LatentCompass: T2I Diffusion Steering via Orthogonal Attribute Spaces for Debiasing, Concept Erasure, and Red Teaming

Anonymous authors
Paper under double-blind review

Abstract

Text-to-image (T2I) diffusion models suffer from biased results stemming from entangled generative priors and a lack of accurate control over outputs. Current mitigation attempts rely on imprecise, adversarially-vulnerable prompt and text embedding interventions, or they require prohibitive and invasive fine-tuning. Further, text-based methods can only control *descriptive* attributes, i.e., what an image depicts, but not *evaluative* attributes, i.e., how it is perceived by an external judge. We propose LatentCompass, an exemplar-based approach that enables disentangled and controllable generation for both descriptive and evaluative concepts in a training-free manner. LatentCompass steers the generative trajectory by **(a)** constructing a nonlinear, low-dimensional, and orthogonal attribute space via a closed-form solution that explicitly isolates desired concepts, **(b)** defining a target-class intervention in this space, and **(c)** reflecting the corresponding intervention directly in the T2I latent space. Extensive evaluations demonstrate that LatentCompass on average **(i)** mitigates generative stereotypes in medical professions by $\mathbf{100}\%$, **(ii)** reduces unsafe concept generation to $\mathbf{1.3}\%$, **(iii)** enhances aesthetic score by $\mathbf{27}\%$, **(iv)** boosts red-teaming success rates against Deepfake detectors by $\mathbf{40}\%$, and **(v)** enables high-fidelity attribute editing without leakage. *CAUTION: This paper includes content that may be inappropriate or offensive.*

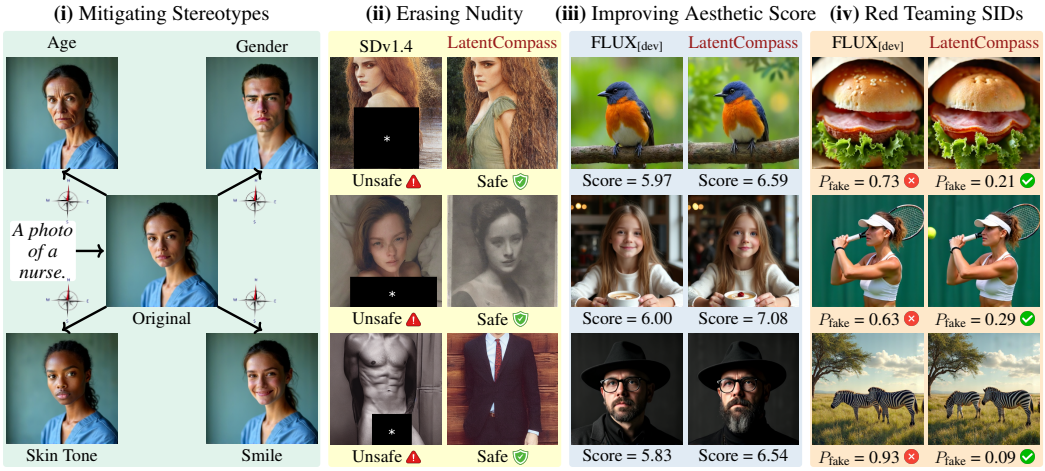


Figure 1: **Controlling descriptive and evaluative concepts in T2I models**: **(i)** Facial attribute steering for debiasing profession generation, **(ii)** improving content safety by nudity erasure, **(iii)** enhancing aesthetic score of synthesized images, **(iv)** red teaming, by finding examples that can deceive synthetic image detectors (SID).

1 Introduction

Controllable generation is essential for overcoming the entangled priors of T2I models, which arise from correlations in their training data (Cho et al., 2023; Friedrich et al., 2023; Naik & Nushi, 2023; Zhang et al., 2023). These entangled priors bias generation toward stereotypical attribute combinations, neglect rare joint conditions, and hinder faithful adherence to user intentions (Zhang

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.

et al., 2023; Dehdashtian et al., 2025b; D'Incà et al., 2024; Vázquez & Garrido-Merchán, 2024). Effective control should override these default priors to synthesize targeted distributions, improve prompt faithfulness, and suppress explicit, harmful, or copyrighted concepts.

A practical control mechanism must satisfy three key requirements. **(R1) Attribute generality**: it must handle both *descriptive attributes*, which are easily captured by natural language (e.g., demographic traits or artistic style), and *evaluative attributes*, which are judge-dependent properties that defy text prompting entirely, such as human aesthetic preference or the "fakeness" boundary learned by a deepfake detector (examples shown in Fig. 1). **(R2) Disentanglement**: modulating a target attribute must leave all others unchanged. **(R3) Efficiency and extensibility**: the mechanism must operate directly on a frozen, pre-trained T2I model without any weight updates, so that new control concepts can be introduced incrementally and at minimal computational cost.

Existing approaches for controlling T2I generation fall into four families. **(1) Training and fine-tuning based methods** (Heng & Soh, 2023; Kumari et al., 2023; Li et al., 2024; Lu et al., 2024; Gandikota et al., 2023) directly optimize model weights to instill new concepts, but their computational cost and lack of modularity do not support R3. **(2) Prompt-based methods** (Zhang et al., 2023; Ramesh et al., 2022; Nguyen et al., 2025; Woolf, 2022) steer generation through text alone, but fail on two fronts: training-data correlations cause models to ignore exact instructions; and evaluative concepts such as perceived aesthetic quality or realness have no natural linguistic expression, making R1 unachievable by construction. **(3) Text-encoder interventions** (Yoon et al., 2025; Ban et al., 2024) operate directly on text embeddings, but remain vulnerable to adversarial prompts that implicitly smuggle in unintended concepts, undermining disentanglement and violating R2. **(4) Classifier guidance** (Dhariwal & Nichol, 2021; Bansal et al., 2023) steers the generative trajectory using gradients from an external classifier, but requires a separate classifier per concept group and produces attribute leakage due to non-orthogonal guidance directions, violating R2.

We propose LatentCompass, an exemplar-based method that satisfies R1–3 by decomposing controllable generation into three subproblems. **(P1)** Given a set of generated images labeled with attributes of interest, LatentCompass constructs, in *closed-form* (satisfies R3), a nonlinear attribute space where each axis encodes an independent control attribute such as gender, style, perceived aesthetic, and content safety, handling both descriptive and evaluative attributes (satisfies R1). Crucially, the axes are explicitly constructed to be **mutually orthogonal**, turning a high-dimensional entangled representation into a compact semantic *compass*; movement along one axis isolates that attribute while minimizing interference with the rest (satisfies R2). **(P2)** Utilizing this geometric framework, the intermediate representation of the image is projected onto the attribute space, where LatentCompass defines a likelihood-based target-class intervention. **(P3)** Finally, this intervention is realized directly in the T2I latent trajectory by solving the pre-image problem, ensuring the final output adheres to the desired control attributes. (satisfies R2).

