YOUR FEW-SHOT DIFFUSION MODELS SECRETLY LOCALIZE TO A SMALL SUBSPACE: AN EXPLANATORY FRAMEWORK IN THE LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot diffusion personalization poses a puzzle: how can a handful of images induce subject or attribute changes without relearning a high-dimensional distribution? We propose *SUDA-MoG*, a shared–unshared framework that models adaptation as selective recalibration of factor-wise Mixture-of-Gaussians laws in frozen pretrained latents. Shared factors inherit source statistics, while unshared factors are calibrated from target examples. Under an ideally decoupled shared–unshared factorization, we derive a generative-error bound whose target-sample term depends on the free active score parameters under stated regularity conditions. For coupled scores, we bound the initial shared gradient through shared-domain mismatch and cross-block coupling. We instantiate the framework with a semantically guided Greedy construction that combines active-block selection with conditional residual completion. Across evaluated tasks, SUDA-MoG achieves target alignment comparable to DreamBooth-LoRA and Custom Diffusion. Matched interventions on DreamBooth-LoRA outputs show that task-aligned coordinates carry part of the adapter's task response under an independent latent evaluator. These results characterize a task-relative functional footprint of few-shot adaptation, connecting inherited structure, selective calibration, and observable task effects.

1 INTRODUCTION

Diffusion models are standard backbones for high-quality generation (Ho et al., 2020; Song et al., 2020). A striking property is their few-shot adaptability: methods such as Textual Inversion, DreamBooth, Custom Diffusion, and Cones induce recognizable subject or attribute changes from only a few target images (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023; Liu et al., 2023). This is useful but mechanistically nontrivial. A tiny support set cannot plausibly specify a new high-dimensional latent distribution from scratch, so personalization must exploit structure already present in the pretrained model. We ask:

What exactly changes during few-shot adaptation?

Low-dimensional analyses study how intrinsic structure affects score learning and sampling (Chen et al., 2023; De Bortoli, 2022; Wang et al., 2024; Li & Yan, 2024; Liang et al., 2025). Within a latent-variable representation, a map from latent coordinates to observations and a probability law on those coordinates jointly determine the data distribution. This suggests two complementary perspectives: the *map side*, concerning changes in the mapping, and the *latent-distribution side*, concerning changes in the latent law. The few-shot analysis of Yang et al. (2024) exemplifies the former through target-specific encoder–decoder adjustments. At the implementation level, personalization methods specify token or weight updates (Gal et al., 2022; Ruiz et al., 2023; Xiang et al., 2023). These accounts leave implicit which latent statistics are inherited and which are recalibrated. We study this complementary latent-law question within a fixed pretrained representation. Multi-subspace MoG modeling (Yang et al., 2026) provides a convenient factor-wise interface for making these latent laws explicit. Figure 1 summarizes this perspective.

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.

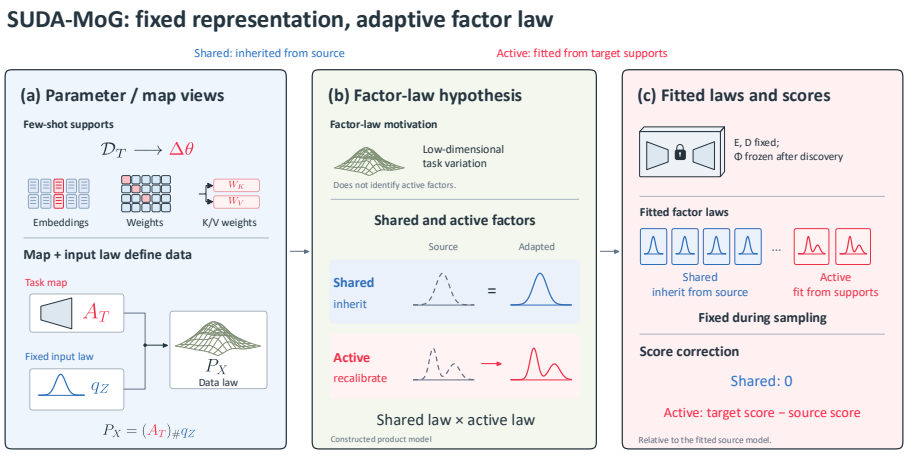


Figure 1: **Positioning SUDA-MoG.** Few-shot adaptation can be viewed through changes in a map or in its latent input law. SUDA-MoG fixes the representation, inherits shared factor laws, and calibrates active laws from target supports.

Two observations motivate a localized statistical account. In controlled AFHQ Dog→Cat transfer, rank 1 attains the highest mean CLIP-T among the tested DreamBooth-LoRA ranks (Table 1). Under full fine-tuning without an explicit low-rank constraint, a median of eight singular components captures 98% of the update energy across the analyzed weight matrices (Fig. 2). Together, these observations motivate the hypothesis that compact parameter updates may correspond to a task-specific functional footprint in latent statistics.

We propose *SUDA-MoG*, a shared-unshared framework for frozen-latent few-shot adaptation. SUDA-MoG models candidate latent factors with MoGs and assigns adaptation roles: shared factors preserve transferable source statistics, whereas unshared factors absorb target-specific updates. This complements the question of how many samples suffice with a structural one: which statistics must adapt, which can be inherited, and how imperfect decoupling allows active changes to influence shared factors.

We develop the theory around two questions: whether the active factor-law parameters can be estimated from few target examples, and when the shared factors remain stable. Under exact shared–unshared decoupling, the score objective separates across factor blocks, so the target-sample-dependent estimation term is governed by the active parameters while invariant shared factors remain stationary. Under imperfect decoupling, residual cross-block coupling induces a controlled shared-gradient response. A mean-only shrinkage estimator provides a direct finite-sample specialization.

Empirically, we instantiate this factor-map with the Greedy construction. With 32 supplied labeled CelebA supports per fit, the complete conditional map and its calibrated laws produce positive mean responses for three jointly targeted attributes. Across several few-shot transfer tasks, SUDA-MoG achieves target alignment comparable to DreamBooth-LoRA and Custom Diffusion. Joint-score experiments test the predicted coupling-dependent shared response. On matched Dog-to-Cat DreamBooth-LoRA outputs, replacing a single base task coordinate with its matched LoRA value recovers 42.7% of the full adapter's gain under an independently trained latent evaluator; recovery decreases under rotations toward shuffled-label directions.

Our contributions are threefold.

- **An explanation of few-shot adaptation.** We introduce SUDA-MoG, which assigns task-relative shared and unshared roles to latent MoG factors. Inherited shared laws and recalibrated unshared laws yield an explicit source–target score correction, connecting reuse of pretrained structure to the statistics estimated from target examples.
- **Learnability and stability under exact and imperfect decoupling.** We show that under an ideally decoupled shared–unshared factorization, the score objective separates across

factor blocks, so the target-sample-dependent estimation term is governed by the free active score parameters while invariant shared factors remain stationary at the source optimum. We further relax exact decoupling and bound the shared-gradient response by shared-domain mismatch and cross-block coupling to the active shift.

- **Instantiation and functional evidence for the latent adaptation footprint.** We instantiate the factor-map interface with a semantically guided Greedy construction. Controlled role, alignment, multi-attribute, and coupling experiments test the predicted selective behavior of the resulting factors, while matched DreamBooth-LoRA interventions show that task-aligned coordinates capture a task-aligned component of real adapter behavior.

2 Related Work

Low-dimensional diffusion theory. Ambient-dimensional score analyses (Lee et al., 2022; Chen et al., 2022; Koehler et al., 2022; Bruno et al., 2023) have been refined for linear, union-of-subspace, intrinsic-dimensional, and mixture structures (Chen et al., 2023; Wang et al., 2024; Li & Yan, 2024; Liang et al., 2025; Shah et al., 2023; Chen et al., 2024). SUDA-MoG builds on the factor-wise mixture statistics and closed-form scores of Yang et al. (2026), assigning shared–unshared adaptation roles in frozen pretrained latents.

Few-shot personalization. Textual Inversion, DreamBooth, Custom Diffusion, and Cones adapt diffusion models from a few images (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023; Liu et al., 2023). Parameter-efficient methods update tokens, attention, weights, or experts (Hu et al., 2022; Xiang et al., 2023; Farhadzadeh et al., 2025; Xiao et al., 2025; Zhuang et al., 2025). Yang et al. (2024) analyzes few-shot map adaptation under low-dimensional structure, while disentanglement and latent-control work seeks meaningful or editable directions (Higgins et al., 2017; Chen et al., 2016; Shen et al., 2020; Härkönen et al., 2020; Shen & Zhou, 2021; Kwon et al., 2022). SUDA-MoG identifies inherited and recalibrated factor laws and tests their functional roles through interventions.

3 Preliminaries

3.1 Score-Based Generative Models

Score-based models perturb $\boldsymbol{x}_0 \sim p_0$ into noisy marginals $\{p_t\}_{t\in[0,1]}$ (Ho et al., 2020; Song et al., 2020). We use variance-exploding notation $\boldsymbol{x}_t = \boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and score $s_t(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. For the forward SDE $d\boldsymbol{x} = f(\boldsymbol{x}, t)\, dt + g(t)\, d\boldsymbol{w}_t$, the reverse-time SDE is (Song et al., 2020)

$$d\boldsymbol{x} = \big(f(\boldsymbol{x}, t) - g(t)^2 s_t(\boldsymbol{x})\big)\, dt + g(t)\, d\bar{\boldsymbol{w}}_t. \tag{1}$$

Structured priors provide analytic score forms alongside neural approximations.

3.2 Mixture of Gaussians Prior

For $\boldsymbol{y} \in \mathbb{R}^d$, let $p_0(\boldsymbol{y}) = \sum_{j=1}^{J} w_j \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_j, \Sigma_j)$. After Gaussian perturbation, the marginal remains a MoG with $\widetilde{\Sigma}_{j,t} = \Sigma_j + \sigma(t)^2 I$ and score

$$s_t(\boldsymbol{y}_t) = \sum_{j=1}^{J} r_j(\boldsymbol{y}_t, t) \widetilde{\Sigma}_{j,t}^{-1} (\boldsymbol{\mu}_j - \boldsymbol{y}_t), \tag{2}$$

where the posterior responsibility is

$$r_j(\boldsymbol{y}_t, t) = \frac{w_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_j, \widetilde{\Sigma}_{j,t})}{\sum_{\ell=1}^{J} w_\ell \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_\ell, \widetilde{\Sigma}_{\ell,t})}.$$

For a factor map $\Phi : \mathcal{Z} \to \mathcal{M}_1 \times \cdots \times \mathcal{M}_K$, write $\mathbf{m} = \Phi(\boldsymbol{z}) = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$. The product model $p_t^{\text{model}}(\mathbf{m}_t) = \prod_k p_t^{(k)}(\mathbf{m}_{k,t})$ has the concatenated score

$$S_t(\mathbf{m}_t) = \big[s_t^{(1)}(\mathbf{m}_{1,t})^\top, \ldots, s_t^{(K)}(\mathbf{m}_{K,t})^\top\big]^\top, \tag{3}$$

with each block given by Eq. (2).

4 MULTI-SUBSPACE FACTORIZATION OF FROZEN LATENTS

Empirical motivation. We begin by examining the concentration of parameter updates and the organization of task-relevant information in pretrained latents. Rank 1 gives the highest mean CLIP-T among tested DreamBooth-LoRA ranks on AFHQ Dog→Cat (Table 1); a median of eight singular components captures 98% of full-fine-tuning update energy (Fig. 2). These diagnostics motivate a compact functional account in latent statistics.

LoRA rank	CLIP-T
1	$\mathbf{0.2806 \pm 0.0097}$
2	0.2789 ± 0.0082
16	0.2752 ± 0.0125
64	0.2757 ± 0.0104

Table 1: DreamBooth-LoRA rank sweep on AFHQ Dog→Cat. Mean CLIP-T with between-seed variation (±).

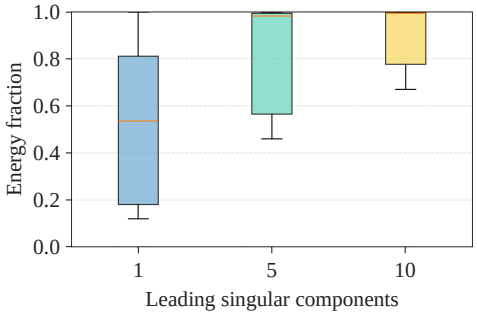


Figure 2: Full-fine-tuning update energy retained by the leading 1, 5, and 10 singular components across the analyzed weight matrices.

SUDA-MoG specifies a factor map, shared–unshared roles, and calibrated laws to model compact active changes with inherited complementary statistics.

4.1 FACTOR MAPS AND ADAPTATION ROLES

Let (E, D) be a frozen pretrained encoder–decoder and $z = E(x) \in \mathcal{Z}$. Source examples from P_S provide reusable statistics; target supports $\mathcal{D}_T = \{x_i\}_{i=1}^n$ from P_T provide calibration.

Definition 4.1 (Factor map and ideal decoupling)**.** A factor map is an invertible coordinate map onto its image,

$$\Phi : \mathcal{Z} \longrightarrow \mathcal{M} \subseteq \prod_{k=1}^{K} \mathcal{M}_k, \qquad \Phi(z) = (\mathbf{m}_1, \ldots, \mathbf{m}_K), \tag{4}$$

where $\mathcal{M}_k \subseteq \mathbb{R}^{d_k}$ and $\mathcal{M}$ contains the attainable tuples. In the smooth setting, Φ is a diffeomorphism onto the embedded coordinate manifold $\mathcal{M}$. For external observations $h_j : \mathcal{Z} \to \mathcal{A}_j$, ideal decoupling means that every admissible change of factor i, with the others fixed, leaves $h_j \circ \Phi^{-1}$ unchanged for all $j \neq i$.

We model factor laws in the ambient space $\mathcal{V} = \prod_k \mathbb{R}^{d_k}$. External observations are specified separately from coordinate extractors. The cross-factor sensitivity $\mathcal{S}_{i,j}(\Phi)$ measures the average local response of observation j to changes in factor i. Appendix A.6 defines this diagnostic and relates vanishing sensitivity to ideal decoupling under product-domain conditions. For each task, partition $\{1, \ldots, K\}$ into disjoint sets $\mathcal{I}_{\mathrm{sh}}$ and $\mathcal{I}_{\mathrm{un}}$: shared laws are inherited, and unshared (active) laws receive target updates.

4.2 GENERAL PARAMETER ADAPTATION

The factor-coordinate laws $Q_S = (\Phi \circ E)_{\#} P_S$ and $Q_T = (\Phi \circ E)_{\#} P_T$ are supported on $\mathcal{M}$. Approximate each source marginal $Q_{S,k}$ by

$$p_k^S(\mathbf{m}_k) = \sum_{j=1}^{J_k} w_{k,j}^S \mathcal{N}(\mathbf{m}_k; \mu_{k,j}^S, \Sigma_{k,j}^S), \qquad \eta_k = \{w_{k,j}, \mu_{k,j}, \Sigma_{k,j}\}_{j=1}^{J_k}. \tag{5}$$

The product $p^{S,\mathrm{model}}(\mathbf{m}) = \prod_k p_k^S(\mathbf{m}_k)$ on $\mathcal{V}$ is a tractable joint surrogate; its approximation error enters the generative bound. Adaptation follows

$$\eta_k^T = \begin{cases} \eta_k^S, & k \in \mathcal{I}_{\mathrm{sh}}, \\ \widehat{\eta}_k(\mathcal{D}_T), & k \in \mathcal{I}_{\mathrm{un}}. \end{cases} \tag{6}$$

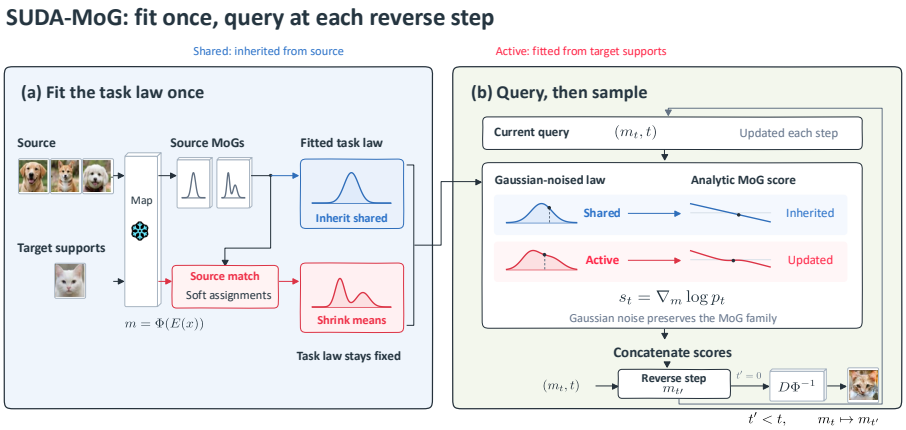


Figure 3: **Analytic sampling after task-law fitting.** Responsibility-weighted mean calibration illustrates one allowed update. With fitted laws fixed, each reverse step evaluates the factor-wise MoG scores at the current noisy coordinate and time, then concatenates them for the reverse update.

The adaptation state $\mathcal{A}_{S\to T} = (\Phi, \mathcal{I}_{\mathrm{sh}}, \mathcal{I}_{\mathrm{un}}, \eta^S, \eta^T)$ records intervention locations and inherited or recalibrated statistics. External observations h_j assess its functional footprint after recomposition.

