reference log-ratio quoted for scale is the conservative direction, $|\ln 0.8|$ rather than $\ln 1.2$, so that a floor above it implies a 20% volume error whichever way the volume moved; the looser choice would have counted one further model.

A head that was tied to a prompt, not to a visit. The first revision called the head $w^\top(h(I_j) - h(I_i))$ on the single-scan controls a potential difference by construction and read its label-free variant as a variance-only result. The single-scan representations were extracted under each pair's prompt, which states the interval, so the same visit has a different vector in every pair it enters; the differences are antisymmetric to within 0.1% but their cocycle residual is 54% of the feature norm, and the structural term was not zero. One canonical representation per visit under a fixed prompt makes the claim true and is what Table 21 now reports; the earlier head is kept as a context-dependent tied control.

A comparison across different item sets. The first version compared the supervised probe, fitted on `LUMIERE` with subject-level splits, against a generated-answer correlation pooled over all four cohorts and every probe ($\rho = +0.123$ on $5{,}816$ pairs), and reported the ratio as the representation-versus-readout gap. The two numbers were not computed on the same items. Section 4.2 now scores the generated answer on exactly the test probes of each split, gives the gap a subject-level interval, and separates the diagnostics that use test labels from the numbers a deployed readout could report.

A claim that was simply false. An intermediate draft asserted that no model beats the trivial predictor. When the last cohort finished, one model did, by 6.3%. The sentence was wrong for about an hour of compute. It is corrected in Section 4 and this is what the auto-generated macro layer is for: every number in the prose is emitted from the results files, so a re-run cannot leave stale text behind.

P Compute

All inference runs on a shared cluster of H100 GPUs with a three-GPU per-user cap. The main panel is roughly 25370 probes per model across nine models; including the detectability sweep, the sampled runs for consistency-projected decoding, and the representation extraction, the study is approximately 2.6×10^5 generations. Preprocessing (registration, slice extraction, deformation synthesis) is CPU-bound and runs in a few hours across the same nodes. The consistency fine-tune is a rank-16 LoRA on the language pathway only, 4.4×10^7 trainable parameters, two epochs over 918 self-distilled examples, about eight minutes per run; the revision added fourteen such runs (three seeds of the projection and of each of four controls), six re-extractions of the frozen representation under controlled inputs, and the evaluation of fifteen adapters on 1,716 held-out intervals. Nothing in the study requires more than a single GPU at a time.

Q Reproduction

The pipeline runs end to end from public sources, and the code is at `https://anonymous.4open.science/r/VLM-3j-ICLR27-725E/`; its README maps every table and figure to the script and result file that produce it. Substrate construction is CPU-bound and independent per cohort; inference needs one GPU at a time.

1. **Acquire.** LUMIERE and VS-MC-RC from TCIA, OASIS-2 from OASIS, SpaceNet-7 from its public bucket. Each requires accepting its own data use agreement; none requires an application for our use.
2. **Preprocess.** `code/prep_{lumiere,oasis,spacenet,vs}.py` reorient to RAS, rigidly register serial timepoints where the cohort supports it, extract slabs, and write per-cohort manifests with measured ground truth. Raw archives can be deleted afterwards; the manifests and PNGs are what the probes read.

