BRAINSTEER: PROBING AND REDUCING THE NEURAL EVIDENCE UTILIZATION GAP IN SINGLE-TRIAL FMRI TIME-SERIES-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Does equivalence under a trained neural readout imply equivalence for generative reconstruction? We introduce BrainSteer, a diagnostic and closed-loop steering framework on frozen DynaDiff. BrainSteer decomposes six-TR evidence into a retained component visible to the trained temporal readout and an Exact-Null component annihilated by it, then injects TRAIN-centered evidence through independent zero-preserving, state-conditioned paths. It adds 947,264 trainable parameters (0.947M). On Shared1000, four-subject PixCorr improves from 0.21766 to 0.24590 and SSIM from 0.30864 to 0.35637; non-development relative gains are 12.51% and 14.32%. On frozen VAL64, CORRECT exceeds SHUFFLED structurally in every subject, and five paired ablations favor Full on both structural metrics. Changing Exact-Null identity changes reconstruction while preserving the frozen readout. Time-Only retains about 0.5% of Full structural specificity; the rollout-initial hidden summary preserves most specificity but remains below Full feedback. A 20-state audit finds lower target predicted-x_0 error under matched evidence throughout. Matched Exact-Null evidence reduces target-relative error, but neither null-only nor coherent retained/null donor transfer reliably recovers donor identity under fixed recipient Official conditioning. Perceptual/semantic metrics do not improve uniformly. For this studied readout, readout-equivalent evidence need not be reconstruction-equivalent when an independent zero-preserving evidence path is available; its residual utility is context-dependent.

1 INTRODUCTION

Generative reconstruction from brain activity tests whether learned representations suffice for downstream generation. Large-scale 7T fMRI resources such as the Natural Scenes Dataset (NSD) (Allen et al., 2022) and latent diffusion models (Rombach et al., 2022) enable increasingly natural and semantically coherent reconstructions (Takagi & Nishimoto, 2023; Chen et al., 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024). DynaDiff retains continuously evolving fMRI rather than a static estimate, using shifted temporal windows to expose decodable visual content (Careil et al., 2025).

Yet image quality and neural evidence utilization are not the same quantity. An image prior can produce plausible samples from a small neural bottleneck; BrainBits showed that high reconstruction scores can survive severe neural compression (Mayo et al., 2024). Consequently, poor reconstruction does not identify a unique failure mode. Relevant evidence may be absent from the measurement, destroyed by the decoder, preserved but poorly transferred to the generator, or present yet useful only at particular generative states. Conventional decoder-vs-decoder comparisons confound these possibilities.

This ambiguity concerns information limitation versus utilization limitation. We ask: *if two pre-aggregation neural evidence tensors are identical under a trained decoder's temporal readout, must they also be equivalent for reconstruction?* We answer this for a frozen DynaDiff pipeline rather than making a universal claim about all decoders. We treat the learned temporal aggregation as a measurable operator. Let $U_{\mathrm{raw}} \in \mathbb{R}^{d\times 6}$ denote the six-TR pre-aggregation representation and $D = U_{\mathrm{raw}} - U_0$ the brain-dependent evidence. For the learned temporal readout $w \in \mathbb{R}^6$, define

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.

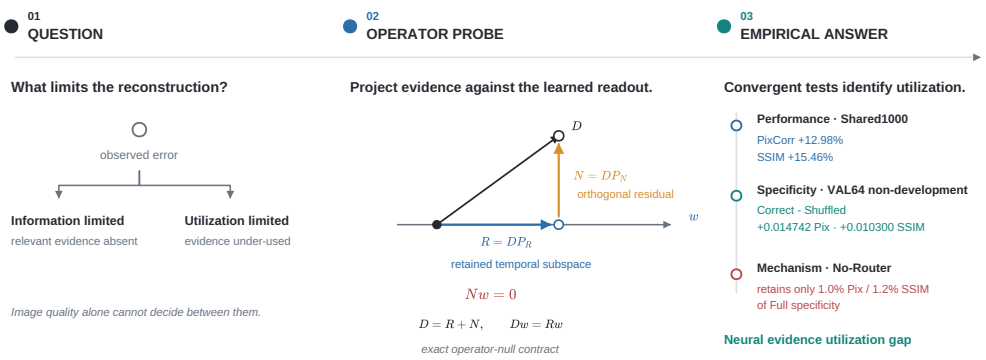


Figure 1: **From failure ambiguity to an operational sufficiency test.** (a) Reconstruction error alone cannot distinguish missing information from under-utilized evidence. (b) DynaDiff's learned temporal readout defines decoder-induced equivalence classes: Exact-Null changes satisfy $Nw = 0$ and are therefore invisible to the frozen readout. (c) Shared1000 performance, frozen VAL64 stimulus specificity, and the No-Router ablation provide the first empirical chain. SUB02/05/07 are non-development subjects.

readout equivalence by $D_1 \sim_w D_2$ iff $D_1 w = D_2 w$. We then construct an exact orthogonal decomposition $D = R + N$ with $RP_R = R$ and $Nw = 0$. Any change confined to N therefore leaves the frozen temporal readout exactly unchanged: $D \sim_w D + N'$ for every $N'w = 0$. Under an operational reconstruction-sufficiency hypothesis, such readout-equivalent evidence should not support systematic stimulus-specific reconstruction differences. We call the reconstruction-relevant variation that survives inside these decoder-induced equivalence classes, together with readout-visible evidence that is preserved but under-exploited downstream, the *neural evidence utilization gap*.

Accessing additional features is not sufficient evidence by itself: an extra adapter could improve images through generic regularization or a stronger image prior. BrainSteer therefore imposes two controls. First, the operational intervention is zero-preserving: if both TRAIN-centered evidence components are zero, the added evidence responses and hidden-state update are exactly zero. Second, the principal scientific contrast is CORRECT $-$SHUFFLED . The two conditions use identical parameters and diffusion machinery, differing only in whether neural evidence belongs to the current stimulus. A positive CORRECT $-$SHUFFLED therefore supports stimulus-identity-dependent reconstruction utility, without implying neurophysiological causality.

Evidence availability does not ensure utility. BrainSteer conditions each branch on the current U-Net state, the evidence, and diffusion time, and re-evaluates this compatibility across 20 scheduler states. A 20-state audit shows an early target-relative redirection that persists to the terminal latent. Cross-stimulus transplantation finds matched Exact-Null utility without reliable donor-identity transfer under fixed recipient Official conditioning.

Our contributions are threefold. (1) We test operational reconstruction sufficiency for a trained temporal readout: readout-equivalent evidence need not be reconstruction-equivalent, revealing a *neural evidence utilization gap*. (2) We combine exact operator-relative retained/null geometry with a zero-preserving, state-conditioned dual-path controller, adding 947,264 trainable parameters to a frozen baseline. (3) Shared1000/VAL64, direct ablations, identity, feedback, trajectory, and donor interventions establish stimulus-specific structural utility in both streams, test how state-conditioned control expresses it, and bound its portability.

2 RELATED WORK

Generative reconstruction from neural representations. Classical work reconstructed natural movies from visual cortex with encoding models and image priors (Nishimoto et al., 2011); modern systems map fMRI to VAE, CLIP, or diffusion representations (Takagi & Nishimoto, 2023; Chen et al., 2023; Ozcelik & VanRullen, 2023). MindEye/MindEye2 improve retrieval and cross-subject alignment (Scotti et al., 2023; 2024), while NeuroPictor and Brain-IT use fMRI-derived seman-

tic/structural controls (Huo et al., 2024; Zalcher et al., 2026). BrainSteer asks a different question: after a strong neural readout is frozen, what stimulus-specific evidence does it still fail to use?

Temporal and state-dependent neural guidance. DynaDiff preserves fMRI time series and exposes temporal evolution through shifted reconstructions (Careil et al., 2025). MindHier argues that static neural guidance can mismatch the coarse-to-fine needs of generation (Zhang et al., 2026). Our novelty is not dynamic guidance alone: evidence is partitioned by the exact retained/null geometry of a trained temporal readout, then used through a zero-preserving, state-conditioned intervention with matched-vs-shuffled specificity controls.

Decoders as probes of evidence utilization. Task-evoked fMRI can actively steer internal representations of language models (Xiao et al., 2026), and NeuralDiffuser uses fMRI-derived visual features to guide diffusion reconstruction (Li et al., 2025); we therefore do not claim the first use of brain signals for generative guidance. BrainBits asks how much recorded brain information reconstruction systems actually use (Mayo et al., 2024), and visual-imagery work adapts a pretrained perception decoder through latent functional alignment and compares decoding across cortical regions (Spera et al., 2026). BrainSteer uses a frozen brain-to-image readout to define equivalence classes and tests stimulus-specific reconstruction utility within them.

Operator null spaces and frozen generative control. Null-space reasoning is established in inverse problems such as DDNM (Wang et al., 2023). ControlNet adds spatial conditioning to frozen diffusion backbones (Zhang et al., 2023), while IP-Adapter introduces a lightweight image-prompt adapter with decoupled cross-attention (Ye et al., 2023). These mechanisms are architectural precedents; they are not our claimed contribution. BrainSteer uses the learned temporal readout of a trained brain decoder itself to define the null space, preserves the Official condition, and treats the added path as an intervention for testing readout-equivalent evidence.

3 METHOD

3.1 SETUP, ZERO REFERENCE, AND FROZEN BASELINE

For each stimulus, the frozen DynaDiff brain encoder produces pre-aggregation evidence $U_{\text{raw}} \in \mathbb{R}^{d\times 6}$ from six consecutive fMRI volumes, with $d = 1552$, and a learned temporal readout $w \in \mathbb{R}^6$. We define

$$D = U_{\text{raw}} - U_0, \qquad U_0 = U(x = 0). \tag{1}$$

We use U_0 as an interpretable no-input functional origin rather than a tuned hyperparameter. More generally, for any fixed encoder-space reference B, TRAIN-only branch centering cancels the reference:

$$\begin{aligned} \widetilde{R}_B &= (U_{\text{raw}} - B)P_R - \mathbb{E}_{\text{tr}}[(U_{\text{raw}} - B)P_R], \\ \widetilde{N}_B &= (U_{\text{raw}} - B)P_N - \mathbb{E}_{\text{tr}}[(U_{\text{raw}} - B)P_N], \end{aligned} \tag{2}$$

so both are TRAIN-centered projections of U_{raw}, algebraically invariant to any fixed encoder-space reference; U_0 gives the uncentered increment an interpretable functional origin.

The subject-specific brain encoder, official condition pathway, Versatile Diffusion backbone (Xu et al., 2023), U-Net/LoRA, sampler, and scheduler remain frozen. Only the BrainSteer side-car is optimized.

3.2 EXACT RETAINED/NULL TEMPORAL DECOMPOSITION

Let $u = w/\|w\|_2$, $P_R = uu^\top$, and $P_N = I - P_R$. We define

$$R = DP_R, \qquad N = DP_N, \qquad D = R + N, \tag{3}$$

so that

$$Nw = 0, \qquad Dw = Rw. \tag{4}$$

The split is one retained plus five null dimensions in the six-dimensional *temporal coordinate system*; it is not a one-TR-vs-five-TR split, and both branches remain combinations of all six time points. "Null" is strictly relative to the learned DynaDiff aggregation operator.

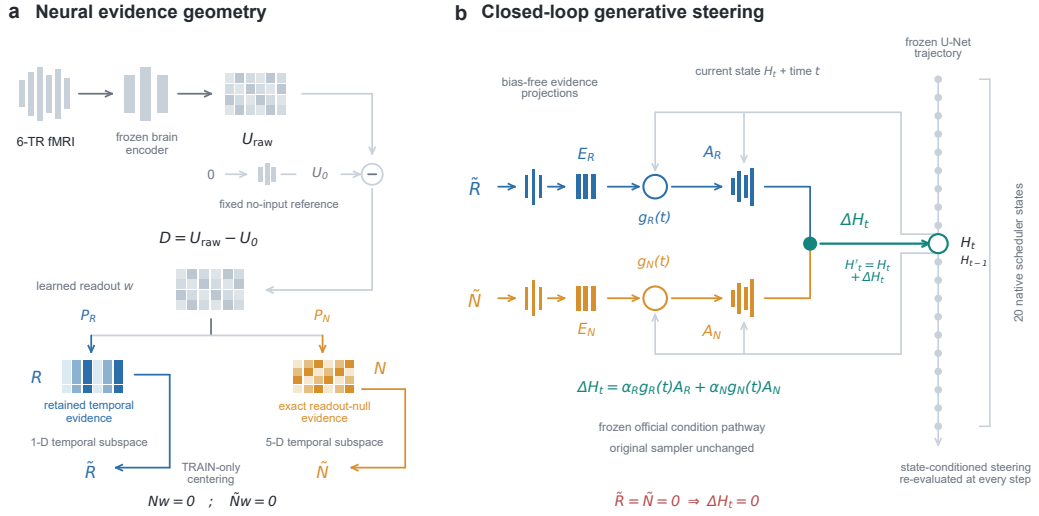


Figure 2: **BrainSteer-v2.** (a) Six-TR pre-aggregation evidence is referenced to the zero-input encoder response, split into retained and exact readout-null temporal components, and TRAIN-centered. (b) The centered branches form independent evidence paths whose state-conditioned responses merge into ΔH_t inside the frozen U-Net. The right rail shows the 20 native scheduler states; the same side-car is re-evaluated at every state.

