CALIBRATE BEFORE YOU MEASURE: CONDITIONING INTERFACES DECIDE WHAT A GENERATIVE PROBE REPORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen diffusion models serve neural decoding as renderers and, through conditional likelihood, as instruments that measure how much a brain recording reveals about a stimulus. Both uses route the signal through a conditioning interface that is treated as a neutral pipe. We show that it is not. We construct a chain of stimulus channels ordered by the data-processing inequality, so that any valid instrument must report a non-increasing curve, and use this ground truth to audit four conditioning interfaces of one diffusion backbone. The single-token image-embedding interface adopted by most EEG and fMRI pipelines scores the true stimulus below a deliberately mismatched condition on spatial layout, whereas a 257-token readout of the same encoder pass reads layout correctly. The gap persists across a second backbone and two disjoint stimulus sets, and a far more bottlenecked caption interface reads layout correctly, so the failure cannot be anticipated from an interface's description and must be measured. A dose–response identifies conditioning strength as the mechanism: raising it improves the axes an interface can express while pushing the axes it cannot below chance. Deployed decoders exhibit the same over-pushing: their decoded vector is a worse estimate of the true embedding than a constant in all ten subjects and is optimal at one third of its deployed strength, and two independent decoders are overconfident by a conformally guaranteed margin that rendering at the optimal strength narrows but does not close. We release the phantom as a pre-flight check for any conditioned generative probe.

1 INTRODUCTION

A frozen diffusion model is an attractive measuring device. It is pretrained, it needs no task-specific fitting, and its conditional likelihood provides a natural score for "how much does this signal tell me about this image". Brain-to-image pipelines condition a diffusion model on a decoded embedding and report reconstruction quality (Scotti et al., 2024; Li et al., 2024), and conditional-likelihood scores are used as a discriminative readout in their own right (Li et al., 2023).

Every one of these uses routes the signal through a *conditioning interface*: the adapter, projection, or cross-attention pathway by which a vector becomes something the denoiser can attend to. The interface is normally treated as plumbing. This paper shows it is the measuring instrument, and that it can be miscalibrated in a way that no amount of care on the neuroscience side would reveal.

The obstacle to noticing this is that brain data has no ground truth about its own information content. If a probe reports that EEG carries no spatial layout, there is no independent fact to check it against; the report is simultaneously a claim about the brain and a claim about the instrument, and the two are not separable. Our central move is to remove the brain from the loop. We construct a chain of stimulus channels

$$x_0 \;\to\; x_{64} \;\to\; x_{32} \;\to\; x_{16} \;\to\; x_8$$

by repeated box averaging, chosen so that each stage is exactly a block average of the previous one. This makes the chain Markov by construction, so the data-processing inequality (Cover & Thomas,

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.

2006) guarantees, for *any* target Y and with no estimation anywhere,

$$I(x_8; Y) \;\le\; I(x_{16}; Y) \;\le\; I(x_{32}; Y) \;\le\; I(x_{64}; Y) \;\le\; I(x_0; Y).$$

A valid instrument must produce a non-increasing curve along this chain. Any instrument that produces an increasing one is not under-reporting; it is reporting the wrong sign.

Our findings are as follows.

Instrument validity is interface-specific and not predictable. Across four conditioning interfaces of the same SDXL backbone, two read spatial layout correctly, one reads it weakly, and one is anti-correlated with it. Nothing in their descriptions—token count, modality, nominal dimensionality—predicts which. The single-token image embedding that dominates the brain-decoding literature is the pathological one.

The empirical null is load-bearing, not a refinement. Unnatural foils are harder to denoise unconditionally, on 84–100% of trials, which biases any two-way likelihood comparison toward the foil regardless of the condition. We quantify this offset and show it is a property of the foil, not of any interface.

Conditioning strength is the mechanism. Sweeping the adapter scale, the layout inversion deepens monotonically and essentially vanishes at low strength, while the axes the interface *can* express improve over the same range. An interface with spatial capacity shows no inversion at any dose.

Deployed decoders exhibit the same over-pushing, twice. Independently of any diffusion computation, the decoded conditioning vector used at full strength is farther from the truth than a constant in every one of the ten THINGS-EEG subjects, and wants to be shrunk by roughly a factor of three. Independently again, the images that come out are overconfident by a conformally guaranteed margin, in two separately trained decoders.

Fixing the strength is necessary and not sufficient. Re-rendering at the optimal conditioning strength measurably widens the renderer's posterior, but the emitted image remains overconfident on every trial. We report this negative result because it locates the residual problem in the renderer rather than in the estimate.

We do not claim that generative probes are useless, and we are careful to show they are not: the same instrument that inverts on layout tracks category and blur correctly. The claim is narrower and, we think, more actionable: validity is a property of the interface–axis pair, it is cheap to check, and it is currently not checked.