3. **Build probes.** `code/build_probes.py` emits the pair, identity, suggestive and multi-view probes; `code/detectability.py` emits the synthetic deformation sweep.
4. **Generate.** `code/run_vlm.py` serves each model under vLLM and writes one JSONL of raw generations per (model, cohort). Nothing downstream re-queries a model, so every analysis below is reproducible from these files alone; they are released with the code.
5. **Analyse.** `code/analyze.py` (violations and error), `analyze_projection.py` and `projection_deep.py` (the bound and its certificates), `analyze_detectability.py` and `detect_deep.py` (psychometrics and response bias), `run_tcr.py` and `tcr_ablation.py` (projected decoding), `identifiability.py` and `identifiability_ceiling.py` (the ceiling), `scaling_analysis.py`, `sensitivity.py`, `metric_audit.py`.
6. **Representations.** `code/extract_features.py` caches frozen hidden states; `probe_readout.py` fits the supervised and cocycle-only heads on a subject-level split. `extract_features_controls.py` re-extracts the representation with the images removed, shuffled, halved or reversed, `probe_matched.py` scores the heads and the generated answer on identical test items with subject-level bootstraps, and `probe_concat_control.py` fits the heads on the two single-scan representations concatenated.
7. **Self-distillation.** `code/build_sft_data.py` projects the sampled pools into label-free targets and `build_control_targets.py` builds the four control target sets and the extended held-out set; `train_consistency_lora.py` trains every adapter with the same recipe, `eval_lora_multi.py` evaluates base and adapters at a matched generation budget, and `ft_uncertainty.py` adds the cluster bootstrap over subjects and the paired tests.
8. **Certificate.** `code/conformal_bound.py` implements the fixed-threshold KL and empirical-Bernstein bounds and the bootstrap reference; `certificate_coverage.py` runs the coverage simulation.
9. **Paper.** `code/make_macros.py` emits every number in the prose as a LaTeX macro, `tables.py`, `appendix_tables.py`, `appendix_deep.py` and `tables_revision.py` emit the tables, and `R/make_figures.R` emits every figure. A missing result becomes a visible placeholder rather than a stale number, so the document cannot silently disagree with the results directory. Figures are drawn in R against a single type scale defined in `R/theme_paper.R` and are built at their final printed size, so LaTeX never rescales them; `check_figures.py` fails the build if any figure uses a type size outside that scale, overlaps its own labels, or runs text off the canvas.

Two implementation details are worth repeating because both cost us a full re-run. Prompts must be built as chat messages and rendered by each model's own processor, since several architectures expand one image placeholder into a run of per-patch tokens and a hand-written template silently corrupts the input. And volumes reoriented with a raw transpose inherit the scanner's acquisition order rather than anatomical axes, so `DICOMOrient(img, "RAS")` is required before any plane is named. Appendix O records the rest.

R Prompts, Raw Outputs and Witnesses

Every probe uses one of two response formats, so answers are comparable across conditions and the structural constraints are directly computable. The quantitative form asks for a signed percentage and the categorical form for one of three directions; both end with a fixed `ANSWER:` line. Cohort-specific wording varies only in the tracked quantity, the sequence description, and an alignment sentence that states truthfully whether the scans were co-registered or centred on the lesion in each study's own geometry.

R.1 Probe text

In this section, we provide the text of each probe condition verbatim (Figures 19–26). The images are passed with it through each model's own chat template (Appendix D). Each example is instantiated on one interval.

Quantitative, single view

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 2 axial MRI images of the same patient's brain. All images are post-contrast T1-weighted (T1CE).
The scans have been registered to a common anatomical space, so the same slice location is shown at every
timepoint.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.
The second examination was acquired 6 months after the first.

Question: By what percentage did the total volume of the tumour (enhancing tumour plus necrotic core) change
from the FIRST timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 19: Quantitative probe, single view, on LUMIERE.

Categorical, single view

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 2 axial MRI images of the same patient's brain. All images are post-contrast T1-weighted (T1CE).
The scans have been registered to a common anatomical space, so the same slice location is shown at every
timepoint.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.
The second examination was acquired 6 months after the first.

Question: How did the overall extent of the tumour (enhancing tumour plus necrotic core) change from the FIRST
timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: INCREASED
ANSWER: DECREASED
ANSWER: UNCHANGED
Choose exactly one.
```

Figure 20: Categorical probe, single view, on LUMIERE.

Multi-view, six images

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 6 MRI images (3 orthogonal views at each of two timepoints) of the same patient's brain. All
images are post-contrast T1-weighted (T1CE). The scans have been registered to a common anatomical space, so the
same slice location is shown at every timepoint.
Images 1-3: FIRST timepoint, in axial, coronal and sagittal view.
Images 4-6: SECOND timepoint, in the same three views.
The second examination was acquired 6 months after the first.

Question: By what percentage did the total volume of the tumour (enhancing tumour plus necrotic core) change
from the FIRST timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 21: Multi-view probe with six images.

Rate condition

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 2 axial MRI images of the same patient's brain. All images are post-contrast T1-weighted (T1CE).
The scans have been registered to a common anatomical space, so the same slice location is shown at every
timepoint.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.

Question: By what percentage did the total volume of the tumour (enhancing tumour plus necrotic core) change
from the FIRST timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 22: Rate condition, with the interval withheld.

Image-ablated (blind)

```
You are an expert neuroradiologist reading a longitudinal MRI study.