The implementation then subtracts subject-specific TRAIN-only branch means,

$$\widetilde{R} = R - \mu_R^{\mathrm{tr}}, \qquad \widetilde{N} = N - \mu_N^{\mathrm{tr}}, \tag{5}$$

where $\mu_R^{\mathrm{tr}} = \mathbb{E}_{\mathrm{tr}}[R]$ and $\mu_N^{\mathrm{tr}} = \mathbb{E}_{\mathrm{tr}}[N]$ are the corresponding subject-specific empirical branch means computed once on TRAIN512 and reused unchanged at validation and test time. Linearity preserves the operator contracts: $\widetilde{R}P_R = \widetilde{R}$ and $\widetilde{N}w = 0$.

3.3 Zero-preserving evidence paths

The canonical implementation transposes each $[B, 1552, 6]$ branch to $[B, 6, 1552]$ and applies a bias-free $1552 \to 128$ projection, non-affine layer normalization, GELU, and multiplicative temporal/type modulation. Bias-free SharedKV maps the 128-D evidence tokens to K/V for the hooked cross-attention dimensions; frozen Official queries and output weights are reused, with no output bias in the added path. BrainSteer hooks 32 `attn2` modules. Hence zero operational evidence cannot create an affine residual:

$$\widetilde{R} = \widetilde{N} = 0 \Longrightarrow E_R = E_N = 0 \Longrightarrow \Delta H_t = 0. \tag{6}$$

Across the tokenizer, K/V projections, routers, and per-module residual scales, the side-car contains 947,264 trainable parameters; the frozen architecture archive reports this count for the canonical BrainSteer-v2 implementation.

3.4 State–evidence compatibility and internal attention

At scheduler state t, let H_t be the current internal U-Net state. Independent routers produce

$$\begin{aligned} g_R(t) &= \tanh\phi_R(H_t, E_R, t), \\ g_N(t) &= \tanh\phi_N(H_t, E_N, t). \end{aligned} \tag{7}$$

In code, hidden/evidence summaries are projected bias-free to a 32-D compatibility space and the timestep acts through multiplicative modulation, so there is no state-only or timestep-only additive gate path. Branch-specific attention responses A_R, A_N remain decoupled and the hook applies

$$H'_t = H_t + \alpha_R g_R(t) A_R + \alpha_N g_N(t) A_N. \tag{8}$$

Here α_R and α_N are learned per-module residual scales. H'_t propagates through the remaining frozen U-Net. Re-evaluating Eq. (8) at all 20 native scheduler states yields closed-loop steering. We interpret the router only as a model-level compatibility mechanism, not a neural timing law.

3.5 SPECIFICITY-AWARE TRAINING

Training pairs matched (Correct) and mismatched (Shuffled) evidence at the same native scheduler state. Let $x_{\text{off}}, x_{\text{cor}}, x_{\text{shuf}}$ be the corresponding x_0 predictions, z_0 the ground-truth latent, and $s_t > 0$ the frozen TRAIN-only normalization from `baseline_scale()`. With $\ell_t(a, b) = \text{MSE}(a, b)/s_t$,

$$\begin{aligned} \mathcal{L}_{\text{cor}} &= \ell_t(x_{\text{cor}}, z_0), \\ \mathcal{L}_{\text{shuf}\to\text{gt}} &= \ell_t(x_{\text{shuf}}, z_0), \\ \mathcal{L}_{\text{preserve}} &= \ell_t(x_{\text{shuf}}, \text{sg}(x_{\text{off}})), \\ \mathcal{L}_{\text{rank}} &= \text{ReLU}(\mathcal{L}_{\text{cor}} - \mathcal{L}_{\text{shuf}\to\text{gt}}). \end{aligned} \tag{9}$$

The code additionally uses $\mathcal{L}_{\text{gate}} =$ `shuffled_gate_l2(`b=1) and $\mathcal{L}_\alpha =$ `alpha_l2()`. The frozen objective is

$$\mathcal{L} = \mathcal{L}_{\text{cor}} + 10\mathcal{L}_{\text{preserve}} + \mathcal{L}_{\text{rank}} + 0.1\mathcal{L}_{\text{gate}} + 10^{-4}\mathcal{L}_\alpha. \tag{10}$$

Thus Shuffled is explicitly kept near detached Official and discouraged from opening large gates; No-Specificity removes the specificity constraints while retaining the steering architecture.

3.6 COUNTERFACTUAL EVALUATION

At evaluation, OFFICIAL uses the frozen baseline, CORRECT uses stimulus i evidence, and SHUFFLED uses evidence from $j \neq i$ with the same steering model, initial noise, and scheduler randomness. We distinguish

$$\begin{aligned} \Delta_{\text{perf}} &= C - O, \qquad \Delta_{\text{spec}} = C - S, \\ \Delta_{\text{mech}} &= (C - S)_{\text{full}} - (C - S)_{\text{ablation}}. \end{aligned} \tag{11}$$

$C - O$ can include generic trajectory correction; $C - S$ is the primary stimulus-identity contrast; Δ_{mech} is tested by direct paired bootstrap. To isolate branch identity within the frozen full model, we additionally use a 2×2 retained/null identity factorial. Using R_i, N_i as shorthand for the operational retained/null evidence associated with stimulus i, the four arms are $CC = (R_i, N_i)$, $CS = (R_i, N_{\pi(i)})$, $SC = (R_{\pi(i)}, N_i)$, and $SS = (R_{\pi(i)}, N_{\pi(i)})$. We define

$$\begin{aligned} \Delta_N &= \tfrac{1}{2}[(CC - CS) + (SC - SS)], \\ \Delta_R &= \tfrac{1}{2}[(CC - SC) + (CS - SS)], \\ I_{R\times N} &= CC - CS - SC + SS. \end{aligned} \tag{12}$$

Here Δ_N and Δ_R test whether matching the stimulus identity of each evidence stream improves reconstruction after averaging over the identity of the other stream; $I_{R\times N}$ measures departure from additivity.

4 EXPERIMENTS

4.1 DATA, TRAINING, AND METRICS

We follow the frozen NSD (Allen et al., 2022)/DynaDiff pipeline on SUB01/02/05/07 with the same six-TR window (TR $= 4/3$ s). SUB01 is the development subject; after design choices were frozen, SUB02/05/07 were excluded from further architecture/hyperparameter development and are termed *non-development subjects*. TRAIN512 supplies TRAIN-only branch statistics and optimization data, VAL64 is the frozen paired counterfactual set, and NSD Shared1000 is the final-scale evaluation. BrainSteer never retrains DynaDiff.

Only the 947,264-parameter BrainSteer side-car is optimized; all decoder/generator weights remain frozen. A mechanical audit across all eight production checkpoints gives the same 947,264 count, versus 1.578–1.583B parameters for the frozen subject-specific Official DynaDiff pipelines, i.e., 0.0598–0.0600% overhead (mean 0.05995%). Training uses 3000 updates, AdamW at 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.99)$, zero weight decay, gradient clipping at 1.0, CFG 3.5, and the frozen 20-state DDPM schedule of our Official implementation (Ho et al., 2020). Loss weights are $\lambda_{\text{preserve}} = 10$, $\lambda_{\text{rank}} = 1$, $\lambda_{\text{gate}} = 0.1$, and $\lambda_\alpha = 10^{-4}$. Validation is never used for optimization. Exact-null numerical and mechanism integrity checks accompany the corresponding experiments.

Following the frozen DynaDiff evaluation pipeline (Careil et al., 2025), we report PixCorr and SSIM (Wang et al., 2004) as primary structural metrics, plus AlexNet (Krizhevsky et al., 2012)/Inception-v3 (Szegedy et al., 2016)/CLIP (Radford et al., 2021) identification, EfficientNet-B (Tan & Le, 2019)/SwAV (Caron et al., 2020) distances, mIoU, and DreamSim distance (Fu et al., 2023). Higher is better except EfficientNet-B, SwAV, and DreamSim; DreamSim specificity is therefore reported as $S - C$.

4.2 Frozen counterfactual protocols

VAL64 uses paired per-image $C - S$ contrasts and bootstrap intervals. The direct paired mechanism ablation compares Full against R-only, N-only, Single-D, No-Router, and No-Specificity by direct paired bootstrap. The retained/null identity factorial uses $CC = (R_i, N_i)$, $CS = (R_i, N_{\pi(i)})$, $SC = (R_{\pi(i)}, N_i)$, and $SS = (R_{\pi(i)}, N_{\pi(i)})$, with main effects and interaction from Eq. (12). Router controls keep the same checkpoints and stochastic path: Frozen-State reuses the rollout-initial hidden summary, whereas Time-Only replaces online routing with a fixed module/branch/timestep mean-gate schedule estimated from Full VAL64 gate telemetry. We treat Time-Only as a secondary, evaluation-distribution-matched (transductive) open-loop control, not as an independent confirmatory comparison. The 20-state trajectory audit compares Correct and Shuffled target predicted-x_0 error with pointwise paired-bootstrap intervals.

The final portability audit asks whether Exact-Null utility transfers donor identity. With recipient Official condition C_i fixed, the Exact-Null donor transplantation holds R_i fixed and substitutes donor N_j for $N = 0$; donor retrieval is among 63 non-recipient targets (chance Top-1 $= 1/63$, Top-5 $= 5/63$, MRR $= 0.075052$). The coherent retained/null donor factorial tests coherent transfer with $II = (R_i, N_i)$, $IJ = (R_i, N_j)$, $JI = (R_j, N_i)$, and $JJ = (R_j, N_j)$. Each recipient uses four pre-specified donor derangements, two model seeds, three generation seeds, the same 20-state rollout, and fixed initial/transition noise. Endpoint definitions, permutations, integrity checks, and complete tables are provided in the appendix.

For Fig. 5, seed 6911 and generation seed 2026083101 are fixed. Examples are first restricted to positive PixCorr/SSIM $C - S$ trials, then selected for qualitative interpretability; each ROI uses identical coordinates across GT/Official/Shuffled/Correct and is visualization-only.

5 Results

5.1 Shared1000: structural fidelity improves at scale

Table 1: **Shared1000 aggregate performance.** BrainSteer averages frozen seeds 6911/6912. Bold is better within each split; EffNet, SwAV, and DreamSim are distances (↓), while the other metrics are higher-is-better.

		All four		Non-dev.	
Group	Metric	Official	BrainSteer	Official	BrainSteer
Low-level	PixCorr↑	.2177	**.2459**	.2084	**.2345**
	SSIM↑	.3086	**.3564**	.3076	**.3517**
	Alex2↑	**.9409**	.9400	**.9351**	.9342
	Alex5↑	**.9719**	.9710	**.9696**	.9691
High-level	Incep↑	**.9104**	.9091	**.9117**	.9103
	CLIP↑	.8983	**.9001**	.8982	**.9002**
	EffNet↓	**.6952**	.7018	**.6926**	.6970
	SwAV↓	**.3718**	.3742	**.3719**	.3732
Additional	mIoU↑	**.1263**	.1261	**.1251**	.1248
	Dream↓	**.4982**	.5063	**.4965**	.5033

Across all subjects, PixCorr/SSIM rise by 0.028241 (12.98%) and 0.047728 (15.46%); non-development gains remain 0.026070 (12.51%) and 0.044064 (14.32%). Every subject improves

on both. CLIP changes only $+0.001792$, while AlexNet/Inception decrease slightly, EffNet/SwAV distances worsen, DreamSim distance increases by 0.008161, and mIoU decreases by 0.000274. The gain is therefore concentrated in pixel/spatial structure.

