BEYOND ALIGNMENT: TEST-TIME NEURAL CONSISTENCY FOR CROSS-SUBJECT VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject visual decoding aims to reuse pretrained brain decoders across individuals with limited subject-specific data. Existing approaches mainly address functional alignment, enabling neural representations to transfer across subjects, but they do not explicitly constrain the subsequent generative trajectory to remain faithful to the neural evidence observed at test time. We propose **Dual-Space Data Consistency (DS-DC)**, a training-free inference framework that reintroduces complementary neural and structural evidence into two stages of cross-subject generation. During diffusion-prior sampling, **Measurement-Space Consistency (MSC)** feeds the observed source-subject fMRI back into the latent trajectory through a frozen learned neural forward operator. During image enhancement, **Query-Guided Structural Consistency (QSC)** uses the fMRI-derived low-level estimate as a reliability-weighted multi-scale structural reference. A key challenge arises in MSC: unlike physical inverse problems with known forward operators, the visual-to-fMRI mapping is only available through an imperfect learned surrogate. We therefore use weak step-wise guidance together with *finite* latent refinement rather than exact measurement fitting. Our analysis shows that such refinement simultaneously reduces latent estimation error and propagates surrogate-model residuals, inducing an implicit spectral regularization that preferentially corrects well-observed directions while limiting error amplification in weakly constrained modes. On the Natural Scenes Dataset, DS-DC consistently improves a pretrained MindAligner decoder across 12 directed cross-subject transfers without updating its parameters. MSC reduces source-fMRI measurement MSE by 24.45%, while the full framework improves all reported low- and high-level decoding metrics. These results show that test-time consistency provides a principled complement to cross-subject alignment, while learned neural constraints are most effective when enforced finitely rather than aggressively.

1 INTRODUCTION

Visual decoding from functional magnetic resonance imaging (fMRI) seeks to recover the visual content associated with distributed neural responses. Recent progress in multimodal representation learning and generative modeling has substantially improved reconstruction quality, with modern systems combining low-level visual estimates, contrastive representations, and diffusion priors Ozcelik & VanRullen (2023); Scotti et al. (2023); Lu et al. (2023); Li et al. (2025). Most high-performing decoders, however, rely on substantial paired data and subject-specific optimization. Cross-subject decoding addresses this limitation by transferring a decoder learned from data-rich subjects to a new individual with limited paired measurements. Existing approaches therefore focus primarily on reducing inter-subject variability through shared representations, functional alignment, or lightweight adaptation Wang et al. (2024); Scotti et al. (2024); Zhou et al. (2024); Gong et al. (2025). MindAligner, for example, aligns measurements from a new subject to the functional space of a reference subject so that a pretrained decoder can be reused Dai et al. (2025).

Alignment, however, addresses *representation transfer* rather than *inference-time faithfulness*. Once the aligned representation enters a powerful generative decoder, the subsequent trajectory is largely governed by the pretrained visual prior. Because fMRI-to-image decoding is severely underdetermined, such a prior is essential for producing plausible images, but plausibility alone does not guarantee compatibility with the neural evidence observed for the source subject at test time. In

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.

this work, *source-subject fMRI* denotes this directly observed test-time measurement, rather than a response predicted in the aligned reference space. The distinction is consequential: two reconstructions may both be plausible under the generative prior while differing in how well their visual representations remain supported by the available neural measurement. Related diffusion inverse problems similarly separate prior-driven generation from explicit agreement with observations Song et al. (2024); Mohajer Hamidi et al. (2025). This suggests that cross-subject decoding requires not only alignment before generation, but also *explicit consistency during generation*. Figure 1 illustrates this distinction: conventional decoding uses the observed fMRI only for conditioning, whereas a closed-loop formulation can reuse the same neural evidence to correct the generative trajectory.

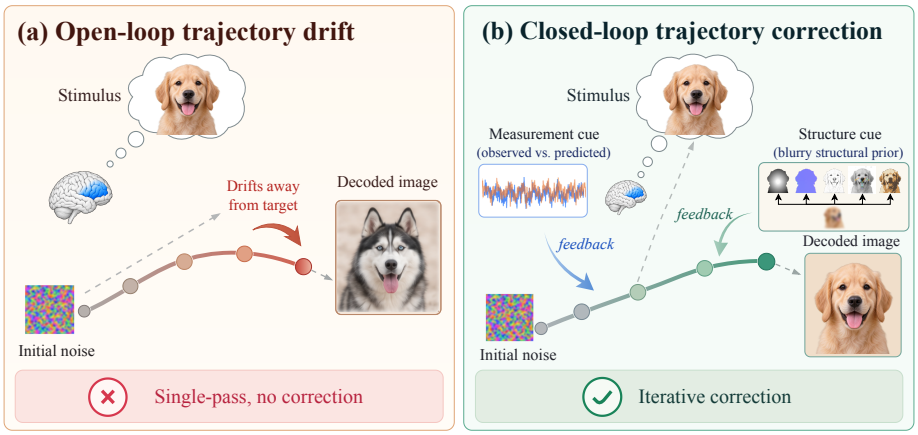

Figure 1: Open-loop versus closed-loop fMRI decoding. Conventional decoding uses the observed fMRI only for initial conditioning, whereas DS-DC reuses the same neural evidence as measurement and structural feedback to iteratively correct the generation trajectory.

Introducing such consistency is nevertheless nontrivial. Unlike classical inverse problems, the visual-to-fMRI forward process is unknown and must be approximated by a learned neural encoding model. Consequently, driving the generated representation to exactly minimize a learned measurement discrepancy can also fit residual errors of the surrogate forward model. This creates a *correction–bias trade-off*: measurement feedback can suppress latent error components that are constrained by the learned operator, while overly aggressive fitting can transfer forward-model residuals back into the latent representation. Our local analysis makes this trade-off explicit. For an idealized finite refinement of the learned measurement objective, the finite-step error decomposes into a retained initial-error component, a forward-residual-induced component, and a bounded Jacobian-variation remainder. The linearized component further exhibits an early-stopping spectral structure: finite refinement limits residual transfer along directions that are weakly constrained by the learned measurement operator, whereas aggressive surrogate fitting approaches an inverse-like residual amplification regime. This analysis motivates finite rather than converged learned measurement fitting.

Motivated by these observations, we introduce Dual-Space Data Consistency (DS-DC), a training-free inference framework that reintroduces test-time evidence at two stages of a pretrained cross-subject decoder. During diffusion-prior sampling, Measurement-Space Consistency (MSC) uses the observed source-subject fMRI as measurement evidence. A frozen learned neural forward operator maps the current visual–semantic estimate back to the source-subject measurement space; weak step-wise guidance and finite latent refinement then feed the resulting discrepancy back into the sampling trajectory. During subsequent image enhancement, Query-Guided Structural Consistency (QSC) uses the fMRI-derived blurry estimate as a complementary structural reference. It retrieves local multi-scale features from this reference and downweights unreliable matches, thereby constraining intermediate enhancement without forcing uncertain fine-scale correspondence. MSC and QSC act sequentially at the stages where their respective evidence is naturally available, and neither requires updating the pretrained decoder.

We evaluate DS-DC on the Natural Scenes Dataset (NSD) Allen et al. (2022) using pretrained MindAligner across 12 directed cross-subject transfers. DS-DC improves the pretrained baseline across all reported low- and high-level metrics without parameter updates. MSC reduces source-fMRI

measurement MSE by 24.45%, while an independently pretrained forward model shows smaller but directionally consistent cross-operator changes, indicating that the observed improvement is not restricted entirely to the guidance operator. Ablations further reveal complementary roles of the two consistency spaces: MSC primarily strengthens measurement- and high-level agreement, whereas QSC restores structural similarity during image enhancement. Together, these results support test-time consistency as a complement to cross-subject alignment rather than a replacement for it.

Our main contributions are summarized as follows:

- We identify a distinct limitation of current cross-subject visual decoding: functional alignment enables representation transfer but does not explicitly constrain the subsequent generative trajectory to remain compatible with the source-subject neural evidence observed at test time.
- We propose DS-DC, a training-free inference framework that introduces complementary measurement-space and structural consistency through MSC and QSC while keeping the pretrained cross-subject decoder fixed.
- We analyze finite learned measurement refinement under an imperfect neural forward operator, deriving a finite-step correction–residual decomposition and a linearized spectral characterization that motivates finite rather than aggressive surrogate measurement fitting.
- Across 12 directed NSD transfers, DS-DC consistently improves the pretrained MindAligner baseline. Measurement-space, cross-operator, and component-wise analyses further characterize the distinct effects of MSC and QSC.

2 METHOD

2.1 PROBLEM FORMULATION

Let $\mathbf{y}_s \in \mathbb{R}^{v_s}$ denote the observed fMRI response from source subject s, and let $\mathbf{x}^\star \in \mathcal{X}$ denote the corresponding visual stimulus. The observed response is modeled as

$$\mathbf{y}_s = \mathcal{H}_s(\mathbf{x}^\star) + \boldsymbol{\epsilon}_s, \tag{1}$$

where $\mathcal{H}_s : \mathcal{X} \to \mathcal{Y}_s$ is the unknown subject-specific visual-to-fMRI encoding process and $\boldsymbol{\epsilon}_s$ captures measurement noise and unmodeled variation.

Given $\mathbf{y}_s$, the pretrained cross-subject decoder provides a conditioning representation $\mathbf{c}_s$ and an fMRI-derived blurry estimate $\mathbf{x}_\mathrm{b}$. Its frozen diffusion prior generates

$$\mathbf{z}_0 \sim p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}_s), \tag{2}$$

where $\mathbf{z}_0 \in \mathcal{Z}$ denotes the latent representation. DS-DC introduces inference-time constraints from $\mathbf{y}_s$ and $\mathbf{x}_\mathrm{b}$ while leaving the pretrained decoder unchanged.

2.2 DUAL-SPACE DATA CONSISTENCY FRAMEWORK

Figure 2 summarizes DS-DC. Measurement-Space Consistency (MSC) promotes agreement between the observed fMRI and its prediction under a frozen neural forward operator during diffusion-prior sampling. Query-Guided Structural Consistency (QSC) regularizes subsequent enhancement using the fMRI-derived blurry estimate as a multi-scale structural reference. The two stages therefore introduce complementary constraints in measurement and feature spaces without updating pretrained model parameters.