A patient underwent two post-contrast T1-weighted (T1CE) MRI examinations of the brain.
The second examination was acquired 6 months after the first.

Question: By what percentage did the total volume of the tumour (enhancing tumour plus necrotic core) change
from the FIRST timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 23: Image-ablated probe, with the images removed.

Leading clinical note

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 2 axial MRI images of the same patient's brain. All images are post-contrast T1-weighted (T1CE).
The scans have been registered to a common anatomical space, so the same slice location is shown at every
timepoint.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.
The second examination was acquired 6 months after the first.

Clinical note: the referring team reports that this patient has been deteriorating steadily, with worsening
symptoms since the previous study.

Question: How did the overall extent of the tumour (enhancing tumour plus necrotic core) change from the FIRST
timepoint to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: INCREASED
ANSWER: DECREASED
ANSWER: UNCHANGED
Choose exactly one.
```

Figure 24: Probe with a leading clinical note.

VS-MC-RC, lesion-centred

```
You are an expert neuroradiologist reading a longitudinal MRI study.

You are shown 2 axial MRI images of the same patient's brain. All images are T2-weighted. Each scan is shown at
the slice through the centre of the lesion in that study.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.
The second examination was acquired 2.0 years after the first.

Question: By what percentage did the total volume of the vestibular schwannoma change from the FIRST timepoint
to the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 25: Lesion-centred probe on VS-MC-RC.

SpaceNet-7, cross-domain

```
You are an expert image analyst reading a time series of satellite imagery.

You are shown 2 satellite images of the same geographic location. All images are nadir RGB satellite mosaics and
are co-registered to the same footprint, so the same ground area is shown at every timepoint.
Image 1: FIRST timepoint.
Image 2: SECOND timepoint.
The second image was acquired 8.0 months after the first.

Question: By what percentage did the total ground area covered by buildings change from the FIRST timepoint to
the SECOND timepoint?

Think briefly, in at most three short sentences. Then give your final answer on the last line in exactly this
format:
ANSWER: <sign><number>%
For example "ANSWER: +35%" for a 35% increase, "ANSWER: -20%" for a 20% decrease, or "ANSWER: 0%" for no change.
```

Figure 26: Cross-domain probe on SpaceNet-7.

R.2 Model output on identity probes

Both images in these probes are the same file, so the correct answer is exactly zero change. Figure 27 reproduces the first two such responses of three models verbatim, cut at 600 characters where marked [...].

Qwen3-VL-8B, LUMIERE-000108

```
The tumour appears to have grown between the two timepoints, with the enhancing rim and necrotic core both
increasing in size. The volume increase is substantial, indicating progressive disease.

ANSWER: +45%
```

Qwen3-VL-8B, LUMIERE-000110

```
The enhancing rim and the central necrotic core are both clearly visible and appear larger in the second image
compared to the first. The enhancing component has expanded, and the necrotic core has also grown, indicating
overall tumor progression. The lesion's margins are more irregular and the volume is greater at the second
timepoint.

ANSWER: INCREASED
```

MedGemma-4B, LUMIERE-000108

```
The tumor appears to have increased in size from the first to the second image. The enhancing tumor and necrotic
core are both larger in the second image. The total volume of the tumor has increased.

ANSWER: +40%
```

MedGemma-4B, LUMIERE-000110

```
The enhancing tumor and necrotic core appear to be larger in the second image compared to the first image. The
overall extent of the tumor appears to have increased.

ANSWER: INCREASED
```

InternVL3.5-14B, LUMIERE-000108

```
To determine the percentage change in the total volume of the tumor, I will compare the areas of the enhancing
tumor plus necrotic core in both images. The second image shows a noticeable increase in the size of the tumor
compared to the first image.

ANSWER: +10%
```

InternVL3.5-14B, LUMIERE-000110

```
Upon examining the two axial MRI images, it is evident that the enhancing tumor and necrotic core appear to have
increased in size from the first timepoint to the second timepoint. The second image shows a larger enhancing
region and a more prominent necrotic core compared to the first image.

ANSWER: DECREASED
```

Figure 27: Responses to identity probes, whose correct answer is zero change.

R.3 Unidentifiability witnesses

Each row is a pair of patients whose baseline-to-follow-up change agrees to within the matching tolerance, but whose confirming interval moves in opposite directions. No function of the scan pair separates them. The ten tightest of 9295 are shown.