Related work. Diffusion models are used as classifiers by comparing conditional denoising losses across candidate conditions (Li et al., 2023; Clark & Jaini, 2023); our probe is this construction with paired noise, applied to a condition that is a vector rather than a class label. The interfaces we audit are the standard ways of conditioning a frozen denoiser on such a vector: IP-Adapter's decoupled cross-attention with a pooled or a patch-grid image embedding (Ye et al., 2023), and native text cross-attention. The adapter scale we sweep plays the role that guidance strength plays in sampling (Ho & Salimans, 2022), and our dose–response is an argument that it should be set by measurement rather than by convention. Brain-to-image work is where these interfaces are used as instruments. fMRI pipelines (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024) and EEG pipelines (Li et al., 2024; Zhang et al., 2025; Yu et al., 2026) all decode into a CLIP image space and render through a pooled-embedding adapter; several add a second pathway—a low-level image (Li et al., 2024), depth and text experts (Zhang et al., 2025)—to supply layout the pooled embedding lacks. We do not rank these systems. Our point is that the pooled-embedding interface they share is the one that fails our phantom on layout, so a layout claim made through it should be accompanied by the calibration in Section 4. Calibrated uncertainty for reconstructions has been reported rarely; we use split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) because its coverage holds without assumptions on the decoder.

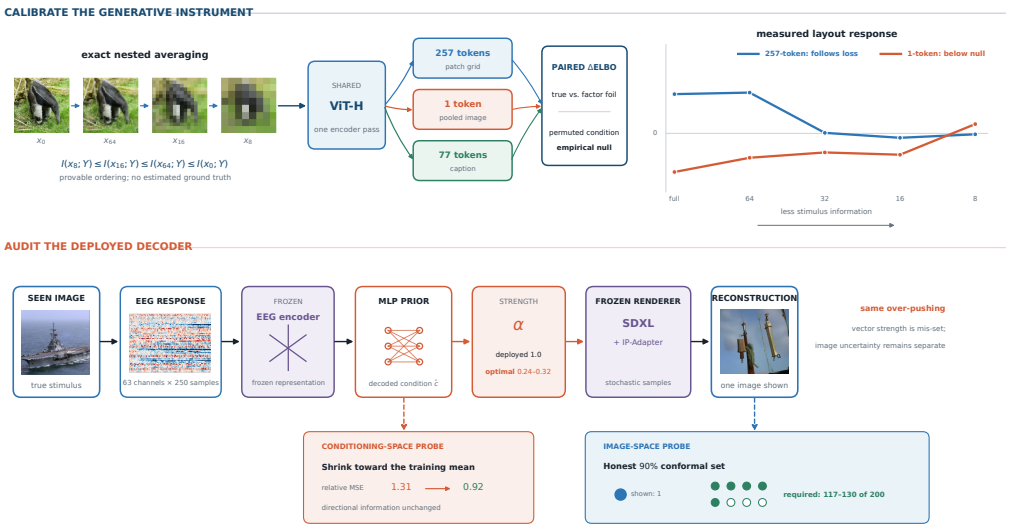

Figure 1: **Method overview: a concrete phantom and the principle it tests.** We first remove the brain from the loop and pass real stimuli through an exactly ordered channel chain. The same encoder pass is exposed through three conditioning interfaces, then scored with paired ΔELBO against a factor foil and an empirical null. A valid readout follows the guaranteed loss of information; the pooled single-token interface instead falls below its own null on layout. In deployed decoders, the same conditioning strength controls both the decoded vector and the frozen renderer. We therefore calibrate the vector by shrinkage and the emitted image separately by split conformal prediction.

2 Setup

2.1 A paired likelihood probe

We use the standard conditional-ELBO contrast. For a candidate image x with latent z, a condition c, and the diffusion model's noise predictor ϵ_θ,

$$s(x; c) \;=\; \mathbb{E}_{t,\epsilon}\big[\, \|\epsilon_\theta(z_t, t, \varnothing) - \epsilon\|^2 \;-\; \|\epsilon_\theta(z_t, t, c) - \epsilon\|^2 \big]\,, \tag{1}$$

which is the reduction in denoising error attributable to the condition, and a monotone proxy for the conditional ELBO gap (Ho et al., 2020; Kingma et al., 2021). The two terms share the same (t, ϵ) draws, so the estimator is paired and the variance from sampling t and ϵ largely cancels.

Pairing is easy to break in implementation. If the noise realisation ϵ is seeded from a position *within a batch* rather than from the image's global index, then the unconditional pass—typically computed once over all images in large batches—and the conditional pass—typically computed per trial over a handful of candidates—receive different noise for the same image, and the difference in Eq. 1 is no longer paired. All results here seed ϵ per image from its global index, so the pairing holds regardless of batching.

2.2 Conditioning interfaces

We compare four ways of delivering a condition to the same frozen SDXL (Podell et al., 2024) denoiser.

- 1**-token**: IP-Adapter (Ye et al., 2023) with the 1024-d projected pooled CLIP ViT-H (Radford et al., 2021) embedding: a single token. This is the interface used by the bulk of published EEG- and fMRI-to-image work.
- 257**-token**: IP-Adapter-plus, which routes the 257×1280 penultimate ViT-H patch grid through a Resampler, giving the condition spatial extent.

- **caption**: no adapter at all. A caption of the image, encoded by SDXL's own text encoders into 77×2048 tokens and entering through native cross-attention. Captions come from Qwen2.5-VL-3B (Bai et al., 2025) under greedy decoding, so the map remains deterministic.
- **layout-explicit caption**: the same, with the captioner instructed to state where things are ("left", "top", "centre").

The 1-token and 257-token interfaces are two readouts of a *single* ViT-H forward pass, which makes that pair a controlled comparison: any difference between them is attributable to the interface and not to the encoder. The text arms are an architecturally independent replication attempt—different modality, different encoder, no adapter.