2.3 MEASUREMENT-SPACE CONSISTENCY

At reverse step t, let $\mathbf{z}_t$ denote the current noisy latent and $\hat{\mathbf{z}}_{0,t}$ its clean latent estimate. Since the true encoding process $\mathcal{H}_s$ is unavailable during decoding, MSC employs a frozen learned surrogate $\mathcal{F}_s : \mathcal{Z} \to \mathcal{Y}_s$,

$$\hat{\mathbf{y}}_{s,t} = \mathcal{F}_s(\hat{\mathbf{z}}_{0,t}), \tag{3}$$

which provides differentiable measurement feedback from the latent space to the observed fMRI space.

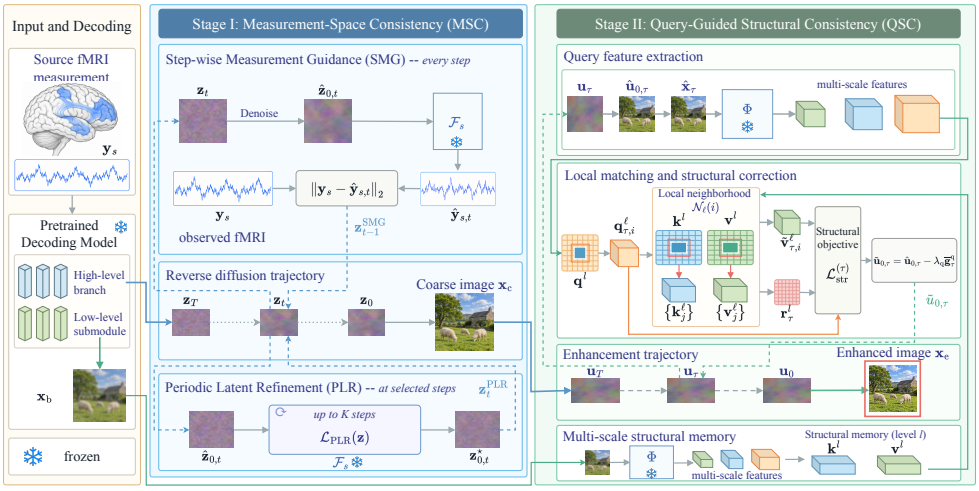

Figure 2: Overview of Dual-Space Data Consistency (DS-DC). MSC introduces measurement-consistency guidance during diffusion-prior sampling, while QSC regularizes image enhancement through local matching to a multi-scale structural memory derived from the fMRI-based blurry estimate. All pretrained components remain frozen during inference.

2.3.1 Step-wise Measurement Guidance

SMG evaluates the measurement discrepancy at each reverse step and uses its gradient to guide the reverse update:

$$\mathcal{L}_{\mathrm{SMG}}^{(t)} = \|\mathbf{y}_s - \hat{\mathbf{y}}_{s,t}\|_2, \qquad \mathbf{g}_t = \nabla_{\mathbf{z}_t} \mathcal{L}_{\mathrm{SMG}}^{(t)}, \qquad \mathbf{z}_{t-1}^{\mathrm{SMG}} = \mathbf{z}_{t-1}^{\mathrm{prior}} - \eta_{\mathrm{m}} \mathbf{g}_t. \tag{4}$$

Here, $\mathbf{z}_{t-1}^{\mathrm{prior}}$ denotes the unmodified reverse-diffusion update and η_{m} controls the guidance strength. This correction is applied throughout the reverse trajectory.

2.3.2 Periodic Latent Refinement

SMG is complemented by finite-step latent refinement at selected reverse steps. Starting from the current clean estimate, PLR minimizes the same measurement discrepancy in the clean latent space:

$$\mathcal{L}_{\mathrm{PLR}}(\mathbf{z}) = \mathrm{MSE}(\mathcal{F}_s(\mathbf{z}), \mathbf{y}_s), \qquad \mathbf{z}_{0,t}^{(0)} = \hat{\mathbf{z}}_{0,t}. \tag{5}$$

A finite number of refinement iterations yields $\mathbf{z}_{0,t}^{\star}$, which is returned to timestep t through diffusion-consistent resampling,

$$\mathbf{z}_t^{\mathrm{PLR}} = \mathcal{R}_t\left(\mathbf{z}_t, \mathbf{z}_{0,t}^{\star}\right), \tag{6}$$

where $\mathcal{R}_t$ denotes the resampling map induced by the original diffusion schedule. Reverse diffusion then continues from $\mathbf{z}_t^{\mathrm{PLR}}$.

2.4 Query-Guided Structural Consistency

After diffusion-prior sampling, QSC refines the coarse reconstruction $\mathbf{x}_{\mathrm{c}}$ along the enhancement trajectory. The fMRI-derived blurry estimate $\mathbf{x}_{\mathrm{b}}$ provides a fixed structural reference, while the current enhancement estimate supplies step-dependent query features.

Multi-scale structural memory. A frozen feature extractor Φ encodes $\mathbf{x}_{\mathrm{b}}$ into multi-scale reference features. For each level $\ell \in \mathcal{L}$,

$$\begin{aligned} \mathbf{F}_{\mathrm{b}}^{\ell} &= \mathrm{Tok}(\Phi_\ell(\mathbf{x}_{\mathrm{b}})) \in \mathbb{R}^{N_\ell \times C_\ell}, \qquad \mathbf{f}_{\mathrm{b},j}^{\ell} = [\mathbf{F}_{\mathrm{b}}^{\ell}]_{j,:}^{\top}, \\ \mathbf{k}_j^{\ell} &= \frac{\mathbf{f}_{\mathrm{b},j}^{\ell}}{\|\mathbf{f}_{\mathrm{b},j}^{\ell}\|_2 + \epsilon_{\mathrm{n}}}, \qquad \mathbf{v}_j^{\ell} = \mathbf{f}_{\mathrm{b},j}^{\ell}. \end{aligned} \tag{7}$$

Here, $\mathrm{Tok}(\cdot)$ converts a feature map into spatial tokens, while $\epsilon_{\mathrm{n}} > 0$ ensures numerical stability. The normalized keys $\mathbf{k}_j^{\ell}$ and values $\mathbf{v}_j^{\ell}$ form the structural memory.

Query feature extraction. At enhancement step τ, let $\mathbf{u}_\tau$ denote the current latent, $\hat{\mathbf{u}}_{0,\tau}$ its denoised prediction, and $\hat{\mathbf{x}}_\tau$ the corresponding image estimate. Using the same feature space,

$$\mathbf{Q}_\tau^\ell = \mathrm{Tok}(\Phi_\ell(\hat{\mathbf{x}}_\tau)), \qquad \mathbf{q}_{\tau,i}^\ell = [\mathbf{Q}_\tau^\ell]_{i,:}^\top, \qquad \bar{\mathbf{q}}_{\tau,i}^\ell = \frac{\mathbf{q}_{\tau,i}^\ell}{\|\mathbf{q}_{\tau,i}^\ell\|_2 + \epsilon_{\mathrm{n}}}. \tag{8}$$

Local matching and structural correction. For each query location i, QSC constructs a spatially regularized similarity within the local neighborhood $\mathcal{N}_\ell(i)$:

$$\begin{aligned} s_{\tau,ij}^\ell &= \frac{\langle \bar{\mathbf{q}}_{\tau,i}^\ell, \mathbf{k}_j^\ell \rangle}{T_{\mathrm{m}}} - \lambda_{\mathrm{p}} \Delta_{ij}^\ell, \\ a_{\tau,ij}^\ell &= \frac{\exp(s_{\tau,ij}^\ell)}{\sum_{j' \in \mathcal{N}_\ell(i)} \exp(s_{\tau,ij'}^\ell)}, \qquad \tilde{\mathbf{v}}_{\tau,i}^\ell = \mathrm{sg}\left[\sum_{j \in \mathcal{N}_\ell(i)} a_{\tau,ij}^\ell \mathbf{v}_j^\ell \right]. \end{aligned} \tag{9}$$

Here, $j \in \mathcal{N}_\ell(i)$, T_{m} denotes the matching temperature, λ_{p} controls the spatial penalty, Δ_{ij}^ℓ is the normalized squared spatial offset, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. A matching-confidence weight $r_{\tau,i}^\ell$ derived from the local affinity distribution attenuates ambiguous correspondences.

The normalized structural residual and its robust penalty are

$$\mathbf{d}_{\tau,i}^\ell = \frac{\mathbf{q}_{\tau,i}^\ell - \tilde{\mathbf{v}}_{\tau,i}^\ell}{\sigma_{\mathrm{b}}^\ell}, \qquad \rho(\mathbf{d}) = \frac{1}{C_\ell} \sum_{c=1}^{C_\ell} \sqrt{d_c^2 + \epsilon_{\mathrm{char}}^2}. \tag{10}$$

Here, $\sigma_{\mathrm{b}}^\ell = \mathrm{RMS}(\mathbf{F}_{\mathrm{b}}^\ell)$ normalizes the reference feature scale, and $\epsilon_{\mathrm{char}} > 0$ is the Charbonnier constant.

The resulting structural objective aggregates reliable correspondences across feature levels:

$$\mathcal{L}_{\mathrm{str},\ell}^{(\tau)} = \frac{\sum_i r_{\tau,i}^\ell \rho(\mathbf{d}_{\tau,i}^\ell)}{\max(\sum_i r_{\tau,i}^\ell, \epsilon_{\mathrm{r}})}, \qquad \mathcal{L}_{\mathrm{str}}^{(\tau)} = \sum_{\ell \in \mathcal{L}} \omega_\ell(\tau) \mathcal{L}_{\mathrm{str},\ell}^{(\tau)}. \tag{11}$$

Here, $\epsilon_{\mathrm{r}} > 0$ stabilizes the normalization and $\omega_\ell(\tau)$ denotes the feature-level weight.

Finally, the structural objective provides a guidance direction for the clean prediction:

$$\mathbf{g}_\tau^{\mathrm{q}} = \nabla_{\mathbf{u}_\tau} \mathcal{L}_{\mathrm{str}}^{(\tau)}, \qquad \tilde{\mathbf{u}}_{0,\tau} = \hat{\mathbf{u}}_{0,\tau} - \lambda_{\mathrm{q}} \bar{\mathbf{g}}_\tau^{\mathrm{q}}. \tag{12}$$

Here, $\bar{\mathbf{g}}_\tau^{\mathrm{q}}$ denotes the normalized structural gradient and λ_{q} controls its strength. The corrected clean prediction is subsequently used by the enhancement trajectory.