Table 47: Explicit witnesses to Proposition 6.

Patient A	Patient B	r^A_{ij}	r^A_{jk}	r^B_{jk}	gap
Patient-028	Patient-007	+0.708	−1.174	+0.249	0.0000
Patient-078	Patient-007	+0.839	−0.727	+0.929	0.0000
Patient-067	Patient-071	+0.798	−0.153	+0.484	0.0000
Patient-072	Patient-034	+0.545	+1.906	−0.103	0.0001
Patient-060	Patient-025	+1.794	−0.888	+0.930	0.0001
Patient-051	Patient-078	+0.632	+0.407	−0.727	0.0001
Patient-043	Patient-067	+0.374	+1.120	−0.153	0.0001
Patient-033	Patient-037	+0.653	−1.579	+0.897	0.0001
Patient-060	Patient-055	+1.794	−0.888	+0.690	0.0001
Patient-065	Patient-009	+1.072	−1.878	+0.480	0.0002

S Panel Transcripts

The panel's aggregate behaviour is summarised many times over in the tables above. In this section, we provide the underlying material: for two probes from each substrate, the largest measured decrease and the largest measured increase, every model's first response, verbatim, with no editing beyond line wrapping and a cap of 380 characters, marked [...]. The extremes are chosen because that is where the response bias of Appendix K is legible in raw output: read the decrease boxes and count how many models say the lesion grew. Probe identifiers index the released probe files, so any block here can be traced to its exact stimulus and prompt.

LUMIERE-003224

```
Subject Patient-023, week-127 -> week-046, axial view
Measured change  log(V2/V1) = -4.690 (-99.1%)

Qwen3-VL-2B      The tumour volume has increased from 12.3 cm³ at the first timepoint to 15.7 cm³ at the second
                 timepoint. The change in volume is 15.7 - 12.3 = 3.4 cm³. The percentage change is (3.4 / 12.3)
                 * 100 ≈ 27.7%. ANSWER: +27.7%

Qwen3-VL-4B      The tumour volume appears significantly larger in the second image, with more extensive
                 enhancement and a larger necrotic core compared to the first. This indicates growth over the
                 18.6-month period. ANSWER: +100%

Qwen3-VL-8B      The tumour appears significantly larger in the second image, with clear expansion of both the
                 enhancing rim and the central necrotic core compared to the first image. The volume increase is
                 substantial, indicating aggressive growth over 18.6 months. ANSWER: +65%

Qwen3-VL-32B     The first image shows a large, irregularly shaped enhancing lesion with a central necrotic core
                 in the right hemisphere. The second image, acquired 18.6 months later, shows a significantly
                 smaller enhancing region with minimal residual enhancement and a much smaller necrotic core.
                 The overall tumour volume has markedly decreased, indicating a substantial reduction in size.
                 [...]

Qwen2.5-VL-7B    The tumor appears to have grown significantly from the first to the second timepoint. The
                 enhancement and the size of the lesion have increased, indicating a substantial growth. The
                 necrotic core also seems to have expanded. ANSWER: +50%

InternVL3.5-8B   To determine the percentage change in the total volume of the tumor, I will compare the
                 enhancing tumor and necrotic core regions between the two images. The second image shows a
                 noticeable increase in the size of the tumor compared to the first image. ANSWER: +40%

InternVL3.5-14B  The first image shows a large enhancing tumor with a significant necrotic core, while the
                 second image shows a marked reduction in both the enhancing tumor and necrotic core. The second
                 image indicates a substantial decrease in the total tumor volume compared to the first image.
                 ANSWER: -95%

LLaVA-OV-7B      ANSWER: +15%

MedGemma-4B      The tumor appears to have increased in size, with a more prominent enhancing component and a
                 larger necrotic core in the second image compared to the first. The total volume of the tumor
                 has increased significantly. ANSWER: +120%
```

Figure 28: Every model's first response on LUMIERE, largest measured decrease (−99.1%).