Active weights, means, and covariances may vary subject to normalization and positive definiteness; fixing any gives a restricted instance. One calibration objective is denoising score matching:

$$\widehat{\eta}_{\mathrm{un}} \in \arg\min_{\eta_{\mathrm{un}} \in \Theta_{\mathrm{un}}} \widehat{\mathcal{L}}_{\mathrm{DSM}}(\eta^S_{\mathrm{sh}}, \eta_{\mathrm{un}}; \mathcal{D}_T). \tag{7}$$

4.3 Structured Scores and Generation

The calibrated target surrogate is

$$p_T^{\mathrm{SUDA}}(\mathbf{m}) = \prod_{k=1}^{K} p_k^T(\mathbf{m}_k), \qquad p_k^T(\mathbf{m}_k) = p(\mathbf{m}_k; \eta_k^T). \tag{8}$$

Independent Gaussian perturbations preserve the MoG marginals, yielding closed-form scores from Eq. (2):

$$S_t^{\mathrm{SUDA}}(\mathbf{m}_t) = \left[s_{1,t}^T(\mathbf{m}_{1,t})^\top, \ldots, s_{K,t}^T(\mathbf{m}_{K,t})^\top\right]^\top. \tag{9}$$

At the same noisy coordinate, ordering blocks by role gives the source–target correction

$$S_t^{\mathrm{SUDA}} - S_t^{S,\mathrm{model}} = \left[0_{\mathrm{sh}}^\top, (S_{\mathrm{un},t}^T - S_{\mathrm{un},t}^S)^\top\right]^\top. \tag{10}$$

Thus all allowed active updates affect only active score blocks in the surrogate.

Analytic reverse-time sampling uses S_t^{SUDA} with fixed fitted laws and evolving state $\mathbf{m}_t$ and scores (Fig. 3). Provided Φ^{-1} is defined on the sampler's output domain, $D \circ \Phi^{-1}$ recomposes and decodes the final state. For image-conditioned editing, a carrier supplies initial coordinates and retained complementary structure. Active updates use responsibility-weighted transport in the rank-one setting or score-difference optimization for joint attributes. Appendix F.3 gives these rules and the matched-coordinate interventions used to evaluate real adapters.

5 Theoretical Analysis

We analyze finite-sample calibration of active score parameters and the population stationarity and initial gradient response of shared parameters.

5.1 Learnability of Unshared Parameters

To quantify the sample cost of selective calibration, fix a factor map, role assignment, and source fits independently of the n i.i.d. target calibration examples. With concatenated factor-wise scores and disjoint parameter blocks,

$$\mathcal{L}_{\mathrm{DSM}}(\eta; P_T) = \sum_{k=1}^{K} \mathcal{L}_k(\eta_k; Q_{T,k}),$$

where $Q_{T,k}$ is the target factor marginal. Inheriting η_{sh}^S leaves an active class with d_{un} free coordinates among its weights, means, and covariances. We analyze finite-grid calibration in Eq. (7), assuming invariant shared marginals and source-stationary shared parameters. The analysis assumes finite-grid approximation, a Bernstein excess-loss condition, and curvature controlling score error. A positive noise floor, fixed stable sampler, and Lipschitz recomposition and decoding transfer it to image-space W_2.

Theorem 5.1 (Generative error under active-score calibration)**.** *Under the full hypotheses of Theorem C.8, let $\widehat{\eta}_{\mathrm{un}}$ minimize empirical DSM over the grid in Assumption C.5, set $\widehat{\eta} = (\eta_{\mathrm{sh}}^S, \widehat{\eta}_{\mathrm{un}})$, and let $G_{\widehat{\eta}}$ be the fixed sampler's decoded output law. For $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$W_2(G_{\widehat{\eta}}, P_T) \le E_{\mathrm{approx}} + E_{\mathrm{VAE}} + C_{\mathrm{score}} \sqrt{\frac{d_{\mathrm{un}} \log(An) + \log(2/\delta)}{n}}.$$

Here $A > 1$ is the grid-size constant and C_{score} depends on fixed regularity and sampling constants, independently of n, δ. The errors $E_{\mathrm{approx}}, E_{\mathrm{VAE}}$ capture fixed model–sampler discrepancy and encoder–decoder reconstruction. The explicit target-sample term depends on the d_{un} free active score parameters, quantifying the calibration cost of selective adaptation.

Example 1 (Responsibility-weighted mean calibration)**.** Mean-only calibration inherits weights, covariances, and shared statistics. For target coordinates $X_{k,i}$, fixed source responsibilities $r_{k,j}^S$, and $\tau > 0$, set

$$N_{k,j} = \sum_{i=1}^{n} r_{k,j}^S(X_{k,i}), \qquad \widehat{\mu}_{k,j} = \frac{\tau \mu_{k,j}^S + \sum_{i=1}^n r_{k,j}^S(X_{k,i}) X_{k,i}}{\tau + N_{k,j}}.$$

Its population estimand is $\mu_{k,j}^\circ = \mathbb{E}_T[r_{k,j}^S(X_k)X_k] / \mathbb{E}_T[r_{k,j}^S(X_k)]$ when the denominator is positive. With bounded coordinates and independent calibration, Corollary G.3 gives simultaneous $\|\widehat{\mu}_{k,j} - \mu_{k,j}^\circ\|_2 \le b_{k,j}$ with probability at least $1-\delta$, separating shrinkage bias and sampling error. If recomposition followed by decoding is defined and L-Lipschitz on the relevant supports, the decoded fitted product-mixture law satisfies

$$W_2(G_{\widehat{\mu}}, P_T) \le W_2(G^\circ, P_T) + L \left(\sum_{k \in \mathcal{I}_{\mathrm{un}}} \sum_j w_{k,j}^S b_{k,j}^2 \right)^{1/2}.$$

Here G° uses population means and inherits the other source statistics; its discrepancy captures model and inherited-law bias. Appendix G proves the bound; Corollary G.6 covers deterministic edits of an independently fixed carrier law.

5.2 Stability of Shared Parameters

All gradients use fixed free statistical-parameter coordinates. Write $\mathcal{L}_T(\eta) = \mathcal{L}_{\mathrm{DSM}}(\eta; P_T)$.

Factorized reference. If $Q_{S,k} = Q_{T,k}$ for each shared factor and shared source parameters are stationary, shared block losses agree across domains. Hence

$$\nabla_{\eta_{\mathrm{sh}}} \mathcal{L}_T(\eta_{\mathrm{sh},S}^\star, \eta_{\mathrm{un}}) = 0$$

for every admissible η_{un} (Theorem D.1): active updates preserve shared stationarity.

Coupled-score extension. We extend the additive model's block-diagonal parameter Hessian to a joint differentiable score with interacting blocks (Proposition E.2), using $\mathcal{L}_T$ for its population target DSM objective. Let $\eta_S^\star$ be the source optimum and $\eta_T^\star$ a target stationary point with both blocks free, in the same coordinates. Set

$$\delta_{\mathrm{sh}} = \eta^\star_{\mathrm{sh},S} - \eta^\star_{\mathrm{sh},T}, \qquad \delta_{\mathrm{un}} = \eta^\star_{\mathrm{un},S} - \eta^\star_{\mathrm{un},T}, \qquad \Delta_{\mathrm{un}} = \|\delta_{\mathrm{un}}\|_2.$$

Theorem 5.2 (Initial shared-gradient response)**.** *Suppose $\mathcal{L}_T$ is C^2 on a convex neighborhood containing the segment from $\eta_T^\star$ to $\eta_S^\star$, with $\nabla\mathcal{L}_T(\eta_T^\star) = 0$ and $\|\delta_{\mathrm{sh}}\|_2 \le \epsilon$. Throughout this segment, assume*

$$\|\mathbf{H}_{\mathrm{sh,sh}}(\eta)\|_{\mathrm{op}} \le L_{\mathrm{Hess}}, \qquad \|\mathbf{H}_{\mathrm{sh,un}}(\eta)\|_{\mathrm{op}} \le \gamma, \quad \mathbf{H} = \nabla^2\mathcal{L}_T.$$

Then

$$\|\nabla_{\eta_{\mathrm{sh}}}\mathcal{L}_T(\eta_S^\star)\|_2 \le \underbrace{L_{\mathrm{Hess}}\epsilon}_{\textit{shared self-shift}} + \underbrace{\gamma\Delta_{\mathrm{un}}}_{\textit{cross-block leakage}}.$$

For the target Hessian averaged along the source–target segment, stationarity gives

$$\nabla_{\eta_{\mathrm{sh}}}\mathcal{L}_T(\eta_S^\star) = \overline{\mathbf{H}}_{\mathrm{sh,sh}}\delta_{\mathrm{sh}} + \overline{\mathbf{H}}_{\mathrm{sh,un}}\delta_{\mathrm{un}}.$$

Bounding each block yields the result (Assumption E.1, Theorem E.3). The two terms quantify shared mismatch and active displacement through score coupling. Section 6.1 tests this parameter response and the observation-level sensitivity $\mathcal{S}_{i,j}(\Phi)$.

6 Experiments

We test factor construction, selective editing, and functional correspondence to real LoRA outputs. Appendices F and H specify discovery resources, calibration data, and execution operators.

6.1 Greedy Construction and Diagnostics

We first introduce a greedy instantiation aimed at finding the direction of greatest difference. Greedy uses task supervision to select a compact active block and conditional residuals to retain complementary information. For an invertible base chart $\Psi(z) = (m, x)$ and orthogonal basis $[B_r\ B_{r,\perp}]$ with r task-aligned directions, define

$$\begin{aligned} u &= B_r^\top x, & v &= B_{r,\perp}^\top x, \\ \widetilde{m} &= m - c(u), & \widetilde{v} &= \Lambda(u)^{-1}[v - a(u)]. \end{aligned} \tag{11}$$

Here c, a are fitted conditional locations and Λ is a positive diagonal scale. The map $\Phi(z) = (\widetilde{m}, u, \widetilde{v})$ retains conditional residuals during active replacement; recomposition restores locations and scales before applying Ψ^{-1}. Binary role tests use a rank-one moment implementation retaining stored channel moments and an orthogonal residual complement. Appendices H.4 and F.1 give both constructions. Feature-block probes support this role allocation: species information is more readable from normalized residuals, and color information from channel statistics (Appendix F.5).

Cross-factor sensitivity. We symmetrically perturb the active coordinate and recompute shared features after recomposition, following $\mathcal{S}_{i,j}$ in Lemma A.6. Source-normalized finite-difference responses are 0.0135, 0.0128, and 0.0172 for Dog→Cat, Glasses, and Young→Old (Appendix E.1).

Coupling-dependent shared gradients. Across 64 fitted two-dimensional joint-DSM models, calibrated coupling tracks held-out shared gradients with mean within-shift Spearman correlation 0.551. At matched active displacement, high-coupling cases have $2.570\times$ larger shared gradients than low-coupling cases, supporting Theorem 5.2's predicted response (Appendix H.6).

6.2 Few-Shot Editing and Selective Adaptation

Cross-method alignment. On Dog→Cat, factor-law calibration achieves CLIP-T close to DreamBooth-LoRA and Custom Diffusion (Table 2).

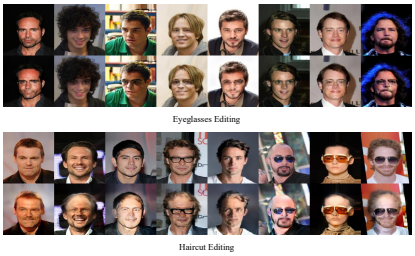
Table 2: **Dog→Cat comparison.** All methods use five target images, AFHQ dog source data (also used for SUDA-MoG map discovery), three seeds, and a common CLIP evaluator. Direction and role ablations are separate experiments.

Method	Adaptation object	CLIP-T ↑
Textual Inversion	Token	0.255
Custom Diffusion	Weights	0.264
DreamBooth-LoRA	LoRA weights	0.264
SUDA-MoG (Greedy)	MoG statistics	0.265

Three-attribute adaptation from 32 supports. From a single set of 32 labeled CelebA supports per fit, we construct the Greedy map in Eq. (11), including its active directions and conditional functions, and fit its factor laws. The supports are balanced over all 16 states of Smile, Bangs, Mouth-open, and Male. The three targets share a three-dimensional active block with single-Gaussian source and target laws; Male is untargeted. Across ten fixed support draws and forty common held-out carriers, all targets have positive mean CLIP-T responses (Fig. 4(c)). Mean paired-base DINO similarity is 0.9769 and LPIPS is 0.0386. The macro CLIP-T gain over shuffled labels is 0.00288; Appendix H.5 reports controls, image-change costs, and independent HOG evaluation.

Direction and role controls. Rank-one LDA exceeds PCA, random, and shuffled-label directions in CLIP-T and cat score; the controls' higher CLIP-I accompanies weaker category change (Fig. 4).

(a) Qualitative attribute editing



(b) Task-aligned factor extraction

Method	CLIP-T ↑	CLIP-I ↑	Cat score ↑
LDA (ours)	0.2817	0.7168	0.8986
PCA	0.2089	0.9038	0.1704
Random	0.2075	0.9380	0.1348
Shuffled LDA	0.2064	0.9429	0.1320

(c) 32-support CelebA editing

Attribute	ΔCLIP-T ↑
Smile	$+0.00283 \pm 0.00116$
Bangs	$+0.00246 \pm 0.00111$
Mouth-open	$+0.00160 \pm 0.00122$
Male (untargeted)	-0.00002 ± 0.00057

Figure 4: **Attribute editing and controls.** (a) Eyeglasses and haircut edits. (b) Dog→Cat direction controls: LDA increases target response; controls retain more source similarity. (c) Joint editing from 32 supports, with Male untargeted. Mean ± sample standard deviation over ten support draws; each draw averages paired gains on forty common held-out carriers. Appendix H.5 reports full CLIP-T and HOG results.

Unshared-only (U) updates exceed shared-only (S) updates in CLIP-T and target accuracy across all three tasks (Table 3). Relative to updating all factors, U gives equal or higher accuracy and higher LAB similarity; All has slightly higher CLIP-T on Dog→Cat and Glasses. Appendix H.1 provides further age, style, and color examples.

6.3 Functional Validation on Real LoRA Outputs

To test correspondence with a separately trained adapter, a LoRA trained on ten target supports and its SD1.5 base generate matched RGB outputs. A frozen DC-AE encoder supplies external analysis latents, and Greedy's frozen rank-one moment setting gives $\Phi(z) = (s, u, q_\perp)$. Replacing only the base task coordinate u with its matched LoRA value, while retaining base moments and residual complement, measures its contribution under an independently trained latent evaluator g (held-out AUC 0.9979). For paired base and LoRA latents z_i^0, z_i^L and the coordinate replacement $z_i^{u\leftarrow L}$, recovery is the ratio of aggregate gains:

$$R_u = \frac{\sum_i [g(z_i^{u\leftarrow L}) - g(z_i^0)]}{\sum_i [g(z_i^L) - g(z_i^0)]}. \tag{12}$$

Table 3: **Role-specific editing.** U updates unshared factors, S shared factors, and All both. MoG component counts are 3 (Dog→Cat), 1 (Glasses), and 5 (Young→Old).

Task	Update	CLIP-T ↑	Target acc. ↑	LAB sim. ↑
Dog→Cat	U	0.2669	1.000	0.831
	S	0.2277	0.333	0.937
	All	0.2716	1.000	0.821
Glasses	U	0.2745	0.567	0.823
	S	0.2263	0.033	0.915
	All	0.2774	0.533	0.810
Young→Old	U	0.2459	0.767	0.795
	S	0.2243	0.167	0.912
	All	0.2424	0.600	0.782

Across 240 pairs, ranks 1 and 16, and three seeds, mean gains are 0.04980 for coordinate replacement and 0.11670 for full LoRA. Unrounded aggregates give 42.68% recovery, with a descriptive 95% crossed-bootstrap interval of $[29.70, 60.80]\%$. This measures the coordinate's contribution under g in the external DC-AE representation. It exceeds all 99 shuffled-label controls ($p = 0.01$) and declines under rotation toward shuffled directions at fixed LoRA endpoints and evaluator (Table 4). Appendix H.2 details the interventions, data splits, and ratio definition.

Table 4: **Task-coordinate recovery under rotation (Dog→Cat).** Ratios of aggregate coordinate gains to full LoRA gain, with fixed LoRA endpoints and independent latent evaluator.

Mean alignment with original axis	1.00	0.73	0.43	0.07
Task-coordinate recovery	42.7%	22.8%	8.0%	0.5%

7 Conclusion

We presented SUDA-MoG as a model of selective few-shot diffusion adaptation in frozen pretrained latents, making explicit which source statistics are inherited and which are estimated from target examples. Under the stated assumptions, the explicit target-sample term in our learnability bound counts only free active score parameters. The stability analysis establishes shared stationarity in the factorized reference model and bounds the initial shared gradient under domain mismatch and score coupling. Role ablations and multi-attribute experiments demonstrate task-directed effects of the assigned updates, while matched DreamBooth-LoRA interventions identify a task-aligned component of the adapter response under an independently trained latent evaluator. Together, these results characterize a task-relative functional footprint of adaptation, connecting selective statistical estimation, factor-specific control, and observable task effects.

Limitations. The Greedy instantiation relies on task-dependent, linearly readable attribute information in pretrained latents. Discovery data and the candidate direction class constrain the active coordinates and rank. The estimation guarantees condition on the map, roles, and source fits being fixed independently of target calibration; they do not cover the full factor-discovery process. Finite MoG capacity may limit the representation of fine-grained structure and rare modes, while the product-law approximation omits residual dependencies between factor blocks.

Future work. Future work should develop more systematic factor discovery and calibration while retaining explicit shared–unshared roles and intervention-based validation. Richer within-factor densities and hierarchical dependence models could extend the current product-MoG approximation. Broader benchmarks, additional pretrained backbones, and multi-concept customization would test the transferability of the statistical account and the scope of its functional correspondence.