3 Local Analysis of Finite Measurement Refinement

MSC uses a learned neural forward operator $\mathcal{F}_s$ to evaluate whether the current visual representation is compatible with the observed source-subject fMRI $\mathbf{y}_s$. Unlike physical inverse problems with a known forward model, $\mathcal{F}_s$ is only a surrogate for the unknown visual-to-neural encoding process. Consequently, minimizing the learned measurement discrepancy need not be equivalent to recovering the ground-truth visual representation. We characterize this discrepancy locally and analyze the role of finite measurement refinement.

Let $\mathbf{x}^\star$ denote the stimulus that elicited $\mathbf{y}_s$, and let

$$\mathbf{z}^\star = E_{\mathrm{vis}}(\mathbf{x}^\star) \tag{13}$$

be its deterministic representation in the latent space of $\mathcal{F}_s$. For a current estimate $\mathbf{z}$, define

$$\mathbf{e} = \mathbf{z} - \mathbf{z}^\star, \qquad \mathbf{d}^\star = \mathcal{F}_s(\mathbf{z}^\star) - \mathbf{y}_s, \tag{14}$$

where $\mathbf{d}^\star$ denotes the effective forward residual at the ground-truth representation. It absorbs any discrepancy between the learned forward prediction and the observed fMRI response, including surrogate approximation error, measurement noise, representation mismatch, and calibration error.

We consider the squared measurement objective

$$\mathcal{L}_{\mathrm{m}}(\mathbf{z};\mathbf{y}_s) = \frac{1}{2}\left\|\mathcal{F}_s(\mathbf{z}) - \mathbf{y}_s\right\|_2^2. \tag{15}$$

Let

$$\mathbf{J}_\star = \mathbf{J}_F(\mathbf{z}^\star), \qquad \sigma_{\max} = \|\mathbf{J}_\star\|_2. \tag{16}$$

Assume that there exists a convex neighborhood $\mathcal{N}$ of $\mathbf{z}^\star$ in which $\mathbf{J}_F$ is β_F-Lipschitz:

$$\|\mathbf{J}_F(\mathbf{z}_1) - \mathbf{J}_F(\mathbf{z}_2)\|_2 \le \beta_F\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \qquad \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{N}. \tag{17}$$

Proposition 1 (Local correction–residual decomposition)**.** *For $\mathbf{z} = \mathbf{z}^\star + \mathbf{e} \in \mathcal{N}$,*

$$\nabla_{\mathbf{z}}\mathcal{L}_{\mathrm{m}}(\mathbf{z};\mathbf{y}_s) = \mathbf{J}_\star^\top\mathbf{J}_\star\mathbf{e} + \mathbf{J}_\star^\top\mathbf{d}^\star + \boldsymbol{\xi}_F(\mathbf{e}), \tag{18}$$

where the Jacobian-variation remainder satisfies

$$\|\boldsymbol{\xi}_F(\mathbf{e})\|_2 \le R_F(\mathbf{e}) := \frac{3}{2}\sigma_{\max}\beta_F\|\mathbf{e}\|_2^2 + \beta_F\|\mathbf{e}\|_2\|\mathbf{d}^\star\|_2 + \frac{1}{2}\beta_F^2\|\mathbf{e}\|_2^3. \tag{19}$$

The first term in Eq. 18 depends on the current latent error, whereas the second is induced by the effective forward residual. In particular,

$$\nabla_{\mathbf{z}}\mathcal{L}_{\mathrm{m}}(\mathbf{z}^\star;\mathbf{y}_s) = \mathbf{J}_\star^\top\mathbf{d}^\star. \tag{20}$$

Hence, whenever $\mathbf{J}_\star^\top\mathbf{d}^\star \neq 0$, the ground-truth representation is not a stationary point of the learned measurement objective.

We next consider an idealized gradient-refinement dynamics associated with the PLR measurement subproblem,

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \gamma\nabla_{\mathbf{z}}\mathcal{L}_{\mathrm{m}}(\mathbf{z}_k;\mathbf{y}_s), \qquad k = 0,\dots,K-1. \tag{21}$$

Assume $\sigma_{\max} > 0$, all iterates $\mathbf{z}_0,\dots,\mathbf{z}_K$ remain in $\mathcal{N}$, and

$$0 < \gamma \le \frac{1}{\sigma_{\max}^2}. \tag{22}$$

Let

$$\mathbf{J}_\star = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \tag{23}$$

be a singular value decomposition. For each $\sigma_i > 0$, define

$$a_i^{(k)} = \mathbf{v}_i^\top(\mathbf{z}_k - \mathbf{z}^\star), \qquad c_i = \mathbf{u}_i^\top\mathbf{d}^\star, \qquad q_i = 1 - \gamma\sigma_i^2. \tag{24}$$

Here, $a_i^{(k)}$ is the signed component of the latent estimation error along the ith right singular direction of the learned measurement operator.

Theorem 1 (Finite-step bias–correction dynamics)**.** *For every $\sigma_i > 0$,*

$$a_i^{(K)} = q_i^K a_i^{(0)} - \frac{1-q_i^K}{\sigma_i}c_i - \gamma\sum_{j=0}^{K-1} q_i^{K-1-j}\xi_{i,j}, \tag{25}$$

where $\xi_{i,j} = \mathbf{v}_i^\top\boldsymbol{\xi}_F(\mathbf{e}_j)$. If

$$R_K = \max_{0\le j<K} R_F(\mathbf{e}_j), \tag{26}$$

then

$$\left|\gamma\sum_{j=0}^{K-1} q_i^{K-1-j}\xi_{i,j}\right| \le \frac{1-q_i^K}{\sigma_i^2}R_K \le K\gamma R_K. \tag{27}$$

Eq. 25 separates the finite-step error into the retained component of the initial error, a residual-induced term, and a Jacobian-variation remainder. To characterize the spectral behavior of the first two terms, define

$$h_K(\sigma) = \frac{1-(1-\gamma\sigma^2)^K}{\sigma}. \tag{28}$$

Corollary 1 (Finite-step behavior of the linearized component). *For* $0 < \sigma \le \sigma_{\max}$,

$$h_K(\sigma) \le \min\left\{K\gamma\sigma, \frac{1}{\sigma}\right\} \le \sqrt{K\gamma}. \tag{29}$$

Moreover, for every nonzero singular mode,

$$\left|(1 - q_i^K)a_i^{(0)}\right| \le K\gamma\sigma_i^2\|\mathbf{e}_0\|_2, \tag{30}$$

and

$$|h_K(\sigma_i)c_i| \le K\gamma\sigma_i\|\mathbf{d}^\star\|_2. \tag{31}$$

Thus, at the level of the linearized component, finite refinement produces smaller measurement corrections and smaller residual transfer along directions that are weakly constrained by the learned measurement operator. The complete nonlinear dynamics also contain the Jacobian-variation remainder in Eq. 27, which is bounded but need not vanish as σ_i becomes small. Accordingly, the spectral interpretation applies to the linearized component rather than to the full nonlinear dynamics.

For the linearized model, fixing $\sigma_i > 0$ and letting $K \to \infty$ yields

$$q_i^K \to 0, \qquad h_K(\sigma_i) \to \frac{1}{\sigma_i}, \tag{32}$$

and hence

$$a_i^{(\infty)} = -\frac{c_i}{\sigma_i}. \tag{33}$$

Finite refinement therefore avoids immediately approaching the aggressive surrogate-fitting limit in the linearized component and limits residual-induced amplification along weakly constrained directions. This provides a local rationale for using finite PLR rather than minimizing the learned measurement objective to convergence. Detailed proofs and the relation to SMG are given in Appendix C.

Remark. The analysis is local and characterizes an idealized gradient-descent dynamics for the PLR measurement subproblem, rather than the exact adaptive optimizer or the full diffusion trajectory. The singular spectrum of $\mathbf{J}_\star$ refers to sensitivity of the learned forward operator rather than biological neural observability, and the spectral interpretation applies only to the linearized component, with nonlinear deviations controlled by the Jacobian-variation remainder in Theorem 1. The analysis concerns MSC and does not identify QSC with null-space recovery.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on the Natural Scenes Dataset (NSD), which provides fMRI responses to natural images from MS-COCO. Following the cross-subject protocol of MindAligner, we use Subjects 1, 2, 5, and 7, resulting in 12 directed source-to-target transfer settings. We adopt the same preprocessing and evaluation split as MindAligner. For each source subject, results are averaged over its three target subjects, and "All" denotes the macro-average over all 12 transfer directions.

Evaluation metrics. Following standard fMRI visual decoding protocols, we evaluate decoded images using both low- and high-level metrics. The low-level group includes PixCorr, SSIM, Alex(2), and Alex(5), while the high-level group includes Inception, CLIP, EfficientNet, and SwAV. Higher values are preferred for PixCorr, SSIM, Alex(2), Alex(5), Inception, and CLIP, whereas lower values are preferred for EfficientNet and SwAV. Alex(2), Alex(5), Inception, and CLIP are reported as percentages.

Default configuration. Unless otherwise specified, all experiments use the pretrained MindAligner decoding pipeline without parameter updates. MSC uses a fixed NeuroFlow neural forward operator, while QSC constructs its multi-scale structural memory from VGG layers 1–3 with $\lambda_{\mathrm{q}} = 0.2$. All pretrained decoding, forward, and feature-extraction components remain frozen during DS-DC inference.

Forward-operator provenance. The neural forward operator used by MSC is pretrained independently of DS-DC and remains fixed throughout inference. Its subject-specific pretraining supervision is treated separately from the cross-subject alignment budget used by MindAligner.

4.2 Main Results

Table 1 compares DS-DC with the frozen pretrained MindAligner decoder. On the overall macro-average, DS-DC improves all eight evaluation metrics, including PixCorr from 0.206 to 0.218, CLIP from 81.47% to 83.06%, and EfficientNet distance from 0.814 to 0.798. The gains are broadly observed across all four source subjects, despite minor metric-specific trade-offs, showing that test-time dual-space consistency improves the frozen pretrained decoder without parameter updates.

Table 1: Cross-subject visual decoding using the pretrained MindAligner backbone. Results for each source subject are averaged over its three target subjects, while "All" denotes the macro-average over all 12 directed transfer settings.