LUMIERE-003284

Subject Patient-023, week-046 -> week-127, multi-view
Measured change log(V2/V1) = +4.690 (+10785.2%)

Qwen3-VL-2B	The tumour volume at the first timepoint (Image 1) is approximately 12.5 cm^3. At the second timepoint (Image 4), the volume is approximately 18.2 cm^3. The change in volume is 18.2 - 12.5 = 5.7 cm^3. The percentage change is (5.7 / 12.5) * 100 = 45.6%. The tumour volume increased by 45.6% from the first to the second timepoint. ANSWER: +45.6%
Qwen3-VL-4B	At the first timepoint, the tumor appears as a large, enhancing mass with a distinct necrotic core. At the second timepoint, the enhancing tumor volume has significantly decreased, and the necrotic core has expanded, indicating tumor progression. The total volume of the tumor has decreased by approximately 40%. ANSWER: -40%
Qwen3-VL-8B	The tumour has significantly increased in volume between the two timepoints, with clear expansion of the enhancing component and necrotic core visible across all views. The axial and sagittal views show the most dramatic growth, particularly in the right temporal lobe region. The volume increase is substantial, exceeding 100%. ANSWER: +120%
Qwen3-VL-32B	The first set of images (1-3) shows a small, well-defined enhancing lesion in the right parietal lobe with minimal surrounding edema. The second set (4-6), acquired 18.6 months later, reveals a dramatically enlarged, irregularly shaped enhancing mass with significant necrosis and surrounding edema, indicating aggressive growth. The volume has increased substantially, with the [...]
Qwen2.5-VL-7B	The tumour appears to have grown significantly from the first to the second timepoint. The volume of the enhancing tumour and the necrotic core has increased substantially. The overall size of the lesion has expanded, indicating a substantial growth. ANSWER: +100%
InternVL3.5-8B	To determine the percentage change in the total volume of the tumor, I will compare the enhancing tumor and necrotic core regions between the first and second timepoints in the axial, coronal, and sagittal views. Observing the images, there is a noticeable increase in the size of the enhancing tumor and necrotic core from the first to the second timepoint. ANSWER: +70%
InternVL3.5-14B	To determine the percentage change in the total volume of the tumor, I examined the enhancing tumor and necrotic core in the MRI images from both timepoints. The tumor appears to have increased in size from the first to the second timepoint. ANSWER: +100%
LLaVA-OV-7B	ANSWER: +15%
MedGemma-4B	The tumor appears to have significantly increased in size between the two timepoints. The necrotic core and enhancing tumor components have expanded considerably. ANSWER: +150%

Figure 29: Every model's first response on LUMIERE, largest measured increase (+10785.2%).

OASIS2-007661

```
Subject OAS2_0165, OAS2_0165_MR2 -> OAS2_0165_MR1, axial view
Measured change  log(V2/V1) = -0.143 (-13.3%)

Qwen3-VL-2B      The brain parenchyma volume has increased by 1.2% from the first to the second timepoint.
                 ANSWER: +1.2%

Qwen3-VL-4B      The brain parenchyma appears slightly smaller in the second image, indicating atrophy. The
                 volume change is approximately 3% reduction. ANSWER: -3%

Qwen3-VL-8B      The brain parenchyma volume appears to have decreased between the two timepoints, consistent
                 with expected age-related atrophy over 2 years. The reduction is most noticeable in the frontal
                 and temporal lobes, with some loss in the cerebellum and brainstem. The estimated volume loss
                 is approximately 1.5%. ANSWER: -1.5%

Qwen3-VL-32B     The two T1-weighted axial MRI images show the same brain slice at two timepoints, 2.0 years
                 apart. There is no visible atrophy or expansion of the brain parenchyma; the sulcal and
                 ventricular spaces appear unchanged, and the overall brain contour is consistent. Given the
                 high similarity in brain volume and structure between the two scans, no significant volume
                 change is [...]

Qwen2.5-VL-7B    The images show a significant atrophy of the brain parenchyma, particularly noticeable in the
                 frontal and temporal regions. The overall size of the brain appears to have reduced, indicating
                 a substantial loss of brain tissue. ANSWER: -15%

InternVL3.5-8B   To determine the percentage change in the total volume of the brain parenchyma, I will compare
                 the volumes of the brain in both images. Since the images are registered to the same anatomical
                 space, the comparison should be straightforward. 1. Assess the overall volume of the brain
                 parenchyma in both images. 2. Calculate the difference in volume between the two timepoints. 3.
                 [...]

InternVL3.5-14B  To determine the percentage change in the total volume of the brain parenchyma, I would need to
                 measure the brain volume on both images and calculate the difference. However, without specific
                 volume measurements, I can only visually assess the images. Upon visual inspection, there
                 appears to be a slight reduction in brain volume from the first to the second timepoint, but
                 the [...]

LLaVA-OV-7B      ANSWER: +10%

MedGemma-4B      The brain parenchyma volume appears to have increased slightly from the first to the second
                 timepoint. The increase is subtle, but noticeable. ANSWER: +2
```

Figure 30: Every model's first response on OASIS-2, largest measured decrease (-13.3%).