AI USE STATEMENT

Generative AI tools were used to assist with refining the conceptual presentation of the shared–unshared adaptation framework, checking and clarifying mathematical statements and proof exposition, and providing feedback on experimental design and interpretation. Generative AI was also used for manuscript editing, including improving organization, clarity, and readability, and for summarizing and comparing relevant literature.

All theoretical claims, derivations, experimental protocols, numerical results, and citations were reviewed and verified by the authors. The authors made the final decisions regarding the formulation of the method, theoretical statements, experimental design, and interpretation of the results, and take full responsibility for the content of this work.

REFERENCES

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Farzad Farhadzadeh, Debasmit Das, Shubhankar Borse, and Fatih Porikli. Zero-shot adaptation of parameter-efficient fine-tuning in diffusion models. *arXiv preprint arXiv:2506.04244*, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1532–1540, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024.

Chendong Xiang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. A closer look at parameter-efficient tuning in diffusion models. *arXiv preprint arXiv:2303.18181*, 2023.

Jinqi Xiao, Qing Yan, Liming Jiang, Zichuan Liu, Hao Kang, Shen Sang, Tiancheng Zhi, Jing Liu, Cheng Yang, Xin Lu, et al. Instructmole: Instruction-guided mixture of low-rank experts for multi-conditional image generation. *arXiv preprint arXiv:2512.21788*, 2025.

Ruofeng Yang, Bo Jiang, Cheng Chen, Baoxiang Wang, Shuai Li, et al. Few-shot diffusion models escape the curse of dimensionality. *Advances in Neural Information Processing Systems*, 37: 68528–68558, 2024.

Ruofeng Yang, Yongcan Li, Bo Jiang, Cheng Chen, and Shuai Li. Multi-subspace multi-modal modeling for diffusion models: Estimation, convergence and mixture of experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MPWIM6rxxU.

Shaobin Zhuang, Yiwei Guo, Yanbo Ding, Kunchang Li, Xinyuan Chen, Yaohui Wang, Fangyikang Wang, Ying Zhang, Chen Li, and Yali Wang. Timestep master: Asymmetrical mixture of timestep lora experts for versatile and efficient diffusion models in vision, 2025. URL https://arxiv.org/abs/2503.07416.

A MATHEMATICAL FOUNDATIONS OF DECOUPLING MAPPINGS

Assumption A.1 (Regularity)**.** All manifolds $\mathcal{Z}, \mathcal{M}_i$ are smooth and embedded in Euclidean spaces. The maps ϕ_i and Φ are C^2. The external observation maps h_i introduced below are C^1. The prior $p(\boldsymbol{z})$ has finite second moments.

Definition A.2 (Decomposition Mapping)**.** Let $\mathcal{Z}$ denote the original latent representation space (e.g., the output of a pretrained VAE encoder). We define a *decomposition mapping* Φ as:

$$\Phi : \mathcal{Z} \longrightarrow \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K$$
$$\Phi(\boldsymbol{z}) = \big(\phi_1(\boldsymbol{z}), \phi_2(\boldsymbol{z}), \ldots, \phi_K(\boldsymbol{z})\big)$$

where each $\phi_i : \mathcal{Z} \to \mathcal{M}_i$ is a *extraction operator*.

For the product-domain results below, we use the following stronger specialization of the factor map in Definition 4.1.

Definition A.3 (Isomorphic Decomposition Mapping)**.** A decomposition mapping $\Phi : \mathcal{Z} \to \prod_{i=1}^{K} \mathcal{M}_i$ is called an *isomorphic mapping* if Φ is a **diffeomorphism** (i.e., a bijective mapping with $\Phi \in C^1$ and $\Phi^{-1} \in C^1$). This ensures that the latent space $\mathcal{Z}$ and the product of statistical coordinate sub-manifolds are informationally equivalent:

$$\boldsymbol{z} = \Phi^{-1}\big(\phi_1(\boldsymbol{z}), \phi_2(\boldsymbol{z}), \ldots, \phi_K(\boldsymbol{z})\big), \quad \forall \boldsymbol{z} \in \mathcal{Z}$$

The inverse reconstructs latents from factor coordinates. The following observation functions characterize the effects of controlled coordinate interventions, and the statistical model specifies their factor laws.

Definition A.4 (External observation functions)**.** For each intended factor, let $h_j : \mathcal{Z} \to \mathcal{A}_j$ denote an external observation function or probe. It may be a physical statistic computed in latent space, a linear attribute probe, or a decoder-composed attribute evaluator. Importantly, h_j is not the coordinate extraction map ϕ_j used inside Φ; otherwise $h_j \circ \Phi^{-1}$ would reduce to the trivial coordinate projection $\mathbf{m} \mapsto \mathbf{m}_j$.

Definition A.5 (Decoupling Mapping)**.** An isomorphic decomposition mapping Φ is a *decoupling mapping* with respect to observation functions $\{h_j\}_{j=1}^{K}$ if changing factor i does not alter the observation associated with any other factor j. Formally, for any $i \neq j$ and any local perturbation $\Delta \in \mathcal{M}_i$,

$$h_j\big(\Phi^{-1}(m_1, \ldots, m_i + \Delta, \ldots, m_K)\big) = h_j\big(\Phi^{-1}(m_1, \ldots, m_i, \ldots, m_K)\big)\,.$$

Lemma A.6 (Criterion for Decoupled Representations)**.** *Let Φ be an isomorphic decomposition mapping onto a product of convex coordinate domains. Assume that the factor law has full support on this domain and that $G_j = h_j \circ \Phi^{-1}$ has continuous first derivatives . Then Φ is decoupled with respect to $\{h_j\}$ if and only if the **Cross-Block Sensitivity** vanishes:*

$$\mathcal{S}_{i,j}(\Phi) := \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\|\nabla_{\mathbf{m}_i} G_j(\mathbf{m})\|_F\right] = 0, \qquad \forall i \neq j.$$

Proof. By Definition A.5, decoupling requires $G_j(m_1, \ldots, m_i + \Delta, \ldots, m_K) = G_j(\mathbf{m})$ for all local perturbations $\Delta \in \mathcal{M}_i$ and all $i \neq j$. Differentiating this identity with respect to $\mathbf{m}_i$ gives $\nabla_{\mathbf{m}_i} G_j(\mathbf{m}) = \mathbf{0}$ almost everywhere. Since the Frobenius norm is nonnegative, the expectation defining $\mathcal{S}_{i,j}$ vanishes if and only if these off-block Jacobians vanish almost everywhere. Continuity and full support extend this equality to every point: a nonzero derivative would remain nonzero on an open set of positive probability. Integrating the zero derivative along each coordinate segment in the convex domain gives the decoupling identity. □

Remark A.7 (Task information and controlled response). For a linear probe with accuracy $\alpha_{i,j}^{\text{linear}}$ and chance or majority baseline $\mathcal{B}_j$, the statistic $\mathcal{C}_{i,j} = |\alpha_{i,j}^{\text{linear}} - \mathcal{B}_j|$ measures readable task information under the observed factor law. The sensitivity $\mathcal{S}_{i,j}$ measures the response of h_j to a change in factor i, with all other coordinates held fixed. These diagnostics guide factor discovery and intervention-based role validation, respectively.

Example 2. In the moment-based setting, the initial extraction records channel statistics , and subsequent operators act on normalized residual coordinates. Their task roles are assessed through external observations and controlled interventions.

B Problem Setup

B.1 Problem Formulation

Definition B.1 (Few-Shot Adaptation Task)**.** Let $\mathcal{X} \subset \mathbb{R}^D$ be the observation space. We assume access to a source generative model G_S trained on distribution P_S. Our goal is to adapt G_S to a target distribution P_T, given only a few-shot dataset $\mathcal{D}_T = \{\boldsymbol{x}_i\}_{i=1}^n$ (where $n \ll 100$).

B.2 Structural Modeling via Geometric Decoupling

To tackle the data scarcity in the few-shot regime, we leverage the geometric inductive bias of the pretrained generator. Based on the **Decoupling Mapping** Φ defined in Appendix A.5, we transform the entangled latent code $\boldsymbol{z}$ into statistical coordinates $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_k) = \Phi(\boldsymbol{z})$.

This approximate factor separability motivates our **statistical independence assumption** in the probabilistic model:

$$p(\mathbf{m}) = \prod_{\ell=1}^{k} p(\mathbf{m}_\ell). \tag{13}$$

This defines a product model with a sum of factor-wise parameter counts; departures of the encoded data from this model contribute to approximation error .

Under the factorization hypothesis, we assume the prior factorizes across decoupled coordinates,

$$p_0(\mathbf{m}_1, \ldots, \mathbf{m}_k) = \prod_{\ell=1}^{k} p_{\ell,0}(\mathbf{m}_\ell), \tag{14}$$

and each factor prior is modeled by a Mixture of Gaussians (MoG)

$$p_{\ell,0}(\mathbf{m}_\ell) = \sum_{k=1}^{K_\ell} w_{\ell,k}\, \mathcal{N}(\mathbf{m}_\ell; \boldsymbol{\mu}_{\ell,k}, \Sigma_{\ell,k}), \tag{15}$$

where ℓ indexes the subspace and k indexes the component.

We interpret each component as a statistical *expert*. During adaptation, we freeze the mapping Φ (preserving the decoupling structure) and only update the expert parameters $(w_{\ell,k}, \boldsymbol{\mu}_{\ell,k}, \Sigma_{\ell,k})$.

B.3 Time-Dependent Distributions and Closed-Form Scores

Under the forward diffusion process with noise schedule σ_t^2, each factor follows $\mathbf{m}_{\ell,t} = \mathbf{m}_{\ell,0} + \sigma_t\, \boldsymbol{\epsilon}$ and the time-marginal remains a MoG:

$$p_{\ell,t}(\mathbf{m}_{\ell,t}) = \sum_{k=1}^{K_\ell} w_{\ell,k}\, \mathcal{N}(\mathbf{m}_{\ell,t}; \boldsymbol{\mu}_{\ell,k}, \Sigma_{\ell,k} + \sigma_t^2 I), \tag{16}$$

where ℓ indexes the subspace.

The corresponding score admits a closed form:

$$s_{\ell,t}(\mathbf{m}_{\ell,t}; \eta_\ell) = \nabla_{\mathbf{m}_{\ell,t}} \log p_{\ell,t}(\mathbf{m}_{\ell,t}) = \sum_{k=1}^{K_\ell} r_{\ell,k}(\mathbf{m}_{\ell,t}, t)\, \widetilde{\Sigma}_{\ell,k,t}^{-1}(\boldsymbol{\mu}_{\ell,k} - \mathbf{m}_{\ell,t}), \tag{17}$$

where $\widetilde{\Sigma}_{\ell,k,t} = \Sigma_{\ell,k} + \sigma_t^2 I$ and

$$r_{\ell,k}(\mathbf{m}_{\ell,t}, t) = \frac{w_{\ell,k}\, \mathcal{N}(\mathbf{m}_{\ell,t}; \boldsymbol{\mu}_{\ell,k}, \widetilde{\Sigma}_{\ell,k,t})}{\sum_{k'=1}^{K_\ell} w_{\ell,k'}\, \mathcal{N}(\mathbf{m}_{\ell,t}; \boldsymbol{\mu}_{\ell,k'}, \widetilde{\Sigma}_{\ell,k',t})}. \tag{18}$$

The final generative distribution is denoted G_η, representing the distribution obtained by reverse diffusion in factor space $(\mathbf{m}_{1,t}, \ldots, \mathbf{m}_{k,t})$, mapping back via Φ^{-1}, and decoding with g_ψ.

B.4 Optimization Objective

With Φ fixed, the theoretical reference estimator optimizes the statistical parameters $\eta = \{\eta_1, \ldots, \eta_k\}$. We define its population DSM objective and factor-wise empirical form below; the practical shrinkage estimator is specified in Appendix F .

Global Denoising Score Matching in Statistical Coordinate Space First, we define the adaptation objective on the global statistical coordinates. Let $\mathbf{m}_0 = \Phi(E(\boldsymbol{x}))$ be the latent statistical code of a target sample $\boldsymbol{x} \in \mathcal{D}_T$, where E is the inversion encoder. We perturb $\mathbf{m}_0$ via the forward diffusion process to obtain $\mathbf{m}_t \sim p_{t|0}(\cdot|\mathbf{m}_0) = \mathcal{N}(\mathbf{m}_0, \sigma_t^2 \mathbf{I})$.

The global Denoising Score Matching (DSM) objective seeks to minimize the Fisher divergence between the model score $S_\eta(\mathbf{m}_t, t)$ and the ground-truth perturbation kernel score:

$$\mathcal{L}_{\text{Global}}(\eta) = \mathbb{E}_{t\sim\mathcal{U}[0,T]} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_T} \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \Big[\big\| S_\eta(\mathbf{m}_t, t) - \nabla_{\mathbf{m}_t} \log p_{t|0}(\mathbf{m}_t|\mathbf{m}_0) \big\|_2^2 \Big], \tag{19}$$

where $\mathbf{m}_t = \mathbf{m}_0 + \sigma_t \boldsymbol{\epsilon}$.

Factorized Objective via Subspace Decoupling Independent Gaussian perturbations give the conditional kernel $p_{t|0}(\mathbf{m}_t \mid \mathbf{m}_0) = \prod_\ell p_{t|0}(\mathbf{m}_{\ell,t} \mid \mathbf{m}_{\ell,0})$. The product MoG model has the concatenated factor-wise score

$$S_\eta(\mathbf{m}_t, t) = \big[s_1(\mathbf{m}_{1,t}, t; \eta_1)^\top, \ldots, s_k(\mathbf{m}_{k,t}, t; \eta_k)^\top \big]^\top. \tag{20}$$

By substituting this decomposition into the global objective, the squared L_2 norm splits into a summation of subspace losses. The optimization problem thus decouples into m parallel subproblems:

$$\mathcal{L}_{\text{DSM}}(\eta) = \sum_{\ell=1}^{k} \mathcal{L}_\ell(\eta_\ell), \tag{21}$$

where each subspace objective is defined as:

$$\mathcal{L}_\ell(\eta_\ell) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{t,\epsilon_\ell} \Big[\big\| s_{\ell,t}(\mathbf{m}_{\ell,t}, t; \eta_\ell) - \nabla_{\mathbf{m}_{\ell,t}} \log p_{t|0}(\mathbf{m}_{\ell,t}|\mathbf{m}_{\ell,0}^{(i)}) \big\|_2^2 \Big]. \tag{22}$$

Here, $\mathbf{m}_{\ell,0}^{(i)}$ denotes the ℓ-th statistical component of the i-th target sample, and the target score simplifies to $\nabla_{\mathbf{m}_{\ell,t}} \log p_{t|0} = -(\mathbf{m}_{\ell,t} - \mathbf{m}_{\ell,0})/\sigma_t^2$.

For this disjoint-block score parameterization, each summand depends only on η_ℓ, hence $\nabla^2_{\eta_i,\eta_j} \mathcal{L}_{\text{DSM}} = 0$ for $i \neq j$, even for correlated target factors. The joint-score extension is defined separately in Appendix E. The default practical update recalibrates means with inherited weights and covariances . In the analysis below, η is treated as an index of the induced score field. Equivalent MoG parameterizations that induce the same score field, such as permutations of mixture components, are identified.

C Generative Error Analysis

In this section, we provide a rigorous bound on the generative error of the adapted latent-diffusion model, bridging score-risk estimation with stochastic differential equation (SDE) stability analysis.

The proof avoids global recovery of a unique MoG parameter vector. Instead, it controls the score field induced by the adapted unshared statistics, then uses reverse-SDE stability to propagate this score-level error to the generated distribution.

C.1 Assumptions and Notation

We collect the regularity and statistical assumptions used throughout.

Assumption C.1 (Factorized Model and Parameter Regularity)**.** Let the latent space decompose into m subspaces, where each marginal $p_{\ell,t}$ is a Mixture of Gaussians. We assume the parameter vector $\eta = \{w_{\ell,k}, \boldsymbol{\mu}_{\ell,k}, \Sigma_{\ell,k}\}_{\ell,k}$ lies in a **compact admissible set** Θ. Specifically, there exist uniform constants $R, \lambda_{\min}, w_{\min} > 0$ such that for all subspaces ℓ and components k:

$$\|\boldsymbol{\mu}_{\ell,k}\| \le R, \quad \Sigma_{\ell,k} \succeq \lambda_{\min} I, \quad \text{and} \quad w_{\ell,k} \ge w_{\min}. \tag{23}$$

Assumption C.2 (Pipeline Stability and Objective Regularity)**.** Fix a common terminal initialization law, diffusion coefficients, reverse-integration interval, and integration rule. Write $Q_\eta = G_\eta^{\mathrm{lat}}$ for the resulting factor-space output law and $G_\eta = (D \circ \Phi^{-1})_{\#} Q_\eta$ for its decoded image law. All comparisons use this same sampler specification.

The decoder–recomposition map $g_\psi \circ \Phi^{-1}$ is L_g-Lipschitz . The DSM time interval has positive noise floor $\sigma(t) \ge \sigma_{\min} > 0$; the same interval is used for the score metric and sampler comparison. The reverse-SDE map admits $C_{\mathrm{stab}}(T)$ such that for the score fields under consideration:

$$W_2(Q_s, Q_{\tilde{s}}) \le C_{\mathrm{stab}}(T) \sup_{t\in[0,T]} \mathbb{E}_{\mathbf{m}_t \sim P_{T,t}} \|s_t(\mathbf{m}_t) - \tilde{s}_t(\mathbf{m}_t)\|_2. \tag{24}$$

Assumption C.3 (Shared-Unshared Subspace Factorization)**.** The statistical coordinates $\mathbf{m}$ decompose into two disjoint sets of subspaces based on index partition $\mathcal{I} = \mathcal{I}_{\mathrm{sh}} \cup \mathcal{I}_{\mathrm{un}}$. The global prior factorizes as:

$$p(\mathbf{m}) = \left(\prod_{\ell \in \mathcal{I}_{\mathrm{sh}}} p_\ell(\mathbf{m}_\ell; \eta_\ell) \right) \left(\prod_{\ell \in \mathcal{I}_{\mathrm{un}}} p_\ell(\mathbf{m}_\ell; \eta_\ell) \right). \tag{25}$$

Crucially, we assume the marginal distributions on the shared subspaces are domain-invariant, i.e., $P_S(\mathbf{m}_\ell) = P_T(\mathbf{m}_\ell)$ for all $\ell \in \mathcal{I}_{\mathrm{sh}}$, and the pretrained source parameters $\{\eta_\ell^*\}_{\ell \in \mathcal{I}_{\mathrm{sh}}}$ are first-order stationary points of the population DSM objective on P_S in free parameter coordinates .