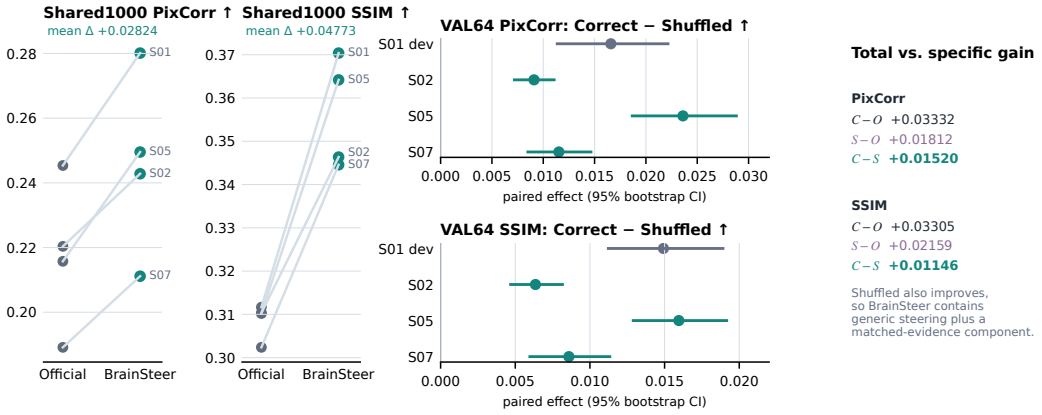


Figure 3: **Performance and stimulus specificity.** (a,b) Shared1000 subject-wise PixCorr/SSIM. (c,d) VAL64 CORRECT $-$SHUFFLED effects with 95% paired bootstrap intervals. SUB01 is development; SUB02/05/07 are non-development. All structural intervals are above zero.

5.2 VAL64: THE GAIN DEPENDS ON EVIDENCE IDENTITY

The all-subject mean CORRECT $-$SHUFFLED is 0.015202 PixCorr / 0.011456 SSIM; on non-development subjects it is 0.014742 / 0.010300. The effect is positive in every subject, with each interval above zero. Because CORRECT and SHUFFLED share the same added network and generation process, this supports stimulus-identity-dependent structural utility rather than merely the presence of a steering adapter. Boundary metrics do not follow the same pattern: non-development mIoU $C - S = -0.000245$ and DreamSim $S - C = -0.002895$ (lower DreamSim is better).

5.3 MECHANISM ABLATIONS

The five ablations are independently trained variants. Single-D preserves $D = R + N$ but removes dual-stream structure; No-Router retains evidence paths, attention, and residual scales; No-Specificity removes rank and shuffled-gate losses while retaining the architecture and remaining losses. The appendix reports full values and secondary DreamSim/mIoU outcomes; Fig. 4 summarizes structural results.

R-only retains 47.6%/55.9% of full PixCorr/SSIM specificity, supporting under-utilization of readout-visible evidence. N-only retains 76.8%/79.8%, showing recoverable stimulus-specific utility in the exact readout-null branch; these percentages are not information-theoretic comparisons between R and N. Direct paired Full-minus-ablation effects are positive for every ablation on both structural metrics. The largest is No-Router: $+0.014598$ PixCorr [0.012524, 0.016735] and $+0.010177$ SSIM [0.008685, 0.011707]. No-Specificity also reduces structural specificity, whereas DreamSim/mIoU do not uniformly favor Full.

5.4 COUNTERFACTUAL MECHANISM INTERVENTIONS

Retained/Exact-Null identity: does it matter? Within the frozen Full policy, matching Exact-Null identity yields $\Delta_N = +0.00825$ PixCorr (95% CI $[+0.00699, +0.00957]$) and $+0.00577$ SSIM ($[+0.00486, +0.00672]$); matching Retained identity yields $\Delta_R = +0.00649$ ($[+0.00543, +0.00757]$) and $+0.00453$ ($[+0.00376, +0.00533]$). Both are positive in all three non-development subjects. Holding R_i fixed while replacing only N_i gives $CC - CS = +0.00775$ PixCorr and $+0.00522$ SSIM. Because both N_i and $N_{\pi(i)}$ are exactly null to the frozen temporal readout, this provides an operational counterexample to reconstruction sufficiency: readout-equivalent

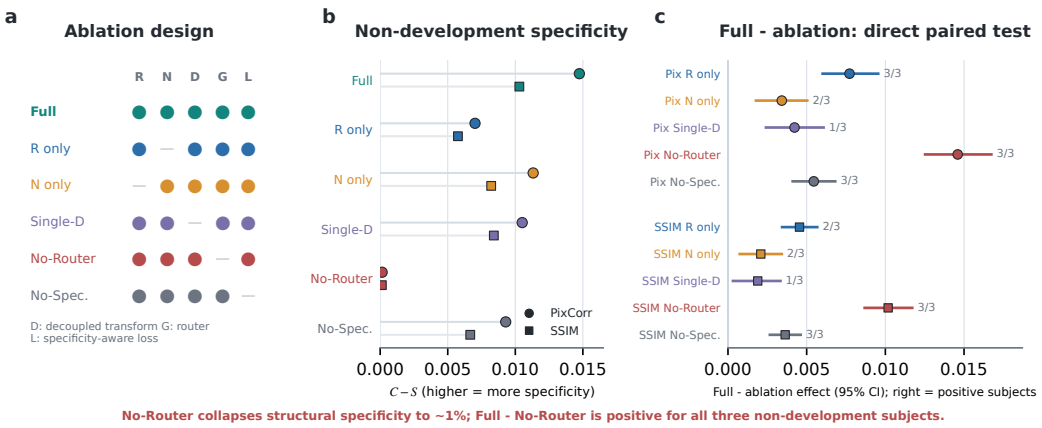


Figure 4: **Mechanism evidence.** (a) Ablation component map. (b) Non-development PixCorr/SSIM specificity. (c) Direct paired Full-minus-ablation effects: dot = Δ_{mech}, bar = 95% CI, right label = number of non-development subjects with a positive effect (not significant-subject count). All aggregate structural CIs exclude zero; No-Router nearly eliminates specificity.

evidence is not reconstruction-equivalent. The negative interaction (PixCorr -0.00100; SSIM -0.001086) indicates partially overlapping, sub-additive contributions.

State feedback: how is evidence used? Full specificity is $+0.014742$ PixCorr / $+0.010300$ SSIM. Frozen-State preserves 89.5%/94.1% but remains below Full by $+0.001549$ PixCorr (95% CI $[+0.000845, +0.002261]$) and $+0.000611$ SSIM ($[+0.000025, +0.001199]$), with 3/3 subjects in the same direction. Time-Only reduces specificity to $+0.000075/+0.000052$, about 0.5% of Full. Thus the router is not a fixed denoising-time gain curve: online evidence/state compatibility drives most specificity, while tracking evolving H_t adds a smaller reproducible benefit.

Trajectory: when does it act? Matched evidence has lower target predicted-x_0 error at all 20 scheduler states, with every pointwise 95% bootstrap interval above zero and all three non-development subjects positive throughout. The relative advantage is $+6.57\%$ at $t = 951$, peaks early at $+7.96\%$ at $t = 901$, and remains $+4.56\%$ at $t = 1$ (95% CI $[+4.04\%, +5.06\%]$). This supports early redirection followed by a persistent target-relative advantage, not monotonic accumulation.

Exact-Null transplantation: is donor identity portable? Exact-Null evidence remains effective in its own recipient context: restoring N_i relative to $N = 0$ reduces Full target-relative error by 6.77% over the trajectory (95% CI $[6.50\%, 7.04\%]$) and 6.59% at the terminal state ($[6.36\%, 6.84\%]$). Yet transplanting N_j does not reliably redirect toward donor j: donor-direction AUC advantage is $+0.001665$ ($[-0.000090, +0.003370]$), MRR is 0.078646 versus chance 0.075052, and donor-label permutation $p = 0.173$. A weak donor-attraction distance effect ($+0.000602$, $[+0.000172, +0.001025]$) likewise lacks reliable donor attribution (MRR 0.080779, permutation $p = 0.083$).

Coherent donor transfer: could a retained/null pair suffice? Coherent (R_j, N_j) transfer likewise yields no reliable donor-specific attribution on the tested interaction-based endpoint: the Full coherence interaction is negative (-0.001733, $[-0.002153, -0.001339]$), while true-donor specificity is near zero with an interval spanning zero ($+0.000013$, $[-0.000073, +0.000098]$; MRR 0.070751, permutation $p = 0.802$). The interaction tests departure from additivity on the chosen error scale; positive interaction is not necessary for donor portability. Under fixed recipient Official conditioning, Exact-Null utility is therefore context-dependent rather than a reliably portable donor-content code.



Figure 5: Four frozen positive cases; matched-coordinate GT/Official/Shuffled/Correct zooms. Per-image $C - S$ values are not population estimates. Full views and selection: appendix.

6 Discussion and Limitations

What the experiments establish. The studied frozen DynaDiff temporal readout defines equivalence classes that do not guarantee reconstruction equivalence. Exact-Null identity changes structural utility without changing the readout, while R-only exposes downstream under-utilization of readout-visible evidence. State-feedback and trajectory audits connect this residual utility to state-conditioned generative control from the earliest audited state through the terminal latent.

Utility is conditional, not portable content. The transplantation tests bound this interpretation. Matched N_i is effective relative to $N = 0$, but N_j does not reliably recover donor identity and coherent (R_j, N_j) transfer does not yield reliable donor-specific attribution on the tested endpoints while the recipient Official condition C_i is fixed. The negative coherence interaction is therefore not evidence for anti-donor coding; with donor specificity near zero, cross-stimulus incompatibility is the more conservative interpretation. BrainSteer is better viewed as exploiting context-dependent corrective evidence than as decoding a standalone semantic vector transferable between stimuli.

Relation, efficiency, and scope. Task-fMRI can steer pretrained language representations (Xiao et al., 2026), and prior brain-to-image systems such as NeuroPictor use fMRI-derived generative controls (Huo et al., 2024); our narrower novelty is operator-level intervention on an exact decoder-induced null space. BrainSteer adds 0.947M trainable parameters while keeping the Official DynaDiff weights frozen, so a substantial trainable-capacity increase is an implausible sole explanation, although parameter count does not imply negligible FLOPs or latency. The main effect is structural: small CLIP changes and the lack of improvement in AlexNet/Inception, EffNet/SwAV, DreamSim, and mIoU rule out universal semantic/perceptual improvement. The null construction is tied to one learned readout, only four NSD subjects are studied, and image-wise bootstrap does not replace larger subject samples. The transplantation result is conditional on fixed recipient Official conditioning and does not prove intrinsic non-portability under every possible conditioning intervention. Trajectory CIs are pointwise, not family-wise, and none of these model-level interventions establishes cortical, neurophysiological, or information-theoretic causality.

7 Conclusion

BrainSteer tests reconstruction sufficiency relative to a frozen dynamic fMRI readout. Exact-Null identity interventions show that readout-equivalent evidence need not be reconstruction-equivalent when an independent zero-preserving path is available. Retained-branch results reveal complementary downstream under-utilization; state-conditioned controls and trajectory audits connect residual evidence to generative correction. Matched Exact-Null utility does not imply portable donor content under fixed recipient Official conditioning. In this studied pipeline, temporal-readout invisibility does not imply reconstruction irrelevance: residual evidence supports context-dependent corrective control.

AI USE STATEMENT

Generative AI tools were used as research assistants for brainstorming and refining research questions, conceptual framing, hypotheses, methodology, experimental design, candidate controls, ablations, and evaluation strategies, and for assisting with interpretation and stress-testing of experimental results. They also assisted with research-method implementation, mathematical claims or proofs, and scientific figure creation. They were used to assist with drafting and revising portions of the manuscript, language editing, manuscript organization, formatting, and consistency and citation checks. AI-generated suggestions were treated as non-authoritative inputs. All adopted methodological decisions, experimental protocols, implementations, mathematical arguments, statistical analyses, numerical results, figures, interpretations, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The Method and Experiments sections specify the frozen baseline, retained/null construction, zero-preserving controller, objective, training settings, and counterfactual design. The appendix records the inference units, within-stimulus seed aggregation, bootstrap and donor-label randomization procedures, complete mechanism tables, scheduler states, parameter audit, qualitative-selection disclosure, and source boundaries.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Marlène Careil, Yohann Benchetrit, and Jean-Rémi King. Dynadiff: Single-stage decoding of images from continuously evolving fMRI. *arXiv preprint arXiv:2505.14556*, 2025.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22710–22720, 2023.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2208.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. NeuroPictor: Refining fMRI-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision (ECCV)*, volume 15109 of *Lecture Notes in Computer Science*, pp. 56–73, 2024. doi: 10.1007/978-3-031-72983-6_4.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012.

Haoyu Li, Hao Wu, and Badong Chen. NeuralDiffuser: Neuroscience-inspired diffusion guidance for fMRI visual reconstruction. *IEEE Transactions on Image Processing*, 34:552–565, 2025. doi: 10.1109/TIP.2025.3526051.

David Mayo, Christopher Wang, Asa Harbin, Abdulrahman Alabdulkareem, Albert Shaw, Boris Katz, and Andrei Barbu. BrainBits: How much of the brain are generative reconstruction methods using? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1723.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011. doi: 10.1016/j.cub.2011.08.031.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13:15666, 2023. doi: 10.1038/s41598-023-42891-8.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 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*, volume 36, 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, 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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024.