Source	Method	Low-Level				High-Level			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Subj1	MindAligner	0.223	0.417	87.83%	92.68%	83.94%	81.83%	0.815	0.461
	DS-DC (Ours)	**0.240**	**0.421**	**89.36%**	**93.75%**	**85.48%**	**83.43%**	**0.796**	**0.452**
Subj2	MindAligner	0.216	**0.429**	87.65%	93.64%	83.84%	83.14%	0.802	0.452
	DS-DC (Ours)	**0.220**	0.424	**88.01%**	**93.70%**	**86.07%**	**84.17%**	**0.790**	**0.448**
Subj5	MindAligner	0.196	0.411	84.37%	91.99%	84.54%	83.20%	0.795	0.452
	DS-DC (Ours)	**0.219**	**0.415**	**85.62%**	**92.97%**	**87.43%**	**85.91%**	**0.769**	**0.437**
Subj7	MindAligner	0.187	0.410	81.66%	88.34%	80.06%	77.71%	0.845	0.485
	DS-DC (Ours)	**0.192**	**0.413**	**82.17%**	**89.19%**	**81.21%**	**78.70%**	**0.838**	**0.479**
All	MindAligner	0.206	0.417	85.38%	91.66%	83.09%	81.47%	0.814	0.462
	DS-DC (Ours)	**0.218**	**0.418**	**86.29%**	**92.40%**	**85.05%**	**83.06%**	**0.798**	**0.454**

4.3 Ablation and Analysis

Complementary roles of MSC and QSC. Table 2 isolates the two consistency spaces. MSC mainly improves feature-based and high-level metrics, while reducing SSIM from 0.417 to 0.398. QSC exhibits a complementary profile, increasing SSIM to 0.435 but providing limited gains on the high-level metrics. Combining both restores SSIM to 0.418 while retaining much of the high-level improvement from MSC and achieves the highest PixCorr (0.218). These results support applying measurement and structural constraints at different stages of the decoding process.

Table 2: Ablation of the two consistency spaces, macro-averaged over all 12 directed transfer settings. MSC acts during diffusion-prior sampling and QSC during image enhancement.

Method	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
MindAligner	0.206	0.417	85.38%	91.66%	83.09%	81.47%	0.814	0.462
+ MSC	0.213	0.398	**87.04%**	**93.05%**	**85.08%**	**83.52%**	**0.787**	**0.444**
+ QSC	0.208	**0.435**	84.56%	90.61%	83.16%	81.05%	0.826	0.474
DS-DC (Ours)	**0.218**	0.418	86.29%	92.40%	85.05%	83.06%	0.798	0.454

Measurement-space validation. We further evaluate whether MSC improves consistency with the observed source-subject fMRI in the learned measurement space. As shown in Table 3, MSC reduces measurement MSE by 24.45% and increases Pearson correlation from 0.358 to 0.541 over all 12 transfer directions. On the six directions supported by both forward models, the improvement is substantially larger when the evaluator matches the guidance operator, whereas smaller but directionally favorable changes remain under cross-operator evaluation in both directions. These results indicate that MSC primarily improves consistency under its guidance surrogate, with part of the change also detectable by an independently pretrained forward model. Detailed component-wise, NMSE, and paired analyses are provided in Appendix B.2.

Table 3: Compact measurement-space validation of MSC. The first row reports the result over all 12 transfer directions; the remaining rows use the six directions supported by both forward operators. Arrows denote the change from unguided sampling to MSC under a fixed evaluator.

Guidance → Evaluator	**MSE ↓**	**Pearson ↑**	**MSE Red. ↑**
NeuroFlow → NeuroFlow (12)	0.374 → 0.282	0.358 → 0.541	**24.45%**
NeuroFlow → NeuroFlow (6)	0.376 → 0.289	0.365 → 0.538	23.17%
NeuroFlow → SynBrain (6)	0.460 → 0.452	0.372 → 0.378	1.77%
SynBrain → SynBrain (6)	0.460 → 0.321	0.372 → 0.512	30.23%
SynBrain → NeuroFlow (6)	0.376 → 0.372	0.365 → 0.370	0.98%

Measurement-space refinement. To examine the intended role of PLR, we vary the refinement depth K while keeping all other components fixed. The analysis uses 1,000 paired source-1-to-target-2 samples, with $K = 0$ denoting the SMG-only setting.

As shown in Figure 3, increasing K monotonically reduces measurement MSE and increases response correlation with the observed fMRI. This behavior supports PLR as an additional measurement-space refinement mechanism beyond step-wise guidance.

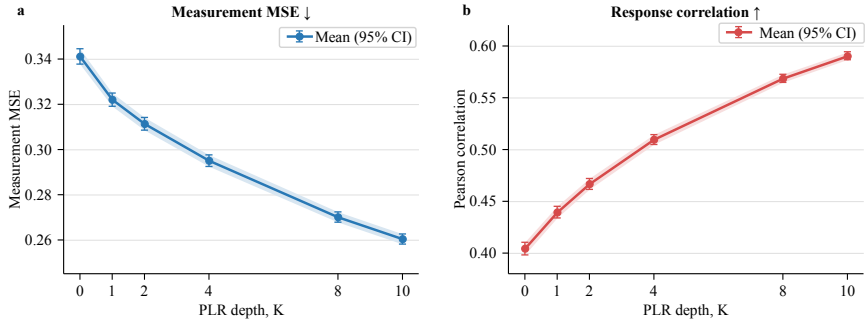

Figure 3: Effect of PLR depth on measurement-space consistency using 1,000 paired source-1-to-target-2 samples. $K = 0$ denotes SMG-only. Error bars indicate 95% paired bootstrap confidence intervals.

5 Conclusion

We introduced Dual-Space Data Consistency (DS-DC), a training-free inference framework for cross-subject fMRI visual decoding. DS-DC complements cross-subject alignment by reusing test-time neural evidence in two complementary spaces: MSC constrains diffusion-prior sampling through measurement feedback and finite latent refinement, while QSC introduces fMRI-derived structural guidance during image enhancement. Experiments on NSD across 12 directed transfer settings show that DS-DC improves the frozen MindAligner baseline, with ablations confirming the complementary roles of the two consistency mechanisms and measurement-space analysis supporting the intended behavior of MSC. Although evaluated with MindAligner, the formulation is not architecture-specific and may extend to related pretrained decoding pipelines with compatible inference states.

AI Use Statement

In this work, we used generative AI tools for language refinement, proofreading, and improving the clarity and organization of the manuscript. We have not used generative AI tools for generating experimental results, designing the proposed methodology, performing scientific analysis, or making research decisions, and the rest of the required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for improving the presentation of figures, tables, and manuscript formatting. We have reviewed all AI-assisted work. Specifically, all AI-assisted text,

mathematical formulations, and visual artifacts were manually checked and revised by the authors. Experimental results, conclusions, and scientific claims were independently developed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

This work involves the analysis of publicly available human fMRI data. The experiments are conducted on the Natural Scenes Dataset (NSD), which was collected and released following appropriate ethical procedures. We use the dataset only for scientific research on brain-vision representation and visual decoding.

Our study focuses on population-level neural representation analysis and does not aim to identify individuals, infer private information, or provide clinical diagnosis. All experiments follow the data usage guidelines and research practices associated with the dataset.

Reproducibility Statement

We provide details necessary to reproduce our experiments in the main paper and supplementary materials. The proposed DS-DC framework, including the Measurement-Space Consistency and Query-Guided Structural Consistency modules, is described in detail, together with the experimental settings, evaluation protocols, and implementation details.

All experiments are conducted using the publicly available Natural Scenes Dataset (NSD). The source code and configuration files will be released after the review process once they have been fully organized and prepared, to facilitate reproduction and further research.

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. URL https://doi.org/10.1038/s41593-021-00962-x.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fMRI data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1W2WlYRq0K.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 14247–14255, 2025.

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.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 5899–5908, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3613832. URL https://doi.org/10.1145/3581783.3613832.

Shayan Mohajer Hamidi, Ben Liang, and EN-HUI YANG. Coupled data and measurement space dynamics for enhanced diffusion posterior sampling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 55554–55592. Curran Associates, Inc., 2025. doi: 10.52202/085713-1857. URL

https://proceedings.neurips.cc/paper_files/paper/2025/file/503acf8f49c96341cd5ae8458942a55c-Paper-Conference.pdf.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. doi: 10.1038/s41598-023-42891-8. URL https://doi.org/10.1038/s41598-023-42891-8.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24705–24728. Curran Associates, Inc., 2023. doi: 10.52202/075280-1073. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4ddab70bf41ffe5d423840644d3357f4-Paper-Conference.pdf.

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 *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=65XKBGH5PO.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 7624–7654, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1f09e1ee5035a4c3fe38a5681cae5815-Paper-Conference.pdf.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342, 2024. doi: 10.1109/CVPR52733.2024.01077.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2246–2257, 2025. doi: 10.1109/CVPR52734.2025.00215.

Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. Clip-mused: Clip-guided multi-subject visual neural information semantic decoding. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 52776–52792, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e962ee491a659f957f0d435f7a133e49-Paper-Conference.pdf.

A Related Work

A.1 Generative fMRI Visual Decoding

fMRI visual decoding has evolved from predicting predefined stimulus attributes and deep visual representations to synthesizing natural images with pretrained generative models. Modern approaches typically map fMRI activity into image-related representations and use generative priors to translate these representations into visual outputs. Brain-Diffuser first reconstructs a low-level image estimate with a VDVAE and then refines it using latent diffusion conditioned on decoded CLIP visual and textual features Ozcelik & VanRullen (2023). MindEye instead maps fMRI activity into CLIP image space and combines the predicted representation with a diffusion prior, while a separate low-level pathway provides complementary information for image-to-image generation Scotti et al. (2023).

Subsequent work has explored additional forms of fMRI-derived guidance during generative decoding. MindDiffuser incorporates decoded CLIP visual features to refine structural information in a second generation stage Lu et al. (2023), whereas NeuralDiffuser introduces primary-visual feature guidance to provide detail-sensitive gradients during diffusion Li et al. (2025). Related work has also investigated uncertainty-aware blur priors that reduce the influence of visual details weakly supported by neural measurements Wu et al. (2025). Together, these studies indicate that high-level semantic representations and fMRI-derived low-level visual cues provide complementary information for generative visual decoding.