2.3 What the guarantee does and does not cover

The data-processing inequality applies to the *channel*, that is to the degraded stimulus. The encoder—ViT-H, or the captioner—is a fixed apparatus applied at each stage of the chain, and monotonicity does not automatically pass through a nonlinear map. Our claim is therefore about the pipeline *as a whole*—encoder, interface, and critic—used as an instrument for the channel, which is exactly how such pipelines are used in practice. Within that claim, the 1-token/257-token pair isolates the interface, because the encoder pass is literally shared.

3 A phantom with provable information ordering

3.1 The channel chain

Every stimulus is resized to 512×512 and then box-averaged (area interpolation) to $k \times k$ and back to 512 with nearest-neighbour upsampling. Because 512 is an integer multiple of every k we use, averaging to 8 is *exactly* the 2×2 block average of the 16 result, and likewise up the chain; the nesting is exact rather than approximate, which is what licenses the inequality. We also include an off-chain channel, a *patch shuffle* that permutes 8×8 blocks: it destroys long-range layout while preserving local texture and the colour histogram exactly, and serves as a ground-truth negative for layout specifically.

3.2 Factors, foils and the empirical null

For each trial we form a two-way test between the true stimulus and a foil that differs along one factor: *category* (a different image), *colour*, *flip* (mirror), *blur*, and three layout manipulations—*scramble* (8×8 patch shuffle, $\sim$ 112 seams), *roll2d* (a two-dimensional half-roll, 2 seams), and *swaph* (a vertical half swap, 1 seam). We introduced the low-seam probes because *scramble* confounds layout with seam artefacts that a spatially uniform condition can also detect. We quote both low-seam probes throughout: Appendix B shows that on a single stimulus set neither one alone is decisive.

Accuracy alone is not interpretable here, for a reason we quantify in Section 4.2. We therefore report every cell against its own empirical null, obtained by re-running the identical computation with conditions permuted away from their trials, and summarise with

$$\text{efficiency} \;=\; \frac{\text{acc} - \text{null}}{1 - \text{null}}, \tag{2}$$

so that 0 means "indistinguishable from a mismatched condition" and 1 means perfect. Negative efficiency means the correct condition performs *worse* than a random one.

4 Instrument validity is interface-specific

All results in this section are subject 08 of THINGS-EEG (Gifford et al., 2022), $n = 200$ trials, $T = 10$ noise draws in the fine band, with the corrected pairing.

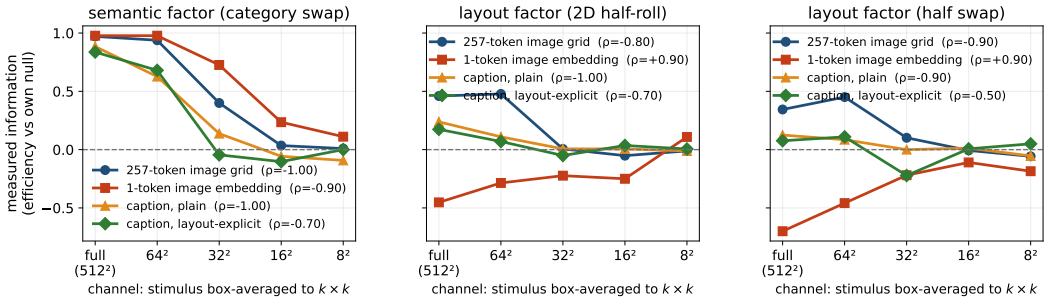

Figure 2: Instrument response curves. Channel information provably decreases from left to right, so a valid instrument must decrease with it. All four interfaces track the semantic axis correctly (left). On both low-seam layout axes (centre, right) the 1-token image embedding runs the wrong way: it sits far below its own null when the channel is richest and rises toward zero as information is provably destroyed.

Table 1: Spearman correlation between chain position and measured information. Negative passes; positive means the probe reports *more* information as the channel provably carries less. Starred cells are inverted.

factor	257-token grid	1-token image	caption, plain	caption, layout
category	-1.00	-0.90	-1.00	-0.70
blur	-1.00	-1.00	-1.00	-0.70
scramble	-0.90	-0.40	-0.60	-0.60
colour	-0.90	$+\mathbf{0.70}^{*}$	-0.20	-0.30
roll2d	-0.80	$+\mathbf{0.90}^{*}$	-1.00	-0.70
swaph	-0.90	$+\mathbf{0.90}^{*}$	-0.90	-0.50

4.1 Response curves

Figure 2 shows the curves and Table 1 the rank correlations between chain position and measured score. Negative values mean the instrument tracks the provable loss.

The result is the opposite of what an argument from bandwidth would predict. The 77-token caption interface is far more semantically bottlenecked than a 1024-d CLIP vector—a sentence simply cannot carry much layout—yet it reads layout correctly, both in sign at the informative end (roll2d 0.24, swaph 0.12) and in ordering along the chain ($\rho = -1.00, -0.90$). The single image token, with more nominal capacity, sits far below its own null when the channel is richest (roll2d -0.45, swaph -0.70) and its curve runs the wrong way along the chain.

The two statistics are not equally well determined, and it matters which one the argument rests on. A parametric bootstrap over the binomial trial counts (Appendix B) puts the 1-token layout ρ at $+0.90$ $[+0.10, +1.00]$: a five-point rank correlation barely excludes zero even at its most extreme, and on two redrawn stimulus sets its interval contains zero outright. Efficiency at the undegraded stimulus is far tighter—roll2d -0.45 $[-0.78, -0.20]$ against $+0.46$ $[+0.33, +0.57]$ for the 257-token readout—and the gap between the two interfaces reproduces on every set we tried. We therefore treat that gap as the result and ρ as a diagnostic that needs dynamic range to mean anything.