Applications of LatentCompass include steering facial attributes to mitigate stereotypical biases concerning perceived gender and skin tone in T2I generation of medical professions (§ 4.1), improving content safety by erasing toxic concepts such as nudity (§ 4.3), steering images to better match human aesthetic preferences (§ 4.5), and red teaming of synthetic image detectors (SID) by shifting images toward their failure modes (§ 4.4). We show examples of these applications in Fig. 1. Due to its exact solution, LatentCompass can efficiently incorporate additional concepts and applications at little cost.

Summary of Contributions.

- **Exemplar-Based Steering Framework.** We propose LatentCompass, an exemplar-based approach for steering T2I diffusion models. By deriving a *nonlinear and orthogonal* attribute space, our method enables precise control over both *descriptive* and *evaluative* semantics **(R1)** while ensuring *disentangled* attribute shifting **(R2)** on both categorical and continuous labels.
- **Efficient & Extensible Approach.** Construction of this attribute space is computationally efficient and does not require iterative training, as it has a *closed-form* solution. Further, new concepts can be extended in a modular way using only a minimal labeled exemplar set **(R3)**.
- **Broad Applications.** We demonstrate the efficacy of LatentCompass across diverse tasks by, on average, reducing generative stereotypes to $\mathbf{0}\%$, reducing biases to $\mathbf{1}\%$, reducing unsafe concept generation to $\mathbf{1.3}\%$, enhancing aesthetic score by $\mathbf{27}\%$, and boosting red-teaming success rate against black-box synthetic image detectors by $\mathbf{40}\%$.

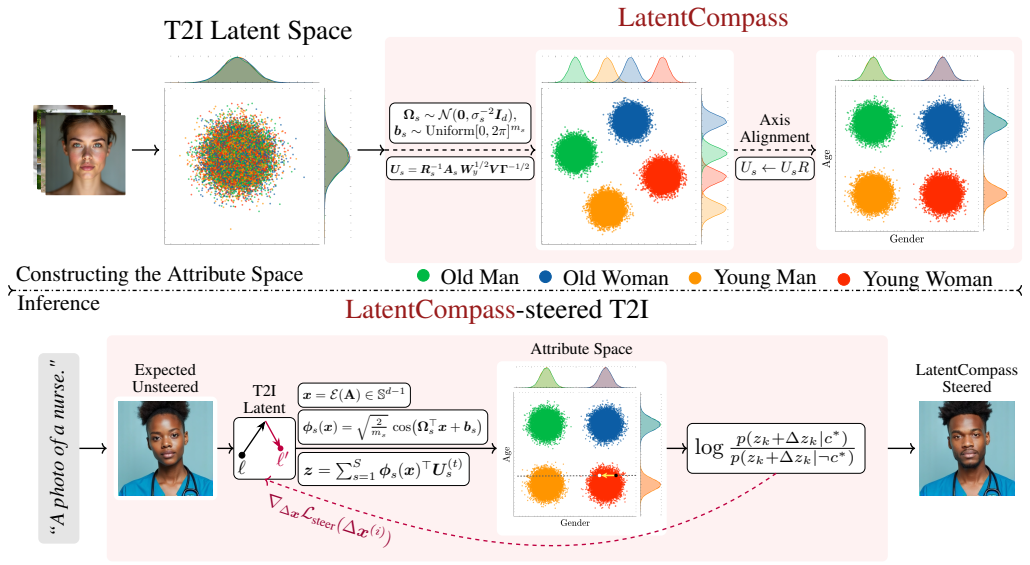


Figure 2: **Overview of LatentCompass: (Top)** LatentCompass uses labeled examples to construct an attribute space where each axis controls a target attribute. **(Bottom)** At inference, LatentCompass steers the T2I latent space to shift the image into the desired attribute region.

2 Problem Setup: Controllable Generation in T2I Models

Notation. Scalars are denoted by lowercase letters (e.g., λ, σ), vectors by bold lowercase (e.g., $\boldsymbol{x}$, $\boldsymbol{z}$), and matrices by bold uppercase (e.g., $\boldsymbol{X}$, $\boldsymbol{U}$). Operators and sets are represented using calligraphic letters (e.g., $\mathcal{E}$, $\mathcal{D}$, $\mathcal{K}$). For a square matrix $\boldsymbol{K}$, we write $\mathrm{Tr}\{\boldsymbol{K}\}$ for its trace.

Problem Formulation. Let $G : \mathscr{L} \times \mathcal{C} \to \mathcal{A}$ denote a text-to-image (T2I) diffusion model that produces an image $\mathbf{A} = G(\boldsymbol{\ell}_T, \boldsymbol{c})$ from an initial noise latent $\boldsymbol{\ell}_T \in \mathscr{L}$ and a text prompt condition $\boldsymbol{c} \in \mathcal{C}$. We associate each generated image with K *concepts of interest*, each captured by a scalar attribute $y_k \in \mathbb{R}$, $k = 1, \ldots, K$. These attributes span two broad categories: **(i) Descriptive attributes**, which are visually interpretable properties that can be easily described by text prompts, such as, facial attributes, nudity, or violence, and, **(ii) Evaluative attributes**, which lack a direct visual correlate but are measurable by an external judge, such as the *realness* score of a synthetic image detector (SID) or a human *aesthetic preference* score.

Concept-guided generation is then the task of *steering* a T2I model along a chosen subset of the attributes. Given a target subset $\mathcal{K} \subseteq \{1, \ldots, K\}$, target values $\{y_k^\star\}_{k \in \mathcal{K}}$, and a measurement function $Y : \mathcal{A} \to \mathbb{R}^K$, we seek a *steering mechanism* $\mathcal{S}$ that produces a modified image $\hat{\mathbf{A}} = \mathcal{S}(G, \boldsymbol{c}, \boldsymbol{\ell}_T, \mathcal{K})$ whose attributes on $\mathcal{K}$ match the targets. Formally:

Problem 1. Concept-Guided Generation

$$\mathcal{S}^\star \;=\; \arg\min_{\mathcal{S}} \; \mathbb{E}_{k \sim \mathcal{K}} \Big[\big\| Y_k\big(\mathcal{S}(G, \boldsymbol{c}, \boldsymbol{\ell}_T, \mathcal{K})\big) - y_k^\star \big\|^2 \Big].$$

Prob. 1 alone is insufficient, as it can drive the target attributes to their desired values while also perturbing the remaining concepts, i.e., attribute leakage. For example, shifting *age* in a face model may inadvertently change the *expression* if the two attributes are correlated in the data. An ideal steering mechanism must satisfy a *preservation constraint*: for every non-target concept $j \notin \mathcal{K}$, its value on the steered image must remain within an ϵ-neighborhood of its value on the unsteered image,

$$\big| Y_j\big(\mathcal{S}(G, \boldsymbol{c}, \boldsymbol{\ell}_T, \mathcal{K})\big) - Y_j\big(G(\boldsymbol{\ell}_T, \boldsymbol{c})\big) \big| \;\le\; \epsilon, \qquad \forall\, j \notin \mathcal{K}. \tag{1}$$

3 LatentCompass: An Approach for Concept-Guided Generation

Naive Approach. The most direct way to control an attribute in a T2I model is through the text prompt. However, prompt-based control is fundamentally limited. **(L1)** Evaluative attributes cannot be expressed in natural language, as they are external to the generative process (e.g., a T2I model has no notion of what makes an image appear *real* to a particular SID). **(L2)** T2I models do not reliably

follow prompts due to training-data correlations (Black et al., 2024; Lee et al., 2023). For instance, prompting $\text{FLUX}_{\text{[dev]}}$ with *"a photo of an old man with a smooth, clean-shaven face"* occasionally yields bearded faces, as the model has learned a strong co-occurrence between *old*, *male*, and *beard*.