Fabrizio Spera, Tommaso Boccato, Michal Olak, Sara Cammarota, Matteo Ciferri, Michelangelo Tronti, Nicola Toschi, and Matteo Ferrante. Seeing the imagined: a latent functional alignment in visual imagery decoding from fMRI data. *arXiv preprint arXiv:2604.15374*, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, 2019.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Mingqing Xiao, Kai Du, and Zhouchen Lin. Beyond representational alignment with brain-guided language models for robust reasoning. *Nature Machine Intelligence*, 8:1275–1289, 2026. doi: 10.1038/s42256-026-01278-w.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7754–7765, 2023.

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.

Amit Zalcher, Roman Beliy, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. In *International Conference on Learning Representations (ICLR)*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Xu Zhang, Ruijie Quan, Wenguan Wang, and Yi Yang. Moving beyond diffusion: Hierarchy-to-hierarchy autoregression for fMRI-to-image reconstruction. In *International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX

A SCOPE AND INFERENCE UNIT

This appendix reports frozen numerical details referenced by the main paper. No new model or scientific claim is introduced here. SUB01 is the development subject; SUB02/05/07 are non-development subjects.

For the primary VAL64 specificity analysis, each stimulus-level $C - S$ value first averages the two model seeds (6911/6912) and three generation seeds (2026083101/02/03), so the six replicates are not treated as independent observations. Each subject therefore contributes 64 paired stimulus values; 50,000 stimulus-bootstrap resamples give percentile 95% intervals. The direct paired mechanism ablations use the same within-stimulus averaging, form $(C - S)_{\text{Full}} - (C - S)_{\text{Ablation}}$ on each matched stimulus, and independently resample the 64 paired stimulus differences within SUB02, SUB05, and SUB07. Each bootstrap draw then equally averages the three fixed-subject means; its reported 95% intervals are percentile intervals. These aggregate intervals quantify stimulus-level uncertainty conditional on the three fixed subjects, not population-level subject uncertainty.

For the mechanism interventions below, model-seed and generation-seed replicates are likewise averaged within each subject/stimulus (or subject/recipient) before stimulus/recipient-level paired bootstrap; experiment-specific aggregation is stated in the corresponding sections. For mIoU, we reuse the frozen DynaDiff evaluator on raw/unblurred reconstructions with `compute_miou(..., eval_res=512)`; its segmentation stage uses the ViT-Adapter/Mask2Former path with a 172-class metric.

Unless stated otherwise, these mechanism interventions use the two frozen BrainSteer model seeds 6911/6912 and generation seeds 2026083101/02/03. The native 20-state scheduler sequence is

$$951, 901, 851, 801, 751, 701, 651, 601, 551, 501,$$
$$451, 401, 351, 301, 251, 201, 151, 101, 51, 1.$$

B MODEL-SIZE AUDIT

BrainSteer-v2 contains 947,264 trainable parameters (0.947264M). The count agrees between serialized state tensors and the instantiated runtime model for all eight production checkpoints (SUB01/02/05/07 $\times$ seeds 6911/6912). The corresponding frozen subject-specific Official DynaDiff pipelines contain 1.578–1.583B parameters, so the added trainable-parameter overhead is 0.0598–0.0600% (mean 0.05995%). This is a parameter-count statement only; it does not imply negligible FLOPs, latency, or memory overhead.

Table 2: Mechanical parameter audit. BrainSteer count is 947,264 for every production checkpoint.

Subject	Official DynaDiff params	BrainSteer params	Overhead (%)
SUB01	1,582,924,866	947,264	0.059843
SUB02	1,580,680,674	947,264	0.059928
SUB05	1,578,757,746	947,264	0.060001
SUB07	1,578,203,682	947,264	0.060022
Mean	1,580,141,742	947,264	0.05995

C DIRECT PAIRED MECHANISM ABLATIONS

For each ablation, we test

$$\Delta_{\text{mech}} = (C - S)_{\text{Full}} - (C - S)_{\text{Ablation}}$$

directly. Positive values favor Full. Intervals use 50,000 paired bootstrap resamples. Full non-development specificity is 0.014742 PixCorr and 0.010300 SSIM. These direct contrasts establish

contribution relative to the specified variants; they do not imply that every component is necessary for any nonzero specificity.

Each direct paired mechanism ablation is an independently trained variant for SUB02/05/07 and model seeds 6911/6912 (30 ablation checkpoints in total), rather than an inference-only toggle of the Full checkpoint. The frozen interventions are: R-only keeps R and sets $N = 0$; N-only sets $R = 0$ and keeps N; Single-D keeps the same total evidence $D = R + N$ but uses D as the retained stream and sets the null stream to zero; No-Router fixes $g_R = g_N = 1$ while retaining the tokenizer, R/N evidence paths, attention, and residual scales; and No-Specificity sets the rank and shuffled-gate loss weights to zero while retaining the steering architecture and the remaining losses.

Table 3: Non-development direct paired mechanism-ablation specificity. DreamSim is a distance, so its positive-specificity orientation is $S - C$; the other columns use $C - S$. Retained percentages are relative to Full on the two primary structural metrics.

Metric	Full	R-only	N-only	Single-D	No-Router	No-Specificity
PixCorr $C - S$	.014742	.007019	.011322	.010512	.000144	.009290
SSIM $C - S$	.010300	.005755	.008215	.008411	.000123	.006666
DreamSim $S - C$	-.002895	+.000519	-.000682	-.000373	-.000260	-.002122
mIoU $C - S$	-.000245	-.000136	-.000659	+.000743	+.001114	-.000897
Pix retained	100.0%	47.6%	76.8%	71.3%	1.0%	63.0%
SSIM retained	100.0%	55.9%	79.8%	81.7%	1.2%	64.7%

Table 4: Direct Full-minus-ablation structural contrasts.

Metric	Ablation spec.	Full−Abl.	95% CI
	PixCorr		
R-only	.007019	.007723	[.006016,.009544]
N-only	.011322	.003420	[.001778,.005039]
Single-D	.010512	.004230	[.002413,.006074]
No-Router	.000144	.014598	[.012524,.016735]
No-Specificity	.009290	.005452	[.004112,.006814]
	SSIM		
R-only	.005755	.004545	[.003439,.005672]
N-only	.008215	.002085	[.000747,.003419]
Single-D	.008411	.001889	[.000316,.003335]
No-Router	.000123	.010177	[.008685,.011707]
No-Specificity	.006666	.003634	[.002663,.004617]

D RETAINED/EXACT-NULL IDENTITY FACTORIAL

The four arms are

$$CC = (R_i, N_i), \qquad CS = (R_i, N_{\pi(i)}),$$
$$SC = (R_{\pi(i)}, N_i), \quad SS = (R_{\pi(i)}, N_{\pi(i)}).$$

The same permutation and stochastic path are shared across arms. For each subject/stimulus/arm, results are averaged over 2 model seeds $\times 3$ generation seeds before 50,000 paired bootstrap resamples. The main effects and interaction are

$$\Delta_N = \tfrac{1}{2}[(CC - CS) + (SC - SS)],$$
$$\Delta_R = \tfrac{1}{2}[(CC - SC) + (CS - SS)].$$
$$I_{R\times N} = CC - CS - SC + SS.$$

Both branch-identity main effects are positive in all three non-development subjects. The interaction is negative on both metrics, indicating sub-additivity/overlap rather than positive synergy. In particular,

Table 5: Retained/Exact-Null identity-factorial PixCorr results. "Non-dev." is the equal-weight aggregate over SUB02/05/07.

Contrast	SUB02	SUB05	SUB07	Non-dev.
CC	.663595	.660561	.657755	.660637
CS	.659127	.647689	.651836	.652884
SC	.659250	.651537	.653158	.654648
SS	.654495	.636956	.646235	.645895
Δ_N	+.004612	+.013727	+.006421	+.008253
Δ_R	+.004488	+.009879	+.005100	+.006489
Interaction	-.000288	-.001709	-.001004	-.001000
CC−CS	+.004468	+.012872	+.005919	+.007753
SC−SS	+.004756	+.014581	+.006923	+.008753
CC−SC	+.004344	+.009025	+.004598	+.005989
CS−SS	+.004632	+.010733	+.005602	+.006989

Table 6: Retained/Exact-Null identity-factorial SSIM results.

Contrast	SUB02	SUB05	SUB07	Non-dev.
CC	.394111	.398858	.395537	.396169
CS	.391122	.390627	.391087	.390945
SC	.391183	.392889	.392464	.392178
SS	.387758	.382902	.386947	.385869
Δ_N	+.003207	+.009109	+.004983	+.005766
Δ_R	+.003146	+.006847	+.003607	+.004533
Interaction	-.000437	-.001755	-.001067	-.001086
CC−CS	+.002988	+.008232	+.004450	+.005223
SC−SS	+.003425	+.009986	+.005517	+.006309
CC−SC	+.002928	+.005970	+.003073	+.003990
CS−SS	+.003364	+.007724	+.004140	+.005076

Table 7: Non-development retained/Exact-Null identity-factorial paired-bootstrap intervals.

Contrast	PixCorr [95% CI]	SSIM [95% CI]
Δ_N	.008253 [.006994,.009567]	.005766 [.004856,.006724]
Δ_R	.006489 [.005427,.007569]	.004533 [.003758,.005330]
Interaction	-.001000 [-.001494,-.000525]	-.001086 [-.001472,-.000736]
CC−CS	.007753 [.006569,.008999]	.005223 [.004386,.006116]
SC−SS	.008753 [.007383,.010170]	.006309 [.005299,.007364]
CC−SC	.005989 [.005023,.006978]	.003990 [.003288,.004711]
CS−SS	.006989 [.005787,.008217]	.005076 [.004196,.005986]

the direct $CC - CS$ contrast changes only Exact-Null identity while both alternatives remain null to the frozen temporal readout.

E ROUTER-MODE / STATE-FEEDBACK CONTROLS

FROZEN-STATE reuses the rollout-initial hidden summary while retaining evidence, timestep, weights, and attention. TIME-ONLY uses a fixed module×branch×timestep mean-gate schedule estimated from Full VAL64 telemetry pooled over Correct/Shuffled trajectories and generation seeds. It is therefore a secondary, evaluation-distribution-matched (transductive) open-loop control, not an independent confirmatory comparison.

Table 8: Router-mode / state-feedback specificity. DreamSim is a distance, so positive specificity is $S - C$.

Metric	Mode	SUB02	SUB05	SUB07	Non-dev.
PixCorr $C - S$	Full	.009100	.023605	.011520	.014742
	Frozen-State	.007903	.021978	.009697	.013193
	Time-Only	.000082	.000032	.000111	.000075
SSIM $C - S$	Full	.006352	.015956	.008590	.010300
	Frozen-State	.005657	.015580	.007828	.009688
	Time-Only	.000009	.000073	.000073	.000052
DreamSim $S - C$	Full	-.001216	-.005363	-.002106	-.002895
	Frozen-State	-.001085	-.003463	-.002588	-.002379
	Time-Only	+.000067	-.000303	-.000078	-.000105

Table 9: Direct Full-minus-router-control contrasts.

Metric	Control	Effect	95% CI
PixCorr	Frozen-State	+.001549	[+.000845,+.002261]
PixCorr	Time-Only	+.014667	[+.012637,+.016801]
SSIM	Frozen-State	+.000611	[+.000025,+.001199]
SSIM	Time-Only	+.010248	[+.008776,+.011780]
DreamSim	Frozen-State	-.000516	[-.001171,+.000155]
DreamSim	Time-Only	-.002790	[-.004035,-.001544]

The reconstruction integrity audit reports 3,456/3,456 expected reconstructions for each mode (10,368 total) and 0/4,608 Official SHA mismatches. Frozen-State therefore removes evolving-state feedback without changing the frozen Official path; it retains most primary specificity but remains below Full. Time-Only retains only about 0.5% of Full primary specificity.

F DENOISING TRAJECTORY AUDIT

At each native denoising state, the audit records the scheduler's predicted-x_0 and measures target MSE in the frozen latent coordinates. The effect is

$$E_t = \frac{\bar{d}_{S,t} - \bar{d}_{C,t}}{\bar{d}_{S,t}},$$

so a positive value means Correct is closer to the target. Slot 0 is the *first* denoising iteration (scheduler timestep $t = 951$); slot 19 is the final iteration ($t = 1$). Thus "earliest" refers to rollout order, not the numerical ordering of the timestep label. Within each subject/stimulus/timestep, six trajectories (2 model seeds × 3 generation seeds) are averaged before 50,000 paired stimulus-bootstrap resamples. The same bootstrap resampling is reused across timesteps; intervals are pointwise, not simultaneous bands.