A.2 Cross-Subject fMRI Visual Decoding

Most high-fidelity fMRI visual decoders are trained separately for individual subjects because voxel organization, response distributions, and functional correspondences vary substantially across brains. This subject-specific training requires substantial paired neural–visual data for each participant and limits the reuse of pretrained decoders. Cross-subject methods address this limitation by learning shared neural representations, functional mappings, or lightweight subject-specific adaptations.

MindEye2 maps measurements from multiple participants into a shared subject space before predicting CLIP representations for subsequent image generation Scotti et al. (2024). MindAligner instead learns an explicit brain transfer matrix together with multi-level functional alignment, allowing neural activity from a new subject to be mapped into the functional space expected by a pretrained decoder Dai et al. (2025). MindTuner combines multi-subject pretraining with lightweight subject-specific adaptation and semantic correction to support data-efficient transfer Gong et al. (2025). These approaches show that inter-subject variability can be partially addressed by transforming neural measurements into representations compatible with pretrained decoding models.

Cross-subject alignment, however, primarily addresses *representation transfer*: it determines how measurements from a new subject can be expressed in a space suitable for an existing decoder. This objective is distinct from *inference-time consistency*. Once the transferred representation enters a pretrained generative model, subsequent sampling is governed jointly by the transferred conditioning and the learned generative prior, without explicitly re-evaluating intermediate estimates against the observed source-subject measurement. DS-DC is therefore complementary to cross-subject alignment: it preserves the underlying transfer mechanism while introducing additional consistency constraints during generative-prior sampling and subsequent image enhancement.

A.3 Consistency-Guided Generation and Structural Refinement

Diffusion models provide expressive image priors for inverse problems, but samples favored by the generative prior need not satisfy the corresponding observations. Diffusion-based inverse solvers address this mismatch by incorporating measurement information into posterior sampling or by introducing explicit data-consistency updates during the reverse process Song et al. (2024); Mohajer Hamidi et al. (2025). In conventional imaging problems, such consistency is commonly defined through a known forward process. For fMRI visual decoding, however, the visual-to-neural transformation is unknown and subject dependent, making direct measurement-space constraints less

straightforward. This motivates the use of learned neural forward mappings to relate intermediate visual representations to the observed neural response.

Structural guidance provides a complementary source of information during generative visual decoding. Prior fMRI decoders have incorporated decoded visual features or low-level image estimates into diffusion-based generation and refinement Ozcelik & VanRullen (2023); Lu et al. (2023); Li et al. (2025). Uncertainty-aware blur priors further illustrate the importance of reducing the influence of structural details that are weakly supported by neural measurements Wu et al. (2025). These studies suggest that low-level fMRI-derived cues can complement high-level semantic conditioning, particularly when spatial structure and fine visual details are not fully preserved by the generative prior.

DS-DC combines these two forms of consistency within a pretrained cross-subject decoder. During diffusion-prior sampling, MSC introduces measurement-space feedback from the observed source-subject fMRI through a learned neural forward operator. During image enhancement, QSC uses the fMRI-derived blurry estimate as a structural reference for local, reliability-aware feature matching. The two mechanisms operate at different inference stages while preserving the underlying cross-subject transfer mechanism and keeping the pretrained components fixed.

B Additional Experimental Results

B.1 MSC Component Analysis

Table 4 further separates MSC into Step-wise Measurement Guidance (SMG) and Periodic Latent Refinement (PLR). Both components improve the baseline, but their metric profiles differ: SMG provides stronger gains on several feature-based metrics, whereas PLR yields a larger improvement in PixCorr. Combining them gives the strongest overall performance among the evaluated MSC variants, including the best or tied-best results on most high-level metrics. This empirical complementarity motivates retaining both the continuous step-wise correction and the periodic latent refinement in the full MSC formulation.

Table 4: Component-wise ablation of Measurement-Space Consistency (MSC). SMG performs step-wise measurement guidance, while PLR performs periodic latent refinement. Results are averaged over three target subjects for each source subject and over all 12 directed transfer settings.

Source	Method	Low-Level				High-Level			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Subj1	MindAligner	0.223	**0.417**	87.83%	92.68%	83.94%	81.83%	0.815	0.461
	+ SMG	0.228	0.398	89.59%	**94.32%**	**85.57%**	83.72%	0.787	0.442
	+ PLR	**0.234**	0.402	89.88%	94.10%	85.28%	83.77%	0.786	0.442
	+ MSC	**0.234**	0.401	**90.05%**	94.24%	85.39%	**83.98%**	**0.784**	**0.441**
Subj2	MindAligner	**0.216**	**0.429**	87.65%	93.64%	83.84%	83.14%	0.802	0.452
	+ SMG	0.208	0.402	**88.78%**	**94.31%**	85.76%	84.33%	0.783	0.439
	+ PLR	0.215	0.405	88.63%	94.24%	85.80%	**84.64%**	0.781	0.437
	+ MSC	**0.216**	0.405	88.76%	**94.31%**	**86.15%**	84.56%	**0.778**	**0.436**
Subj5	MindAligner	0.196	**0.411**	84.37%	91.99%	84.54%	83.20%	0.795	0.452
	+ SMG	0.206	0.391	86.29%	93.33%	87.08%	85.94%	0.760	0.430
	+ PLR	0.213	0.396	**86.46%**	**93.57%**	87.20%	86.30%	0.758	0.427
	+ MSC	**0.214**	0.395	86.39%	93.55%	**87.42%**	**86.39%**	**0.757**	**0.426**
Subj7	MindAligner	**0.187**	**0.410**	81.66%	88.34%	80.06%	77.71%	0.845	0.485
	+ SMG	0.180	0.390	**83.51%**	89.93%	80.82%	**79.20%**	**0.830**	0.473
	+ PLR	**0.187**	0.393	82.97%	**90.11%**	81.07%	**79.20%**	0.831	**0.472**
	+ MSC	**0.187**	0.392	82.96%	90.08%	**81.36%**	79.15%	**0.830**	**0.472**
All	MindAligner	0.206	**0.417**	85.38%	91.66%	83.09%	81.47%	0.814	0.462
	+ SMG	0.206	0.395	**87.04%**	92.97%	84.81%	83.30%	0.790	0.446
	+ PLR	0.212	0.399	86.99%	93.00%	84.84%	83.48%	0.789	**0.444**
	+ MSC	**0.213**	0.398	**87.04%**	**93.05%**	**85.08%**	**83.52%**	**0.787**	**0.444**

B.2 MEASUREMENT-SPACE VALIDATION

Evaluation protocol. We evaluate measurement-space consistency using the final diffusion-prior representation produced by Stage I. Given a fixed representation, a pretrained subject-specific neural forward operator predicts the corresponding source-subject fMRI response, which is compared with the observed measurement. All forward operators remain frozen during evaluation; the evaluator neither modifies the latent representation nor participates in visual decoding.

Table 5 reports two complementary analyses. Panel (a) evaluates the effects of SMG and PLR under the primary NeuroFlow operator over all 12 directed transfer settings. Panel (b) performs bidirectional cross-operator evaluation over the six directions supported by both NeuroFlow and SynBrain.

Direct measurement-space consistency. As shown in Table 5(a), both SMG and PLR reduce the discrepancy between the forward-predicted response and the observed source-subject fMRI, while their combination yields the strongest measurement consistency. MSC reduces MSE from 0.374 to 0.282 (24.45%), decreases NMSE from 0.878 to 0.669, and increases Pearson correlation from 0.358 to 0.541.

These results directly support the intended role of MSC as an inference-time measurement-space constraint. Importantly, the evaluation is defined with respect to a learned neural forward operator and should not be interpreted as direct agreement with the unknown biological encoding process.

Cross-operator evaluation. Because the direct analysis uses the same learned operator for guidance and evaluation, we additionally evaluate the fixed final representations using the alternative pretrained forward model. Table 5(b) shows that the largest gains occur when the evaluator matches the guidance operator. For NeuroFlow-guided representations evaluated by SynBrain, MSE decreases from 0.460 to 0.452 and Pearson correlation increases from 0.372 to 0.378. In the reverse SynBrain-to-NeuroFlow setting, MSE decreases from 0.376 to 0.372 and Pearson correlation increases from 0.365 to 0.370.

Thus, cross-operator changes are substantially smaller than the corresponding guidance-matched improvements. This pattern indicates that MSC primarily optimizes a guidance-specific learned measurement objective, while a smaller component of the resulting change remains detectable under an independently pretrained operator. The result therefore serves as a robustness check rather than evidence of operator-independent measurement consistency.

Paired analysis. We further examine cross-operator changes at the individual transfer-direction level using paired bootstrap analysis. For NeuroFlow-guided representations evaluated by SynBrain, the paired MSE difference is negative in all six transfer settings. For SynBrain-guided representations evaluated by NeuroFlow, the corresponding trend holds in five of the six settings. We therefore interpret the cross-operator result as a modest but reproducible robustness signal, rather than evidence that the learned forward operators are equivalent or that either represents the unknown true neural encoding process.

B.3 QSC HYPERPARAMETER ANALYSIS

Table 6 reveals a clear and monotonic trend over the evaluated guidance range. Increasing the QSC scale improves PixCorr and SSIM, whereas several high-level metrics gradually decrease. Thus, stronger structural guidance does not uniformly improve all aspects of decoding, but shifts the operating point toward greater low-level similarity. We choose 0.2 as the default scale because it provides an intermediate trade-off between structural and high-level performance rather than maximizing a single metric.

Table 7 shows that increasing the depth of the structural memory changes the same balance from a different perspective. Using VGG layers 1–4 improves several feature-based metrics, but decreases PixCorr and SSIM relative to layers 1–3. This suggests that including deeper features changes the information emphasized by QSC, whereas the shallower memory is better aligned with the structural role assigned to QSC in DS-DC. We therefore use VGG layers 1–3 in the default configuration.

Table 5: Measurement-space validation of MSC. (a) Component-wise evaluation under the primary NeuroFlow forward operator over all 12 transfer directions. (b) Bidirectional cross-operator evaluation over the six directions supported by both NeuroFlow and SynBrain. Arrows denote the change from unguided sampling to MSC under a fixed evaluator. Raw metric values are compared only within the same evaluator.