The raw numbers make the pathology concrete. For the 1-token interface on the undegraded stimulus under *swaph*, accuracy is 0.32 with the correct condition and 0.60 with a mismatched one: supplying the true condition makes the instrument *worse* than supplying a random one. Crucially the same instrument tracks category ($\rho = -0.90$) and blur ($\rho = -1.00$) correctly, so this is not a broken critic. The failure is specific to the axis the interface cannot represent.

Two secondary observations. First, instructing the captioner to describe position does not help and slightly hurts (roll2d $0.24 \to 0.17$, swaph $0.12 \to 0.08$, every ρ weaker), which locates the text arm's layout ceiling in SDXL's binding of position words rather than in the caption's vocabulary.

Table 2: Efficiency at the fully informative end of the chain (the undegraded stimulus).

factor	257-token grid	1-token image	caption, plain	caption, layout
category	0.97	0.98	0.88	0.84
blur	0.81	0.86	0.42	0.38
scramble	0.85	0.03	0.46	0.41
colour	0.50	−0.30	−0.02	0.09
roll2d	0.46	**−0.45**	0.24	0.17
swaph	0.34	**−0.70**	0.12	0.08

Table 3: Image-side bias: mean $m_u(\text{foil}) - m_u(\text{true})$ in units of 10^{-3} ϵ-MSE, with the percentage of trials on which the foil is unconditionally harder. Positive values push the two-way test toward the foil whatever the condition says.

pathway	category	flip	scramble	blur	swaph	roll2d
1-token	−2.7 (50%)	−0.2 (48%)	+30.9 (100%)	−80.4 (0%)	+3.5 (84%)	+6.1 (91%)
257-token	−3.1 (48%)	−0.6 (42%)	+47.1 (100%)	−87.6 (0%)	+10.7 (94%)	+16.2 (98%)
caption	−3.3 (48%)	−0.5 (44%)	+47.8 (100%)	−86.1 (0%)	+13.8 (96%)	+18.7 (98%)

Second, the 257-token interface passes an additional ground-truth check: with the patch shuffle as the channel, so that layout is destroyed by construction, it reports layout at $-0.01/-0.02$ while retaining category at 0.42. It correctly says there is no layout here.

4.2 Why the empirical null is load-bearing

Rearranging the two-way decision in Eq. 1,

$$s(x_{\text{true}}) > s(x_{\text{foil}}) \iff \underbrace{m_u(x_{\text{true}}) - m_u(x_{\text{foil}})}_{\text{condition never sees this}} > m_c(x_{\text{true}}) - m_c(x_{\text{foil}}),$$

where m_u, m_c are the unconditional and conditional errors. The unconditional term is a fixed per-trial offset. Table 3 measures it directly from the cached unconditional tables.

The design validates itself in three ways. *Category* and *flip* are unbiased (48–50%), as they must be since both foils are natural images. Every layout foil is unconditionally harder on 84–100% of trials, which is why raw layout accuracies sit far below 0.5 everywhere. *Blur* is biased the other way on 100% of trials. None of this is a property of an interface; it is a property of the foil. A two-way ΔELBO discrimination reported without an empirical null is therefore not interpretable, and this offset is large enough to manufacture or erase an effect on its own.

It also settles *flip*: with an image-side bias near zero and near-zero measured signal in all four interfaces, mirror-flip invisibility is a genuine null rather than a masked effect.

4.3 Conditioning strength is the knob

Table 4 varies the adapter scale, which controls how hard the condition pushes the denoiser. One knob does two things at once: turning the condition up improves measurement on the axes the interface can express (category $0.94 \to 1.00$, blur $0.70 \to 0.91$) and drives it negative on every axis it cannot (colour, scramble, roll2d, swaph). At scale 0.2 the inversion is essentially gone. The 257-token control shows no inversion at any dose, so this is not a generic consequence of pushing a condition hard—it is the consequence of pushing hard *along an axis the interface cannot represent*.

5 The same over-pushing in deployed decoders

The phantom concerns measurement. We now show that the quantity it implicates—how hard an under-informative condition is pushed—is also mis-set in systems built for reconstruction, and that this is visible without running a diffusion model at all.

Table 4: Dose–response in adapter scale on the undegraded stimulus. For the 1-token interface the layout inversion deepens monotonically with strength and nearly vanishes at 0.2, while the axes it can express improve over the same range. The 257-token interface shows no inversion at any dose.

scale	category	blur	scramble	colour	roll2d	swaph
1-token image embedding						
0.2	0.94	0.70	0.47	−0.05	**+0.03**	**−0.15**
0.4	0.93	0.67	0.32	−0.22	−0.20	−0.40
0.8 (deployed)	0.98	0.86	0.03	−0.30	−0.45	−0.70
1.0	1.00	0.91	0.05	−0.13	−0.32	−0.71
257-token grid (control)						
0.2	0.94	0.82	0.85	0.44	0.54	0.33
0.8	0.97	0.81	0.85	0.50	0.46	0.34
1.0	0.95	0.88	0.87	0.59	0.53	0.45

Table 5: Shrinkage in conditioning space, subject 08. Relative MSE is squared error divided by that of predicting $\bar{c}$ for every trial, so 1.0 means "no better than a constant". Subscript c denotes retrieval after subtracting $\bar{c}$ from both sides.