Table 10: All 20 native scheduler states in the denoising trajectory audit. Positive effect means lower target predicted-x_0 MSE under Correct evidence.

Slot	t	Non-dev. effect	95% CI	SUB02	SUB05	SUB07
0	951	.065740	[.057377,.073997]	.045581	.108131	.043508
1	901	.079553	[.070358,.088844]	.059324	.125432	.053902
2	851	.072651	[.064507,.080578]	.055804	.110533	.051615
3	801	.075478	[.066641,.084162]	.057171	.113461	.055802
4	751	.074764	[.066091,.083156]	.055419	.111232	.057642
5	701	.072568	[.064205,.080837]	.051994	.111595	.054115
6	651	.066036	[.058080,.073996]	.046989	.097885	.053233
7	601	.061848	[.054627,.068944]	.042564	.094221	.048759
8	551	.057662	[.051074,.064221]	.039176	.088773	.045037
9	501	.053458	[.047462,.059463]	.037190	.081442	.041744
10	451	.052508	[.046514,.058493]	.035870	.082453	.039202
11	401	.051395	[.045601,.057189]	.035192	.079309	.039684
12	351	.050056	[.044419,.055650]	.033539	.078103	.038525
13	301	.049414	[.043917,.054876]	.033322	.076693	.038228
14	251	.049038	[.043584,.054426]	.032913	.076352	.037848
15	201	.048273	[.042895,.053594]	.032355	.075116	.037348
16	151	.047367	[.042059,.052608]	.031567	.073852	.036682
17	101	.046346	[.041148,.051468]	.030706	.072488	.035844
18	51	.045635	[.040502,.050704]	.030169	.071365	.035370
19	1	.045552	[.040429,.050612]	.030103	.071261	.035293

All 20 pointwise intervals exclude zero, and all three non-development subjects remain positive at every state. The effect appears at the first audited iteration, peaks at $t = 901$, then decays while remaining positive through the terminal latent. This is not monotonic accumulation and does not identify a separate local causal effect for each denoising step. Official model-seed trajectory max error is exactly zero for SUB02/05/07.

G EXACT-NULL DONOR TRANSPLANTATION

The Exact-Null donor transplantation fixes recipient Official condition C_i and retained evidence R_i, comparing

$$\begin{aligned} \text{MATCH} &= (R_i, N_i), \quad N0 = (R_i, 0), \\ \text{TRANSPLANT} &= (R_i, N_j), \qquad j \neq i. \end{aligned}$$

Each recipient uses four pre-specified 64-way donor derangements generated from base seed 2026083104 with offsets 0, 2, 7, and 14; the first map exactly reuses the retained/Exact-Null identity-factorial permutation. Donors are not selected by outcome. The same initial noise and DDPM transition noise are used across matched counterfactual arms.

For each candidate target k, direction alignment and attraction are

$$A_{ijk,t} = \cos\Big(x_{i,t}^{\text{changed}} - x_{i,t}^{\text{base}},\; z_k - x_{i,t}^{\text{base}}\Big),$$

$$G_{ijk,t} = \frac{\text{MSE}(x_{i,t}^{\text{base}}, z_k) - \text{MSE}(x_{i,t}^{\text{changed}}, z_k)}{\text{MSE}(x_{i,t}^{\text{base}}, z_k)}.$$

The true donor is ranked among 63 non-recipient candidates: chance Top-1 $= 1/63 = 0.015873$, Top-5 $= 5/63 = 0.079365$, and MRR $= 0.075052$. The executed analysis first averages the 2 model seeds $\times$ 3 generation seeds at the full candidate-score-vector level for each subject, recipient, donor map, and scheduler state. For each of the four pre-registered donor maps, donor-relative endpoints are then computed separately and only those endpoint values are averaged within recipient. For a score vector s, donor advantage is s_j minus the mean score of the other 62 non-recipient, non-donor candidates; the donor rank is among all 63 non-recipient candidates with larger score ranked better

and exact ties assigned midrank. "AUC" denotes the unweighted arithmetic mean over the 20 native scheduler slots, not a timestep-weighted trapezoidal integral; terminal endpoints are computed from the stored terminal candidate scores. Thus trajectory MRR is the mean reciprocal rank over the four donor maps and 20 scheduler slots. Intervals use 50,000 recipient bootstrap resamples, independently resampling the 64 recipients within each fixed subject and equally weighting SUB02/05/07 in the aggregate; reported 95% intervals are percentile intervals.

Table 11: Exact-Null donor-transplantation direction alignment. ADV is true-donor advantage over non-donor candidates.

Endpoint	Full	Frozen-State	Time-Only
AUC ADV	+.001665	+.001488	+.001352
95% CI	[-.000090,+.003370]	[-.000285,+.003185]	[-.000563,+.003204]
AUC MRR	.078646	.078548	.078056
AUC perm. p	.172641	.182391	.209140
Term. ADV	+.002034	+.001672	+.001901
95% CI	[-.000139,+.004158]	[-.000369,+.003707]	[-.000406,+.004198]
Term. MRR	.081329	.081895	.081356
Top-1	.016927	.019531	.014323
Top-5	.096354	.100260	.091146
Term. perm. p	.112344	.094745	.111744

Table 12: Exact-Null donor-transplantation attraction endpoint. A small positive Full/Frozen-State distance tendency is present, but donor-label attribution remains unreliable.

Endpoint	Full	Frozen-State	Time-Only
AUC ADV	+.000602	+.000540	+.000388
95% CI	[+.000172,+.001025]	[+.000127,+.000945]	[-.000068,+.000839]
AUC MRR	.080779	.079868	.078162
AUC perm. p	.082596	.119244	.225189
Term. ADV	+.000768	+.000658	+.000607
95% CI	[+.000197,+.001344]	[+.000128,+.001191]	[-.000020,+.001231]
Term. MRR	.082267	.080875	.077434
Top-1	.016927	.016927	.011719
Top-5	.102865	.093750	.089844
Term. perm. p	.081096	.130593	.311484

Table 13: Matched Exact-Null positive control in the donor-transplantation experiment: target-error reduction of matched N_i relative to $N = 0$.

Mode	AUC	95% CI	Terminal	95% CI
Full	.067693	[.065047,.070356]	.065931	[.063567,.068390]
Frozen-State	.065709	[.063135,.068329]	.060533	[.058217,.062941]
Time-Only	.054882	[.053664,.056150]	.056887	[.055450,.058374]

The donor-label permutation tests in Tables 11 and 12 are one-sided MRR tests, not tests of ADV. They keep each six-run-averaged candidate-score landscape fixed. For every subject, recipient, and donor-map slot, a pseudo donor is drawn uniformly from the 63 non-recipient candidates; for trajectory AUC the same pseudo label is used across all 20 states. For each possible label, reciprocal rank is computed among the 63 candidates (using the same midrank rule), and trajectory values are averaged over the 20 slots. Each null replicate then averages the four donor slots and 64 recipients within subject and equally averages the three fixed subjects. The reported p-value is $\{1 + \#(\mathrm{MRR}_{\mathrm{perm}} \geq \mathrm{MRR}_{\mathrm{obs}})\}/(20{,}000 + 1)$; the terminal test uses the analogous terminal rank.

Table 14: Direct router-mode contrasts for AUC donor advantage/MRR in the Exact-Null donor-transplantation experiment.

Endpoint	Contrast	ADV Δ [95% CI]	MRR Δ [95% CI]
Direction	Full− Frozen-State	+.000177 [-.000073,+.000428]	+.000098 [-.001832,+.002036]
Direction	Full− Time-Only	+.000314 [-.000445,+.001084]	+.000590 [-.003293,+.004498]
Attraction	Full− Frozen-State	+.000062 [-.000002,+.000127]	+.000911 [-.001373,+.003438]
Attraction	Full− Time-Only	+.000214 [-.000022,+.000450]	+.002617 [-.001373,+.006700]

The permutation base seed is 2026090801. This distinction permits a positive ADV interval together with a non-significant donor-label MRR test, because they test different statistics and nulls.

The positive control shows that matched Exact-Null evidence is effective in its recipient context. Donor identity, however, is not reliably recovered: Full direction-alignment intervals cross zero and permutation tests are not significant; the attraction endpoint shows a weak distance tendency without reliable donor attribution. Exact-null swap audits give maximum frozen-readout deviations of 4.47×10^{-8} (SUB02) and 5.96×10^{-8} (SUB05/SUB07).

H COHERENT RETAINED/NULL DONOR FACTORIAL

The coherent retained/null donor factorial fixes recipient Official condition C_i and uses the same donor maps as the Exact-Null donor transplantation, comparing

$$II = (R_i, N_i), \quad IJ = (R_i, N_j),$$
$$JI = (R_j, N_i), \quad JJ = (R_j, N_j).$$

Initial noise and the DDPM stochastic path are shared across all four arms. For candidate target k, the primary coherence interaction is

$$I_k = \frac{d_{JI,k} + d_{IJ,k} - d_{JJ,k} - d_{II,k}}{d_{II,k}}.$$

The interaction measures departure from additivity on the normalized candidate-distance scale. A positive I_j would indicate a super-additive coherent-pairing effect at the true donor, but positive donor transfer does not require $I_j > 0$. We therefore interpret I_j and the interaction-derived specificity/retrieval endpoints as coherence/non-additivity tests rather than necessary conditions for portability. The executed analysis first averages the 2 model seeds $\times$ 3 generation seeds at the candidate-MSE level. It then computes the factorial endpoints separately for each of the four donor maps and averages those endpoint values within recipient. "AUC" is the unweighted arithmetic mean over the 20 native scheduler slots. Recipient-level intervals use 50,000 paired bootstrap resamples, independently resampling 64 recipients within each fixed subject and equally weighting SUB02/05/07; reported 95% intervals are percentile intervals.

True-donor specificity is

$$\mathrm{SPEC} = I_j - \frac{1}{62} \sum_{k \neq i,j} I_k.$$

For the coherent-factorial donor-label retrieval test, higher candidate-specific interaction score is ranked better among the 63 non-recipient candidates. The displayed AUC MRR and its permutation p first average each candidate's interaction score over the 20 native scheduler slots and then rank that 63-candidate vector; the terminal test ranks the terminal interaction vector. The endpoint code uses midrank for exact ties, while the frozen permutation routine uses a stable descending sort of the same candidate vector, deterministically preserving candidate order for exact ties. The randomization keeps each interaction-score landscape fixed and, independently for each subject–recipient–donor-map row,

Table 15: True-donor coherence interaction in the coherent retained/null donor factorial. Negative Full/Frozen-State interactions do not establish anti-donor coding because donor specificity and retrieval remain near zero/chance.

Endpoint	Full	Frozen-State	Time-Only
AUC interaction	-.001733	-.001442	+.000001
95% CI	[-.002153,-.001339]	[-.001824,-.001089]	[-.000010,+.000011]
Terminal interaction	-.001354	-.001050	+.000008
95% CI	[-.001725,-.001000]	[-.001360,-.000756]	[-.000015,+.000032]
N_j effect $\mid R_i$	-.019393	-.017425	-.000043
N_j effect $\mid R_j$	-.021126	-.018868	-.000042

replaces the true donor by a uniformly drawn non-recipient candidate. The test statistic is mean reciprocal rank over all such rows, with one-sided $p = \{1+\#(\mathrm{MRR}_{\mathrm{perm}} \geq \mathrm{MRR}_{\mathrm{obs}})\}/(20{,}000+1)$. The permutation base seed is 2026090902. The reported AUC Top-1/Top-5 values are instead the arithmetic mean of the per-state, per-donor Top-1/Top-5 indicators computed with the endpoint midrank convention.

Table 16: True-donor specificity and retrieval in the coherent retained/null donor factorial. Chance MRR is 0.075052.

Endpoint	Full	Frozen-State	Time-Only
AUC SPEC	+.000013	+.000028	+.000002
95% CI	[-.000073,+.000098]	[-.000055,+.000112]	[-.000005,+.000008]
Term. SPEC	-.000029	+.000040	+.000005
95% CI	[-.000148,+.000092]	[-.000072,+.000155]	[-.000009,+.000018]
AUC MRR	.070751	.077660	.070443
AUC perm. p	.802110	.297585	.810659
Term. MRR	.067473	.078774	.079684
Term. perm. p	.935603	.231338	.178691
AUC Top-1	.013411	.015820	.015885
AUC Top-5	.077018	.083138	.078190

Table 17: Donor/recipient direction sanity check for the coherent retained/null donor factorial.