(a) Direct measurement-space consistency (12 directions)

Method	MSE ↓	NMSE ↓	Pearson ↑	MSE Red. ↑
MindAligner	0.374	0.878	0.358	–
+ SMG	0.302	0.716	0.488	19.21%
+ PLR	0.327	0.770	0.462	12.52%
MSC	**0.282**	**0.669**	**0.541**	**24.45%**

(b) Bidirectional cross-operator evaluation (6 common directions)

Guidance	Evaluator	MSE ↓	NMSE ↓	Pearson ↑	MSE Red. ↑
NeuroFlow	NeuroFlow	0.376 → 0.289	0.876 → 0.678	0.365 → 0.538	23.17%
NeuroFlow	SynBrain	0.460 → 0.452	1.083 → 1.063	0.372 → 0.378	1.77%
SynBrain	SynBrain	0.460 → 0.321	1.083 → 0.757	0.372 → 0.512	30.23%
SynBrain	NeuroFlow	0.376 → 0.372	0.876 → 0.867	0.365 → 0.370	0.98%

Table 6: Sensitivity analysis of QSC guidance strength. Results are averaged over three target subjects for each source subject and over all 12 transfer directions.

Source	**QSC Scale**	**Low-Level**				**High-Level**			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Subj1	MindAligner	0.223	0.417	87.83%	92.68%	83.94%	81.83%	0.815	0.461
	MSC only	0.234	0.401	**90.05%**	**94.24%**	85.39%	**83.98%**	**0.784**	**0.441**
	0.1	0.237	0.411	89.76%	94.05%	85.41%	83.75%	0.789	0.445
	0.2	0.240	0.421	89.36%	93.75%	**85.48%**	83.43%	0.796	0.452
	0.3	**0.242**	**0.429**	88.88%	93.31%	85.34%	83.19%	0.803	0.461
Subj2	MindAligner	0.216	0.429	87.65%	93.64%	83.84%	83.14%	0.802	0.452
	MSC only	0.216	0.405	**88.76%**	**94.31%**	**86.15%**	**84.56%**	**0.778**	**0.436**
	0.1	0.218	0.415	88.43%	94.06%	86.14%	84.38%	0.784	0.440
	0.2	0.220	0.424	88.01%	93.70%	86.07%	84.17%	0.790	0.448
	0.3	**0.223**	**0.432**	87.54%	93.31%	85.99%	83.89%	0.798	0.457
Subj5	MindAligner	0.196	0.411	84.37%	91.99%	84.54%	83.20%	0.795	0.452
	MSC only	0.214	0.395	**86.39%**	**93.55%**	87.42%	**86.39%**	**0.757**	**0.426**
	0.1	0.216	0.405	86.05%	93.33%	**87.48%**	86.18%	0.761	0.430
	0.2	0.219	0.415	85.62%	92.97%	87.43%	85.91%	0.769	0.437
	0.3	**0.222**	**0.424**	85.16%	92.45%	87.17%	85.47%	0.777	0.446
Subj7	MindAligner	0.187	0.410	81.66%	88.34%	80.06%	77.71%	0.845	0.485
	MSC only	0.187	0.392	**82.96%**	**90.08%**	81.36%	**79.15%**	**0.830**	**0.472**
	0.1	0.189	0.403	82.63%	89.69%	**81.38%**	78.97%	0.833	0.474
	0.2	0.192	0.413	82.17%	89.19%	81.21%	78.70%	0.838	0.479
	0.3	**0.194**	**0.421**	81.65%	88.58%	81.21%	78.25%	0.843	0.485
All	MindAligner	0.206	0.417	85.38%	91.66%	83.09%	81.47%	0.814	0.462
	MSC only	0.213	0.398	**87.04%**	**93.05%**	85.08%	**83.52%**	**0.787**	**0.444**
	0.1	0.215	0.408	86.72%	92.78%	**85.10%**	83.32%	0.792	0.447
	0.2	0.218	0.418	86.29%	92.40%	85.05%	83.06%	0.798	0.454
	0.3	**0.220**	**0.426**	85.81%	91.91%	84.93%	82.70%	0.805	0.462

B.4 Source-wise Analysis of the Two Consistency Spaces

The main paper reports the macro-averaged ablation of MSC and QSC over all 12 directed transfer settings. Here, we provide the corresponding source-wise results in Table 8 to examine whether the observed behavior is consistent across different source subjects.

For each source subject, results are averaged over its three target subjects, while "All" denotes the macro-average over all 12 directed transfer directions.

Table 7: Ablation of the VGG feature depth used to construct the structural memory in QSC. The QSC scale is fixed to 0.2.

Source	QSC Memory	Low-Level				High-Level			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Subj1	MindAligner	0.223	0.417	87.83%	92.68%	83.94%	81.83%	0.815	0.461
	VGG 1–3	**0.240**	**0.421**	89.36%	93.75%	**85.48%**	83.43%	0.796	0.452
	VGG 1–4	0.237	0.411	**89.70%**	**94.07%**	85.42%	**83.68%**	**0.789**	**0.444**
Subj2	MindAligner	0.216	**0.429**	87.65%	93.64%	83.84%	83.14%	0.802	0.452
	VGG 1–3	**0.220**	0.424	88.01%	93.70%	86.07%	84.17%	0.790	0.448
	VGG 1–4	0.218	0.415	**88.31%**	**94.06%**	**86.28%**	**84.36%**	**0.783**	**0.439**
Subj5	MindAligner	0.196	0.411	84.37%	91.99%	84.54%	83.20%	0.795	0.452
	VGG 1–3	**0.219**	**0.415**	85.62%	92.97%	87.43%	85.91%	0.769	0.437
	VGG 1–4	0.217	0.406	**85.90%**	**93.28%**	**87.75%**	**86.08%**	**0.761**	**0.429**
Subj7	MindAligner	0.187	0.410	81.66%	88.34%	80.06%	77.71%	0.845	0.485
	VGG 1–3	**0.192**	**0.413**	82.17%	89.19%	81.21%	78.70%	0.838	0.479
	VGG 1–4	0.190	0.403	**82.54%**	**89.66%**	**81.58%**	**78.88%**	**0.832**	**0.473**
All	MindAligner	0.206	0.417	85.38%	91.66%	83.09%	81.47%	0.814	0.462
	VGG 1–3	**0.218**	**0.418**	86.29%	92.40%	85.05%	83.06%	0.798	0.454
	VGG 1–4	0.215	0.409	**86.61%**	**92.77%**	**85.26%**	**83.25%**	**0.791**	**0.446**

Table 8: Full source-wise ablation of the two consistency spaces in DS-DC. MSC is applied during diffusion-prior sampling, whereas QSC is applied during subsequent image enhancement. Results for each source subject are averaged over its three target subjects, while "All" denotes the macro-average over all 12 directed transfer settings. The corresponding macro-averaged ablation is reported in the main paper.

Source	Method	Low-Level				High-Level			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Subj1	MindAligner	0.223	0.417	87.83%	92.68%	83.94%	81.83%	0.815	0.461
	+ MSC	0.234	0.401	**90.05%**	**94.24%**	85.39%	**83.98%**	**0.784**	**0.441**
	+ QSC	0.229	**0.435**	87.04%	92.11%	83.79%	81.79%	0.825	0.470
	DS-DC (Ours)	**0.240**	0.421	89.36%	93.75%	**85.48%**	83.43%	0.796	0.452
Subj2	MindAligner	0.216	0.429	87.65%	93.64%	83.84%	83.14%	0.802	0.452
	+ MSC	0.216	0.405	**88.76%**	**94.31%**	**86.15%**	**84.56%**	**0.778**	**0.436**
	+ QSC	0.216	**0.445**	86.46%	92.53%	84.41%	82.84%	0.814	0.465
	DS-DC (Ours)	**0.220**	0.424	88.01%	93.70%	86.07%	84.17%	0.790	0.448
Subj5	MindAligner	0.196	0.411	84.37%	91.99%	84.54%	83.20%	0.795	0.452
	+ MSC	0.214	0.395	**86.39%**	**93.55%**	87.42%	**86.39%**	**0.757**	**0.426**
	+ QSC	0.199	**0.432**	83.67%	90.64%	84.53%	82.55%	0.812	0.465
	DS-DC (Ours)	**0.219**	0.415	85.62%	92.97%	**87.43%**	85.91%	0.769	0.437
Subj7	MindAligner	0.187	0.410	81.66%	88.34%	80.06%	77.71%	0.845	0.485
	+ MSC	0.187	0.392	**82.96%**	**90.08%**	**81.36%**	**79.15%**	**0.830**	**0.472**
	+ QSC	0.186	**0.429**	81.07%	87.16%	79.90%	77.04%	0.855	0.495
	DS-DC (Ours)	**0.192**	0.413	82.17%	89.19%	81.21%	78.70%	0.838	0.479
All	MindAligner	0.206	0.417	85.38%	91.66%	83.09%	81.47%	0.814	0.462
	+ MSC	0.213	0.398	**87.04%**	**93.05%**	**85.08%**	**83.52%**	**0.787**	**0.444**
	+ QSC	0.208	**0.435**	84.56%	90.61%	83.16%	81.05%	0.826	0.474
	DS-DC (Ours)	**0.218**	0.418	86.29%	92.40%	85.05%	83.06%	0.798	0.454

Source-wise behavior. Table 8 shows that the complementary behavior of MSC and QSC is consistently observed across different source subjects, although the magnitude of improvement varies.

MSC generally improves feature-based and high-level semantic metrics while slightly reducing SSIM, reflecting the trade-off introduced by enforcing measurement consistency during diffusion-prior sampling. In contrast, QSC consistently provides stronger structural similarity, as indicated by higher SSIM values compared with MSC.

Combining the two consistency spaces improves PixCorr for all evaluated source subjects and recovers part of the structural similarity reduced by MSC while retaining its semantic improvements. These results further support that MSC and QSC play complementary roles rather than optimizing a single dominant objective.

B.5 Additional PLR Depth Analysis

The main paper analyzes PLR depth from the perspective of measurement-space optimization. Here we further examine whether improved measurement consistency directly translates into latent recovery and final visual quality.

As shown in Table 9, increasing PLR depth substantially improves measurement consistency, while the changes in ground-truth token distance and final visual metrics remain limited.

This observation indicates that measurement consistency alone does not fully determine the recovered visual representation, motivating the structural correction mechanism introduced by QSC.