estimator	rel. MSE	top1	rank	top1$_c$	rank$_c$
prior mean only	1.000	–	–	–	–
decoded, full strength (deployed)	**1.310**	0.180	24.4	0.175	18.7
optimal magnitude rescale, s^*=0.62	1.123	0.180	24.4	0.110	27.6
shrink toward mean, α^*=0.31	**0.923**	0.060	58.3	**0.175**	**18.7**
per-direction shrinkage	0.910	0.065	53.6	0.150	19.3

5.1 Conditioning space: the decoded vector is used at $\approx 3\times$ its useful strength

Take a standard EEG-to-image pipeline: a frozen encoder (Song et al., 2024) followed by an MLP mapping EEG to CLIP ViT-H space, whose output $\hat{c}$ conditions the generator at full strength. Write $\bar{c}$ for the mean of the training targets and consider the shrunk condition $\tilde{c}(\alpha) = \bar{c} + \alpha(\hat{c} - \bar{c})$, a classical bias–variance trade (James & Stein, 1961). Everything below is fitted on one half of the test trials and evaluated on the other, cross-fitted both ways; $\bar{c}$ and the PCA basis come from training targets the decoder already saw.

Three readings, in Table 5. First, **as used, the decoded condition is a worse estimate of the true embedding than a constant**: relative MSE 1.31 on subject 08, and 1.26–1.47 (mean 1.36) across all ten THINGS-EEG subjects. This is not a length problem—rescaling optimally about the origin still leaves 1.12. Second, **it nevertheless carries real information and is merely used at the wrong strength**: top-1 retrieval is 0.18 against a chance of 0.005, and pulling the estimate about 70% of the way to the prior mean takes relative MSE from 1.31 to 0.92—from worse-than-constant to better-than-constant—*at no cost in directional information whatsoever*. The same crossing happens in every subject: raw relative MSE is always above 1.25, the MSE-optimal α^* always sits in $[0.24, 0.32]$, and the shrunk estimator always lands below 0.97. Centred retrieval is unchanged to three decimals, because shrinking toward the mean is a pure rescale in centred coordinates and cosine is scale invariant. The apparent collapse in the uncentred columns ($0.180 \to 0.060$) is an artefact of the uncentred cosine, which mixes the shared mean direction into every comparison; it is worth noting that the field's standard retrieval metric moves this much under an operation that provably changes no directional information. Third, EEG supports only the leading band of CLIP space: fitting a separate α per group of 64 PCA directions gives 0.38 for the top band and ≤ 0.15 elsewhere, with 13 of 16 bands below 0.1 (the same leading-band pattern holds in all ten subjects; see Appendix C). The additional gain of per-direction over global shrinkage is small and comes with a small centred-retrieval cost, so we report it as an honest null on the more elaborate variant. Variance explained by the optimally shrunk decoder is 7.7% on subject 08 and 6.0% on average.

Table 6: 90%-coverage conformal set size out of a 200-image gallery, mean over ten subjects. The system emits one image. NICE and DCTD are two independently trained EEG-to-image pipelines. References bound the axis: the oracle is the true ViT-H embedding rendered through the same pipeline, and the floor is the true stimulus used as its own reconstruction. The overconfidence ratio is from the renderer's own sample spread (subject 08).

representation	NICE	DCTD	oracle ViT-H	true stimulus	overconf. ratio
pixel	179.9	175.5	160.8	0.9	1.05
DINOv2	139.8	121.5	**1.0**	0.9	**2.17**
CLIP	129.5	116.7	**1.1**	0.9	**2.52**

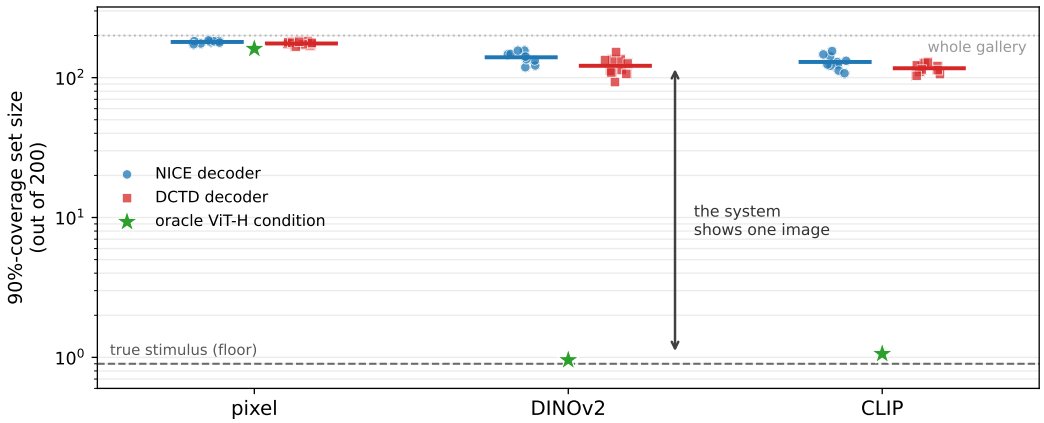

Figure 3: What an honest 90% prediction set costs. Each point is one subject; bars are means. The green star is a perfect condition through the *same* renderer. In the two semantic spaces it is a singleton while both deployed decoders need over a hundred images, a gap of two orders of magnitude—and the systems emit one. In pixel space the star sits inside the cloud, so there the limitation is the renderer rather than the signal.