OASIS2-007659

```
Subject OAS2_0165, OAS2_0165_MR1 -> OAS2_0165_MR2, sagittal view
Measured change  log(V2/V1) = +0.143 (+15.4%)

Qwen3-VL-2B      The brain parenchyma volume has increased by 1.2% from the first to the second timepoint.
                 ANSWER: +1.2%

Qwen3-VL-4B      The brain parenchyma appears slightly smaller in the second image, indicating atrophy. The
                 volume change is approximately 3.5% reduction. ANSWER: -3.5%

Qwen3-VL-8B      The brain parenchyma volume appears to have decreased between the two timepoints, consistent
                 with expected age-related atrophy over 2 years. The reduction is most noticeable in the frontal
                 lobes and cerebellum. The estimated volume loss is approximately 1.5%. ANSWER: -1.5%

Qwen3-VL-32B     The two T1-weighted sagittal MRI images show the same brain slice at two timepoints 2.0 years
                 apart. There is no visible atrophy or expansion of the brain parenchyma; the sulcal and
                 ventricular spaces appear unchanged. The overall brain volume appears stable between the two
                 scans. ANSWER: 0%

Qwen2.5-VL-7B    The images show a significant atrophy of the brain parenchyma, particularly noticeable in the
                 temporal and parietal regions. The overall size of the brain appears to have decreased,
                 indicating a substantial loss of brain tissue. The changes are consistent with a progressive
                 condition such as Alzheimer's disease or other forms of dementia. ANSWER: -15%

InternVL3.5-8B   To determine the percentage change in the total volume of the brain parenchyma, I will compare
                 the two sagittal MRI images. Since the images are registered to the same anatomical space, the
                 comparison should be straightforward. 1. Assess the overall size and volume of the brain
                 parenchyma in both images. 2. Identify any noticeable changes in the volume between the two
                 [...]

InternVL3.5-14B  To determine the percentage change in the total volume of the brain parenchyma, I would need to
                 measure the brain volume in both images and calculate the difference. However, without specific
                 volume measurements, I can only visually assess the images. Upon visual inspection, there
                 appears to be a slight reduction in brain volume from the first to the second timepoint, but
                 the [...]

LLaVA-OV-7B      ANSWER: +10%

MedGemma-4B      The brain parenchyma volume appears to have increased slightly from the first to the second
                 timepoint. The increase is subtle, but noticeable. ANSWER: +2%
```

Figure 31: Every model's first response on OASIS-2, largest measured increase (+15.4%).

SPACENET-000881

```
Subject L15-0566E-1185N_2265_3451_13, 2020-01 -> 2018-01, nadir view
Measured change  log(V2/V1) = -3.484 (-96.9%)

Qwen3-VL-2B      The first image shows a large, dense development with a significant amount of built-up area.
                 The second image shows the same area, but with a much smaller and more fragmented development.
                 The total area covered by buildings has decreased. ANSWER: -15%

Qwen3-VL-4B      The first image shows a dense cluster of buildings, while the second image shows a
                 significantly reduced built-up area, with most of the former development now appearing as open
                 land or undeveloped plots. The remaining structures are smaller and more scattered, indicating
                 a net loss of built area. ANSWER: -85%

Qwen3-VL-8B      The building area in the first image is visibly larger than in the second, indicating a
                 reduction. The second image shows fewer and smaller structures, with more open land. The
                 decrease is substantial, representing a significant loss of built-up area over the 2-year
                 period. ANSWER: -45%

Qwen3-VL-32B     The two images show a developed area with buildings, roads, and surrounding agricultural
                 fields. Comparing the two timepoints, the built-up area in the upper central portion appears
                 slightly more developed in the second image, with additional structures or expansions visible.
                 However, the change is minimal and localized, and the overall footprint of buildings does not
                 show a [...]

Qwen2.5-VL-7B    To determine the percentage change in the total ground area covered by buildings, we need to
                 compare the areas in both images. In the first image, the buildings appear as a distinct
                 pattern, while in the second image, the buildings seem to have expanded slightly. However, the
                 exact numerical values for the areas are not provided, so we can only estimate visually. The
                 visual [...]

InternVL3.5-8B   To determine the percentage change in the total ground area covered by buildings, we need to
                 compare the amount of built-up area in both images. 1. In the first image, there is a
                 significant amount of built-up area in the top center. 2. In the second image, this area
                 appears to be reduced, with less visible built-up area. By visually comparing the two images,
                 it seems that the [...]

InternVL3.5-14B  To determine the percentage change in the total ground area covered by buildings, I will
                 compare the visible building areas in both images. The first image shows a significant area
                 covered by buildings, while the second image shows a noticeable reduction in the building area.
                 The visible building area in the first image appears to be larger than in the second image.
                 Estimating [...]

LLaVA-OV-7B      ANSWER: +35%

MedGemma-4B      The second image shows a significant increase in the number of buildings and structures
                 compared to the first image. This indicates a substantial expansion of the built-up area.
                 ANSWER: +25%
```