Exemplar-Based Guidance. An alternative is to learn concept directions from *examples*. Given a set of images labeled with attributes of interest, one can discover directions in the T2I latent space along which each attribute varies. This resolves both the above-mentioned limitations. Evaluative attributes become accessible the moment a labeling oracle (i.e., an external judge) can annotate examples with the target property (addresses L1). Furthermore, operating directly in the latent space provides control over the internal representation, enabling calibrated edits that bypass the entangled text-conditioning pathway entirely (addresses L2).

LatentCompass steers the image generation process via exemplar-based guidance by decomposing the task into three constituent subproblems: **(P1)** constructing a nonlinear, disentangled representation space from a limited set of examples annotated with either categorical or continuous labels, where individual axes encode independent attributes; **(P2)** defining a target-class intervention in this space; and **(P3)** realizing this intervention in the latent trajectory of the underlying T2I diffusion process.

3.1 P1: Constructing the Attribute Space

Let $\mathcal{E}$ be a frozen pretrained feature extractor that maps an image to a zero-mean, L_2-normalized embedding $\boldsymbol{x} = \mathcal{E}(\mathbf{A}) \in \mathbb{S}^{d-1}$. For each timestep t of the T2I sampling trajectory, we seek an encoder $g_t : \mathbb{R}^d \to \mathbb{R}^K$ that maps $\boldsymbol{x}$ to an attribute coordinate $\boldsymbol{z} = g_t(\boldsymbol{x})$ while satisfying three requirements: **(E1) Nonlinear Expressivity:** the mapping captures the highly nonlinear statistical dependencies between embeddings and attributes, **(E2) Orthogonal Disentanglement:** the axes are mutually orthogonal to help with preventing attribute leakage (Eqs. (1) and (7)), and **(E3) Semantic Alignment:** each axis z_k is maximally dependent on its corresponding attribute y_k.

To satisfy requirements (E1–3), we model g_t as a kernelized supervised projection in a reproducing kernel Hilbert space (RKHS), which captures nonlinear dependencies between the embedding and the attributes (E1) while admitting an orthogonality constraint compatible with the data covariance (E2). For N exemplars, rather than working with the $N \times N$ kernel matrix directly, we approximate the RKHS via Random Fourier Features (RFF) (Rahimi & Recht, 2007), which provide a finite-dimensional map $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ with $\boldsymbol{\phi}(\boldsymbol{u})^\top \boldsymbol{\phi}(\boldsymbol{v}) \approx k(\boldsymbol{u}, \boldsymbol{v})$, reducing the $O(N^3)$ inversion bottleneck to $O(N)$ and keeping g_t differentiable in $\boldsymbol{x}$. A single kernel bandwidth, however, fixes the spatial frequency at which similarity is measured, which is too rigid since attributes are most discriminable at different scales. We therefore extend the kernel construction to S RFF banks. In the following, we describe the encoder fitting at a time step, dropping the subscript t for clarity.

Per-scale supervised covariances (E1). For each scale $s \in \{1, \dots, S\}$, the RFF bank is defined as

$$\boldsymbol{\phi}_s(\boldsymbol{x}) \;=\; \sqrt{\tfrac{2}{m_s}}\,\cos\!\big(\boldsymbol{\Omega}_s^\top \boldsymbol{x} + \boldsymbol{b}_s\big), \qquad \boldsymbol{\Omega}_s \sim \mathcal{N}\big(\mathbf{0}, \sigma_s^{-2}\boldsymbol{I}_d\big),\;\; \boldsymbol{b}_s \sim \mathrm{Unif}[0, 2\pi]^{m_s}, \tag{2}$$

with $\boldsymbol{\Omega}_s \in \mathbb{R}^{d \times m_s}$ and per-bank RFF dimension m_s.

Joint axis construction across scales (E2). Rather than concatenating the S banks into a single high-dimensional feature space and solving one large supervised problem, we treat each bank as an *independent source of evidence* about the attributes and fuse them through a joint $K \times K$ diagonalization. We first define the per-scale, variance-normalized design matrix $\boldsymbol{\Phi}_s \in \mathbb{R}^{N \times m_s}$, the per-scale cross-covariance $\boldsymbol{A}_s := \boldsymbol{\Phi}_s^\top \boldsymbol{Y} \in \mathbb{R}^{m_s \times K}$, where $\boldsymbol{Y} \in \mathbb{R}^{N \times K}$ is the mean-centered label matrix, and the regularized, mean-centered per-scale Gram $\boldsymbol{R}_s := \boldsymbol{\Phi}_s^\top \boldsymbol{H} \boldsymbol{\Phi}_s + \lambda \boldsymbol{I}_{m_s}$, where $\boldsymbol{H} = \boldsymbol{I}_N - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^\top$ is the centering matrix. Then, we seek K axes, i.e., one per attribute, by solving

$$\max_{\{\boldsymbol{U}_s\}_{s=1}^S} \; \mathrm{Tr}\Big\{ \boldsymbol{W}_y^{1/2} \Big(\textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{A}_s \Big)^\top \Big(\textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{A}_s \Big) \boldsymbol{W}_y^{1/2} \Big\}, \;\text{s.t.}\; \textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{R}_s \boldsymbol{U}_s = \boldsymbol{I}_K, \tag{3}$$

where $\boldsymbol{W}_y = \mathrm{diag}(w_1, \dots, w_K) \succ 0$ is an optional per-attribute weighting and $\lambda > 0$ regularizes against rank deficiency in $\boldsymbol{\Phi}_s^\top \boldsymbol{\Phi}_s$ when $N < m_s$. The trace objective maximizes the joint statistical dependence (Gretton et al., 2005; Barshan et al., 2011) between the fused projection $\sum_s \boldsymbol{U}_s^\top \boldsymbol{A}_s$ and the labels (E3); the summed Mahalanobis constraint enforces orthogonality of the recovered axes under the joint feature geometry of the S banks (E2). See the derivation of Eq. (3) in § I.1.