Mode	Donor gain [95% CI]	Recipient gain [95% CI]	Preference interaction
Full	-.037364 [-.041185,-.033671]	-.057923 [-.063399,-.052653]	+.000977
Frozen-State	-.033214 [-.036876,-.029623]	-.051742 [-.056967,-.046583]	+.000802
Time-Only	-.000019 [-.000165,+.000124]	-.000212 [-.000413,-.000010]	-.000009

The Full coherence interaction is negative, while true-donor specificity is near zero with intervals spanning zero and retrieval remains near chance. The negative raw interaction is therefore not evidence for anti-donor coding. Under fixed recipient Official conditioning, coherent cross-stimulus transplantation is more conservatively interpreted as incompatibility/interference than as donor-content injection. This does not establish intrinsic non-portability under conditioning regimes that were not tested. The coherent-factorial exact-null swap deviations match the Exact-Null donor-transplantation audit above. All 54/54 expected coherent-factorial experimental cells completed across the two model seeds.

Table 18: Direct mode contrasts for the coherent retained/null donor factorial.

Contrast	Endpoint	AUC Δ	95% CI	Terminal Δ	95% CI
Full−Frozen-State	interaction	-.000290	[-.000454,-.000123]	-.000303	[-.000500,-.000104]
Full−Frozen-State	specificity	-.000016	[-.000052,+.000021]	-.000069	[-.000137,-.000001]
Full−Time-Only	interaction	-.001733	[-.002155,-.001342]	-.001362	[-.001734,-.001004]
Full−Time-Only	specificity	+.000011	[-.000074,+.000096]	-.000034	[-.000154,+.000085]

I QUALITATIVE SELECTION DISCLOSURE

The compact zoom-only examples shown in Fig. 5 are accompanied here by the complete full-image views and matched-coordinate ROIs in Fig. 6.

The examples in Fig. 5 fix model seed 6911 and generation seed 2026083101. Cases are first restricted to trials with positive full-image PixCorr and SSIM $C - S$, then selected for qualitative interpretability. ROIs are selected only after the full-image case is fixed, and GT/Official/Shuffled/Correct use identical crop coordinates. These examples are visualization-only and are not presented as random or population-representative samples.

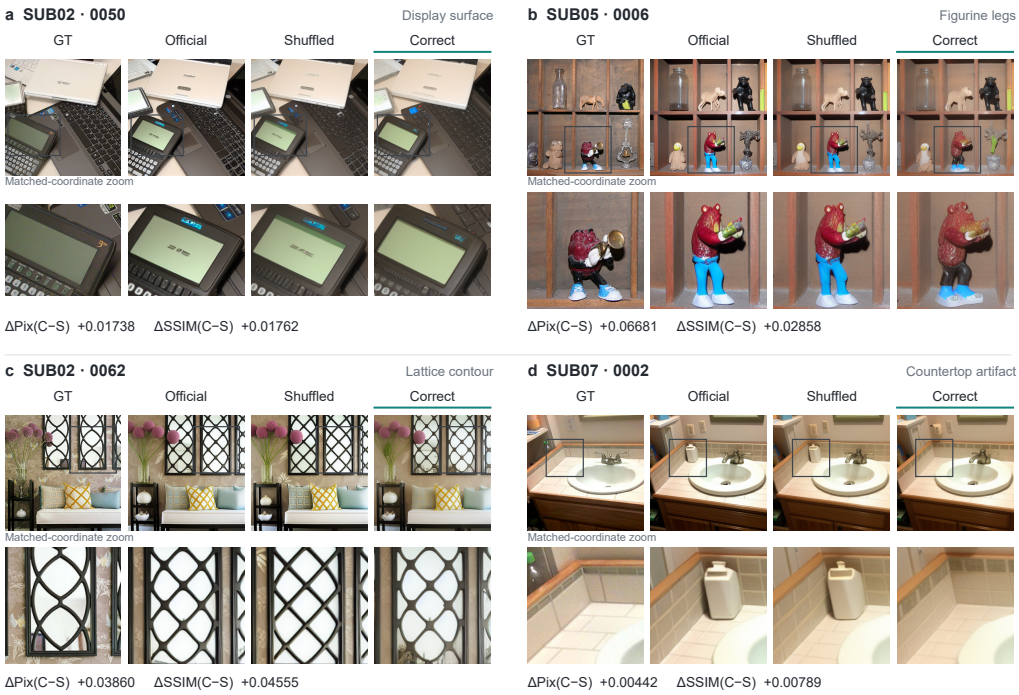


Figure 6: **Full-image context for the qualitative examples.** Full-image views and matched-coordinate zooms for the four frozen positive cases summarized in Fig. 5. GT, Official, Shuffled, and Correct use identical crop coordinates; boxes mark the ROIs. Displayed ΔPix/ΔSSIM values are per-image $C - S$ descriptors, not population estimates. Quantitative conclusions rely on the complete frozen evaluations.

I.1 COMPLETE VAL64 QUALITATIVE GALLERY

For completeness, Figs. 7–15 show GT and BrainSteer Correct reconstructions for all 64 VAL64 stimuli of each non-development subject at the fixed visualization seed. Unlike the four-case visualization above, no qualitative selection is applied in this gallery. These panels provide visual context rather than an additional statistical endpoint; Official/Shuffled comparisons and quantitative inference are reported in the main text and preceding analyses.

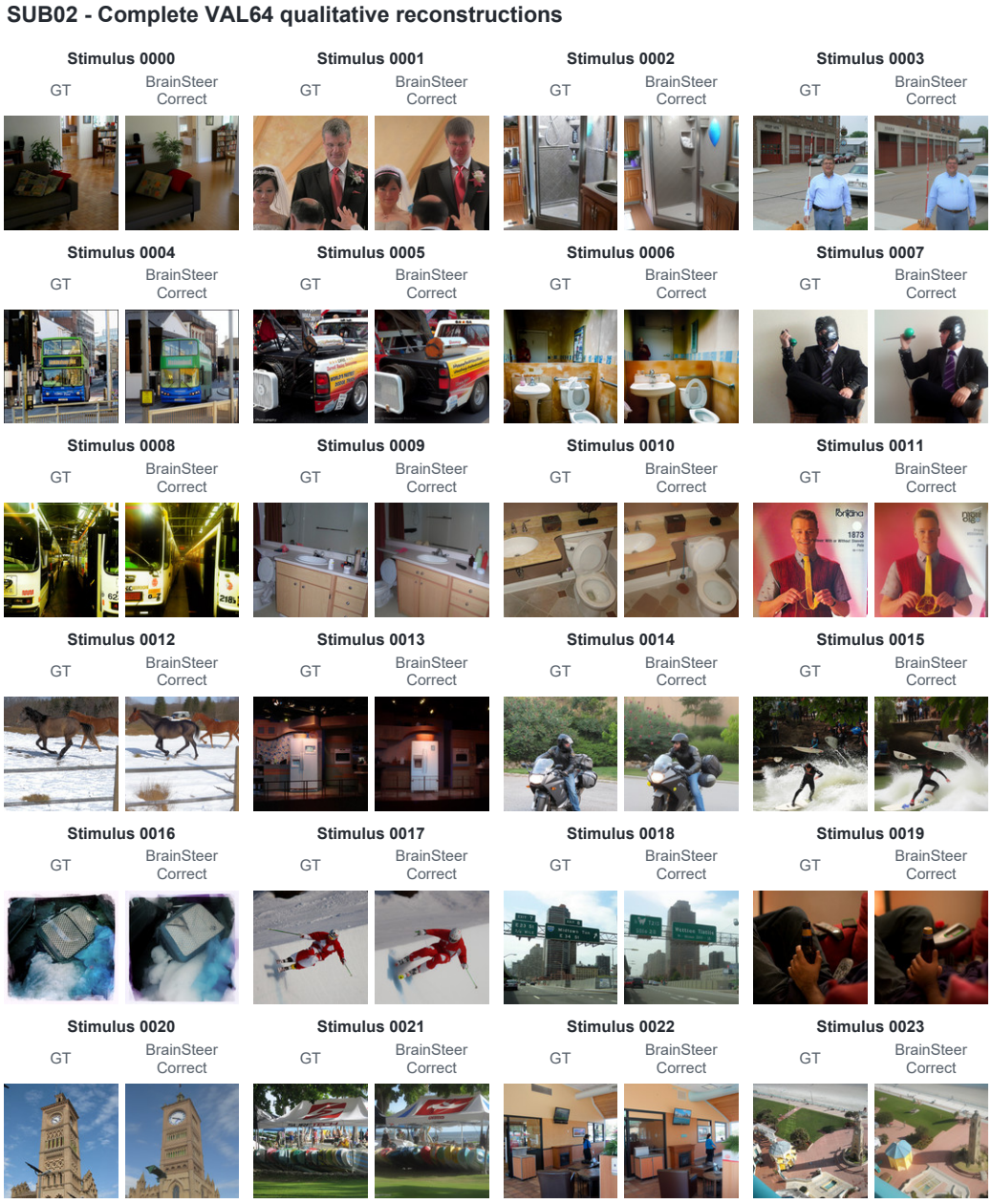


Figure 7: **Complete VAL64 qualitative gallery for SUB02 (1/3).** GT and BrainSteer Correct reconstructions for the first subset of the complete 64-stimulus VAL64 evaluation. No qualitative selection is applied.

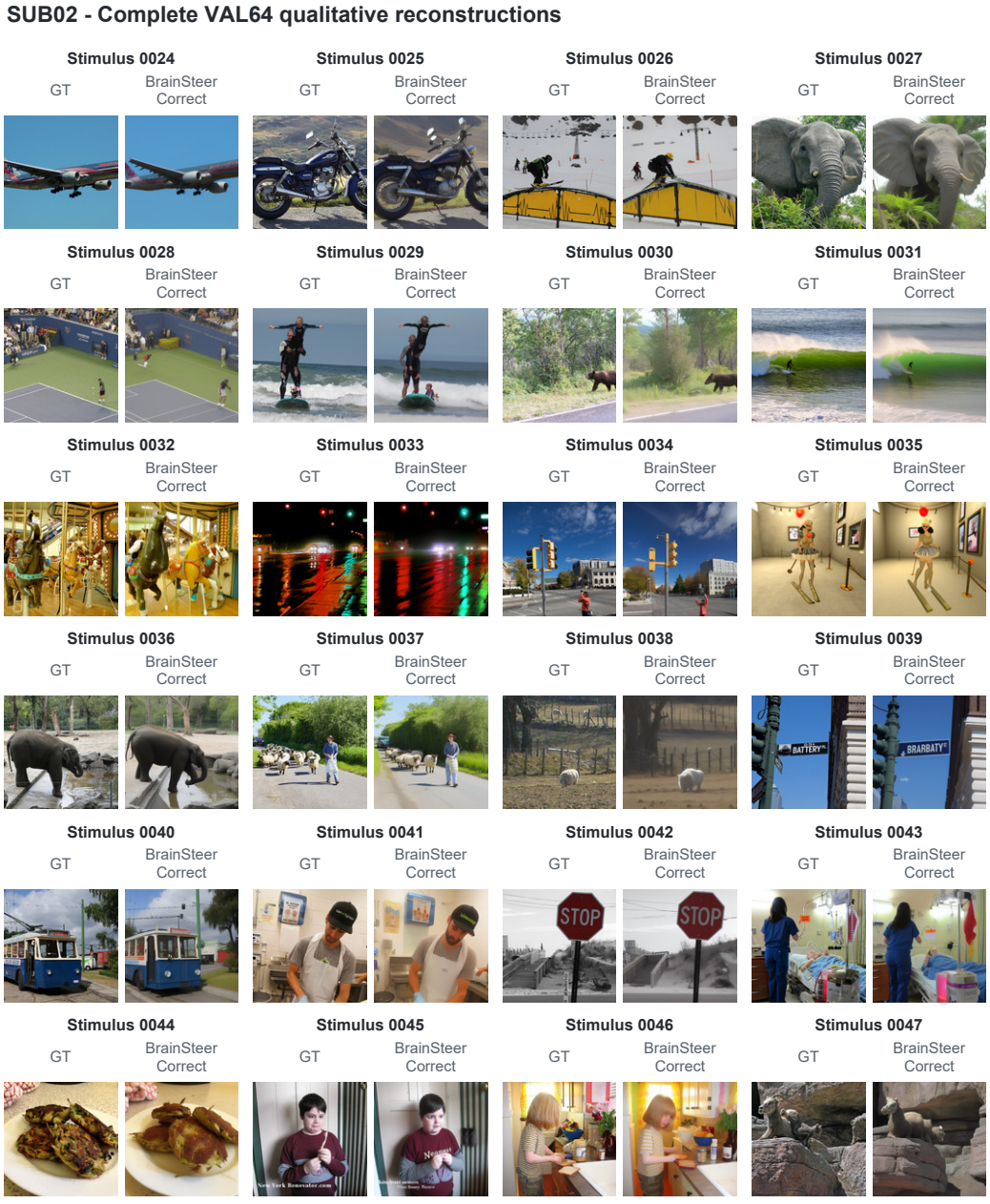


Figure 8: **Complete VAL64 qualitative gallery for SUB02 (2/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