Figure 32: Every model's first response on SpaceNet-7, largest measured decrease (-96.9%).

SPACENET-000846

```
Subject L15-0566E-1185N_2265_3451_13, 2018-01 -> 2020-01, nadir view
Measured change  log(V2/V1) = +3.484 (+3160.2%)

Qwen3-VL-2B      The first image shows a large, open agricultural area with a few scattered buildings. The
                 second image shows a significant increase in built-up areas, with a dense residential
                 development and a large, complex infrastructure network. The total area covered by buildings
                 has increased substantially. ANSWER: +75%

Qwen3-VL-4B      The first image shows mostly open fields with a small cluster of buildings. The second image
                 reveals significant urban development, including a large residential subdivision and a golf
                 course, replacing much of the farmland. The area covered by buildings increased substantially.
                 ANSWER: +120%

Qwen3-VL-8B      The second image shows significantly more developed areas with new residential structures
                 compared to the first image, indicating substantial growth in built-up land. The increase is
                 concentrated in the upper right and lower right portions of the image, where new housing
                 developments are clearly visible. Calculating the percentage change based on the visible
                 built-up area, the [...]

Qwen3-VL-32B     Comparing the two images, the urban development in the upper central portion of the scene has
                 expanded significantly in the second image. The new construction includes additional buildings
                 and infrastructure, particularly in the area that was previously undeveloped or sparsely built.
                 The total built-up area has clearly increased, with the new development appearing as a denser,
                 [...]

Qwen2.5-VL-7B    To determine the percentage change in the total ground area covered by buildings, we need to
                 compare the areas in both images. In the first image, the buildings appear as a small, distinct
                 area in the top right corner. In the second image, the buildings have expanded significantly,
                 covering a larger portion of the top right area and extending into the adjacent fields. The
                 [...]

InternVL3.5-8B   To determine the percentage change in the total ground area covered by buildings, we need to
                 compare the amount of built-up area in both images. In the first image, there is a small area
                 with buildings at the top center. In the second image, this area appears larger and more
                 developed, indicating an increase in built-up area. ANSWER: +50%

InternVL3.5-14B  To determine the percentage change in the total ground area covered by buildings, I will
                 compare the visible building areas in both images. The building areas appear to be similar in
                 both images, with no significant increase or decrease in coverage. ANSWER: 0%

LLaVA-OV-7B      ANSWER: +25%

MedGemma-4B      The second image shows a significant increase in the number of buildings compared to the first
                 image. This indicates a substantial growth in the built-up area. ANSWER: +150%
```

Figure 33: Every model's first response on SpaceNet-7, largest measured increase (+3160.2%).