Closed-form solution. Eq. (3) admits a closed-form solution via a single $K \times K$ eigendecomposition. Define the joint supervised matrix

$$\boldsymbol{M} \;=\; \boldsymbol{W}_y^{1/2} \left(\sum_{s=1}^{S} \boldsymbol{A}_s^\top \boldsymbol{R}_s^{-1} \boldsymbol{A}_s \right) \boldsymbol{W}_y^{1/2} \;\in\; \mathbb{R}^{K \times K}, \tag{4}$$

and let $\boldsymbol{M} = \boldsymbol{V} \boldsymbol{\Gamma} \boldsymbol{V}^\top$ be its eigendecomposition with $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_K) \succ 0$. The per-bank projection matrices are recovered as $\boldsymbol{U}_s \;=\; \boldsymbol{R}_s^{-1} \, \boldsymbol{A}_s \, \boldsymbol{W}_y^{1/2} \, \boldsymbol{V} \, \boldsymbol{\Gamma}^{-1/2} \;\in\; \mathbb{R}^{m_s \times K}$. The derivation of this closed-form solution by Lagrangian analysis is given in § I.2.

Semantic Axis Alignment (E3). Although Eq. (3) recovers the optimal supervised subspace, its basis is rotationally invariant, lacking a 1-to-1 correspondence with target attributes. To resolve this identifiability problem and enforce disentanglement, we apply Orthogonal Procrustes alignment (Schönemann, 1966). By computing the cross-covariance $\boldsymbol{C} = \boldsymbol{Z}^\top \boldsymbol{Y}$ and its SVD $\boldsymbol{C} = \boldsymbol{U}_C \boldsymbol{\Sigma}_C \boldsymbol{V}_C^\top$, we derive the optimal rotation $\boldsymbol{R} = \boldsymbol{U}_C \boldsymbol{V}_C^\top$. Updating the projection matrices via $\boldsymbol{U}_s \leftarrow \boldsymbol{U}_s \boldsymbol{R}$ anchors the representation to a canonical orientation, establishing the practical coordinate system required for precise downstream steering. The optimality of this rotation under the Frobenius criterion is established in § I.3.

Encoder at inference. Given a new embedding $\boldsymbol{x}$ at timestep t, the final attribute coordinate is computed by projecting the multi-scale features through the timestep-specific aligned basis:

$$\boldsymbol{z} = g_t(\boldsymbol{x}) = \textstyle\sum_{s=1}^{S} \boldsymbol{\phi}_s(\boldsymbol{x})^\top \boldsymbol{U}_s^{(t)} \in \mathbb{R}^K. \tag{5}$$

3.2 P2: Defining the Attribute-Space Intervention

Given the disentangled attribute space, we define a target-class intervention objective along a selected subset of attribute axes. By design (E2, E3), each axis independently controls its corresponding attribute, allowing target-class evidence to be expressed independently along each selected axis.

Gaussian Density Estimation. For each attribute $k \in \{1, \ldots, K\}$ and class $c \in \{-1, +1\}$, we model the class-conditional density $p(z_k \mid c) = \mathcal{N}(\mu_{k,c}, \sigma_{k,c}^2)$ under the assumption that the distribution of a low-dimensional projection of high-dimensional data is approximately normal (Diaconis & Freedman, 1984). For continuous labels, we set a threshold at $\mathbb{E}[y_k]$ to obtain binary c, enabling directional target-class supervision (e.g., toward higher or lower values). The sufficient statistics are computed from the projected examples $\mathcal{Z}$ at timestep t: $\mu_{k,c} = \mathbb{E}_{\boldsymbol{z} \in \mathcal{Z}}[z_k \mid c], \sigma_{k,c}^2 = \mathrm{Var}_{\boldsymbol{z} \in \mathcal{Z}}[z_k \mid c]$.

Log-Likelihood Ratio Objective. To steer z_k toward a target class c^*, we maximize the log-likelihood ratio (LLR) between the target class c^* and the opposing class $\neg c^*$ of shifted coordinate $\tilde{z}_k = z_k + \Delta z_k$. Under the Gaussian assumption, this yields a closed-form quadratic objective:

$$\mathcal{L}_{\mathrm{LLR}}^{(c^*)}(\tilde{z}_k) = \log \frac{p(\tilde{z}_k \mid c^*)}{p(\tilde{z}_k \mid \neg c^*)} = \frac{(\tilde{z}_k - \mu_{k,\neg c^*})^2}{2\sigma_{k,\neg c^*}^2} - \frac{(\tilde{z}_k - \mu_{k,c^*})^2}{2\sigma_{k,c^*}^2} + \log \frac{\sigma_{k,\neg c^*}}{\sigma_{k,c^*}}. \tag{6}$$

Attribute-Space Intervention Objective. Given the disentangled attribute space, we characterize the desired intervention by increasing the evidence for the target classes along the selected attribute axes. Specifically, for a target subset $\mathcal{K}$, the attribute-space objective is $\mathcal{J}_{\mathrm{attr}}(\Delta \mathbf{z}) = \sum_{k \in \mathcal{K}} \mathcal{L}_{\mathrm{LLR}}^{(c_k^*)}(z_k + \Delta z_k)$. This objective characterizes the desired intervention in attribute space. We do not explicitly optimize it to compute Δz^*; instead, P3 realizes the corresponding target-class intervention directly in the diffusion latent space through the optimization in Eq. (7).

3.3 P3: Propagating the Shift into the Generative Trajectory

While the steering objective is formulated in the K-dimensional attribute space, the actual intervention must be applied on the T2I diffusion latent $\boldsymbol{\ell}_t$. Since our encoder $g_t(\boldsymbol{x})$ is constructed from differentiable operations (RFFs and linear projections), we can propagate the steering signal directly back to the image latent via gradient-based optimization.

Latent Optimization. We seek a latent perturbation $\Delta \boldsymbol{\ell}$ such that the edited latent $\boldsymbol{\ell}' = \boldsymbol{\ell} + \Delta \boldsymbol{\ell}$ achieves the desired attribute profile for targets $\mathcal{K}$, while preserving the original class assignments for all non-target attributes $j \notin \mathcal{K}$. We project the perturbed latent to the clean data manifold via

the Tweedie estimate $\hat{\boldsymbol{\ell}}_0(\boldsymbol{\ell}_t + \Delta\boldsymbol{\ell})$, decode it to pixel space via VAE decoder $\mathcal{D}$, and extract its zero-mean, L_2-normalized embedding $\boldsymbol{x}' = \mathcal{E}\big(\mathcal{D}(\hat{\boldsymbol{\ell}}_0)\big) \in \mathbb{S}^{d-1}$. On this hypersphere, we optimize $\Delta\boldsymbol{\ell}$ to maximize target-class evidence c_k^* while preserving the classes of non-target attributes c_j^{orig}:

$$\mathcal{L}_{\text{steer}}(\Delta\boldsymbol{\ell}) = \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}\Big(\hat{z}_k(\Delta\boldsymbol{\ell})\Big) + \frac{1}{K-|\mathcal{K}|}\sum_{j\notin\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_j^{\text{orig}})}\Big(\hat{z}_j(\Delta\boldsymbol{\ell})\Big), \tag{7}$$

where $\hat{z}_m(\Delta\boldsymbol{\ell}) = \big[g_t(\boldsymbol{x}')\big]_m$ denotes the m-th shifted coordinate.