Assumption C.4 (Score-Level Estimation Regularity)**.** Let $\mathcal{S}_{\mathrm{un}} = \{S_\eta : \eta_{\mathrm{sh}} = \eta_{\mathrm{sh}}^S, \eta_{\mathrm{un}} \in \Theta_{\mathrm{un}}\}$ be the class of score fields obtained by varying only the unshared MoG statistics. Define the score metric

$$\rho(\eta, \eta') := \sup_{t\in[0,T]} \left(\mathbb{E}_{\mathbf{m}_t \sim P_{T,t}} \|S_\eta(\mathbf{m}_t, t) - S_{\eta'}(\mathbf{m}_t, t)\|_2^2 \right)^{1/2}.$$

We assume that the induced score class has finite parametric complexity

$$\log \mathcal{N}(\varepsilon, \mathcal{S}_{\mathrm{un}}, \rho) \le d_{\mathrm{un}} \log(A/\varepsilon),$$

where $d_{\mathrm{un}} = \dim(\eta_{\mathrm{un}})$, and that the population DSM risk satisfies a local Bernstein, or quadratic-growth, condition around the best unshared score field

$$\eta^\dagger \in \arg\min_{\eta_{\mathrm{un}} \in \Theta_{\mathrm{un}}} \mathcal{L}_T(\eta_{\mathrm{sh}}^S, \eta_{\mathrm{un}}) :$$

$$\mathcal{L}_T(\eta) - \mathcal{L}_T(\eta^\dagger) \ge c_{\mathrm{curv}}\, \rho^2(\eta, \eta^\dagger)$$

for all η in the candidate grid defined below. This condition directly controls the supremum-time metric by excess risk.

Assumption C.5 (Finite-grid estimation conditions)**.** Condition on the factorization, inherited parameters, and role assignment. Let $\ell_\eta(X) = \mathbb{E}_{t,\epsilon} \|S_\eta(\mathbf{m}_t, t) + \epsilon/\sigma_t\|^2$ and $f_\eta = \ell_\eta - \ell_{\eta^\dagger}$, with $e_\eta = \mathbb{E} f_\eta \ge 0$. For constants independent of n:

1. A deterministic grid $\Theta_n \subset \Theta_{\mathrm{un}}$ has $M_n \le (An)^{d_{\mathrm{un}}}$ points, $A > 1$, and contains η_0 with $e_{\eta_0} \le a_n = C_{\mathrm{app}}/n$.
2. For every grid point, the centered excess loss obeys

$$\log \mathbb{E} e^{\lambda(f_\eta - e_\eta)} \le \frac{B e_\eta \lambda^2}{2(1 - b|\lambda|)}, \qquad |\lambda| < 1/b,$$

for fixed $B, b > 0$ and both signs of λ.

Here d_{un} counts the free parameters in the constrained score class, with fixed mixture labels and free simplex coordinates. If only means vary, it counts only mean coordinates. A compact parameter net of radius $1/n$ and a population-risk Lipschitz bound provide the first condition. The moment condition and score curvature are additional statistical assumptions beyond compact MoG parameters.

C.2 KEY LEMMAS

Lemma C.6 (Bounds on Score Norm and Gradient)**.** *Under Assumption C.1, let $s_{\ell,t}(\mathbf{m}_\ell)$ be the score of the ℓ-th subspace. There exist constants C_0, C_1 (depending on $R, \lambda_{\min}$) such that for all $t \geq 0$ and $\mathbf{m}_\ell$:*

$$\|s_{\ell,t}(\mathbf{m}_\ell)\| \leq C_0(1 + \|\mathbf{m}_\ell\|), \qquad \|\nabla_{\mathbf{m}_\ell} s_{\ell,t}(\mathbf{m}_\ell)\|_{\text{op}} \leq C_1(1 + \|\mathbf{m}_\ell\|^2).$$

Proof. To simplify notation within this proof, we fix the subspace index ℓ and drop it where context is clear (e.g., writing $\mathbf{m}$ for $\mathbf{m}_\ell$). Let the time-dependent covariance be $\widetilde{\Sigma}_{\ell,k,t} = \Sigma_{\ell,k} + \sigma_t^2 I$. We define the auxiliary vector function:

$$\mathbf{u}_{\ell,k,t}(\mathbf{m}) = -\widetilde{\Sigma}_{\ell,k,t}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\ell,k}),$$

and let $N_{\ell,k,t}(\mathbf{m}) = \mathcal{N}(\mathbf{m}; \boldsymbol{\mu}_{\ell,k}, \widetilde{\Sigma}_{\ell,k,t})$. The mixture score for the subspace is given by:

$$s_{\ell,t}(\mathbf{m}) = \frac{\sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}(\mathbf{m}) \mathbf{u}_{\ell,k,t}(\mathbf{m})}{\sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}(\mathbf{m})}.$$

We assume Assumption C.1 holds.

(i) Bound on $\|s_{\ell,t}(\mathbf{m})\|$. The score can be rewritten as a convex combination:

$$s_{\ell,t}(\mathbf{m}) = \sum_{k=1}^{K_\ell} r_{\ell,k}(\mathbf{m}) \mathbf{u}_{\ell,k,t}(\mathbf{m}), \quad \text{where} \quad r_{\ell,k}(\mathbf{m}) := \frac{w_{\ell,k} N_{\ell,k,t}(\mathbf{m})}{\sum_{i=1}^{K_\ell} w_{\ell,i} N_{\ell,i,t}(\mathbf{m})}.$$

Here $r_{\ell,k}(\mathbf{m}) \geq 0$ and $\sum_k r_{\ell,k}(\mathbf{m}) = 1$. By the convexity of the norm:

$$\|s_{\ell,t}(\mathbf{m})\| = \left\| \sum_{k=1}^{K_\ell} r_{\ell,k}(\mathbf{m}) \mathbf{u}_{\ell,k,t}(\mathbf{m}) \right\| \leq \sum_{k=1}^{K_\ell} r_{\ell,k}(\mathbf{m}) \|\mathbf{u}_{\ell,k,t}(\mathbf{m})\| \leq \max_{1 \leq k \leq K_\ell} \|\mathbf{u}_{\ell,k,t}(\mathbf{m})\|.$$

For each component k, we have:

$$\|\mathbf{u}_{\ell,k,t}(\mathbf{m})\| = \|\widetilde{\Sigma}_{\ell,k,t}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\ell,k})\| \leq \|\widetilde{\Sigma}_{\ell,k,t}^{-1}\|_{\text{op}} (\|\mathbf{m}\| + \|\boldsymbol{\mu}_{\ell,k}\|).$$

Since $\widetilde{\Sigma}_{\ell,k,t} = \Sigma_{\ell,k} + \sigma_t^2 I \succeq \Sigma_{\ell,k} \succeq \lambda_{\min} I$, we get $\|\widetilde{\Sigma}_{\ell,k,t}^{-1}\|_{\text{op}} \leq 1/\lambda_{\min}$. Using the parameter bound $\|\boldsymbol{\mu}_{\ell,k}\| \leq R$:

$$\|\mathbf{u}_{\ell,k,t}(\mathbf{m})\| \leq \frac{1}{\lambda_{\min}} (\|\mathbf{m}\| + R).$$

Taking the maximum over k yields:

$$\|s_{\ell,t}(\mathbf{m})\| \leq \frac{1}{\lambda_{\min}} (\|\mathbf{m}\| + R).$$

Defining $C_0 := \frac{1+R}{\lambda_{\min}}$, we obtain:

$$\|s_{\ell,t}(\mathbf{m})\| \leq \frac{1+R}{\lambda_{\min}} (1 + \|\mathbf{m}\|) = C_0(1 + \|\mathbf{m}\|),$$

which proves the first bound.

(ii) Bound on $\|\nabla_{\mathbf{m}} s_{\ell,t}(\mathbf{m})\|_{\text{op}}$.

Let $A(\mathbf{m}) := \sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}(\mathbf{m}) \mathbf{u}_{\ell,k,t}(\mathbf{m})$ and $B(\mathbf{m}) := \sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}(\mathbf{m})$, so that $s_{\ell,t} = A/B$. We compute derivatives with respect to $\mathbf{m}$. Note that:

$$\nabla N_{\ell,k,t}(\mathbf{m}) = N_{\ell,k,t}(\mathbf{m}) \mathbf{u}_{\ell,k,t}(\mathbf{m}), \qquad \nabla \mathbf{u}_{\ell,k,t}(\mathbf{m}) = -\widetilde{\Sigma}_{\ell,k,t}^{-1}.$$

Using the product rule:

$$\nabla\big(N_{\ell,k,t} \mathbf{u}_{\ell,k,t}\big) = (\nabla N_{\ell,k,t}) \mathbf{u}_{\ell,k,t}^\top + N_{\ell,k,t} \nabla \mathbf{u}_{\ell,k,t} = N_{\ell,k,t}\big(\mathbf{u}_{\ell,k,t} \mathbf{u}_{\ell,k,t}^\top - \widetilde{\Sigma}_{\ell,k,t}^{-1}\big).$$

Therefore:

$$\nabla A(\mathbf{m}) = \sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}\big(\mathbf{u}_{\ell,k,t}\mathbf{u}_{\ell,k,t}^\top - \widetilde{\Sigma}_{\ell,k,t}^{-1}\big),$$

$$\nabla B(\mathbf{m}) = \sum_{k=1}^{K_\ell} w_{\ell,k} \nabla N_{\ell,k,t} = \sum_{k=1}^{K_\ell} w_{\ell,k} N_{\ell,k,t}\mathbf{u}_{\ell,k,t} = A(\mathbf{m}).$$

By the quotient rule for vector-valued functions ($\nabla(A/B) = B^{-1}\nabla A - B^{-2}A(\nabla B)^\top$):

$$\nabla s_{\ell,t} = \frac{\nabla A}{B} - \frac{AA^\top}{B^2} = \sum_{k=1}^{K_\ell} r_{\ell,k}(\mathbf{m})\big(\mathbf{u}_{\ell,k,t}\mathbf{u}_{\ell,k,t}^\top - \widetilde{\Sigma}_{\ell,k,t}^{-1}\big) - s_{\ell,t}s_{\ell,t}^\top.$$

This gives the decomposition:

$$\nabla_{\mathbf{m}} s_{\ell,t} = \underbrace{\sum_{k=1}^{K_\ell} r_{\ell,k}\mathbf{u}_{\ell,k,t}\mathbf{u}_{\ell,k,t}^\top}_{=:U} - s_{\ell,t}s_{\ell,t}^\top - \underbrace{\sum_{k=1}^{K_\ell} r_{\ell,k}\widetilde{\Sigma}_{\ell,k,t}^{-1}}_{=:M}. \tag{26}$$

We now bound each term:

For U:

$$\|U\|_{\mathrm{op}} \le \sum_{k=1}^{K_\ell} r_{\ell,k}\|\mathbf{u}_{\ell,k,t}\mathbf{u}_{\ell,k,t}^\top\|_{\mathrm{op}} = \sum_{k=1}^{K_\ell} r_{\ell,k}\|\mathbf{u}_{\ell,k,t}\|^2.$$

Using the bound $\|\mathbf{u}_{\ell,k,t}\| \le \frac{1}{\lambda_{\min}}(\|\mathbf{m}\| + R)$:

$$\|U\|_{\mathrm{op}} \le \frac{1}{\lambda_{\min}^2}(\|\mathbf{m}\| + R)^2 \le \frac{2}{\lambda_{\min}^2}(\|\mathbf{m}\|^2 + R^2).$$

For $s_{\ell,t}s_{\ell,t}^\top$:

$$\|s_{\ell,t}s_{\ell,t}^\top\|_{\mathrm{op}} = \|s_{\ell,t}(\mathbf{m})\|^2 \le C_0^2(1 + \|\mathbf{m}\|)^2 \le 2C_0^2(1 + \|\mathbf{m}\|^2).$$

For M: Since $\|\widetilde{\Sigma}_{\ell,k,t}^{-1}\|_{\mathrm{op}} \le \lambda_{\min}^{-1}$ and $\sum r_{\ell,k} = 1$:

$$\|M\|_{\mathrm{op}} \le \frac{1}{\lambda_{\min}}.$$

Combining these into Eq. (26):

$$\begin{aligned}\|\nabla_{\mathbf{m}} s_{\ell,t}(\mathbf{m})\|_{\mathrm{op}} &\le \|U\|_{\mathrm{op}} + \|s_{\ell,t}s_{\ell,t}^\top\|_{\mathrm{op}} + \|M\|_{\mathrm{op}} \\ &\le \frac{2}{\lambda_{\min}^2}(\|\mathbf{m}\|^2 + R^2) + 2C_0^2(1 + \|\mathbf{m}\|^2) + \frac{1}{\lambda_{\min}}.\end{aligned}$$

Collecting terms, there exists a constant C_1 (depending on $R, \lambda_{\min}$) such that:

$$\|\nabla_{\mathbf{m}} s_{\ell,t}(\mathbf{m})\|_{\mathrm{op}} \le C_1(1 + \|\mathbf{m}\|^2).$$

This completes the proof. □

Lemma C.7 (Local Score-Class Lipschitzness)**.** *Fix a radius $R_{\mathrm{Supp}} > 0$. For any $\mathbf{m}_\ell$ with $\|\mathbf{m}_\ell\| \le R_{\mathrm{Supp}}$, the score function is Lipschitz with respect to parameters η_ℓ:*

$$\|s_{\ell,t}(\mathbf{m}_\ell; \eta_\ell) - s_{\ell,t}(\mathbf{m}_\ell; \eta_\ell')\| \le L_\eta(R_{\mathrm{Supp}})\|\eta_\ell - \eta_\ell'\|.$$

Proof. To simplify notation within this proof, we fix the subspace index ℓ and write $\mathbf{m}$ for $\mathbf{m}_\ell$, K for K_ℓ, and η for η_ℓ. Let the time-dependent covariance be $\widetilde{\Sigma}_{k,t} = \Sigma_{\ell,k} + \sigma_t^2 I$. We define the auxiliary

vector $\mathbf{u}_{k,t}(\mathbf{m}) = -\widetilde{\Sigma}_{k,t}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\ell,k})$ and the Gaussian component $N_{k,t}(\mathbf{m}) = \mathcal{N}(\mathbf{m}; \boldsymbol{\mu}_{\ell,k}, \widetilde{\Sigma}_{k,t})$. The score function is given by:

$$s_{\ell,t}(\mathbf{m};\eta) = \sum_{k=1}^{K} r_k(\mathbf{m};\eta)\mathbf{u}_{k,t}(\mathbf{m};\eta), \quad \text{where } r_k(\mathbf{m};\eta) = \frac{w_{\ell,k}N_{k,t}(\mathbf{m})}{\sum_{i=1}^{K} w_{\ell,i}N_{i,t}(\mathbf{m})}.$$

We assume Assumption C.1 holds, implying parameter boundedness

$$w_{\ell,k} \geq w_{\min}, \quad \|\boldsymbol{\mu}_{\ell,k}\| \leq R, \quad \widetilde{\Sigma}_{k,t} \succeq \lambda_{\min} I. \tag{27}$$

Fix any radius $R_{\text{Supp}} > 0$ and restrict to $\|\mathbf{m}\| \leq R_{\text{Supp}}$. To prove that $s_{\ell,t}(\mathbf{m};\eta)$ is Lipschitz continuous with respect to η on this bounded region, it suffices to show that the gradient $\nabla_\eta s_{\ell,t}(\mathbf{m};\eta)$ is bounded uniformly over the admissible parameter space and over $\|\mathbf{m}\| \leq R_{\text{Supp}}$. By the Mean Value Theorem, if $\|\nabla_\eta s_{\ell,t}(\mathbf{m};\eta)\| \leq L_\eta(R_{\text{Supp}})$, then $\|s_{\ell,t}(\mathbf{m};\eta) - s_{\ell,t}(\mathbf{m};\eta')\| \leq L_\eta(R_{\text{Supp}})\|\eta - \eta'\|$.

Consider the gradient with respect to a generic parameter $\theta \in \eta$:

$$\frac{\partial s_{\ell,t}}{\partial \theta} = \sum_{k=1}^{K} \left[\frac{\partial r_k}{\partial \theta}\mathbf{u}_{k,t} + r_k \frac{\partial \mathbf{u}_{k,t}}{\partial \theta} \right]. \tag{28}$$

To compute the derivative of the responsibility r_k, let $L_k = \log(w_{\ell,k}N_{k,t}(\mathbf{m}))$. Then $r_k = e^{L_k}/\sum_i e^{L_i}$, and its derivative is given by

$$\frac{\partial r_k}{\partial \theta} = r_k\left(\frac{\partial L_k}{\partial \theta} - \sum_{i=1}^{K} r_i \frac{\partial L_i}{\partial \theta}\right).$$

Substituting this into Eq. (28), we bound the terms for each parameter type. Recall from Lemma C.6 that $\|\widetilde{\Sigma}_{k,t}^{-1}\|_{\text{op}} \leq \lambda_{\min}^{-1}$.