5.2 Image space: the emitted reconstruction is overconfident

A single sharp image is a claim that the stimulus was *that* one. We check the claim with split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021): calibrate a threshold on half the trials so the true stimulus is covered on 90% of them, then report how large the resulting set is. Coverage holds by construction, so the size is the result.

The references in Table 6 are what make the numbers interpretable, and Figure 3 shows why the distinction matters. A perfect conditioning signal through the *same* SDXL and IP-Adapter yields an essentially singleton set (1.0, 1.1), so the axis has full dynamic range and a large set is not an artefact of a weak metric. Either EEG system needs about 120–140 of 200 images to keep its promise and shows one. The decomposition is also informative: at the pixel level the limitation is the renderer (oracle 160.8 versus EEG 176–180—even perfect conditioning cannot pin down pixels here), whereas at the semantic level it is the signal (oracle 1.0 versus EEG 117–140). The two pipelines agree on the diagnosis, and no individual subject in either approaches the oracle floor.

A second, independent check uses the renderer's own stochasticity. Resampling with different seeds produces an implicit posterior, whose within-trial spread is what the system believes its uncertainty to be. Comparing that spread to the actual error, the ratio is 2.17 in DINOv2 space and 2.52 in CLIP space, and exceeds 1 on 100% of trials. The samples agree with each other roughly 2.5× more than any of them agrees with the truth.

Table 7: Effect of shrinking the condition before rendering. Ratio is error/spread of the K=6 samples; 1.0 is honest. |set| is the 90%-coverage conformal size in an 80-image gallery, taking sample k=0 as the point estimate.

	α=1.0 (deployed)		α=0.5		α=0.3 (MSE-opt.)	
space	ratio	|set|	ratio	|set|	ratio	|set|
pixel	1.05	73.2	1.05	74.6	1.05	74.4
DINOv2	2.19	56.6	1.84	54.7	1.66	53.5
CLIP	2.56	39.6	2.20	38.5	2.00	51.9

5.3 Does shrinking the condition calibrate the renderer?

Section 5.1 is an argument about the *estimate*; it does not by itself establish anything about the *renderer*, and the two can disagree. We therefore render subject 08 at $\alpha \in \{1.0, 0.5, 0.3\}$ (K=6 samples, 80 trials) and re-measure both calibration quantities on the output.

Table 7 is a partial, honest result. Shrinkage does what the estimate-side argument predicted for the *posterior*: in CLIP space the within-trial spread rises ($0.157 \to 0.197$) and the overconfidence ratio falls ($2.56 \to 2.00$); DINOv2 moves the same way ($2.19 \to 1.66$). Pixel space was already nearly honest and does not change. What shrinkage does *not* do is make the emitted image honest. At the MSE-optimal α the samples are still overconfident on 100% of trials in both semantic spaces, and the CLIP conformal set grows ($39.6 \to 51.9$) because the point estimate itself has been pulled toward the prior. Matching conditioning strength to measured information is therefore necessary for the vector and insufficient for the picture: the renderer still collapses around a condition that cannot support a singleton.

6 Recommendations and limitations

Three practices follow. Calibrate the interface before trusting the number: the phantom needs no brain data and one pass over a few hundred stimuli, and an axis that sits below its empirical null is not a measurement; the null is not optional, since Table 3 shows the image-side offset alone can create or destroy an effect. Match conditioning strength to measured information: the deployed setting is both too strong for the estimate (Section 5.1) and the regime in which the measurement inverts (Table 4), and Table 7 shows that fixing it is necessary but leaves the renderer collapsed. Report sets, not images, with the reference columns of Table 6 so a large set is attributed to the signal or to the renderer.

The evidence has known edges. The phantom is one adapter family; Appendix A replicates it on SD1.5 and Appendix B on two disjoint stimulus sets, and in both the interface gap on layout reproduces while the inversion's size and the set of *other* inverted axes do not, so neither should be assumed across pipelines. The inversion is about three times deeper on the THINGS-EEG test images than on either training draw; we ask others to quote its sign, not its magnitude. A second subject is not an independent replication, because every subject sees the same test images. The data-processing guarantee covers the channel, so the claim is about the pipeline as an instrument, isolated to the interface only by the shared-encoder 1-token/257-token pair. Shrinkage is fitted to squared error, which is the estimate's loss and not the generator's, and the render-side experiment is one subject, 80 trials, three values of α.

7 Conclusion

A conditioned generative model used as a measuring device inherits the blind spots of its conditioning interface; on axes the interface cannot represent it does not go quiet but reports the wrong sign, more strongly the harder the condition is pushed. Which axes those are is not predictable from the interface's description, so they must be measured, which a ground-truth phantom makes cheap. The same over-pushing makes deployed reconstructions over-committed in conditioning space and overconfident in image space, by margins that can be quantified with guarantees.

AI USE STATEMENT

AI assistants were used for coding support and manuscript drafting; all experiments and scientific claims were verified by the authors.