Latent Update and Early Stopping Criterion. We initialize the perturbation as $\Delta\boldsymbol{\ell}^{(0)} = \mathbf{0}$ and refine it using projected gradient ascent. To regularize the update, we maximize the joint evidence while explicitly projecting the updated state back onto the hypersphere defined by $\|\boldsymbol{\ell}\|_2$:

$$\widetilde{\Delta\boldsymbol{\ell}} = \Delta\boldsymbol{\ell}^{(i)} + \eta\nabla_{\Delta\boldsymbol{\ell}}\mathcal{L}_{\text{steer}}(\Delta\boldsymbol{\ell}^{(i)}), \qquad \Delta\boldsymbol{\ell}^{(i+1)} = \left(\frac{\boldsymbol{\ell}+\widetilde{\Delta\boldsymbol{\ell}}}{\|\boldsymbol{\ell}+\widetilde{\Delta\boldsymbol{\ell}}\|_2}\|\boldsymbol{\ell}\|_2\right) - \boldsymbol{\ell}, \tag{8}$$

where η is the step size. Rather than running optimization for a fixed number of steps, which frequently leads to attribute over-saturation and artifact generation, we terminate the optimization early once the evidence for all targeted attributes safely exceeds a statistical confidence threshold τ_k:

$$\min_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}\Big(\hat{z}_k(\Delta\boldsymbol{\ell}^{(i+1)})\Big) \geq \tau_k(\mu_{k,c^*}, \sigma_{k,c^*}). \tag{9}$$

Upon reaching the stopping criterion, we resume the image generation using the final edited latent.

4 Experimental Evaluation

We evaluate LatentCompass across multiple domains for both *descriptive* and *evaluative* attributes. For descriptive attributes, we consider the tasks of mitigating bias and stereotype in profession generation (§ 4.1), disentangled face attribute editing (§ 4.2), and improving content safety by erasing nudity (§ 4.3 and § C.1). For evaluative attributes, we perform two experiments: we red team synthetic image detectors (SID) by steering T2I models to produce false negatives (§ 4.4), and we try to improve the aesthetic score of the synthesized images (§ 4.5). We use FLUX$_{\text{[dev]}}$(BlackForestLabs, 2024) and SDv1.4 (Rombach et al., 2022) as the T2I generator.

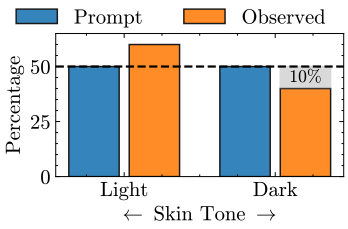


Figure 3: Prompt non-adherence.

4.1 Bias and Stereotype Mitigation via LatentCompass

Motivation. T2I models inherit demographic correlations from their training data: prompts for occupations consistently default to a single race–gender stereotype, even when the prompt itself is demographically neutral (Bianchi et al., 2023; Dehdashtian et al., 2025b; Naik & Nushi, 2023). Explicit prompting often fails to override these priors, occasionally yielding images that contradict the requested attributes (see Fig. 3). LatentCompass mitigates these biases without altering the prompt, semantic meaning, or image quality, as shown in Fig. 4.

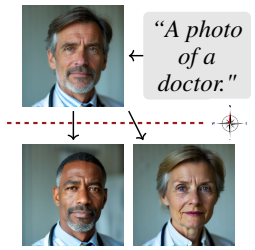


Figure 4: **(Top)** Unsteered **(Bottom)** LatentCompass.

Setup & Metrics. For each concept $C \in \{\textit{Doctor}, \textit{Nurse}\}$ we generate 300 images and measure the two binary attributes $\{\textit{perceived gender}, \textit{skin tone}\}$. LatentCompass is applied to the minority classes for each concept. We report the empirical conditional density $P_{D,C}(y)$ in the generated set D alongside two metrics: **1) Stereotype Score** (Dehdashtian et al., 2025b), $\Psi = \max\big(0,\ P_{D,C}(y) - P^*_{D,C}(y)\big)$, where $P^*_C(y)$ is the real-world occupational density obtained from U.S. census data. **2) Demographic Parity (DP)** as $\text{DP} = \big|\, P_{D,C}(y=0) - P_{D,C}(y=1)\,\big|$, which measures imbalance between the two classes regardless of direction. The two metrics are complementary: stereotype score targets sociologically meaningful violations of the real-world distribution, while DP captures raw distributional imbalance. To ensure debiasing does not degrade image quality, we report CLIP-IQA (Wang et al., 2023a).

Table 1: Evaluation of LatentCompass gender and skin tone stereotype/bias mitigation in FLUX$_{[dev]}$, with separate steering runs targeting census/50-50 distributions, respectively.

	Method	Gender Stereotype			Gender Bias		Skin Tone Stereotype			Skin Tone Bias		CLIP-IQA ↑
		$P_C(y)$	$P_C^*(y)$	$\Psi \downarrow$	$P_C(y)$	DP ↓	$P_C(y)$	$P_C^*(y)$	$\Psi \downarrow$	$P_C(y)$	DP ↓	
Doctor	Unsteered	82.6	61.0	21.6	82.6	65.2	99.0	94.0	5.0	99.0	98.0	73.1 ± 8.4
	LatentCompass	61.0	61.0	**0.0**	49.1	**1.8**	94.0	94.0	**0.0**	50.3	**0.6**	74.6 ± 8.6
Nurse	Unsteered	89.3	88.0	1.3	89.3	68.6	98.0	89.0	9.0	98.0	96.0	72.7 ± 8.9
	LatentCompass	88.0	88.0	**0.0**	50.8	**1.6**	89.0	89.0	**0.0**	50.0	**0.0**	72.5 ± 8.2

Results. Tab. 1 demonstrates how LatentCompass significantly attenuates gender and skin tone stereotypes, reducing Stereotype Score (Ψ) to $\mathbf{0}\%$. Further, we reduce DP from $\mathbf{82}\%$ down to $\mathbf{1}\%$ (compared to DP of 20% via explicit prompting baseline in Fig. 3) on average, demonstrating robust performance in mitigating historical biases (Birhane et al., 2024) inherent in T2I priors. Fig. 4 visualizes this effect: the steered samples preserve the visual cues of the prompted profession (uniform, stethoscope, clinical setting) while diversifying gender, confirming that the orthogonality of the attribute axes prevents leakage into the occupation concept itself.

Remark. LatentCompass aligns the demographics of the generated data with any user-specified reference distribution, mitigating both stereotypes and biases while preserving image quality.