Table 9: Additional analysis of PLR depth. PLR improves measurement consistency, while latent similarity and final visual metrics exhibit limited changes.

K	MSE↓	GT-token L2↓	PixCorr↑	CLIP↑
0	0.3412	1.065624	0.1986	0.8283
10	0.2605	1.065442	0.1998	0.8324

B.6 Residual-conditioned Analysis

We further investigate whether the initial forward residual is associated with the amount of measurement-space correction obtained from PLR refinement.

Samples are divided into low- and high-residual groups according to the median initial forward residual. Figure 4 reports the relative measurement MSE reduction with respect to the $K = 0$ SMG-only setting.

Across all refinement depths, the high-residual group consistently obtains larger measurement-space improvement than the low-residual group. At $K = 10$, PLR reduces the measurement MSE by 22.4% and 24.7% for the low- and high-residual groups, respectively, indicating that samples with larger initial measurement mismatch receive stronger correction in the learned measurement space.

B.7 Qualitative Comparison

Figure 5 provides multi-seed qualitative comparisons between the baseline decoder and DS-DC. The same random seeds are used for both methods. Across different scene categories, DS-DC consistently produces reconstructions with clearer structures and finer visual details, which agrees with the quantitative improvements reported in the main paper.

B.8 Local Linearity Analysis

We further analyze the local behavior of PLR updates using the normalized linearization error ρ_K. As shown in Table 10, the normalized error decreases consistently as the refinement depth increases, indicating that finite-step PLR refinement reduces the deviation from the local linear approximation.

However, the remaining non-negligible approximation error suggests that the optimization trajectory still operates in a nonlinear regime. Therefore, PLR adopts finite-step refinement rather than exhaustive optimization, providing controlled measurement-space correction while avoiding excessive latent displacement.

B.9 Neural Forward Operator Comparison

Table 11 examines whether the observed improvements depend strongly on the specific learned forward operator used by MSC.

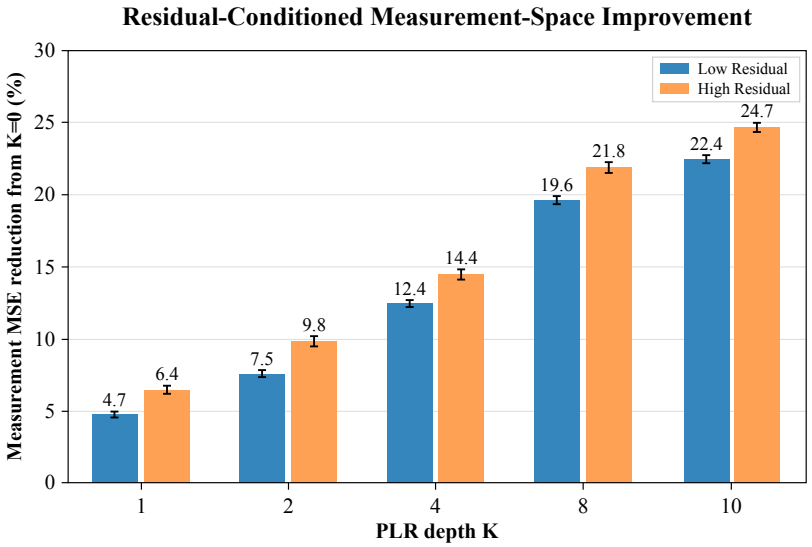

Figure 4: Residual-conditioned measurement-space improvement. Samples are divided into low- and high-residual groups according to the median initial forward residual. Bars indicate measurement MSE reduction relative to the $K = 0$ condition, with error bars showing 95% paired bootstrap confidence intervals.

Table 10: Normalized linearization error ρ_K under different PLR refinement depths.

K	0	2	4	8	10
ρ_K	2.415	1.971	1.845	1.736	1.687

SynBrain and NeuroFlow show consistent improvements over the pretrained baseline in the evaluated $1 \to 2$ transfer setting. The corresponding DS-DC variants exhibit similar trends under the two operator implementations.

This result suggests that the proposed consistency formulation is not restricted to a single evaluated operator implementation. However, since this analysis is limited to one transfer direction, we do not draw stronger conclusions regarding complete operator independence.

Table 11: Comparison of MSC using different neural forward operators on the $1 \to 2$ transfer setting. For DS-DC configurations, QSC uses VGG layers 1–3 with guidance scale 0.2.

Operator	**Configuration**	**Low-Level**				**High-Level**			
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
–	MindAligner	0.221	0.414	87.83%	92.35%	83.82%	82.47%	0.809	0.460
SynBrain	MSC	0.231	0.407	89.65%	94.32%	**85.22%**	**84.52%**	0.779	0.441
	DS-DC	0.236	**0.426**	88.98%	93.72%	84.96%	84.19%	0.792	0.453
NeuroFlow	MSC	0.232	0.400	**90.52%**	**94.43%**	85.12%	84.21%	**0.779**	**0.439**
	DS-DC (Ours)	**0.237**	0.420	89.94%	93.97%	85.01%	83.86%	0.790	0.450

C Detailed Analysis of Learned Measurement Refinement

C.1 Setup and Local Regularity

Let

$$\mathcal{F}_s : \mathcal{Z} \to \mathcal{Y}_s \tag{34}$$



Figure 5: Multi-seed qualitative comparison of cross-subject visual decoding. Each column uses the same random seed for the baseline decoder and DS-DC. DS-DC achieves clearer structures and finer details across diverse scenes.

denote the fixed learned neural forward operator used by MSC, and let $y_s \in \mathcal{Y}_s$ be the observed source-subject fMRI measurement. For the stimulus $x^\star$ associated with y_s, define

$$z^\star = E_{\text{vis}}(x^\star). \tag{35}$$

For any z, let

$$e = z - z^\star \tag{36}$$

and

$$d^\star = \mathcal{F}_s(z^\star) - y_s. \tag{37}$$

No assumption is made that $d^\star = 0$. Rather, $d^\star$ collects the discrepancy between the learned forward prediction at $z^\star$ and the observed neural response.

We study

$$\mathcal{L}_{\mathrm{m}}(z; y_s) = \frac{1}{2} \left\| \mathcal{F}_s(z) - y_s \right\|_2^2, \tag{38}$$

with gradient

$$g_F(z) := \nabla_z \mathcal{L}_{\mathrm{m}}(z; y_s) = J_F(z)^\top \big(\mathcal{F}_s(z) - y_s \big). \tag{39}$$

Assume that there exists an open ball

$$\mathcal{N} = B(z^\star, r) \tag{40}$$

such that $\mathcal{F}_s$ is continuously differentiable on $\mathcal{N}$ and

$$\| J_F(z_1) - J_F(z_2) \|_2 \le \beta_F \| z_1 - z_2 \|_2, \qquad z_1, z_2 \in \mathcal{N}. \tag{41}$$

Since $\mathcal{N}$ is convex, $z^\star + te \in \mathcal{N}$ for every $t \in [0,1]$ whenever $z^\star + e \in \mathcal{N}$.

Let

$$J_\star = J_F(z^\star), \qquad \sigma_{\max} = \| J_\star \|_2. \tag{42}$$

C.2 First-Order Expansion of the Forward Map

Lemma 1. *For every $z = z^\star + e \in \mathcal{N}$,*

$$\mathcal{F}_s(z) = \mathcal{F}_s(z^\star) + J_\star e + r_F(e), \tag{43}$$

where

$$\| r_F(e) \|_2 \le \frac{\beta_F}{2} \| e \|_2^2. \tag{44}$$

Proof. By the fundamental theorem of calculus,

$$\mathcal{F}_s(z^\star + e) - \mathcal{F}_s(z^\star) = \int_0^1 J_F(z^\star + te) e \, dt. \tag{45}$$

Subtracting $J_\star e$ yields

$$r_F(e) = \int_0^1 \big[J_F(z^\star + te) - J_\star \big] e \, dt. \tag{46}$$

Hence

$$\| r_F(e) \|_2 \le \int_0^1 \| J_F(z^\star + te) - J_\star \|_2 \| e \|_2 \, dt \le \int_0^1 \beta_F t \| e \|_2^2 \, dt = \frac{\beta_F}{2} \| e \|_2^2.$$

□

C.3 Local Decomposition of the Measurement Gradient

Proposition 2. *For every $z = z^\star + e \in \mathcal{N}$,*

$$g_F(z) = J_\star^\top J_\star e + J_\star^\top d^\star + \xi_F(e), \tag{47}$$

where

$$\| \xi_F(e) \|_2 \le R_F(e) := \frac{3}{2} \sigma_{\max} \beta_F \| e \|_2^2 + \beta_F \| e \|_2 \| d^\star \|_2 + \frac{1}{2} \beta_F^2 \| e \|_2^3. \tag{48}$$

Proof. Define

$$\Delta J(e) = J_F(z^\star + e) - J_\star. \tag{49}$$

By Eq. equation 41,

$$\|\Delta J(e)\|_2 \le \beta_F \|e\|_2. \tag{50}$$

Lemma 1 gives

$$\mathcal{F}_s(z) - y_s = d^\star + J_\star e + r_F(e). \tag{51}$$

Substituting into Eq. equation 39,

$$\begin{aligned} g_F(z) &= \big(J_\star + \Delta J(e)\big)^\top \big(d^\star + J_\star e + r_F(e)\big) \\ &= J_\star^\top J_\star e + J_\star^\top d^\star + \xi_F(e), \end{aligned} \tag{52}$$

where

$$\xi_F(e) = J_\star^\top r_F(e) + \Delta J(e)^\top J_\star e + \Delta J(e)^\top d^\star + \Delta J(e)^\top r_F(e). \tag{53}$$

The four terms satisfy

$$\|J_\star^\top r_F(e)\|_2 \le \frac{1}{2}\sigma_{\max}\beta_F\|e\|_2^2, \tag{54}$$

$$\|\Delta J(e)^\top J_\star e\|_2 \le \sigma_{\max}\beta_F\|e\|_2^2, \tag{55}$$

$$\|\Delta J(e)^\top d^\star\|_2 \le \beta_F\|e\|_2\|d^\star\|_2, \tag{56}$$

and

$$\|\Delta J(e)^\top r_F(e)\|_2 \le \frac{1}{2}\beta_F^2\|e\|_2^3. \tag{57}$$

Summing these bounds gives Eq. equation 48. □

Remark. The term $\xi_F(e)$ is referred to as a *Jacobian-variation remainder*, rather than a higher-order remainder. Indeed, when $d^\star \neq 0$,

$$\|\Delta J(e)^\top d^\star\|_2 \le \beta_F\|e\|_2\|d^\star\|_2 \tag{58}$$

is first order in $\|e\|_2$.