(i) Term about $\partial w_{\ell,k}$.

The dependence is only through r_k. For $w_{\ell,k}$, we have $\frac{\partial L_k}{\partial w_{\ell,k}} = \frac{1}{w_{\ell,k}}$.

Since $w_{\ell,k} \geq w_{\min}$, we have $|\frac{\partial L_k}{\partial w_{\ell,k}}| \leq \frac{1}{w_{\min}}$. The term $\mathbf{u}_{k,t}$ does not depend on weights. Thus, the gradient is bounded by a linear combination of $\|\mathbf{u}_{k,t}\|$. Since $\|\mathbf{u}_{k,t}(\mathbf{m})\|$ grows linearly with $\|\mathbf{m}\|$, this contribution is bounded by $O(\|\mathbf{m}\|)$.

(ii) Term about $\partial \boldsymbol{\mu}_{\ell,k}$.

Both L_k and $\mathbf{u}_{k,t}$ depend on $\boldsymbol{\mu}_{\ell,k}$.

For the vector term:

$$\nabla_{\boldsymbol{\mu}_{\ell,k}} \mathbf{u}_{k,t}(\mathbf{m}) = \nabla_{\boldsymbol{\mu}_{\ell,k}}[-\widetilde{\Sigma}_{k,t}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\ell,k})] = \widetilde{\Sigma}_{k,t}^{-1}.$$

Its norm is bounded by $\lambda_{\min}^{-1}$. For the log-likelihood term:

$$\nabla_{\boldsymbol{\mu}_{\ell,k}} L_k = \widetilde{\Sigma}_{k,t}^{-1}(\mathbf{m} - \boldsymbol{\mu}_{\ell,k}) = -\mathbf{u}_{k,t}(\mathbf{m}).$$

Substituting back, the gradient involves terms of order $\|\mathbf{u}_{k,t}\|^2$ (from $r_k\mathbf{u}_{k,t}\frac{\partial L_k}{\partial \theta}$) and constant terms (from $\frac{\partial \mathbf{u}}{\partial \theta}$). Since $\|\mathbf{u}_{k,t}\|^2$ grows as $O(\|\mathbf{m}\|^2)$, this contribution is bounded on the compact set.

(iii) Term about $\partial \Sigma_{\ell,k}$.

Here $\widetilde{\Sigma}_{k,t} = \Sigma_{\ell,k} + \sigma_t^2 I$, so $\nabla_{\Sigma_{\ell,k}} \widetilde{\Sigma}_{k,t}$ is the identity tensor.

Using the identity $\partial A^{-1} = -A^{-1}(\partial A)A^{-1}$, we get:

$$D_\Sigma \mathbf{u}_{k,t}[H] = \widetilde{\Sigma}_{k,t}^{-1} H \widetilde{\Sigma}_{k,t}^{-1}(\mathbf{m} - \mu_{\ell,k}) = -\widetilde{\Sigma}_{k,t}^{-1} H \mathbf{u}_{k,t}.$$

For $\|H\|_{\text{op}} = 1$, its norm is bounded by $\lambda_{\min}^{-1}\|\mathbf{u}_{k,t}\| = O(1 + \|\mathbf{m}\|)$.

$$\frac{\partial L_k}{\partial \Sigma_{\ell,k}} = -\frac{1}{2}\widetilde{\Sigma}_{k,t}^{-1} + \frac{1}{2}\mathbf{u}_{k,t}\mathbf{u}_{k,t}^\top.$$

The derivative of the log-Gaussian density yields terms proportional to $\widetilde{\Sigma}_{k,t}^{-1}$ (trace term) and quadratic form terms $\mathbf{u}_{k,t}\mathbf{u}_{k,t}^{\top}$. The norm is dominated by $\|\mathbf{u}_{k,t}\|^2$, scaling as $O(\|\mathbf{m}\|^2)$.

Conclusion. Combining these results, the Jacobian $\nabla_\eta s_{\ell,t}(\mathbf{m};\eta)$ contains terms of degree at most three in $\|\mathbf{u}_{k,t}(\mathbf{m})\|$. From Lemma C.6, we know $\|\mathbf{u}_{k,t}(\mathbf{m})\| \le \frac{1}{\lambda_{\min}}(\|\mathbf{m}\| + R)$. Therefore, for any fixed radius $R_{\text{Supp}} > 0$, there exists a constant $C_{poly}(R_{\text{Supp}})$ such that whenever $\|\mathbf{m}\| \le R_{\text{Supp}}$,

$$\|\nabla_\eta s_{\ell,t}(\mathbf{m};\eta)\| \le C_{poly}(R_{\text{Supp}}).$$

This implies the claimed local Lipschitz property with constant $L_\eta(R_{\text{Supp}}) := C_{poly}(R_{\text{Supp}})$. We use this property only to control the complexity and local covering numbers of the induced score class; the main theorem below does not require global identifiability of η. □

C.3 Main Result: Error Propagation via Score Risk

We now derive the final bound at the level of induced score fields. This avoids identifying a unique MoG parameterization and is invariant to component permutations that leave the score unchanged.

Theorem C.8 (Generative Error Bound via Score Risk)**.** *Conditionally on the frozen factorization and source state, let $\mathcal{D}_T = \{\boldsymbol{x}_i\}_{i=1}^n$ consist of i.i.d. draws from P_T, and let $\hat{\eta} \in \arg\min_{\eta\in\Theta_n} n^{-1}\sum_i \ell_\eta(\boldsymbol{x}_i)$ be the constrained grid DSM estimator, inheriting η_{sh}^S. Let*

$$\eta^\dagger \in \arg\min_{\eta_{\text{un}}\in\Theta_{\text{un}}} \mathcal{L}_T(\eta_{\text{sh}}^S, \eta_{\text{un}})$$

be the best population unshared score field in the SUDA-MoG class. Under Assumptions C.1– C.5 , for any confidence parameter $\delta \in (0,1)$, the following bound holds with probability at least $1-\delta$:

$$W_2(G_{\hat{\eta}}, P_T) \;\le\; E_{\text{approx}} + E_{\text{VAE}} + C_{\text{score}}\sqrt{\frac{d_{\text{un}}\log(An) + \log(2/\delta)}{n}}. \tag{29}$$

Here $d = W_2$ denotes image-space Wasserstein distance and d_{un} is the number of free trainable unshared statistics. The approximation term E_{approx} contains the best-in-class score approximation error of $\eta^\dagger$, including factor-wise MoG misspecification, approximate shared-unshared partition error, ignored nonlinear coupling, and other fixed score-class gaps. The explicit complexity factor counts unshared score parameters. The constants can depend on the representation and sampler; global MoG parameter identifiability is not required .

Proof. The proof avoids parameter recovery and instead controls the score field induced by the adapted MoG statistics.

Let

$$\mathcal{L}_T(\eta) = \mathbb{E}_{\boldsymbol{x}\sim P_T,t,\epsilon}\left[\|S_\eta(\mathbf{m}_t,t) - \nabla_{\mathbf{m}_t}\log p_{t|0}(\mathbf{m}_t|\mathbf{m}_0)\|_2^2\right]$$

be the population DSM objective in factor space, where $\mathbf{m}_0 = \Phi(E(\boldsymbol{x}))$ and $\mathbf{m}_t = \mathbf{m}_0 + \sigma_t\epsilon$. Let $\hat{\mathcal{L}}_T(\eta)$ be its empirical counterpart on $\mathcal{D}_T$. By the denoising score matching identity, the population DSM objective differs from the squared score error to the target marginal score only by a constant independent of η:

$$\mathcal{L}_T(\eta) = C_T + \mathbb{E}_t\mathbb{E}_{\mathbf{m}_t\sim P_{T,t}}\|S_\eta(\mathbf{m}_t,t) - S_T(\mathbf{m}_t,t)\|_2^2.$$

The induced score field S_η is the statistical object in this analysis, and it is invariant to permutations of the mixture-component labels .

Let $u = \log(2M_n/\delta)$ and let $\widehat{e}_\eta$ denote the empirical mean of f_η. The moment condition gives the two-sided Bernstein bound

$$|\widehat{e}_\eta - e_\eta| \le \sqrt{2Be_\eta u/n} + bu/n \le e_\eta/2 + (B+b)u/n$$

simultaneously for all grid points with probability at least $1-\delta$, by a union bound. Write $c_n = (B+b)u/n$. Comparing the empirical minimizer with the approximating point η_0 gives

$$e_{\hat{\eta}}/2 - c_n \le \widehat{e}_{\hat{\eta}} \le \widehat{e}_{\eta_0} \le 3a_n/2 + c_n, \qquad e_{\hat{\eta}} \le 3a_n + 4c_n.$$

Score curvature in Assumption C.4 now gives

$$\rho(\hat{\eta}, \eta^\dagger) \le \sqrt{\frac{3C_{\text{app}} + 4(B+b)[d_{\text{un}} \log(An) + \log(2/\delta)]}{c_{\text{curv}} n}}.$$

Absorbing fixed constants into C_{est} yields the stated rate. This score-level estimation bound accommodates parameter non-identifiability: parameter vectors that differ by a mixture-component permutation induce the same score field and have zero distance under ρ.

Next, by the reverse-SDE stability assumption,

$$W_2(G^{\text{lat}}_{\hat{\eta}}, G^{\text{lat}}_{\eta^\dagger}) \le C_{\text{stab}}(T)\, \rho(\hat{\eta}, \eta^\dagger).$$

Mapping from factor space back to image space through $g_\psi \circ \Phi^{-1}$ and using its Lipschitz constant L_g gives

$$d(G_{\hat{\eta}}, G_{\eta^\dagger}) \le L_g C_{\text{stab}}(T) \rho(\hat{\eta}, \eta^\dagger).$$

Substituting the score-level estimation bound yields

$$d(G_{\hat{\eta}}, G_{\eta^\dagger}) \le L_g C_{\text{stab}}(T) C_{\text{est}} \sqrt{\frac{d_{\text{un}} \log(An) + \log(2/\delta)}{n}}.$$

Finally, the best population score field $G_{\eta^\dagger}$ may still differ from the true target distribution because of the factor-wise MoG approximation, the frozen encoder-decoder reconstruction error, and any misspecification of the shared-unshared factorization. These fixed pipeline errors are collected into E_{approx} and E_{VAE}. Precisely, let $Q_T = (\Phi \circ E)_{\#} P_T$ and define

$$E_{\text{approx}} = L_g W_2(Q_{\eta^\dagger}, Q_T), \qquad E_{\text{VAE}} = W_2((D \circ E)_{\#} P_T, P_T).$$

The first term includes the fixed sampler's initialization, positive-noise endpoint, and any numerical integration discrepancy. Applying the triangle inequality gives

$$d(G_{\hat{\eta}}, P_T) \le E_{\text{approx}} + E_{\text{VAE}} + C_{\text{score}} \sqrt{\frac{d_{\text{un}} \log(An) + \log(2/\delta)}{n}},$$

where $C_{\text{score}} = L_g C_{\text{stab}}(T) C_{\text{est}}$. This proves the theorem. □

D Stability Analysis of Shared Experts

D.1 Optimization under Domain Shift

In this section, we provide the theoretical analysis of the loss landscape under our factorization. Specifically, we show that even if all expert parameters were unfrozen, the target-domain objective provides zero gradient signal on the shared experts at the source optimum. For the additive population objective initialized at the shared stationary point, unshared updates leave the shared gradient zero .

Theorem D.1 (Stability of Shared Experts)**.** *Suppose Assumption C.3 holds. Let $\eta^*_{\text{sh}} = \{\eta^*_\ell\}_{\ell \in \mathcal{I}_{\text{sh}}}$ be the pretrained parameters optimized on the source domain P_S. Then, the gradient of the target-domain DSM objective with respect to these shared parameters vanishes at η^*_{sh}:*

$$\nabla_{\eta_{\text{sh}}} \mathcal{L}_{\text{DSM}}(\eta; P_T)\Big|_{\eta_{\text{sh}} = \eta^*_{\text{sh}}} = \mathbf{0}. \tag{30}$$

Consequently, η^*_{sh} remains a stationary point for the target task, implying that "unfreezing" these parameters would initially yield zero gradients.

Proof. The argument proceeds by exploiting the geometric decoupling of the latent space. First, we show that the product factorization of the prior (Assumption C.3) leads to an additive decomposition of the global DSM loss into independent subspace losses. Second, for any shared subspace, the marginal distribution is invariant between source and target domains ($P_S = P_T$). Therefore, the loss landscape for shared parameters is identical across domains.

(i) Loss Decomposition. The global DSM objective on the target domain is given by:

$$\mathcal{L}_{\mathrm{DSM}}(\eta; P_T) = \tfrac{1}{2}\mathbb{E}_{\mathbf{m}_0,t,\epsilon}\|S_\eta(\mathbf{m}_t, t) + \epsilon/\sigma_t\|_2^2, \quad \mathbf{m}_t = \mathbf{m}_0 + \sigma_t\epsilon.$$

By Assumption C.3, the factor vector is partitioned into coordinate blocks indexed by $\ell \in \{1, \dots, m\}$. The score function $s_t(\mathbf{m})$ is a concatenation of subspace scores $[s_{\ell,t}(\mathbf{m}_\ell)]_{\ell=1}^m$. The squared Euclidean norm of a concatenated vector decomposes additively:

$$\|S_\eta(\mathbf{m}_t, t) + \epsilon/\sigma_t\|_2^2 = \sum_{\ell=1}^{K} \|s_{\ell,t}(\mathbf{m}_{\ell,t}; \eta_\ell) + \epsilon_\ell/\sigma_t\|_2^2.$$

Substituting this into the expectation, the total loss splits into a sum of independent subspace losses:

$$\mathcal{L}_{\mathrm{DSM}}(\eta; P_T) = \sum_{\ell=1}^{m} \mathcal{L}_\ell(\eta_\ell; P_T(\mathbf{m}_\ell)), \tag{31}$$

where $\mathcal{L}_\ell$ depends only on the parameters η_ℓ and the marginal distribution of the ℓ-th subspace.

(ii) Gradient Independence. Consider the gradient with respect to a specific shared parameter block η_k, where $k \in \mathcal{I}_{\mathrm{sh}}$. Due to the decoupled structure in Eq. (31), η_k only affects the k-th term in the summation. Thus:

$$\nabla_{\eta_k}\mathcal{L}_{\mathrm{DSM}}(\eta; P_T) = \nabla_{\eta_k}\mathcal{L}_k(\eta_k; P_T(\mathbf{m}_k)).$$

(iii) Domain Invariance. Crucially, Assumption C.3 states that for any shared subspace $k \in \mathcal{I}_{\mathrm{sh}}$, the marginal distribution is domain-invariant:

$$P_T(\mathbf{m}_k) = P_S(\mathbf{m}_k).$$

Consequently, the expected loss surface for this subspace on the target domain is identical to that on the source domain:

$$\mathcal{L}_k(\eta_k; P_T(\mathbf{m}_k)) \equiv \mathcal{L}_k(\eta_k; P_S(\mathbf{m}_k)).$$

(iv) Stationarity. By definition (Assumption C.3), the pretrained parameters η_k^* are stationary points of the source objective $\mathcal{L}_k(\cdot; P_S)$. Hence the source gradient is zero :

$$\nabla_{\eta_k}\mathcal{L}_k(\eta_k; P_S)\Big|_{\eta_k=\eta_k^*} = \mathbf{0}.$$

Combining this with the identity derived in Step 3, we obtain:

$$\nabla_{\eta_k}\mathcal{L}_k(\eta_k; P_T)\Big|_{\eta_k=\eta_k^*} = \mathbf{0}.$$

Since this holds for all shared subspaces $k \in \mathcal{I}_{\mathrm{sh}}$, the full gradient vector with respect to shared parameters vanishes. □

D.2 Initial Gradient Response to Domain Shift

Theorem D.1 addresses exact invariance of the shared marginals, while active marginals may change. We next allow shared marginals to shift and analyze the initial gradient response of each factor under this relaxation.

Under local curvature bounds, the magnitude of the initial gradient is bounded above and below by the source–target parameter discrepancy . This quantifies the initial response of each factor to a domain shift.

Assumption D.2 (Local regularity around $\eta_{\ell,T}^*$)**.** For each subspace ℓ , the target minimizer is an interior stationary point in free parameter coordinates. There exists a convex neighbourhood $\mathcal{N}_\ell$ of the target minimizer $\eta_{\ell,T}^*$ such that the subspace loss $\mathcal{L}_\ell(\cdot; P_T)$ is twice continuously differentiable on $\mathcal{N}_\ell$ and its Hessian satisfies the uniform spectral bounds

$$\mu I \preceq \nabla^2\mathcal{L}_\ell(\theta) \preceq L_{\mathrm{Hess}} I, \qquad \forall\theta \in \mathcal{N}_\ell,$$

for some constants $0 < \mu \le L_{\mathrm{Hess}} < \infty$. We further assume the initialization (pretrained source parameter) lies in this neighbourhood: $\eta_{\ell,S}^* \in \mathcal{N}_\ell$.