4.2 Attribute Disentanglement in LatentCompass Steering

Motivation. When editing one attribute, others should remain unaffected (R2). To evaluate this disentanglement capability of LatentCompass, we edit four facial attributes, {*age, skin tone, perceived gender, expression*}, and measure the corresponding attribute leakage using an external MLLM.

Setup & Results. We steer a set of 1000 images for $K = 4$ attributes. We measure the success rate of the edits, while also measuring the preservation of the other $K - 1 = 3$ attributes. In Fig. 5 we show the success rate along the diagonal, while the off-diagonal entries denote the leakage rate. On average, LatentCompass achieves a steering success rate of $\mathbf{99.5}\%$, while having a preservation rate of $\mathbf{99.7}\%$.

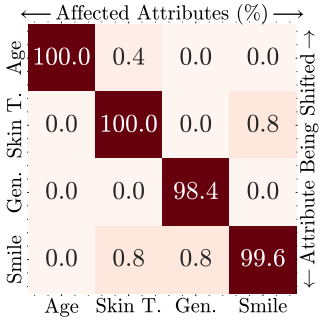


Figure 5: Edit Confusion Matrix.

Remark. LatentCompass precisely edits target attributes while preserving non-target ones.

4.3 Ensuring Content Safety by Erasing Nudity via LatentCompass

Motivation. T2I models can proliferate nudity, violence, or other harmful and not-safe-for-work (NSFW) content. Existing training-free concept erasure approaches suppress NSFW concepts mostly through prompt embedding manipulation, but these methods remain susceptible to adversarial and implicit prompts as they are *input-level* interventions (Yoon et al., 2025). Instead, we explore an *output-level* erasure where the T2I latents containing nudity are steered toward safe subspaces.

Setup & Benchmarks. We use LatentCompass to shift a continuous attribute, *Nudity* $\in [0, 1]$ strictly towards zero. Following the settings in Yoon et al. (2025), we evaluate LatentCompass on removing nudity in SDv1.4 on three adversarial prompt benchmarks: P4D (Chin et al., 2024), Ring-A-Bell (Tsai et al., 2023), and MMA-Diffusion (Yang et al., 2024). To assess erasure effectiveness, we report the **Attack Success Rate (ASR)** of these benchmarks against each erasure method. To measure generation quality, we employ **Frechet Inception Distance (FID)** (Heusel et al., 2017) and **CLIP score** on COCO (Lin et al., 2014). Extended results on FLUX.1 presented in § C.1.

Tab. 2 compares the effectiveness of LatentCompass and baselines in erasing nudity. LatentCompass achieves state-of-the-art erasure by nearly eliminating adversarial vulnerabilities across all benchmarks. Unlike fine-tuning-free baselines that primarily steer prompts and can therefore be bypassed by the adversarial prompts used in these benchmarks, LatentCompass detects unsafe content directly from the generated image representation and steers the latent toward safe content when needed. This output-level intervention yields substantial reductions in Attack Success Rate (ASR), even outperforming fine-tuning-based approaches. Crucially, this robustness does not come at the cost of utility: LatentCompass achieves the best image fidelity (FID) among all evaluated methods while maintaining CLIP alignment on par with the foundational model. Ultimately, these results demonstrate that output-level steering provides a more resilient defense against implicit and adversarial prompts while avoiding the generation-quality degradation typical of concept erasure.

Table 2: Comparison of ASR (↓) for nudity erasure on SD-v1.4, across adversarial prompt benchmarks.

Method	Fine-Tuning Free	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓	COCO FID ↓	COCO CLIP ↑
SD-v1.4 (Rombach et al., 2022)	-	98.7	83.1	95.7	-	31.3
ESD (Gandikota et al., 2023)	×	75.0	52.8	87.3	-	30.7
SA (Heng & Soh, 2023)	×	62.3	32.9	20.5	54.98	30.6
CA (Kumari et al., 2023)	×	92.7	77.3	85.5	40.99	31.2
MACE (Lu et al., 2024)	×	14.6	7.6	18.3	52.24	29.4
SDID (Li et al., 2024)	×	93.3	69.6	90.7	22.99	30.5
UCE (Gandikota et al., 2024b)	✓	66.7	33.1	86.7	31.25	**31.3**
RECE (Gong et al., 2024)	✓	38.1	13.4	67.5	37.60	30.9
SLD-Medium (Schramowski et al., 2023)	✓	93.4	64.6	94.2	31.47	31.0
SLD-Strong (Schramowski et al., 2023)	✓	86.1	62.0	92.0	40.88	29.6
SLD-Max (Schramowski et al., 2023)	✓	74.2	57.0	83.7	50.51	28.5
SAFREE (Yoon et al., 2025)	✓	38.4	11.4	58.5	36.35	31.1
LatentCompass (ours)	✓	**2.7** (-35)	**1.2** (-10)	**0.0** (-58)	**22.08**	31.2

Remark. LatentCompass eliminates nearly all adversarial nudity vulnerabilities by steering latents at the output level, preserving higher image fidelity than both input-level and fine-tuning baselines.

4.4 LatentCompass as a Tool for Red Teaming Synthetic Image Detectors

Table 3: LatentCompass against T2I-specific SIDs at 256×256.

Method	UFD	RINE	FatFormer	AVG ± STD	Time (s) ↓	CLIP-IQA ↑
No Steering	67.6	52.4	55.1	58.4 ± 6.6	∼ 5.0	0.8427
PolyJuice	96.3	81.2	83.2	86.8± 6.6	∼ 198.6	0.8410
LatentCompass	**100.0** (+32)	**99.5** (+47)	**96.4** (+41)	98.6 ± 1.6	∼ 8.2	**0.8640**

Motivation. As T2I models close the gap between synthetic and real imagery, synthetic image detectors (SIDs) have emerged as the primary defense against malicious uses such as disinformation and identity fraud (Shiohara & Yamasaki, 2022; Ojha et al., 2023; Koutlis & Papadopoulos, 2024; Dehdashtian et al., 2025a). Red teaming SIDs unveils their failure modes and is essential for hardening these detectors before deployment. As SIDs are mostly used in proprietary settings, it is important to red team them in a black-box manner. We use LatentCompass to steer an *evaluative* attribute, i.e., detector *realness*, which cannot be described through text.

Setup. Following PolyJuice (Dehdashtian et al., 2025a), we use $\text{FLUX}_{[\text{dev}]}$ as the T2I model and prompts from COCO, and report Attack Success Rate (ASR) against three *black-box* SIDs: UFD (Ojha et al., 2023), RINE (Koutlis & Papadopoulos, 2024), and FatFormer (Liu et al., 2024). We use LatentCompass to flip a binary label *Fakeness* $\in \{0, 1\}$ to zero.