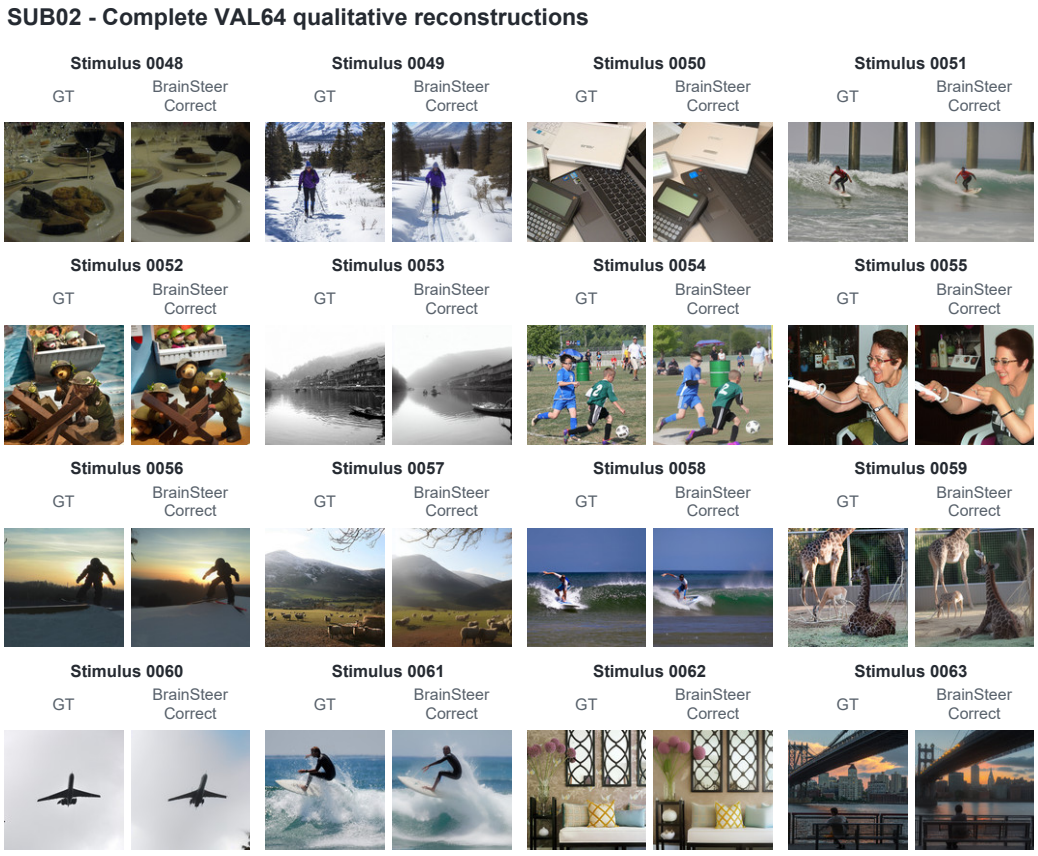


Figure 9: **Complete VAL64 qualitative gallery for SUB02 (3/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

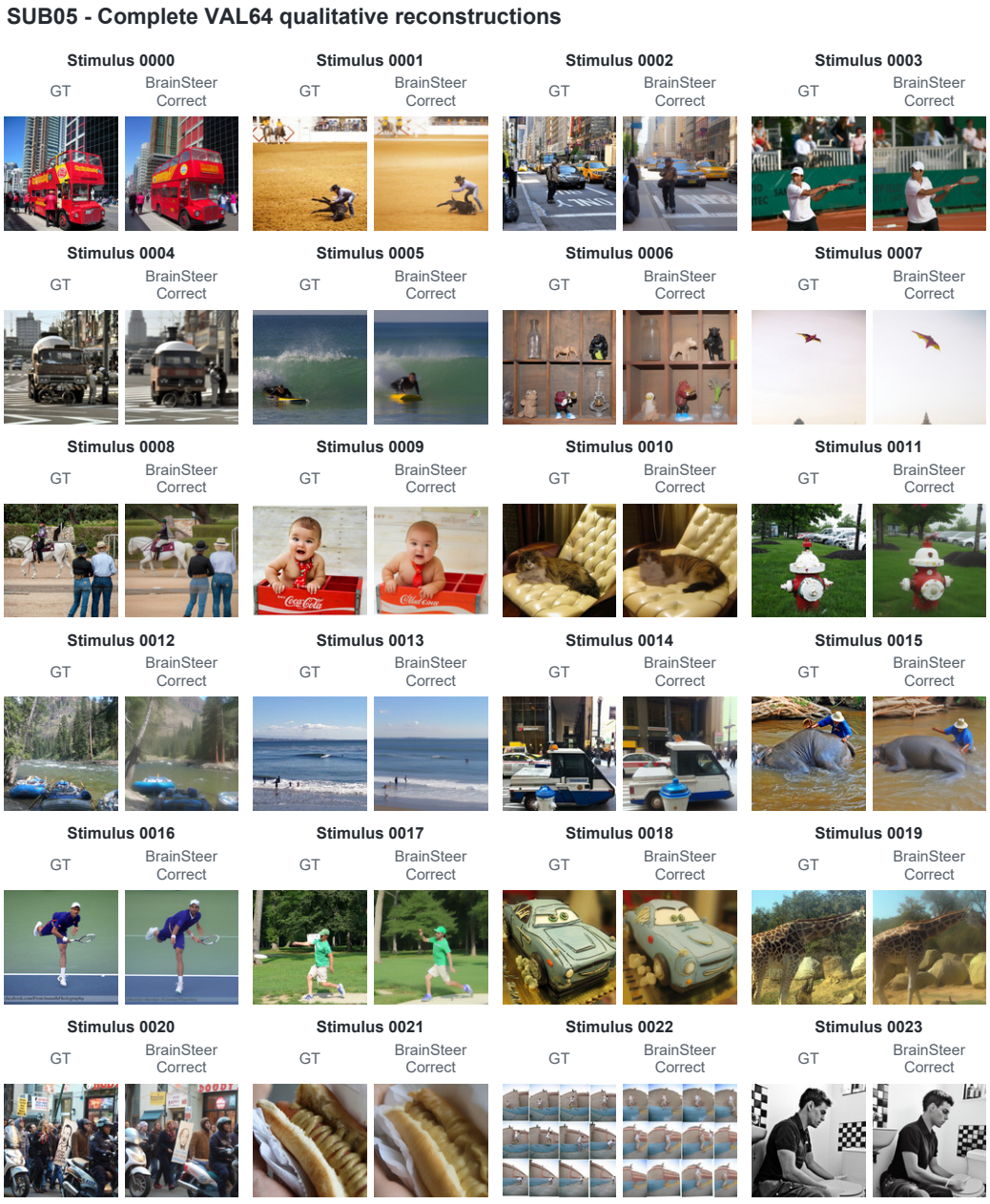


Figure 10: **Complete VAL64 qualitative gallery for SUB05 (1/3).** GT and BrainSteer Correct reconstructions for the first subset of the complete 64-stimulus VAL64 evaluation. No qualitative selection is applied.

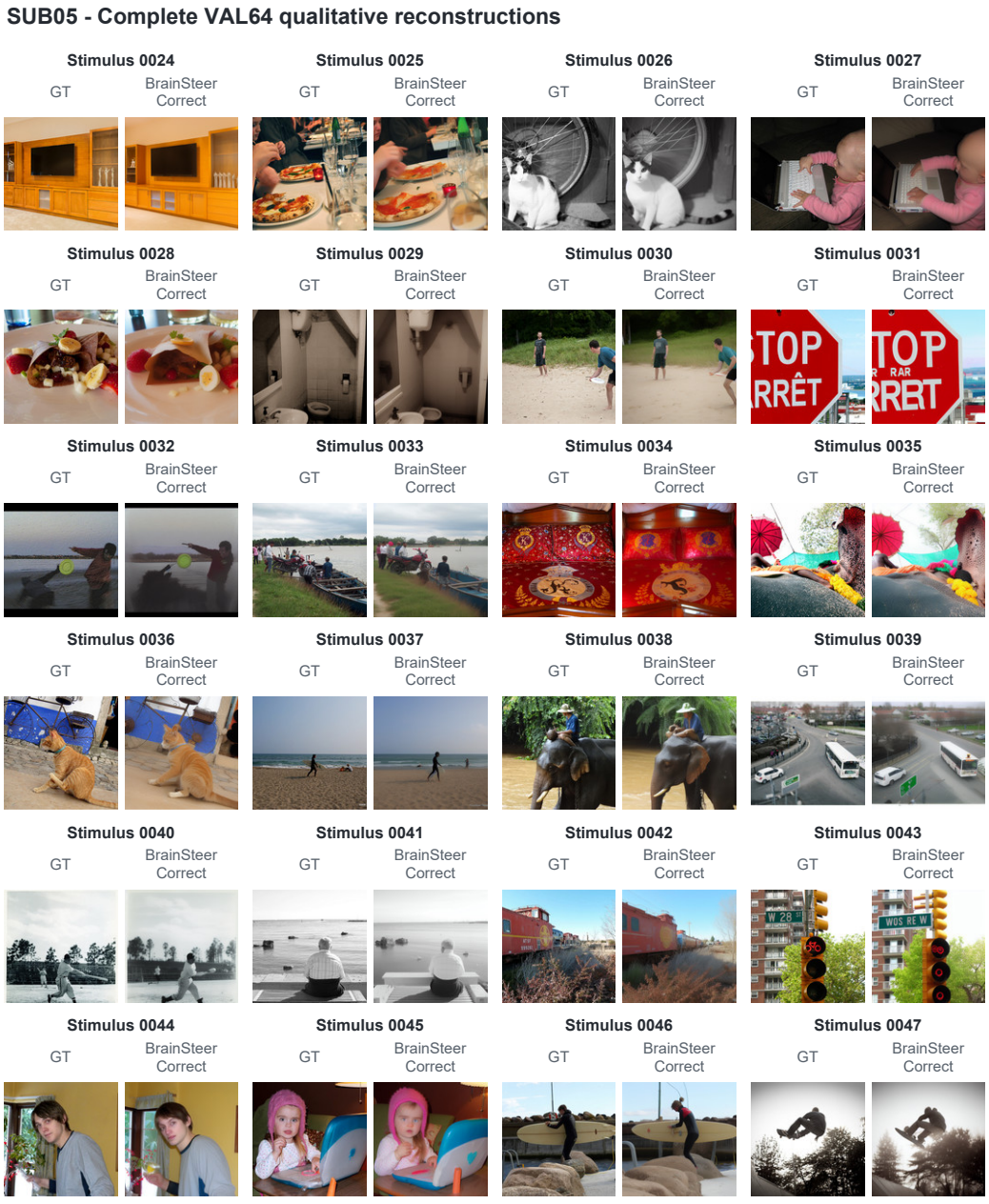


Figure 11: **Complete VAL64 qualitative gallery for SUB05 (2/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

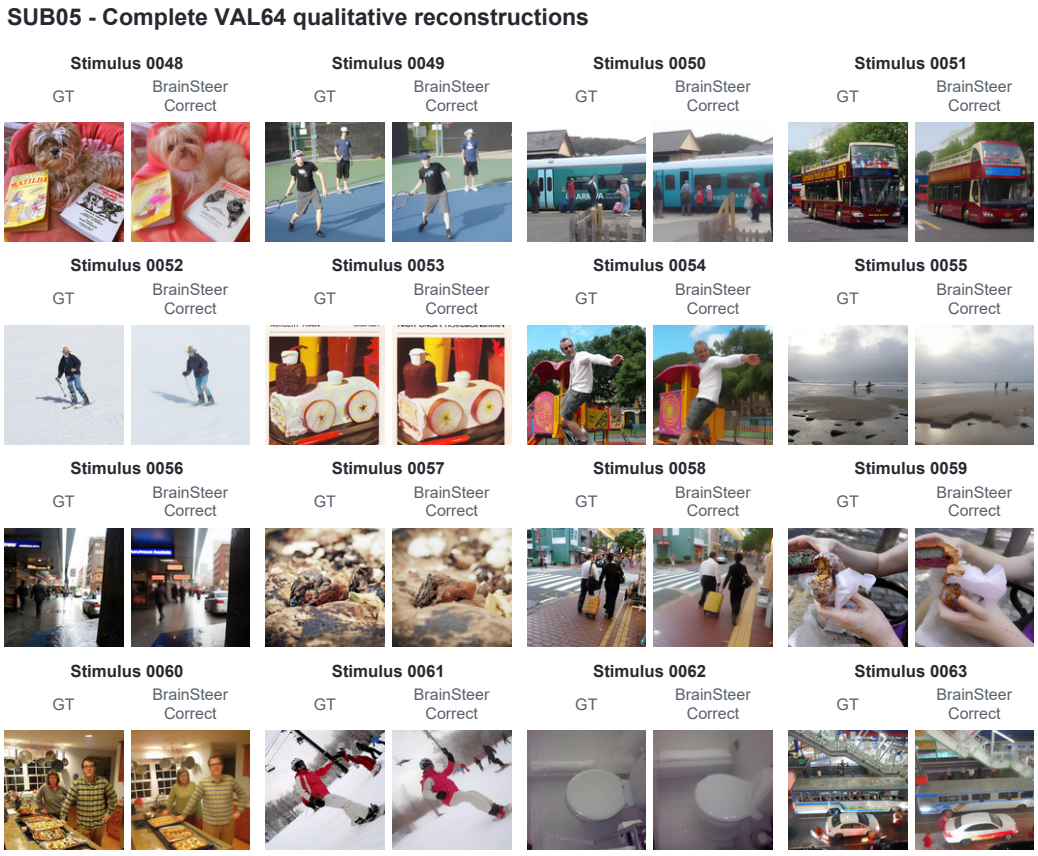


Figure 12: **Complete VAL64 qualitative gallery for SUB05 (3/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