Theorem D.3 (Gradient alignment with parameter shift)**.** *Under Assumption D.2, define the parameter shift*

$$\Delta_\ell := \eta^*_{\ell,S} - \eta^*_{\ell,T}.$$

Then the gradient of the target loss at the initialization satisfies the exact identity

$$\nabla_{\eta_\ell} \mathcal{L}_\ell(\eta^*_{\ell,S}; P_T) = \Big(\int_0^1 \nabla^2 \mathcal{L}_\ell(\eta^*_{\ell,T} + \tau\Delta_\ell)\, d\tau \Big)\, \Delta_\ell,$$

and the following spectral bounds hold:

$$\mu \, \|\Delta_\ell\|_2 \;\le\; \|\nabla_{\eta_\ell} \mathcal{L}_\ell(\eta^*_{\ell,S}; P_T)\|_2 \;\le\; L_{\mathrm{Hess}} \, \|\Delta_\ell\|_2.$$

Proof. Since $\mathcal{L}_\ell$ is C^2 on the convex neighbourhood $\mathcal{N}_\ell$ and both $\eta^*_{\ell,S}$ and $\eta^*_{\ell,T}$ lie in $\mathcal{N}_\ell$, we may apply the fundamental theorem of calculus (integral form of Taylor's theorem) to the gradient map along the line segment joining $\eta^*_{\ell,T}$ and $\eta^*_{\ell,S}$. Concretely,

$$\nabla \mathcal{L}_\ell(\eta^*_{\ell,S}) = \nabla \mathcal{L}_\ell\big(\eta^*_{\ell,T} + \Delta_\ell\big) = \nabla \mathcal{L}_\ell(\eta^*_{\ell,T}) + \int_0^1 \nabla^2 \mathcal{L}_\ell\big(\eta^*_{\ell,T} + \tau\Delta_\ell\big)\, \Delta_\ell \, d\tau.$$

By optimality of $\eta^*_{\ell,T}$ (local minimizer) we have $\nabla \mathcal{L}_\ell(\eta^*_{\ell,T}) = \mathbf{0}$, hence the identity

$$\nabla \mathcal{L}_\ell(\eta^*_{\ell,S}) = \Big(\int_0^1 \nabla^2 \mathcal{L}_\ell(\eta^*_{\ell,T} + \tau\Delta_\ell)\, d\tau \Big) \Delta_\ell.$$

Define the averaged Hessian

$$\bar{H}_\ell := \int_0^1 \nabla^2 \mathcal{L}_\ell(\eta^*_{\ell,T} + \tau\Delta_\ell)\, d\tau.$$

By the spectral bounds in Assumption D.2 we have for every $\tau \in [0,1]$

$$\mu I \preceq \nabla^2 \mathcal{L}_\ell(\eta^*_{\ell,T} + \tau\Delta_\ell) \preceq L_{\mathrm{Hess}} I,$$

and integration preserves the Loewner ordering, therefore

$$\mu I \preceq \bar{H}_\ell \preceq L_{\mathrm{Hess}} I.$$

Taking Euclidean norms and using standard bounds for matrix-vector products gives

$$\|\nabla \mathcal{L}_\ell(\eta^*_{\ell,S})\|_2 = \|\bar{H}_\ell \Delta_\ell\|_2 \le \|\bar{H}_\ell\|_{\mathrm{op}} \, \|\Delta_\ell\|_2 \le L_{\mathrm{Hess}} \|\Delta_\ell\|_2,$$

and likewise for the lower bound, using the smallest eigenvalue of $\bar{H}_\ell$,

$$\|\nabla \mathcal{L}_\ell(\eta^*_{\ell,S})\|_2 \ge \lambda_{\min}(\bar{H}_\ell) \, \|\Delta_\ell\|_2 \ge \mu \|\Delta_\ell\|_2.$$

This completes the proof. □

Corollary D.4 (Initial decrease under gradient descent)**.** *Under Assumption D.2, consider a single gradient-descent step on subspace ℓ with step-size $\alpha \in (0, 2/L_{\mathrm{Hess}}]$:*

$$\eta^+_\ell = \eta^*_{\ell,S} - \alpha \nabla_{\eta_\ell} \mathcal{L}_\ell(\eta^*_{\ell,S}; P_T).$$

Assume the update segment remains in $\mathcal{N}_\ell$. Then the loss decreases by at least

$$\mathcal{L}_\ell(\eta^*_{\ell,S}; P_T) - \mathcal{L}_\ell(\eta^+_\ell; P_T) \;\ge\; \alpha\Big(1 - \frac{\alpha L_{\mathrm{Hess}}}{2}\Big) \|\nabla_{\eta_\ell} \mathcal{L}_\ell(\eta^*_{\ell,S}; P_T)\|_2^2 \;\ge\; \alpha\Big(1 - \frac{\alpha L_{\mathrm{Hess}}}{2}\Big) \mu^2 \|\Delta_\ell\|_2^2.$$

Proof. The first inequality is standard for L-smooth functions (here $L = L_{\mathrm{Hess}}$): for any $\alpha \in (0, 2/L)$,

$$\mathcal{L}_\ell(\eta) - \mathcal{L}_\ell(\eta - \alpha \nabla \mathcal{L}_\ell(\eta)) \ge \alpha\Big(1 - \frac{\alpha L}{2}\Big) \|\nabla \mathcal{L}_\ell(\eta)\|_2^2.$$

Apply this with $\eta = \eta^*_{\ell,S}$ and $L = L_{\mathrm{Hess}}$ to obtain the first displayed inequality. The second inequality follows from the lower bound in Theorem D.3, i.e. $\|\nabla \mathcal{L}_\ell(\eta^*_{\ell,S})\|_2 \ge \mu \|\Delta_\ell\|_2$. Combining yields the stated bound. □

Remark. Theorem D.3 quantifies the initial gradient response, and Corollary D.4 gives the corresponding one-step loss decrease. With common curvature bounds, factor i has a larger initial gradient than factor j whenever $\mu\|\Delta_i\| > L_{\mathrm{Hess}}\|\Delta_j\|$.

E Robustness Analysis under Imperfect Decoupling

The factor-wise objective above has zero cross-block Hessian by construction. We now consider a coupled differentiable score S_η and its joint target DSM objective $\mathcal{L}_T(\eta) = \frac{1}{2}\mathbb{E}\|S_\eta - Y\|^2$, where $Y = -\epsilon/\sigma_t$. Shared and unshared parameter blocks can interact within this score model. The following assumptions quantify that interaction .

Assumption E.1 (ϵ-Approximate Invariance and γ-Weak Coupling)**.** Let $\eta_{\mathrm{sh}} = \{\eta_\ell\}_{\ell\in\mathcal{I}_{\mathrm{sh}}}$ and $\eta_{\mathrm{un}} = \{\eta_\ell\}_{\ell\in\mathcal{I}_{\mathrm{un}}}$ be partitioned , $\eta = [\eta_{\mathrm{sh}}^\top, \eta_{\mathrm{un}}^\top]^\top$. We posit the following regularity conditions:

(i) Minor Domain Shift. The optimal shared parameters on the target domain $\eta^*_{\mathrm{sh},T}$ are close to the pretrained source parameters $\eta^*_{\mathrm{sh},S}$. Specifically, we assume the Euclidean distance is bounded by $\|\eta^*_{\mathrm{sh},S} - \eta^*_{\mathrm{sh},T}\|_2 \le \epsilon$.

(ii) Weak Hessian Coupling. The interaction between subspaces in the loss landscape is bounded by γ. The loss is C^2 on a convex neighborhood containing the segment from η^*_T to η^*_S, and $\nabla\mathcal{L}_T(\eta^*_T) = 0$. Along this segment, assume $\|\mathbf{H}_{\mathrm{sh,un}}\|_{\mathrm{op}} \le \gamma$ and $\|\mathbf{H}_{\mathrm{sh,sh}}\|_{\mathrm{op}} \le L_{\mathrm{Hess}}$.

Proposition E.2 (Parameter coupling in joint DSM)**.** *Let $r = S_\eta - Y$ and $J_i = \partial S_\eta/\partial\eta_i$. Under differentiation beneath the expectation,*

$$H_{i,j} = \mathbb{E}\left[J_i^\top J_j + \sum_q r_q\, \partial^2_{\eta_i,\eta_j} S_{\eta,q}\right].$$

For concatenated factor-wise scores with disjoint parameters, both terms vanish for $i \neq j$.

Proof. Differentiate $\mathcal{L}_T = \frac{1}{2}\mathbb{E}\sum_q r_q^2$ once with respect to η_i and once with respect to η_j . In the factor-wise case, the Jacobians have disjoint output support and the mixed score derivatives are zero. □

The score derivatives in this expression determine the coupling bound γ in Assumption E.1. The experiments evaluate this loss-level response with joint-DSM models and evaluate $\mathcal{S}_{i,j}$ through controlled coordinate interventions.

Theorem E.3 (Stability under Imperfect Decoupling)**.** *Under Assumption E.1 , let $\Delta_{\mathrm{un}} = \|\eta^*_{\mathrm{un},S} - \eta^*_{\mathrm{un},T}\|_2$ denote the magnitude of the shift required for the unshared subspace. The norm of the initial gradient on the shared parameters η_{sh} at the source optimum η^*_S satisfies:*

$$\|\nabla_{\eta_{\mathrm{sh}}}\mathcal{L}_{\mathrm{DSM}}(\eta^*_S; P_T)\|_2 \le \underbrace{L_{\mathrm{Hess}}\cdot\epsilon}_{\textit{Self-Shift Impact}} + \underbrace{\gamma\cdot\Delta_{\mathrm{un}}}_{\textit{Gradient Leakage}} . \tag{32}$$

Proof. Let $\delta = \eta^*_S - \eta^*_T = (\delta_{\mathrm{sh}}, \delta_{\mathrm{un}})$, where $\|\delta_{\mathrm{sh}}\| \le \epsilon$ and $\|\delta_{\mathrm{un}}\| = \Delta_{\mathrm{un}}$. Stationarity at η^*_T and the fundamental theorem of calculus give

$$\nabla\mathcal{L}_T(\eta^*_S) = \left(\int_0^1 H(\eta^*_T + \tau\delta)\, d\tau\right)\delta.$$

Writing the averaged Hessian as $\bar{H}$ and taking its shared block yields

$$\nabla_{\eta_{\mathrm{sh}}}\mathcal{L}_T(\eta^*_S) = \bar{H}_{\mathrm{sh,sh}}\delta_{\mathrm{sh}} + \bar{H}_{\mathrm{sh,un}}\delta_{\mathrm{un}}.$$

The norm of an averaged block is at most the average of its norms. Applying Assumption E.1 and the triangle inequality therefore gives $\|\nabla_{\eta_{\mathrm{sh}}}\mathcal{L}_T(\eta^*_S)\| \le L_{\mathrm{Hess}}\epsilon + \gamma\Delta_{\mathrm{un}}$. □

E.1 Finite-Difference Sensitivity

For source task coordinate u_i, we recompose after perturbations $u_i \pm \delta$ and extract shared features $v^{\pm}_{\mathrm{sh},i}$. With N carriers and source feature mean $\bar{v}^S_{\mathrm{sh}}$, we use

$$\widehat{\gamma}_{\mathrm{FD}} = \frac{1}{N}\sum_i \left\| \frac{v^+_{\mathrm{sh},i} - v^-_{\mathrm{sh},i}}{2\delta} \right\|_2, \quad \delta = 0.5, \qquad \text{FD scale} = \frac{\widehat{\gamma}_{\mathrm{FD}}(\max_i u_i - \min_i u_i)}{\|\bar{v}^S_{\mathrm{sh}}\|_1 + 10^{-8}}.$$

The main experiments report this finite-difference scale in the implemented feature coordinates. Direct score-Hessian and shared-gradient measurements use the joint-DSM protocol in Appendix H.6.

F Implementation and Evaluation Protocol

This section specifies the simple rank-one setting of Greedy used in the main role experiments. Calibration fixes the statistical task state; the execution operators used for each experiment are specified below .

F.1 Rank-One Moment Coordinates and Recomposition

For a latent tensor $\boldsymbol{z} \in \mathbb{R}^{C\times H\times W}$, we first compute channel-wise moments

$$\boldsymbol{s} = (\boldsymbol{\mu}, \boldsymbol{\sigma}), \qquad \boldsymbol{\mu}_c = \frac{1}{HW}\sum_{h,w} z_{c,h,w}, \quad \boldsymbol{\sigma}^2_c = \frac{1}{HW}\sum_{h,w}(z_{c,h,w} - \boldsymbol{\mu}_c)^2.$$

The normalized residual is

$$\boldsymbol{r} = (\boldsymbol{z} - \boldsymbol{\mu})/(\boldsymbol{\sigma} + \delta), \qquad \boldsymbol{q} = \mathrm{vec}(\boldsymbol{r}).$$

Given the LDA direction $\boldsymbol{a}$ in the residual space, we keep both the target axis and its orthogonal complement:

$$u = \boldsymbol{a}^\top \boldsymbol{q}, \qquad \boldsymbol{q}_\perp = (I - \boldsymbol{a}\boldsymbol{a}^\top)\boldsymbol{q}.$$

Thus the practical coordinate map is $\Phi(\boldsymbol{z}) = (\boldsymbol{s}, u, \boldsymbol{q}_\perp)$. The recomposition map R_{greedy} is

$$\boldsymbol{q}' = u'\boldsymbol{a} + \boldsymbol{q}'_\perp, \qquad \boldsymbol{z}' = (\boldsymbol{\sigma}' + \delta) \odot \mathrm{unvec}(\boldsymbol{q}') + \boldsymbol{\mu}'.$$

For unit-norm $\boldsymbol{a}$, this construction reconstructs each encoded latent: $R_{\mathrm{greedy}} \circ \Phi = \mathrm{id}$. An active-coordinate edit is defined by $z' = R_{\mathrm{greedy}}(s, u', q_\perp)$, using the carrier's stored moments and residual complement. Shared-feature responses are measured on the recomposed latent, or on its decoded and re-encoded output, as specified by the evaluation protocol. The finite-difference diagnostic in Appendix E.1 quantifies these responses. Role ablations and matched-output interventions use this stored-coordinate recomposition.

F.2 Few-Shot MoG Recalibration

For each factor k, we fit a source MoG $\eta^S_k = \{w^S_{k,j}, \boldsymbol{\mu}^S_{k,j}, \Sigma^S_{k,j}\}_{j=1}^{J_k}$ using source factor coordinates. Component counts combine a task-informed complexity estimate with Bayesian information criterion (BIC) selection. Task complexity guides the candidate counts; fitted MoGs are compared using

$$\mathrm{BIC}(J_k) = -2\log L(\widehat{\eta}_k) + p_k \log N_k,$$

where $L(\widehat{\eta}_k)$ is the fitted likelihood, p_k is the number of free mixture parameters, and N_k is the number of fitting observations. Lower BIC is preferred. The task configurations use 3 components for Dog→Cat, 1 for Glasses, 3 for Haircut, and 5 for Young→Old. Target observations receive fixed source-component responsibilities:

$$r^S_{i,j} = \frac{w^S_{k,j}\mathcal{N}(\mathbf{m}^T_{k,i}; \boldsymbol{\mu}^S_{k,j}, \Sigma^S_{k,j})}{\sum_\ell w^S_{k,\ell}\mathcal{N}(\mathbf{m}^T_{k,i}; \boldsymbol{\mu}^S_{k,\ell}, \Sigma^S_{k,\ell})}, \qquad N_j = \sum_i r^S_{i,j}.$$

The responsibility-weighted target mean is

$$\bar{\boldsymbol{\mu}}^T_{k,j} = \frac{1}{N_j + \epsilon_0}\sum_i r^S_{i,j}\mathbf{m}^T_{k,i}.$$

For unshared factors, the default few-shot update is a shrinkage mean update

$$\boldsymbol{\mu}_{k,j}^T = (1-\lambda_j)\boldsymbol{\mu}_{k,j}^S + \lambda_j \bar{\boldsymbol{\mu}}_{k,j}^T, \qquad \lambda_j = \frac{N_j}{N_j + \tau_\mu},$$

while shared factors use $\eta_k^T = \eta_k^S$. Mixture weights and covariances are inherited by default in the smallest-shot regime. Optional covariance updates require sufficient component counts. Mixture weights are either inherited as a whole or updated as a complete vector, using

$$w_{k,j}^T = \frac{N_j + \tau_w w_{k,j}^S}{n + \tau_w}, \qquad \Sigma_{k,j}^T = (1-\lambda_j^\Sigma)\Sigma_{k,j}^S + \lambda_j^\Sigma \widehat{\Sigma}_{k,j}^T + \rho I,$$

with the same source-responsibility matching. These shrinkage rules make the few-shot estimator stable and keep the target update localized to the unshared factors.

F.3 Execution of Calibrated Factor Laws

Analytic generation. For additive Gaussian noise, the calibrated factor law has score

$$s_{k,t}^T(\mathbf{m}_k) = \nabla_{\mathbf{m}_k} \log \sum_j w_{k,j}^T \mathcal{N}(\mathbf{m}_k; \mu_{k,j}^T, \Sigma_{k,j}^T + \sigma_t^2 I).$$

Concatenating these scores defines a factor-space reverse sampler; recomposition and decoding produce the output law analyzed in Theorem 5.1.

Image-conditioned editing. A carrier supplies the initial active coordinate and complementary structure. For rank-one mean recalibration, the deterministic update is

$$T_\mu(u) = u + \sum_j r_j^S(u)(\mu_j^T - \mu_j^S).$$

Recomposition retains the stored shared coordinates. For a single Gaussian, this is a translation by the calibrated mean shift. For a mixture, responsibilities blend the component shifts into one edit per carrier; Corollary G.6 bounds calibration error for the resulting edited carrier law. The 32-support joint-attribute experiment updates its complete active vector by maximizing $\log p_T(u) - \log p_S(u)$ with a reconstruction-metric preconditioner, backtracking, and a latent-displacement bound. The fitted laws remain fixed during the at most 32 steps (Appendix H.5).

Matched-adapter interventions. With both adapter outputs and the analysis map frozen, the functional test transplants the task coordinate from a LoRA output to its paired base. The external evaluator measures the contribution of that coordinate to the adapter's response (Appendix H.2).