At $e = 0$,

$$g_F(z^\star) = J_\star^\top d^\star. \tag{59}$$

Thus, if $J_\star^\top d^\star \neq 0$, $z^\star$ is not a stationary point of the learned measurement objective.

C.4 Idealized Finite-Step Refinement

The practical PLR procedure minimizes a voxel-wise MSE using a finite number of adaptive optimization steps. For a fixed measurement dimension, this MSE differs from Eq. equation 38 only by a positive constant. For the idealized gradient-descent analysis below, that constant can be absorbed into the step size γ. No such exact reparameterization is assumed for the adaptive optimizer used in the implementation.

Consider

$$z_{k+1} = z_k - \gamma g_F(z_k), \qquad k = 0, \dots, K-1. \tag{60}$$

Let

$$e_k = z_k - z^\star \tag{61}$$

and assume

$$z_k \in \mathcal{N}, \qquad k = 0, \dots, K. \tag{62}$$

Then Proposition 2 yields

$$e_{k+1} = e_k - \gamma\big(J_\star^\top J_\star e_k + J_\star^\top d^\star + \xi_F(e_k)\big). \tag{63}$$

If $J_\star = 0$, there are no nonzero singular modes and the mode-wise results below are vacuous. We therefore assume

$$\sigma_{\max} = \|J_\star\|_2 > 0. \tag{64}$$

Let

$$J_\star = U\Sigma V^\top. \tag{65}$$

For every $\sigma_i > 0$, define

$$a_i^{(k)} = v_i^\top e_k, \qquad c_i = u_i^\top d^\star, \qquad \xi_{i,k} = v_i^\top \xi_F(e_k). \tag{66}$$

Assume

$$0 < \gamma \le \frac{1}{\sigma_{\max}^2} \tag{67}$$

and set

$$q_i = 1 - \gamma\sigma_i^2. \tag{68}$$

Then $0 \le q_i < 1$ for every $\sigma_i > 0$.

C.5 Finite-Step Bias–Correction Dynamics

Theorem 2. *For every $\sigma_i > 0$, the iteration in Eq. equation 60 satisfies*

$$a_i^{(K)} = q_i^K a_i^{(0)} - \frac{1 - q_i^K}{\sigma_i} c_i \gamma \sum_{j=0}^{K-1} q_i^{K-1-j} \xi_{i,j}. \tag{69}$$

If

$$R_K = \max_{0 \le j < K} R_F(e_j), \tag{70}$$

then

$$\left| \gamma \sum_{j=0}^{K-1} q_i^{K-1-j} \xi_{i,j} \right| \le \frac{1 - q_i^K}{\sigma_i^2} R_K \le K\gamma R_K. \tag{71}$$

Proof. Projecting Eq. equation 63 onto v_i gives

$$\begin{aligned} a_i^{(k+1)} &= a_i^{(k)} - \gamma\sigma_i^2 a_i^{(k)} - \gamma\sigma_i c_i - \gamma\xi_{i,k} \\ &= q_i a_i^{(k)} - \gamma\sigma_i c_i - \gamma\xi_{i,k}. \end{aligned} \tag{72}$$

Iterating this recurrence yields

$$a_i^{(K)} = q_i^K a_i^{(0)} - \gamma\sigma_i \sum_{j=0}^{K-1} q_i^{K-1-j} c_i - \gamma \sum_{j=0}^{K-1} q_i^{K-1-j} \xi_{i,j}.$$

Since

$$\sum_{j=0}^{K-1} q_i^{K-1-j} = \frac{1 - q_i^K}{1 - q_i} \tag{73}$$

and

$$1 - q_i = \gamma\sigma_i^2, \tag{74}$$

we have

$$\gamma\sigma_i \frac{1 - q_i^K}{1 - q_i} = \frac{1 - q_i^K}{\sigma_i}, \tag{75}$$

which proves Eq. 69.

Furthermore,

$$|\xi_{i,j}| \le \|\xi_F(e_j)\|_2 \le R_K. \tag{76}$$

Hence

$$\left| \gamma \sum_{j=0}^{K-1} q_i^{K-1-j} \xi_{i,j} \right| \le \gamma R_K \sum_{j=0}^{K-1} q_i^{K-1-j} = \frac{1 - q_i^K}{\sigma_i^2} R_K.$$

Because

$$0 \le \gamma\sigma_i^2 \le 1, \tag{77}$$

Bernoulli's inequality gives

$$1 - (1 - \gamma\sigma_i^2)^K \le K\gamma\sigma_i^2. \tag{78}$$

Therefore,

$$\frac{1 - q_i^K}{\sigma_i^2} \le K\gamma, \tag{79}$$

which proves Eq. equation 71. □

C.6 Finite-Step Spectral Behavior of the Linearized Component

Theorem 2 contains a Jacobian-variation remainder that is generally not diagonal in the singular-vector basis of $J_\star$. We therefore restrict the following spectral interpretation to the first two terms of Eq. 69.

Define

$$h_K(\sigma) = \frac{1-(1-\gamma\sigma^2)^K}{\sigma}. \tag{80}$$

Corollary 2. *Under Eq. equation 67, for every $0 < \sigma \le \sigma_{\max}$,*

$$1-(1-\gamma\sigma^2)^K \le \min\left\{K\gamma\sigma^2, 1\right\}, \tag{81}$$

and

$$h_K(\sigma) \le \min\left\{K\gamma\sigma, \frac{1}{\sigma}\right\} \le \sqrt{K\gamma}. \tag{82}$$

Consequently, for every nonzero singular mode,

$$\left|(1-q_i^K)a_i^{(0)}\right| \le K\gamma\sigma_i^2\|e_0\|_2, \tag{83}$$

and

$$|h_K(\sigma_i)c_i| \le K\gamma\sigma_i\|d^\star\|_2. \tag{84}$$

Proof. Let

$$x = \gamma\sigma^2. \tag{85}$$

Then $0 < x \le 1$. Bernoulli's inequality gives

$$(1-x)^K \ge 1-Kx, \tag{86}$$

and hence

$$1-(1-x)^K \le Kx. \tag{87}$$

Since $1-(1-x)^K \le 1$, Eq. equation 81 follows.

Dividing by $\sigma > 0$ yields

$$h_K(\sigma) \le \min\left\{K\gamma\sigma, \frac{1}{\sigma}\right\}. \tag{88}$$

For positive a and b, $\min\{a,b\} \le \sqrt{ab}$. Taking

$$a = K\gamma\sigma, \qquad b = \frac{1}{\sigma} \tag{89}$$

gives

$$h_K(\sigma) \le \sqrt{K\gamma}. \tag{90}$$

Finally,

$$|a_i^{(0)}| = |v_i^\top e_0| \le \|e_0\|_2 \tag{91}$$

and

$$|c_i| = |u_i^\top d^\star| \le \|d^\star\|_2. \tag{92}$$

Combining these inequalities with Eqs. equation 81 and equation 82 gives Eqs. equation 83 and equation 84. □

For fixed K, γ, e_0, and $d^\star$, Eqs. equation 83 and equation 84 show that the linearized measurement correction is at most quadratic in σ_i, whereas the residual-transfer term is at most linear in σ_i as σ_i becomes small. In contrast, Theorem 2 provides only

$$\left|\gamma\sum_{j=0}^{K-1} q_i^{K-1-j}\xi_{i,j}\right| \le K\gamma R_K \tag{93}$$

for the Jacobian-variation remainder; this term is not required to vanish with σ_i.

Thus, the early-stopping spectral interpretation applies to the linearized component of finite refinement. The nonlinear dynamics deviate from this mode-wise structure through a bounded Jacobian-variation remainder.

C.7 Linearized Limit of Aggressive Surrogate Fitting

To compare finite refinement with aggressive surrogate fitting, consider the local linearized model obtained by setting

$$\xi_F(e) = 0. \tag{94}$$

Then

$$a_i^{(K)} = q_i^K a_i^{(0)} - h_K(\sigma_i) c_i. \tag{95}$$

For any fixed $\sigma_i > 0$,

$$q_i^K \to 0 \tag{96}$$

and

$$h_K(\sigma_i) \to \frac{1}{\sigma_i} \tag{97}$$

as $K \to \infty$. Therefore,

$$a_i^{(\infty)} = -\frac{c_i}{\sigma_i} = -\frac{u_i^\top d^\star}{\sigma_i}. \tag{98}$$

The same result follows from the stationary condition of

$$\mathcal{L}_{\text{lin}}(e) = \frac{1}{2} \left\| J_\star e + d^\star \right\|_2^2. \tag{99}$$

At a stationary point,

$$J_\star^\top \left(J_\star e + d^\star \right) = 0. \tag{100}$$

Projecting onto v_i gives

$$\sigma_i^2 v_i^\top e + \sigma_i u_i^\top d^\star = 0, \tag{101}$$

and hence

$$v_i^\top e = -\frac{u_i^\top d^\star}{\sigma_i}, \qquad \sigma_i > 0. \tag{102}$$

Accordingly, the relevant residual-induced bias is governed by

$$\frac{|u_i^\top d^\star|}{\sigma_i}, \tag{103}$$

rather than by σ_i alone. A small singular value produces a large potential amplification only when the effective residual has a non-negligible projection onto the corresponding measurement direction.

C.8 Zero-Singular-Value Directions

For a mode with

$$\sigma_i = 0, \tag{104}$$

we have

$$J_\star v_i = 0. \tag{105}$$

Therefore,

$$v_i^\top J_\star^\top J_\star e = 0 \tag{106}$$

and

$$v_i^\top J_\star^\top d^\star = (J_\star v_i)^\top d^\star = 0. \tag{107}$$

Hence the linearized measurement model provides no first-order correction along this direction. Under the full local dynamics,

$$a_i^{(k+1)} - a_i^{(k)} = -\gamma v_i^\top \xi_F(e_k). \tag{108}$$

Thus, changes along a zero-singular-value direction arise solely from local Jacobian variation at the level of the present analysis.