REPRODUCIBILITY STATEMENT

The phantom protocol requires no brain data and is fully specified in Section 3. Code, cached scores, and configurations will be released. Primary runs use THINGS-EEG subject 08 with SDXL and the publicly released ViT-H IP-Adapter and IP-Adapter-plus weights for SDXL; the backbone replication uses SD1.5 with the corresponding SD1.5 weights. Code and cached scores are provided as supplementary material.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

William James and Charles Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379, 1961.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *International Conference on Machine Learning (ICML)*, 2024.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations (ICLR)*, 2024.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang Li. D^2-FOSA: Dual-diffusion guided EEG-to-image reconstruction with frequency-oriented semantic alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

A THE PHANTOM ON A SECOND BACKBONE

Section 4 measures one backbone, so the natural objection is that the interface effect is a quirk of SDXL. We repeated the entire phantom on Stable Diffusion 1.5 (Rombach et al., 2022) with the SD1.5 IP-Adapter (1-token) and IP-Adapter-plus (257-token) weights, at the same subject, the same n=200 trials, the same T=10 fine-band draws, the same adapter scale 0.8 and the same seed.

The comparison is tighter than a typical replication. Both adapter families consume the *same* ViT-H condition—the 1024-d projected vector for the 1-token pathway and the 257×1280 penultimate grid for the Resampler pathway—so the two runs reuse byte-identical condition tensors, channel images, foils and (t, ε) draws. The denoiser and the adapter that writes into it are the only things that differ; the two VAEs are not the same, so latents are re-encoded per backbone.

The interface effect replicates. Tables 8 and 9 give the two headline quantities of Section 4 for both backbones. On SD1.5 the 1-token interface again reads the two low-seam layout axes backwards—roll2d $\rho = +0.70$, swaph $\rho = +0.70$, against $+0.90$ and $+0.90$ on SDXL—while the 257-token readout of the same denoiser again respects the ordering on every axis. At the informative end the sign pattern of Table 8 agrees with SDXL in all six rows, including the collapse of the high-seam *scramble* probe to ≈ 0 for the 1-token interface (0.03 on SDXL, 0.02 on SD1.5). The spatial ablation reproduces as well: pooling the SD1.5 patch grid to its mean drops roll2d to -0.54 and swaph to -0.42, i.e. onto the 1-token interface's own values, so on this backbone too the 1-token pathway behaves like a multi-token pathway with its spatial content removed.

Table 8: Efficiency at the fully informative end of the chain (the undegraded stimulus), both backbones. The sign agrees in all six rows: on both layout axes the 1-token interface is below its own null while the 257-token readout of the same denoiser is well above it.

	SDXL		SD1.5	
factor	257-token	1-token	257-token	1-token
category	0.97	0.98	1.00	0.90
blur	0.81	0.86	0.97	0.78
scramble	0.85	0.03	0.45	0.02
colour	0.50	−0.30	0.79	0.11
roll2d	0.46	**−0.45**	0.60	**−0.25**
swaph	0.34	**−0.70**	0.42	**−0.51**

Table 9: Spearman correlation between chain position and measured information, both backbones. Negative passes. The 257-token interface passes on every axis of both backbones; the 1-token interface is inverted on both low-seam layout axes of both backbones. Which *other* axes invert is backbone-specific (colour, flip).

	SDXL		SD1.5	
factor	257-token	1-token	257-token	1-token
category	−1.00	−0.90	−0.90	−1.00
blur	−1.00	−1.00	−0.90	−1.00
scramble	−0.90	−0.40	−0.70	−0.30
colour	−0.90	**+0.70***	−1.00	−0.70
flip	−0.90	−0.80	−0.70	**+0.40***
roll2d	−0.80	**+0.90***	−0.60	**+0.70***
swaph	−0.90	**+0.90***	−0.30	**+0.70***

Which axes are blind does not transfer. The two backbones disagree on exactly the axes one would have to guess at. SDXL's 1-token interface inverts on colour (+0.70) and SD1.5's does not (−0.70); SD1.5's inverts mildly on flip (+0.40) where SDXL's does not (−0.80). So the blind set is not a property of the interface alone—it is a property of the interface *and* the denoiser it drives. This strengthens rather than weakens the argument of Section 6: if the failure mode were predictable from the adapter's token count, one calibration would serve every pipeline built on it, and it does not.

Magnitudes do not transfer either. SD1.5's 257-token arm passes the ordering test but is harsher about it: roll2d efficiency falls from 0.60 on the undegraded stimulus to −0.34 already at 64×64 and stays near −0.5 below that, whereas SDXL's stays within ±0.05 of its null over the same range. Respecting the ordering therefore does not license trusting a multi-token interface at the bottom of the chain; it only licenses trusting the direction of the change. A phantom run costs a few GPU-hours and answers both questions for the pipeline actually in use, which is the practice we are advocating.

B A DISJOINT STIMULUS SET, AND HOW MUCH THE TWO STATISTICS CARRY

Why this is not a second subject. C1 uses no brain data; the subject only selects which images the phantom runs on. In THINGS-EEG every subject is shown the same 200 test images in the same order, so re-running the phantom on another subject reproduces the previous run exactly. The question that matters—does the effect depend on which stimuli we happened to pick?—requires different images. We therefore drew two further sets (A, B) of 200 images, one per concept, from the 1654 THINGS-EEG *training* concepts. Each shares no image and no category name with the 200 test concepts, and the two share no image with each other. Corpus, preprocessing, foil code, channel chain and seed are unchanged.