Algorithm 1 Greedy SUDA-MoG calibration and execution

Require: Source set $\mathcal{D}_S$, discovery data $\mathcal{D}_{\text{disc}}$, calibration supports $\mathcal{D}_T$, frozen encoder-decoder (E, D)

1: Construct Φ and assign roles using $\mathcal{D}_{\text{disc}}$; encode source and calibration images
2: Fit source MoGs η_k^S on factor coordinates
3: Hold the discovered chart and roles fixed during calibration
4: **for** each unshared factor **do**
5: Compute source responsibilities and effective counts over all target supports
6: Update unshared means by responsibility-weighted shrinkage
7: Inherit weights and covariances by default; optional weights update jointly, while covariance updates require sufficient component counts
8: **end for**
9: Build the adapted factor-wise score field from $\{\eta_k^T\}_{k=1}^K$
10: Execute the analytic sampler or specified carrier-editing rule; recompose and decode

Ensure: Raw adapted samples or image-conditioned transfers

Algorithm 1 organizes SUDA-MoG into factor construction, selective calibration, and generation.

F.4 Prompting and Quantitative Evaluation

Quantitative evaluations use raw outputs before optional filtering or refinement. Each comparison fixes its evaluation phrases and output counts. Image-conditioned transfer produces one output per held-out source image. The evaluation phrase names the target class or attribute: cat for Dog→Cat, the corresponding CelebA attribute, and old age for Young→Old. Appendix H specifies the retained experiments' evaluators, data splits, and generation settings.

F.5 Feature-Block Diagnostics

We characterize how task information is distributed across the encoded feature blocks. Each latent provides channel statistics $s = (\mu, \sigma)$ and a normalized residual $\bar{z}$. For Table 5, linear classifiers are fitted on the training coordinates and evaluated on held-out images. Species is the binary Dog/Cat label; color uses the predefined bins of the color preservation diagnostic. The majority baselines are 50% and 30.8%, respectively. Species information is more readable from the residual, while color information is more readable from channel statistics. This pattern motivates the role allocation in Greedy.

Table 5: **Task information across latent feature blocks.** Held-out linear-probe accuracy (%).

Feature block	Species	Color
Channel statistics (μ, σ)	58.40	59.04
Normalized residual $\bar{z}$	79.40	31.37

Figure 5 provides projection views of these candidate feature blocks. The style panels show channel statistics and the semantic panels show normalized residuals. The displayed color contrast is Black/White; the table reports the feature-block probe on its predefined color bins.

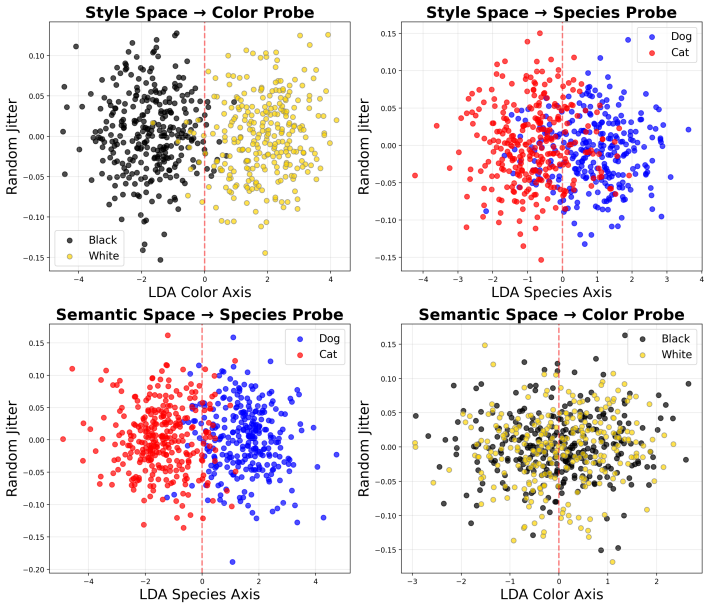


Figure 5: **Attribute structure in candidate feature blocks.** Horizontal coordinates are task-specific LDA projections; vertical jitter separates points for visualization. The panels display species and Black/White color labels in the two feature blocks.

G MEAN-ONLY CALIBRATION: A SPECIAL CASE AND ITS EXECUTION

All statements in this section condition on the chart, task roles, and fitted source mixture parameters. The target calibration examples are independent of this construction. Coordinates within an example may be dependent; independence is required across examples.

G.1 CONCENTRATION OF THE SHRINKAGE UPDATE

Lemma G.1 (Weighted target-mean concentration). *Let $X_1, \dots, X_n$ be iid vectors in $[-R, R]^d$, let $r : \mathbb{R}^d \to [0, 1]$ be fixed, and define $\alpha = \mathbb{E}r(X) > 0$, $\mu^\circ = \mathbb{E}[r(X)X]/\alpha$, and $N = \sum_i r(X_i)$. For every $t > 0$, with probability at least $1 - 2de^{-t}$,*

$$\left\| \sum_i r(X_i)(X_i - \mu^\circ) \right\|_2 \le R\sqrt{d}\big(\sqrt{8n\alpha t} + 4t/3\big).$$

Moreover, if $n\alpha \ge 8t$, then $\Pr(N < n\alpha/2) \le e^{-t}$.

Proof. The coordinates of μ° lie in $[-R, R]$ because it is an average with nonnegative normalized weights. For coordinate h, set $Y_{i,h} = r(X_i)(X_{i,h} - \mu^\circ_h)$. Then $\mathbb{E}Y_{i,h} = 0$, $|Y_{i,h}| \le 2R$, and $\mathbb{E}Y_{i,h}^2 \le 4R^2\mathbb{E}r(X_i)^2 \le 4R^2\alpha$. The two-sided Bernstein inequality bounds $|\sum_i Y_{i,h}|$ by $R\sqrt{8n\alpha t} + 4Rt/3$ except on an event of probability $2e^{-t}$. A union bound over the d coordinates and $\|v\|_2 \le \sqrt{d}\|v\|_\infty$ give the first statement. The multiplicative Chernoff bound for independent variables in $[0, 1]$ gives $\Pr(N < n\alpha/2) \le \exp(-n\alpha/8)$, proving the second. □

Lemma G.2 (Coupling mixtures with inherited covariances). *For two mixtures $p = \sum_j w_j\mathcal{N}(\mu_j, \Sigma_j)$ and $q = \sum_j w_j\mathcal{N}(\nu_j, \Sigma_j)$,*

$$W_2^2(p, q) \le \sum_j w_j\|\mu_j - \nu_j\|_2^2.$$

Proof. Draw the same component $J \sim w$ and noise $\xi \mid J \sim \mathcal{N}(0, \Sigma_J)$ for both variables. The pair $(\mu_J + \xi, \nu_J + \xi)$ has marginals p, q and squared cost $\mathbb{E}\|\mu_J - \nu_J\|_2^2$. The infimum over all couplings is no larger. □

This special case updates active means by source-responsibility shrinkage while inheriting weights and covariances. It uses a different estimator from the finite-grid DSM result. Condition on a chart, role assignment, and source MoGs constructed independently of the n i.i.d. target calibration examples. For active factor k, write its coordinates as $X_{k,i} \in \mathbb{R}^{d_k}$, its source responsibilities as $r^S_{k,j}$, and $\tau = \tau_\mu > 0$. The update is

$$N_{k,j} = \sum_{i=1}^n r^S_{k,j}(X_{k,i}), \qquad \widehat{\mu}_{k,j} = \frac{\tau\mu^S_{k,j} + \sum_{i=1}^n r^S_{k,j}(X_{k,i})X_{k,i}}{\tau + N_{k,j}}. \tag{33}$$

Here $N_{k,j}$ is the target responsibility mass assigned to source component j, while τ controls shrinkage toward its source mean. Components receiving more target mass place more weight on their responsibility-weighted target means. Source weights and covariances are inherited. Define the population responsibility mass and its associated target mean by

$$\alpha_{k,j} = \mathbb{E}_T r^S_{k,j}(X_k) > 0, \qquad \mu^\circ_{k,j} = \frac{\mathbb{E}_T[r^S_{k,j}(X_k)X_k]}{\alpha_{k,j}}. \tag{34}$$

The population estimand $\mu^\circ_{k,j}$ is the target mean weighted by the fixed source responsibilities. Calibration therefore estimates a source-indexed target statistic. The product law p° uses these population means on active factors and inherits all remaining source statistics. Its calibration target is the target-data statistic in Eq. (34); predicting the endpoint law of a trained adapter requires an additional correspondence.

Corollary G.3 (Mean-only parameter calibration)**.** *Suppose* $|(X_k)_h| \le R_k$ *almost surely for every active coordinate. Set* $M = \sum_{k\in\mathcal{I}_{\mathrm{un}}} J_k(d_k+1)$ *and* $t = \log(4M/\delta)$*. With probability at least* $1-\delta$*, simultaneously,*

$$\|\widehat{\mu}_{k,j} - \mu^{\circ}_{k,j}\|_2 \le b_{k,j} := \frac{\tau\|\mu^S_{k,j} - \mu^{\circ}_{k,j}\|_2 + R_k\sqrt{d_k}\big(\sqrt{8n\alpha_{k,j}t} + 4t/3\big)}{\tau + N_{k,j}}. \tag{35}$$

If recomposition followed by decoding is L*-Lipschitz on factor space, the decoded exact product-mixture law* $G_{\widehat{\mu}}$ *satisfies*

$$W_2(G_{\widehat{\mu}}, P_T) \le W_2(G^{\circ}, P_T) + L\Big(\sum_{k\in\mathcal{I}_{\mathrm{un}}}\sum_{j=1}^{J_k} w^S_{k,j} b^2_{k,j}\Big)^{1/2}, \tag{36}$$

where G° *is the decoded law of* p°*.*

The numerator in Eq. (35) separates shrinkage bias from target-sampling error. The bound remains valid at small effective counts because its denominator is the observed $\tau + N_{k,j}$. When $n\alpha_{k,j} \ge 8t$, $N_{k,j} \ge n\alpha_{k,j}/2$ also holds on the same event. The sampling term then scales as $R_k\sqrt{d_k t/(n\alpha_{k,j})}$, with a lower-order $R_k\sqrt{d_k}t/(n\alpha_{k,j})$ term. Only recalibrated factors contribute to the finite-sample estimation term. The population discrepancy $W_2(G^{\circ}, P_T)$ captures the combined effect of the product-model approximation, fixed responsibilities, inherited statistics, and decoding. For deterministic edits, Corollary G.6 gives the corresponding decomposition around the population edited carrier law, with component weights $\beta_{k,j} = \mathbb{E}_Q r^S_{k,j}(X_k)$. It permits dependent carrier factors when Q is fixed independently of calibration and has finite second moments. The proof below combines the preceding concentration and mixture-coupling lemmas; numerical stabilization is addressed afterwards.

Proof of Corollary G.3. Subtracting Eq. (34) from Eq. (33) yields the exact identity

$$\widehat{\mu}_{k,j} - \mu^{\circ}_{k,j} = \frac{\tau(\mu^S_{k,j} - \mu^{\circ}_{k,j}) + \sum_i r^S_{k,j}(X_{k,i})(X_{k,i} - \mu^{\circ}_{k,j})}{\tau + N_{k,j}}.$$

Apply Lemma G.1 to every active component. The sum of the failure probabilities for all numerator and eligible count bounds is at most $\sum_{k,j}(2d_k+1)e^{-t} \le 2Me^{-t} \le \delta$. The triangle inequality gives Eq. (35). Couple the active-factor mixtures using Lemma G.2, independently across factors, and couple each shared factor identically. The resulting product-law squared transport cost is at most $\sum_{k,j} w^S_{k,j} b^2_{k,j}$ over active factors. A Lipschitz map multiplies W_2 by at most L. Finally, the triangle inequality through G° gives Eq. (36). □

Numerical stabilization and sparse components. Equation (33) is well-defined even at $N = 0$ because $\tau > 0$. The implementation form in Appendix F uses $\bar{\mu} = \sum_i r(X_i)X_i/(N+\epsilon_0)$ before shrinkage. For this version the mean-error bound gains at most

$$e = \frac{\epsilon_0 NR\sqrt{d}}{(N+\epsilon_0)(\tau+N)}.$$

This follows by subtracting the two updates and using $\|\sum_i r(X_i)X_i\|_2 \le NR\sqrt{d}$. If a numerical threshold freezes a component at its source mean instead, its bound is simply $\|\mu^S - \mu^{\circ}\|_2$. These component bounds substitute directly into the same mixture coupling. The corollary concerns mean recalibration with inherited weights and covariances; the optional covariance updates use a different estimator.

G.2 An exact product chart

Proposition G.4 (Fixed affine factor coordinates)**.** *Let* P *be the orthogonal projection that subtracts each channel's spatial mean,* $B = I - P$*, and* S *a fixed positive channelwise scaling operator. For a unit* $a \in \mathrm{range}(P)$*, define*

$$s = Bz, \quad q = S^{-1}Pz, \quad u = a^{\top}q, \quad v = (I - aa^{\top})q.$$

This is a linear bijection from the latent space to $\mathrm{range}(B) \times \mathbb{R} \times (\mathrm{range}(P) \cap a^{\perp})$*, with inverse* $z = s + S(au + v)$*. Changing* u *alone preserves* s *and* v *under re-extraction, and* $\|z' - z\|_2 = \|Sa\|_2|u' - u|$*.*

Proof. Channelwise scaling commutes with P and B. Hence q, a, v are in $\mathrm{range}(P)$, $a^\top v = 0$, and $q = au + v$. Applying B and $S^{-1}P$ to $s + S(au + v)$ recovers s and $au + v$; projecting the latter onto a and $a^\perp$ recovers u, v. Conversely $Bz + Pz = z$. The perturbation identity follows from $z' - z = Sa(u' - u)$. □

Here shared coordinate means are invariant by construction. Gaussian factor laws use intrinsic orthonormal coordinates on the displayed linear subspaces. The statistical product prior remains a modeling choice, and external image attributes are assessed by the probes and intervention experiments. The normalized moment chart used in the role and LoRA tests is specified separately in Appendix F.

G.3 Component transport and deterministic carrier edits

Proposition G.5 (Posterior component transport)**.** *Draw $X \sim p^S = \sum_j w_j^S \mathcal{N}(\mu_j^S, \Sigma_j^S)$ independently of the target calibration samples. Draw $J \mid X \sim r^S(X)$ and set $X' = X + \widehat{\mu}_J - \mu_J^S$. Conditional on the fitted target means, $X' \sim \sum_j w_j^S \mathcal{N}(\widehat{\mu}_j, \Sigma_j^S)$.*

Proof. Bayes' rule gives $\Pr(X \in dx, J = j) = w_j^S \mathcal{N}(dx; \mu_j^S, \Sigma_j^S)$. Conditioning on $J = j$ and translating by $\widehat{\mu}_j - \mu_j^S$ yields $\mathcal{N}(\widehat{\mu}_j, \Sigma_j^S)$; summing over j proves the claim. □

Applied independently to active factors of a source product-law draw, this operator samples the fitted target law while retaining every shared coordinate. In the affine chart, these coordinates are also preserved upon re-extraction. The deterministic soft shift is its conditional mean: $T_{\mathrm{soft}}(X) = \mathbb{E}[X' \mid X, \widehat{\mu}]$. The law of total covariance gives

$$\mathrm{Cov}(X' \mid \widehat{\mu}) - \mathrm{Cov}(T_{\mathrm{soft}}(X) \mid \widehat{\mu}) = \mathbb{E}[\mathrm{Cov}(X' \mid X, \widehat{\mu}) \mid \widehat{\mu}] \succeq 0.$$

For source-model inputs, this identity quantifies the variation contributed by posterior component draws. The deterministic update uses their conditional mean as one edit per carrier; the following corollary analyzes its decoded output law.

Corollary G.6 (Calibration error for deterministic carrier edits)**.** *Under Corollary G.3, let Q be a fixed joint law of carrier factors, independent of calibration, with finite second moments. On active factors define $T_{\mu,k}(x) = x_k + \sum_j r_{k,j}^S(x_k)(\mu_{k,j} - \mu_{k,j}^S)$, and leave shared factors unchanged. Set $\beta_{k,j} = \mathbb{E}_Q r_{k,j}^S(X_k)$ and let $G_\mu^{\mathrm{soft},Q}$ be the law after applying T_μ and the L-Lipschitz recomposition–decoder map. On the same event as the calibration corollary,*

$$W_2(G_{\widehat{\mu}}^{\mathrm{soft},Q}, P_T) \le W_2(G_{\mu^\circ}^{\mathrm{soft},Q}, P_T) + L\Big(\sum_{k \in \mathcal{I}_{\mathrm{un}}, j} \beta_{k,j} b_{k,j}^2\Big)^{1/2}.$$

Proof. Couple the two edits using the same carrier $X \sim Q$. Jensen's inequality and $\sum_j r_{k,j}^S = 1$ give $\|T_{\widehat{\mu},k}(X) - T_{\mu^\circ,k}(X)\|_2^2 \le \sum_j r_{k,j}^S(X_k) \|\widehat{\mu}_{k,j} - \mu_{k,j}^\circ\|_2^2$. Sum over active blocks, take expectation, and apply the simultaneous mean bounds. Lipschitz recomposition and the triangle inequality yield the result. □

The bound separates the population editing discrepancy from finite-support calibration error, weighted by the carrier responsibilities $\beta_{k,j}$.

H Protocols for Functional and Supplemental Evidence

The protocols below specify fitting and evaluation for the reported experiments. Matched-adapter interventions hold the discovered map fixed; the 32-support editing experiment jointly fits the map, conditional functions, and factor laws on each support set.