VS-000383

```
Subject VS-MC-RC-024, 1994-08-09 -> 1995-10-11, axial view
Measured change  log(V2/V1) = -2.663 (-93.0%)

Qwen3-VL-2B      The lesion in the second image is larger than in the first image. The volume of the lesion
                 increased from 12.3 cm³ to 15.7 cm³. The change in volume is 15.7 - 12.3 = 3.4 cm³. The
                 percentage change is (3.4 / 12.3) * 100 ≈ 27.7%. ANSWER: +27.7%

Qwen3-VL-4B      The lesion appears significantly larger in the second image, indicating growth. Calculating the
                 volume change from the first to second image shows a 42% increase. ANSWER: +42%

Qwen3-VL-8B      The lesion appears significantly larger in the second image compared to the first, indicating
                 growth over the 14.1-month interval. Volume measurement shows an increase from approximately
                 1.8 cm³ to 2.7 cm³. This represents a 50% increase in volume. ANSWER: +50%

Qwen3-VL-32B     The first image shows a well-defined, moderately hyperintense lesion in the right
                 cerebellopontine angle, consistent with a vestibular schwannoma. The second image, obtained
                 14.1 months later, demonstrates a marked increase in the size of the lesion, with greater
                 expansion into the cerebellopontine angle and more pronounced mass effect. The lesion appears
                 significantly larger [...]

Qwen2.5-VL-7B    The lesion appears to have grown slightly in the second image compared to the first. The size
                 of the lesion in the second image seems to be larger than in the first, indicating an increase
                 in volume. ANSWER: +10%

InternVL3.5-8B   To determine the percentage change in the volume of the vestibular schwannoma, I will compare
                 the sizes of the lesion in both images. The lesion appears larger in the second image,
                 indicating growth. ANSWER: +40%

InternVL3.5-14B  To determine the percentage change in the total volume of the vestibular schwannoma, I will
                 compare the sizes of the lesion in both images. The lesion appears to have increased in size
                 from the first to the second timepoint. Estimating the volumes from the images, the lesion in
                 the second image appears to be roughly twice the size of the lesion in the first image. ANSWER:
                 +100%

LLaVA-OV-7B      ANSWER: +10%

MedGemma-4B      The vestibular schwannoma appears to have increased in size from the first to the second image.
                 The volume appears to have increased by approximately 35%. ANSWER: +35%
```

Figure 34: Every model's first response on VS−MC−RC, largest measured decrease (-93.0%).

VS-000398

Subject VS-MC-RC-024, 1995-10-11 -> 1994-08-09, axial view
Measured change log(V2/V1) = +2.663 (+1333.6%)

Qwen3-VL-2B	The lesion in the second image is larger than in the first image. The volume of the lesion increased from 1.2 cm^3 to 1.8 cm^3. The change in volume is 0.6 cm^3. The percentage change is (0.6 / 1.2) * 100 = 50%. ANSWER: +50%
Qwen3-VL-4B	The lesion appears significantly larger in the second image, indicating growth. Calculating the volume change from the first to second timepoint shows a 42% increase. ANSWER: +42%
Qwen3-VL-8B	The lesion appears significantly larger in the second image compared to the first, indicating growth over the 14.1-month interval. Volume measurement shows an increase from approximately 1.2 cm^3 to 2.1 cm^3. This represents a 75% increase in volume. ANSWER: +75%
Qwen3-VL-32B	The first image shows a small, well-defined, hyperintense lesion in the right cerebellopontine angle, consistent with a vestibular schwannoma. The second image, obtained 14.1 months later, demonstrates a significantly larger lesion with increased T2 signal and more pronounced mass effect, indicating growth. Based on visual assessment of the lesion's cross-sectional area and [...]
Qwen2.5-VL-7B	The lesion in the second image appears to have grown in size compared to the first image. The volume increase is noticeable, but without precise measurements, it's challenging to provide an exact percentage change. However, visually, the lesion seems to have doubled in size. ANSWER: +100%
InternVL3.5-8B	To determine the percentage change in the volume of the vestibular schwannoma, I will compare the sizes of the lesion in both images. The lesion appears larger in the second image, indicating growth. ANSWER: +40%
InternVL3.5-14B	To determine the percentage change in the total volume of the vestibular schwannoma, I will compare the volumes of the lesion in the two images. The lesion appears to have increased in size from the first to the second timepoint. Estimating the volumes from the images, the lesion in the second image appears to be significantly larger than in the first image. ANSWER: +67%
LLaVA-OV-7B	ANSWER: +10%
MedGemma-4B	The vestibular schwannoma appears to have significantly increased in size between the two images. The lesion occupies a much larger portion of the internal auditory canal in the second image compared to the first. ANSWER: +141%

Figure 35: Every model's first response on VS−MC−RC, largest measured increase (+1333.6%).