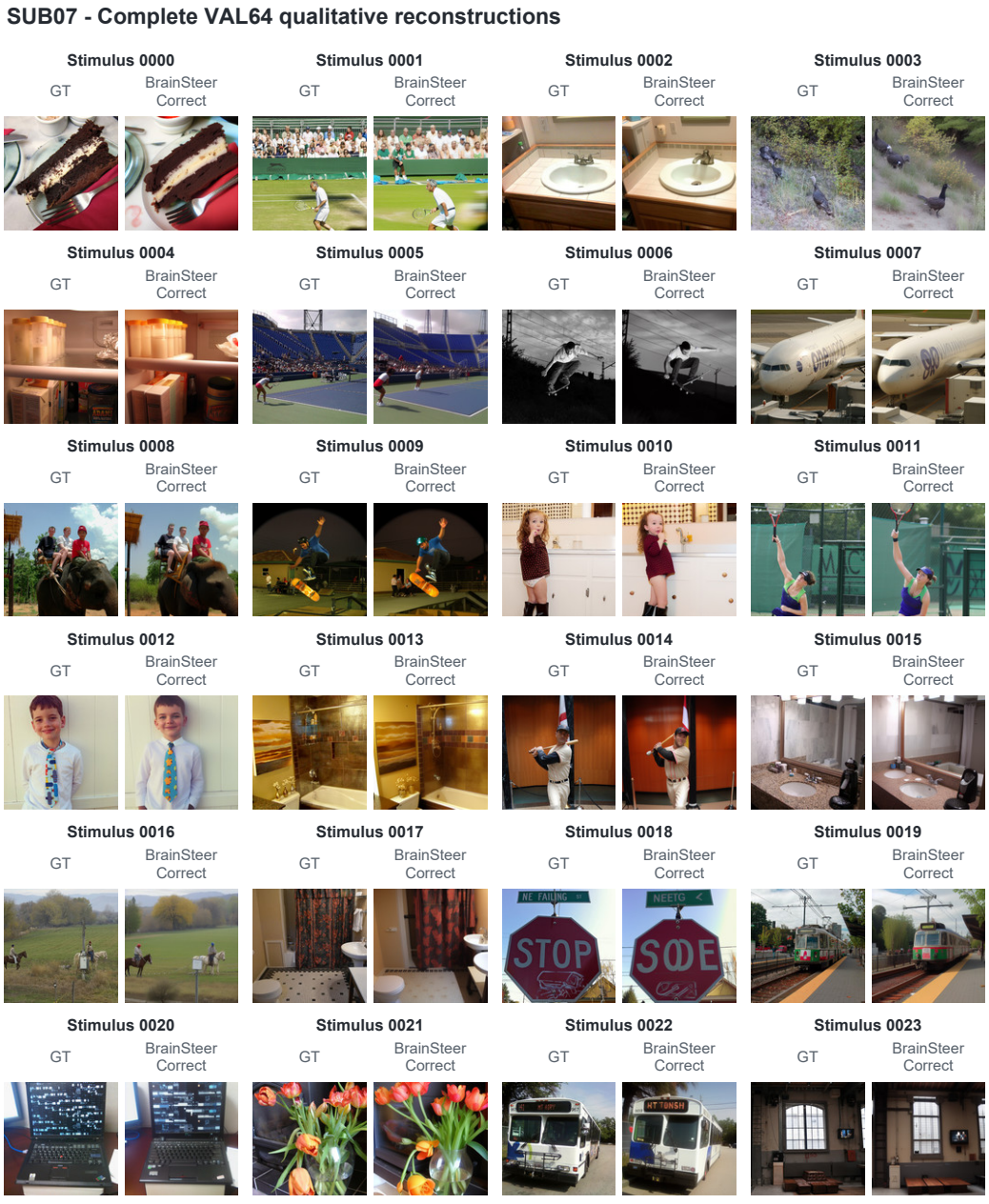


Figure 13: **Complete VAL64 qualitative gallery for SUB07 (1/3).** GT and BrainSteer Correct reconstructions for the first subset of the complete 64-stimulus VAL64 evaluation. No qualitative selection is applied.

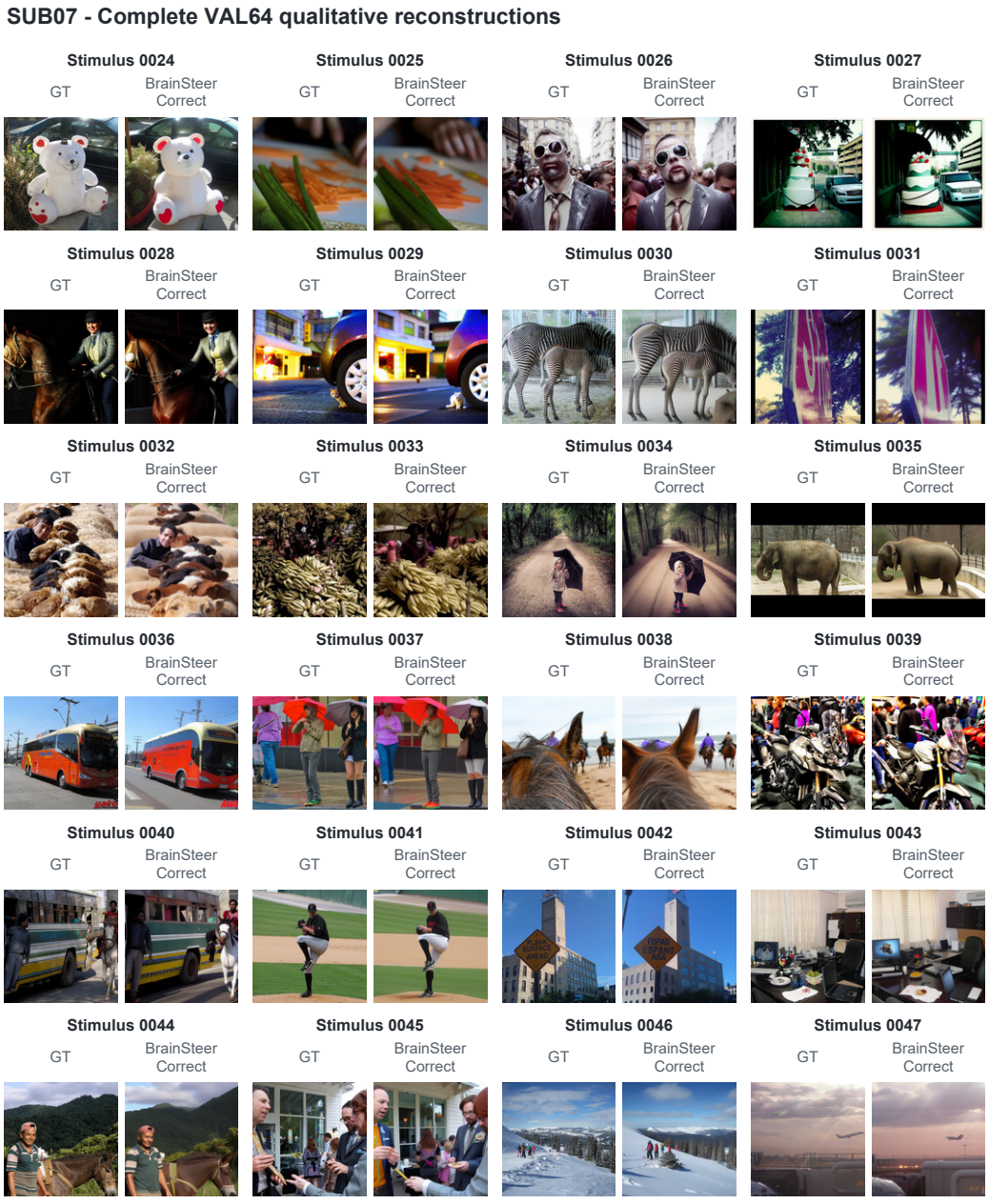


Figure 14: **Complete VAL64 qualitative gallery for SUB07 (2/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

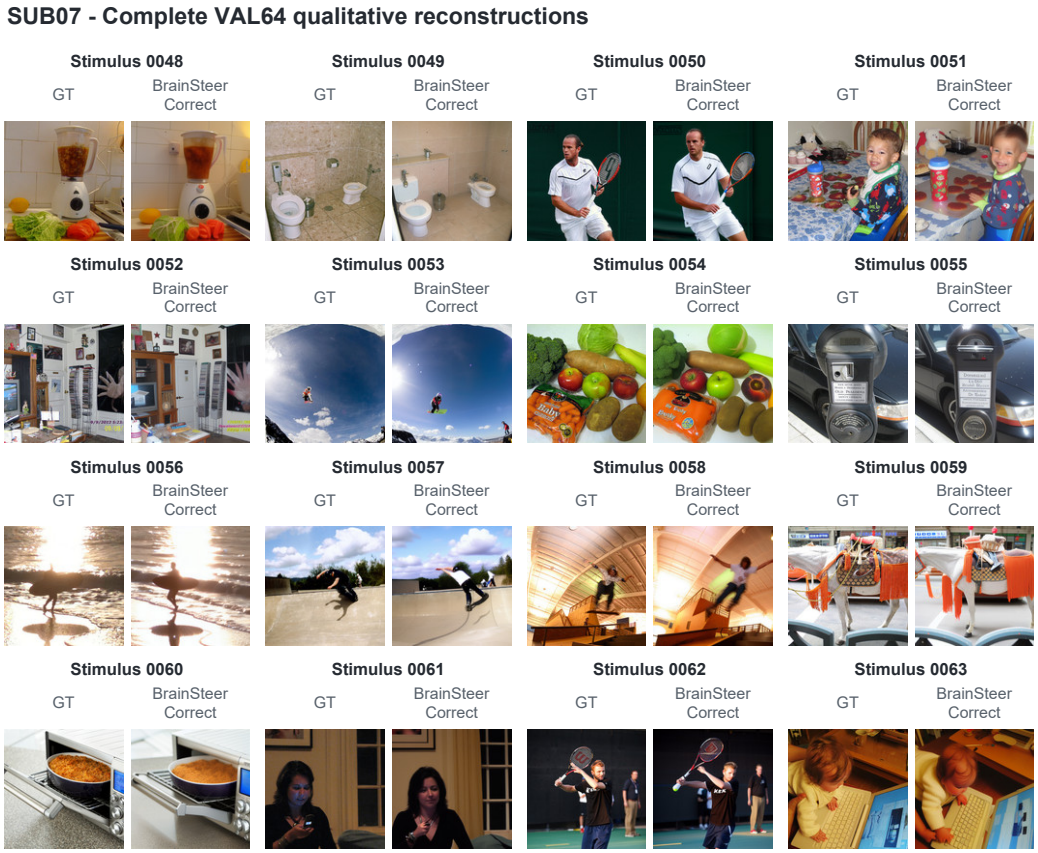


Figure 15: **Complete VAL64 qualitative gallery for SUB07 (3/3).** GT and BrainSteer Correct reconstructions from the same frozen VAL64 evaluation.

J SOURCE BOUNDARIES

The fixed architecture archive reports 947,264 trainable BrainSteer-v2 parameters and 32 hooked cross-attention modules. The model-size audit above is complete for the eight production BrainSteer checkpoints and the four subject-specific Official DynaDiff parameter totals. A separate fixed-reference numerical gauge-invariance diagnostic is not promoted as a frozen performance claim: algebraically, any fixed encoder-space reference cancels after the stated TRAIN-only linear branch centering, while the implementation-level diagnostic remains separate from the performance and mechanism results reported here.