H.1 ADDITIONAL QUALITATIVE EDITS

Figure 6 shows paired source and edited images for age, style, structure, and color changes using Greedy SUDA-MoG. The Young→Old role comparison in Table 3 complements the age examples: U, S, and All obtain CLIP-T scores of 0.2459, 0.2243, and 0.2424, respectively. Their target accuracies are 0.767, 0.167, and 0.600, with LAB similarities 0.795, 0.912, and 0.782. These are task-level role-ablation summaries; the panels provide qualitative examples of the corresponding editing tasks.

(a) Young→Old



(b) Realism→Impressionism



(c) Structure and color editing



(d) Color editing



Figure 6: **Additional qualitative edits with Greedy SUDA-MoG.** Each source row is followed by its edited row. (a) Age editing. (b) Style transfer. (c) Joint structure and color editing. (d) Color editing. Table 3 provides the task-level Young→Old role comparison.

H.2 MATCHED REAL-LORA INTERVENTION

For matched base and LoRA latents z_i^0, z_i^L, the frozen rank-one moment setting of Greedy gives $\Phi(z) = (s, u, q_\perp)$. Define

$$z_i^u = R_{\text{greedy}}(s_i^0, u_i^L, q_{\perp,i}^0),$$
$$R_u = \frac{\sum_i [g(z_i^u) - g(z_i^0)]}{\sum_i [g(z_i^L) - g(z_i^0)]}.$$

The independent latent target evaluator g has held-out AUC 0.9979. Recovery is a ratio of aggregate gains. Factor fitting and LoRA training reuse the same ten raw target supports. The standardized logistic-regression evaluator trains on 96 dog and 95 cat images, with a held-out test of 128 images per class. Its training set excludes discovery images, supports, carriers, and test images. All headline and rotation values use this evaluator with the factor and adapter outputs frozen. Each matched pair shares source image, target-neutral prompt, scheduler, and noise seed. Ranks 1 and 16 and three seeds yield six groups and 240 pairs; reported gain SD is over these six groups: retained gain is 0.04980 ± 0.03037 versus 0.11670 ± 0.06028 for full LoRA. The fitted direction $a \in \mathbb{R}^{2048}$ defines the scalar coordinate $u = a^\top q \in \mathbb{R}$. The active-block dimension is therefore one; a specifies its extraction from the normalized latent. Recovery measures this coordinate's contribution to g on the matched cohort.

A 5000-replicate crossed bootstrap keeps the two ranks as fixed conditions and resamples the three paired training-seed indices and 40 common carrier indices independently. Each sampled seed and carrier is used across both ranks, preserving all paired comparisons. Recomputing the ratio of aggregate gains gives a descriptive 95% interval for R_u of $[0.2970, 0.6080]$, conditional on the frozen chart and evaluator.

The correct-label direction exceeds all 99 shuffled-label controls (null mean/maximum $0.00054/0.00468$). The empirical randomization value is $(1 + \#\{\text{null} \geq \text{observed}\})/100 = 0.01$. Rotation keeps the LoRA model, evaluator, and downstream protocol fixed. Table 4 reports recovery at four factor alignments.

Matched adapter training. The Dog-to-Cat adapters use ten raw target images, ranks 1 and 16, and seeds 9–11. They use Stable Diffusion 1.5 with LoRA on UNet attention query and value projections, scaling parameter equal to rank, zero dropout, and no bias. Training uses 300 AdamW updates, learning rate 10^{-4}, batch size four, noise-prediction MSE, and gradient clipping at 1.0. Images are resized and center-cropped to 512 pixels. The training prompt is "a photo of a zxq animal". The matched outputs use DDIM with 50 steps, guidance 7.5, and strength .75.

H.3 CONDITIONAL RECONSTRUCTION AND COORDINATE PRESERVATION

The conditional Greedy experiments use an invertible base chart $\Psi(z) = (m, x)$ and an orthogonal basis $[B_r\ B_{r,\perp}]$. Their factor map and inverse are

$$\begin{aligned} u = B_r^\top x, \quad v = B_{r,\perp}^\top x, \quad \Phi(z) = (\widetilde{m}, u, \widetilde{v}), \\ \widetilde{m} = m - c(u), \qquad \widetilde{v} = \Lambda(u)^{-1}[v - a(u)]. \end{aligned} \tag{37}$$

$$\Phi^{-1}(\widetilde{m}, u, \widetilde{v}) = \Psi^{-1}(\widetilde{m} + c(u),\ B_r u + B_{r,\perp}[a(u) + \Lambda(u)\widetilde{v}])\,. \tag{38}$$

Here c, a are fitted conditional locations and $\Lambda(u)$ has positive diagonal entries. The active block is u; the retained blocks are $(\widetilde{m}, \widetilde{v})$. Fixed completion holds m, v at their carrier values while using the same edited u.

Lemma H.1 (Conditional coordinate reconstruction)**.** *Suppose Ψ is a bijection onto its base coordinate domain, $[B_r\ B_{r,\perp}]$ is orthogonal, and every diagonal entry of $\Lambda(u)$ is positive. The map in Eq.* (37) *is a bijection onto its image with inverse Eq.* (38)*. For every active replacement u' whose reconstructed base coordinates lie in that domain, reconstruction and re-extraction return $(\widetilde{m}, u', \widetilde{v})$.*

Proof. The residual coordinates uniquely recover $m = \widetilde{m} + c(u)$ and $v = a(u) + \Lambda(u)\widetilde{v}$. Orthogonality recovers $x = B_r u + B_{r,\perp} v$, and Ψ^{-1} then recovers z. Substitution into the extraction equations returns all three blocks, including after an admissible replacement of u by u'. □

This identity establishes preservation of the chosen coordinates. External-observation preservation is measured through $\mathcal{S}_{i,j}(\Phi)$ and the intervention experiments; the product law specifies the statistical approximation on these coordinates.

Corollary H.2 (Completion at a fixed active endpoint)**.** *Under Lemma H.1, let* $(m, x) = \Psi(z)$, $u = B_r^\top x$, $v = B_{r,\perp}^\top x$, *and let* $(\widetilde{m}, \widetilde{v})$ *be the residual coordinates in Eq.* (37)*. For a replacement* u'*, define the fixed-complement and conditional reconstructions in the base chart by*

$$(m_F, x_F) = (m,\ B_r u' + B_{r,\perp} v), \tag{39}$$

$$(m_C, x_C) = (\widetilde{m} + c(u'),\ B_r u' + B_{r,\perp}[a(u') + \Lambda(u')\widetilde{v}]). \tag{40}$$

Assume both (m_F, x_F) *and* (m_C, x_C) *belong to the base coordinate domain. Then* $\Psi^{-1}(m_F, x_F)$ *and* $\Psi^{-1}(m_C, x_C)$ *have active coordinate* u'*. Their difference in the base chart is*

$$m_C - m_F = c(u') - c(u), \tag{41}$$

$$x_C - x_F = B_{r,\perp}[a(u') - a(u) + (\Lambda(u') - \Lambda(u))\widetilde{v}]\,. \tag{42}$$

The conditional reconstruction also retains $(\widetilde{m}, \widetilde{v})$ *under re-extraction.*

Proof. Substitute $m = \widetilde{m} + c(u)$ and $v = a(u) + \Lambda(u)\widetilde{v}$ into the two reconstructions and subtract. Orthogonality gives $B_r^\top x_C = B_r^\top x_F = u'$, and Lemma H.1 gives residual preservation. □

Relative to the original carrier, the conditional change in x adds $B_r(u' - u)$ to Eq. (42); its change in m is Eq. (41). Thus a low-dimensional control can induce changes throughout the complementary coordinates. The comparison at fixed u' isolates their joint completion contribution; decoded task effects and image preservation are measured separately.

H.4 Conditional chart construction from supports

Base chart and active directions. For each latent channel, an orthogonal DCT separates the DC coefficient from the remaining coefficients y. We apply $\rho(y) = \operatorname{asinh}(\|y\|)y/\|y\|$, with $\rho(0) = 0$, followed by a training-fitted orthogonal rotation. The DC and final rotated coordinate form m; the remaining coordinates, standardized using training means and positive scales, form x. The inverse uses $\rho^{-1}(w) = \sinh(\|w\|)w/\|w\|$, with $\rho^{-1}(0) = 0$, and the inverse rotation and DCT. Let W_d contain the fitted PCA directions. For each active attribute, we fit the linear-logistic logit $f_k(x) = \beta_k^\top W_d^\top x + b_k$. The rows of G are $(W_d\beta_k)^\top/\|W_d\beta_k\|$, and its leading right singular vectors define B_r. PCA restricts direction discovery; the complete orthogonal complement is retained in the chart. The 32-support construction uses $d = 31$ and active rank $r = 3$, with fitting parameters specified in Appendix H.5.

Conditional residual coordinates. Write $h(u) = (1, u^\top)^\top$. Ridge regression gives $c(u) = Ch(u)$ and a preliminary structural residual $v - Ah(u)$. For each binary active-attribute state b, we estimate its residual mean μ_b and coordinatewise standard deviation σ_b, floored at 0.10. Additional linear-logistic heads on u give probabilities $\pi_k(u)$ clipped to $[.01, .99]$. The fitted functions are

$$\omega_b(u) = \prod_k \pi_k(u)^{b_k}[1 - \pi_k(u)]^{1-b_k}, \tag{43}$$

$$a(u) = Ah(u) + \sum_b \omega_b(u)\mu_b, \tag{44}$$

$$\Lambda(u) = \operatorname{diag}\left[\exp\left(\sum_b \omega_b(u)\log\sigma_b\right)\right]. \tag{45}$$

Positivity of $\Lambda(u)$ makes each residual transformation invertible. Reconstructing $m = \widetilde{m} + c(u)$ and $v = a(u) + \Lambda(u)\widetilde{v}$, followed by $x = B_r u + B_{r,\perp} v$ and Ψ^{-1}, gives Eq. (38). The nonlinear dependence of the reconstructed latent on u comes from this conditional map and the radial base chart. The fitted discovery logits have constant input gradients. Probability clipping makes the implemented chart a continuous, piecewise-smooth bijection; the differentiable analyses use their stated regularity assumptions.

H.5 COMPLETE CONDITIONAL ADAPTATION FROM 32 SUPPORTS

Supports and fitted state. Each of ten fixed CelebA support draws contains 32 distinct labeled images, two per state of (Smile, Male, Bangs, Mouth-open). The draws contain 320 distinct images in total; all forty evaluation carriers are support-disjoint and are shared across draws. Supports are supplied using existing dataset annotations, so this budget counts the images provided for fitting. Results aggregate all ten method-development draws and all forty carriers. The DCT/radial chart, standardization, 31-dimensional PCA, task heads, conditional regressions, gates, residual moments and factor laws are all fitted using only the 32 supports of that draw. External evaluation heads are trained separately.

The chart uses the construction in Appendix H.4, with rank three fixed by the three specified active tasks. Normalized logistic-head gradients define the complete active block; the heads use balanced logistic regression with $C = .001$. Conditional regressions use ridge penalty 10 and conditional gates use $C = 1$. State-specific residual moments shrink toward the pooled support moments with strength 32, including the between-mean variance term; conditional scales are floored at 0.10. Source state 000 and target state 111 each contain four supports after pooling both Male values. In standardized active coordinates, a one-component Gaussian law uses their means, a target-minus-source mean shrinkage factor $4/(4+4)$, and common covariance $0.5\widehat{\Sigma}_{32} + 0.5I + 10^{-4}I$. This is a specified single-component ($J = 1$) instance of the MoG family.

Update and controls. Let $F(u) = \Phi^{-1}(\widetilde{m}, u, \widetilde{v})$ for a carrier's retained residuals, $J = \partial F/\partial u$, and $\ell(u) = \log p_T(u) - \log p_S(u)$. Each of at most 32 steps follows $[J^\top J + 10^{-3}\,\mathrm{tr}(J^\top J)I/3]^{-1}\nabla_u \ell$, normalized to a predicted latent displacement of $0.15\|z^0\|/32$. Backtracking factors $1, 1/2, \ldots, 1/512$ accept only objective increases with $\|F(u) - z^0\|/\|z^0\| \le 0.15$; an infeasible step leaves the state unchanged. Optimization uses the fitted source and target laws; external heads are reserved for evaluation. The reported arms are paired base, conditional edit, and a jointly shuffled-label conditional control. The shuffled control refits every supervised component from the same supports. The 0.15 displacement bound applies to the conditional arm. Float32 coordinate-closure error is at most 1.87×10^{-7}.

Evaluation and uncertainty. OpenAI CLIP ViT-B/32 uses one fixed target phrase per attribute. The independent HOG-logistic evaluators use 512 training and 512 test images, disjoint from adaptation fitting and carriers. Features use bilinear 128×128 grayscale images, nine orientation bins, 8×8 cells, 2×2 blocks, L2-Hys normalization, training-set standardization, and balanced logistic regression ($C = .2$). HOG scores continuous clamped pixels; CLIP, DINO ViT-S/8 and LPIPS-Alex score the saved PNGs. Gains are edit minus paired base. We use 20,000 crossed bootstrap replicates (seed 20260926), resampling the ten support-draw indices and forty carrier indices independently and applying each sample to every arm and attribute. The 95% percentile intervals describe variation within this evaluated cohort. Figure 4(c) gives the mean $\pm$ sample SD of the ten carrier-averaged draw gains.

Table 6: **Three-attribute editing from 32 supports per fit.** Paired gains over the unedited carrier across all ten support draws and forty carriers; brackets give 95% crossed draw/carrier bootstrap intervals. Male is not an editing target.

Attribute	ΔCLIP-T [95% CI]	ΔHOG probability [95% CI]
Smile	$+0.00283$ $[+0.00132, +0.00436]$	$+0.06192$ $[+0.02553, +0.10633]$
Bangs	$+0.00246$ $[+0.00079, +0.00415]$	$+0.05539$ $[+0.02611, +0.09088]$
Mouth-open	$+0.00160$ $[+0.00005, +0.00314]$	$+0.05304$ $[+0.01153, +0.09713]$
Male (untargeted)	-0.00002 $[-0.00101, +0.00101]$	$+0.00667$ $[-0.01070, +0.02710]$

All three active attributes have positive CLIP gains with 95% interval lower bounds above zero, and each mean HOG gain exceeds 0.05. The macro CLIP gain exceeds the shuffled-label control by $+0.002877$ $[0.001629, 0.004087]$, while untargeted Male absolute HOG drift has an interval upper bound below 0.05. These results meet the four specified criteria for joint attribute response and untargeted-attribute stability.

Table 7: **Shuffled-label control for the 32-support CelebA edit.** Mean gains relative to paired base over all ten draws and forty carriers. Correct-label gains and intervals are in Table 6.

Attribute	ΔCLIP-T	ΔHOG probability
Smile	-0.00111	-0.00243
Bangs	-0.00015	$+0.00796$
Mouth-open	-0.00048	$+0.02913$
Male (untargeted)	$+0.00024$	-0.00305

Conditional DINO similarity is 0.97692 $[0.97362, 0.97990]$ and LPIPS is 0.03857 $[0.03633, 0.04087]$. Untargeted Male absolute HOG drift is 0.02574 $[0.00996, 0.04662]$; its signed response appears in Table 6. The shuffled-label control gives DINO similarity 0.97609 and LPIPS 0.03790. Before clamping, out-of-range pixel fractions are 0.01150 for the conditional edit and 0.01412 for the shuffled control, versus 0.00632 for base. Visual changes are subtle and vary across carriers.

H.6 Direct DSM coupling

For coordinate c, the joint variable is $y = (s_c, u)$ and its model is $p_\theta(y) = \sum_{j=1}^{2} w_j \mathcal{N}(y; (a_j, b_j), \Sigma_j)$. Weights and full covariances are fitted on source calibration data and then fixed; $\theta_{\mathrm{sh}} = (a_1, a_2)$ and $\theta_{\mathrm{un}} = (b_1, b_2)$ are optimized by DSM. At each fixed noise level, $y_t = \alpha_t y + \sigma_t \epsilon$ with $\sigma_t^2 = 1 - \alpha_t^2$, and the loss is $\frac{1}{2}\mathbb{E}\| - \sigma_t \nabla \log p_{\theta,t}(y_t) - \epsilon\|^2$. For a common displacement d in u, the two unshared means shift by d, so $\Delta_{\mathrm{un}} = \sqrt{2}|d|$. The calibrated directional predictor is $\kappa_{\mathrm{cal}}\Delta_{\mathrm{un}}$, where $\kappa_{\mathrm{cal}} = \|H^{\mathrm{cal}}_{\mathrm{sh,un}}(1,1)^\top/\sqrt{2}\|_2$.

We fit 64 separate two-dimensional joint DSM problems, pairing each channel mean or standard deviation with the active coordinate. AFHQ/DC-AE calibration and confirmation splits are disjoint. A calibrated coupling predictor achieves mean within-shift held-out Spearman 0.551, and high-coupling coordinates yield $2.570\times$ larger shared gradients at matched displacement.

I Broader Impact

This work studies few-shot adaptation of diffusion models through a measurement-driven analysis of what latent statistics change during personalization. Potential positive impacts include more efficient personalization, lower adaptation cost, and improved interpretability of generative-model adaptation by making the adapted latent statistics explicit.

The same capabilities also raise potential risks. Because the framework can be applied to facial attributes and subject-level personalization, misuse could include misleading personalized content, unauthorized attribute editing, or deepfake-style manipulation. The method also inherits biases and failure modes from the pretrained backbones and datasets used to construct the frozen latent space. SUDA-MoG provides an analysis and diagnostic framework ; deployment requires application-specific validation . Responsible use should follow dataset and model licenses, obtain consent for subject-specific personalization, and combine generated-content release with appropriate safeguards such as watermarking, provenance tracking, detection tools, or access control when used in sensitive settings.