Table 10: Efficiency on the undegraded stimulus across three mutually disjoint stimulus sets. The 257-token column is stable to within ± 0.11 everywhere; the 1-token column is negative on both layout axes in all three sets, and in none of the six comparisons do the two columns' 95% intervals overlap.

	test set (Section 4)		training draw A		training draw B	
factor	257-tok	1-tok	257-tok	1-tok	257-tok	1-tok
category	0.97	0.98	0.96	0.98	0.96	0.96
blur	0.81	0.86	0.87	0.81	0.85	0.85
scramble	0.85	0.03	0.82	0.00	0.84	0.07
colour	0.50	−0.30	0.47	−0.02	0.39	−0.13
roll2d	0.46	**−0.45**	0.57	**−0.08**	0.47	**−0.12**
swaph	0.34	**−0.70**	0.37	**−0.24**	0.35	**−0.16**

What transfers. The sign pattern of Table 10 agrees in all eighteen cells. The interface gap is the stable object: the 257-token readout reads both layout axes well in every set (roll2d $0.46/0.57/0.47$, swaph $0.34/0.37/0.35$) and the 1-token readout of the same encoder pass is below its own null in every set (roll2d $-0.45/-0.08/-0.12$, swaph $-0.70/-0.24/-0.16$). On draw B, for instance, the intervals are roll2d $0.47\,[0.34, 0.59]$ against $-0.12\,[-0.35, 0.06]$ and swaph $0.35\,[0.19, 0.49]$ against $-0.16\,[-0.40, 0.04]$. The *scramble* collapse reproduces throughout (0.03, 0.00, 0.07), so the high-seam probe stays uninformative for a spatially uniform condition, and the spatial ablation floor barely moves (roll2d $0.21/0.19/0.19$).

What does not transfer is the magnitude. The inversion is about three times deeper on the test images than on either training draw, on both axes, while the two training draws agree closely with each other—so this is a property of the stimulus set and not sampling noise. We have no mechanism to offer for it; the test concepts are a curated split and may simply be more canonically framed. The practical consequence is that an effect size measured on one stimulus set should not be quoted for another, and the colour axis is a sharper warning still: its inversion is -0.30, -0.02, -0.13, i.e. present, absent and marginal across three draws of the same corpus.

The rank correlation carries almost nothing. Bootstrapping the chain ρ shows how little a five-point rank correlation determines. For the 1-token interface it is $+0.90\,[+0.10, +1.00]$, $-0.40\,[-0.90, +0.90]$, $+0.60\,[-0.60, +0.90]$ on roll2d across the three sets and $+0.90\,[+0.10, +1.00]$, $+0.80\,[-0.40, +1.00]$, $+0.10\,[-0.70, +0.90]$ on swaph. Five of these six intervals contain zero and all of them overlap, so the swings in sign are noise in a statistic that was never precise enough to lead with. Efficiency against the empirical null is the quantity that holds up: the probability that the 1-token layout efficiency on the undegraded stimulus is negative is $100/78/90\%$ on roll2d and $100/99/94\%$ on swaph. *swaph* is therefore the steadier of the two low-seam probes and *roll2d* the one with more headroom; neither dominates, which is why we report both. A practitioner running this phantom should read the sign of efficiency against the empirical null, on more than one stimulus set, and should not read a rank correlation over five chain positions at all.

C SHRINKAGE ACROSS TEN SUBJECTS

Table 11 reports the Section 5.1 quantities for every THINGS-EEG subject. Relative MSE of the deployed (unshrunk) prediction is above 1.25 in all ten; the MSE-optimal α^* lies in $[0.24, 0.32]$; after shrinkage, relative MSE is below 0.97 in all ten. Centred top-1 is unchanged by shrinkage in every subject, as the geometry requires. Per-direction α is largest in the leading PCA band (0.32–0.41) and at or below 0.15 in the remaining fifteen bands.

Table 11: Conditioning-space shrinkage, all ten subjects. "raw" is the deployed prediction; "shrunk" uses the cross-fitted α^*. Chance top-1 is 0.005.

sub	raw MSE	α^*	shrunk MSE	top1$_c$	rank$_c$	leading α
1	1.41	0.25	0.95	0.13	31.3	0.32
2	1.39	0.25	0.95	0.10	31.1	0.32
3	1.35	0.28	0.94	0.16	24.9	0.35
4	1.37	0.30	0.92	0.16	18.7	0.36
5	1.26	0.25	0.97	0.08	42.8	0.33
6	1.34	0.28	0.94	0.14	21.1	0.35
7	1.44	0.26	0.94	0.16	22.9	0.32
8	1.31	0.31	0.92	0.18	18.7	0.38
9	1.47	0.24	0.95	0.11	30.6	0.32
10	1.26	0.32	0.92	0.18	17.3	0.41
mean	1.36	0.27	0.94	0.14	25.9	0.35

D COVERAGE ON A SECOND DECODER

The NICE and DCTD reconstructions in Table 6 are independently trained pipelines that share the same renderer family. Per-subject CLIP set sizes range 107–155 (NICE) and 103–128 (DCTD). No subject in either system produces a singleton, and none approaches the oracle floor of 1.1.

E CLIP VERSUS ΔELBO AS A SAMPLE VERIFIER

On 50 trials with $K{=}8$ posterior samples, a CLIP-space nearest-neighbour verifier and the paired ΔELBO verifier of Eq. 1 agree on the selected sample on 4% of trials—indistinguishable from chance at $1/8$. Reconstruction metrics of the two selections are close (CLIP 0.83 vs. 0.82, DINOv2 0.77 vs. 0.79), so the disagreement is not that one verifier is broken. It is that the two scores are answering different questions, which is consistent with the interface-specificity of Section 4: a CLIP-space pick and a diffusion-likelihood pick are two different instruments and should not be treated as interchangeable.