Results. Tab. 3 reports ASR across the three SIDs. Although PolyJuice boosts ASR over the unsteered baseline, it plateaus at 86% on average. LatentCompass closes this gap and tightens the variance across detectors. This improvement is a result of the non-linear geometry of our attribute space, as PolyJuice only uses linear shifts in the T2I latent space. Further, LatentCompass adds a negligible overhead ($\sim 3s$) for generating a false negative compared to a PolyJuice ($\sim 193s$), therefore showing the practicality of our approach in dataset generation. In Tab. 4, we also show that LatentCompass also outperforms transfer-based attacks, such as DiffPGD (Xue et al., 2023), that performs white-box attacks on a proxy detector trained on RINE labels.

Table 4: LatentCompass ASR (%) vs. transfer-based baseline.

Method	ASR (↑)
Unsteered	52.4
DiffPGD^t (x^n)	57.2
DiffPGD^t (x_0^n)	70.1
LatentCompass (ours)	**99.5** (+46)

Remark. LatentCompass on average achieves 98.6% black-box red-teaming success against SIDs.

4.5 Improving Aesthetic Score of T2I Generation via LatentCompass

Motivation. A prime example of an evaluative attribute is the aesthetic quality of an image. Since human preference is intrinsically tied to visual appeal, aligning T2I models to maximize aesthetic scores enhances their practical utility and drives broader adoption across real-world applications.

Setup. We use ASV2 (Schuhmann & LAION, 2023) as a proxy for human preference. LatentCompass is fit on the examples labeled by this predictor and then used to enhance a continuous attribute *Aes* $\in [0, 10]$ for 1000 COCO prompts.

Results. From Fig. 6, we observe that an unsteered $\text{FLUX}_{[\text{dev}]}$ obtains *Aes* of $\sim$ **5.5** on average, while LatentCompass-steered images have an average score of $\sim$ **7**. Interestingly, though LatentCompass is fitted using $\text{FLUX}_{[\text{dev}]}$-generated examples with a maximum *Aes* of **7.4**, we

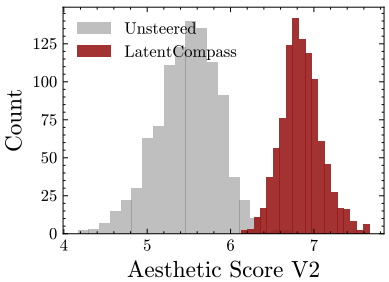


Figure 6: Improving ASV2 of COCO-prompted images via LatentCompass.

find that LatentCompass-steered samples have a maximum *Aes* **> 9**. This suggests that the nonlinear direction found by LatentCompass is highly generalizable and enables extrapolation. We further obtain an unbiased evaluation of aesthetics by using an external multimodal LLM in § C.2. Our aesthetic steering improves the external score from 6.7 ± 1.3 to $\mathbf{7.9 \pm 0.9}$ (Fig. 8).

Remark. LatentCompass significantly improves visual appeal, shifting average aesthetic scores from 5.5 to 7.0 and successfully extrapolating beyond its fitting exemplars.

5 Related Work

Conventional steering approaches typically rely on prompt engineering (Ban et al., 2024; Woolf, 2022; Nguyen et al., 2025) or CLIP-based text guidance (Zhang et al., 2023; Rombach et al., 2022; Ramesh et al., 2022). However, these approaches struggle to override entrenched training-data correlations, and are inherently incapable of steering non-linguistic, evaluative attributes. To bypass text bottlenecks, classifier-guidance (Ban et al., 2024; Wallace et al., 2023; Dhariwal & Nichol, 2021; Zhang et al., 2023) steers latent trajectories using gradients from external classifiers. While operating directly on latents, this approach struggles with attribute leakage (Zhang et al., 2023).

Generative stereotypes and biases are inherent issues in T2I models that cannot be fully resolved by prompt engineering (Schramowski et al., 2023; Zhang et al., 2023). Although open-set auditing of T2I models has improved (D'Incà et al., 2024; Dehdashtian et al., 2025b; Bianchi et al., 2023; Naik & Nushi, 2023; Luccioni et al., 2024; Buolamwini & Gebru, 2018), current surface-level text mitigations fail to explicitly disentangle correlated attributes within generative priors, leaving the latent space structurally biased. LatentCompass complements text-based methods by solving the structural bias in the image latent space, enabling more comprehensive visual bias mitigation.

Red-teaming of SIDs. Diffusion-based adversarial attacks (Xue et al., 2023; Dai et al., 2024; Chen et al., 2024) are effective on SIDs, but they demand white-box access, limiting their utility for commercial black-box detectors. While black-box methods like PolyJuice (Dehdashtian et al., 2025a) red team SIDs via linear latent shifts, unlike LatentCompass, these linear approximations cannot capture the highly non-linear geometry of *fakeness*, necessitating expensive hyperparameter search.

Concept erasure. Training-based unlearning methods alter model weights to suppress concepts, but suffer from prohibitive computational costs and catastrophic forgetting (Heng & Soh, 2023; Kumari et al., 2023; Li et al., 2024; Lu et al., 2024; Gandikota et al., 2023; 2024a). Alternatively, training-free interventions bypass retraining by manipulating classifier-free guidance (Schramowski et al., 2023) or steering text embeddings away from toxic subspaces (Wang et al., 2023b; Gaintseva et al., 2026; Yoon et al., 2025; Brack et al., 2023; Zhang et al., 2024; Biswas et al., 2025). However, since these models are vision-unaware, 1) they struggle with safeguarding against subversive harmful prompts, and 2) their control is limited to linguistically expressible concepts and cannot directly target evaluative attributes. In contrast, LatentCompass presents a complementary approach that intervenes directly on the image latents, addressing the shortcomings of text-only intervention.

6 Concluding Remarks

In this paper, we present LatentCompass, an exemplar-based, training-free method for controllable text-to-image generation that operates directly on the latent space of frozen T2I diffusion models. To overcome entrenched generative stereotypes and predispositions, LatentCompass constructs a closed-form, nonlinear, and explicitly orthogonal attribute subspace from a small set of generated labeled examples, then steers the T2I by propagating the shift back into the diffusion trajectory. Owing to its orthogonal-by-construction design, a single attribute compass supports semantic steering, stereotype mitigation, concept erasure, and black-box red-teaming of synthetic image detectors, with new concepts added by simply refitting the closed-form encoder on a new labeled set.

Limitations & Future Directions. Similar to any training-free steering approach, LatentCompass is constrained to the T2I model's existing generative data manifold and cannot synthesize out-of-distribution concepts. Moreover, it requires a limited search for the steering magnitude, which is standard for all the guidance methods. Finally, because the base model cannot faithfully render continuous demographic spectra, our attribute evaluations for *perceived gender* and *skin tone* (§ 4.1, § 4.2) are restricted to binary categories, presenting an avenue for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing, editing, and revising the writing of parts of the paper. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *ECCV*. Springer, 2024.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 2011.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *ACM Conference on Fairness, Accountability, and Transparency*, 2023.

Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. The dark side of dataset scaling: Evaluating racial classification in multimodal models. FAccT '24, pp. 1229–1244. Association for Computing Machinery, 2024.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

BlackForestLabs. FLUX. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency*, 2018.

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4Debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 8468–8486, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Sepehr Dehdashtian, Mashrur Morshed, Jacob Seidman, Gaurav Bharaj, and Vishnu Naresh Boddeti. PolyJuice Makes It Real: Black-Box, Universal Red-Teaming for Synthetic Image Detectors. In *Neural Information Processing Systems*, 2025a.

Sepehr Dehdashtian, Gautam Sreekumar, and Vishnu Naresh Boddeti. OASIS Uncovers: High-Quality T2I Models, Same Old Stereotypes. In *International Conference on Learning Representations*, 2025b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pp. 793–815, 1984.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. OpenBias: Open-set Bias Detection in Text-to-Image Generative Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614, 2011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024a.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5111–5120, 2024b.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*. Springer, 2005.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 2011.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pp. 394–411. Springer, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22691–22702, 2023.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 2024.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

Viet Nguyen, Anh Nguyen, Trung Dao, Khoi Nguyen, Cuong Pham, Toan Tran, and Anh Tran. Supercharged one-step text-to-image diffusion models with negative prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18004–18013, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. Flowchef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15308–15318, October 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Christoph Schuhmann and LAION. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2023.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Not AI Tech. Nudenet. https://github.com/notAI-tech/NudeNet, 2023.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Adriana Fernández de Caleya Vázquez and Eduardo C Garrido-Merchán. A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence. *arXiv preprint arXiv:2407.01556*, 2024.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7246–7256. IEEE, 2023.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *Advances in Neural Information Processing Systems*, 2023b.

Max Woolf. Stable diffusion 2.0 and the importance of negative prompts for good results, 2022, 2022.

Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36:2894–2921, 2023.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, 2025.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3969–3980, 2023.

Hongxiang Zhang, Yifeng He, and Hao Chen. Steerdiff: Steering towards safe text-to-image diffusion models. *arXiv preprint arXiv:2410.02710*, 2024.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Outline of Appendix

A Ethical Considerations **14**

B Implementation Details **15**

B.1 Generating Examples. 15

B.1.1 Prompt templates. 15

B.1.2 Attribute Labels. 15

B.1.3 T2I Generation Settings 16

B.1.4 Feature Extractors 16

B.2 Fitting Encoders 16

B.3 Steering 16

B.3.1 Estimating The Denoised Latents 16

B.3.2 Gradient Calculation 17

B.4 Generation Hyperparameters 17

B.5 Experiments Compute Resources 17

B.6 Algorithmic Summary of LatentCompass 18

B.7 Threat Model for Red Teaming Experiment 19

C Extended Quantitative Results **21**

C.1 Erasing Nudity via LatentCompass on FLUX.1 21

C.2 Image Aesthetic Evaluation via MLLM-as-a-Judge 21

C.3 Image NSFW Independent Evaluation via MLLM-as-a-Judge 22

D Extended Qualitative Examples **24**

E Summary of Key Notations **27**

F Ablation on the Number of Exemplars **29**

G LatentCompass Attribute Space Visualization **29**

H Potential Defense Mechanisms Against LatentCompass **30**

I Proofs **30**

I.1 Derivation of the Multi-Scale Supervised Axis Objective 30

I.2 Derivation of the Closed-Form Solution to Eq. (3) 32

I.3 Optimality of the Procrustes Axis Alignment 33

A Ethical Considerations

While LatentCompass is designed to mitigate harmful biases and stress-test safety-critical detectors, its red-teaming capability is dual-use and could be misused to evade synthetic image detection in

the wild. We strongly oppose such use and intend LatentCompass solely for responsible auditing, debiasing, and detector hardening; we discuss potential defenses against malicious deployment in Appendix § H.

B Implementation Details

In this section, we provide the implementation details and hyperparameters required for independent reproduction of the results presented in the paper. The official code will be released upon acceptance.

B.1 Generating Examples.

For all our experiments in § 4, we first construct an example set by generating 40K images with the base T2I model. In Fig. 13, we perform an ablation on the effect of the number of examples used to construct the LatentCompass attribute space.

B.1.1 Prompt templates.

(i) Face Attribute Steering: For § 4.1 and § 4.2, we generate a common set of images following the template `"A realistic photo of a <age> <skin tone> <perceived gender> with a <smile> expression."`, where, age $\in$ {young, old}, skin tone $\in$ {light, dark}, perceived gender $\in$ {male, female}, and smile $\in$ {neutral, smiling}. We also append some additional phrases at the end of the prompt to improve the generation quality: `"RAW color photograph, photorealistic, shot on DSLR, 85mm lens, shallow depth of field, natural skin texture, pores visible, realistic lighting, no stylization."`

(ii) Red Teaming SID and Steering Aesthetic Preference: For both § 4.4 and § 4.5, we generate a common set of examples using a subset of text captions from the COCO dataset (Lin et al., 2014). No additional quality prompts are appended. Specifically, we picked random set of 40K prompts for constructing the example set from the COCO training set. The evaluation, in contrast, is done on an unseen subset of 1000 text captions from the test set.

(iii) Content Safety Improvement: We use 1K prompts from the MMA-Diffusion (Yang et al., 2024) benchmark, specifically the 'target_prompt' column that contains toxic prompts detectable by safety filters. We also use a large language model (LLM) to generate a set of 1K 'safe' prompt candidates, and then manually check/correct each prompt with human supervision. We generate 20 random images for the 2K safe + unsafe prompts. Note that, in Tab. 2, when MMADiffusion is an evaluation dataset, the prompts utilized are from the 'adv_prompt' column, which are more subtle adversarial prompts that are completely different from the 'target_prompt' (raw toxic prompts) that we use to fit LatentCompass. Further, the additional datasets, P4D and Ring-A-Bell, also serve the purpose of unbiased evaluation sets. To also verify that LatentCompass is not benefiting from circular evaluation (since the labels are from NudeNet, and the actual benchmark is also NudeNet), we validate our nudity removal results with an independent external judge (Qwen-3.5), in Tab. 7.

B.1.2 Attribute Labels.

(i) Face Attribute Steering: The four binary attributes (age, skin tone, perceived gender, smile) are independently and identically distributed (i.i.d.) according to a Bernoulli distribution with $p = 0.5$. Since we construct the prompts ourselves, we have access to the attributes associated with every text prompt. After image generation, we also classify each face with an external multimodal LLM, Qwen3.5-9B (Qwen Team, 2026). We refer to the attributes specified before generation, and the result attributes labeled by Qwen, as 'prompt' and 'observed' labels respectively. For fitting the encoder, we solely use the 'prompt' labels; Qwen3.5-9B is used for evaluation only.

(ii) Red Teaming SIDs: We label the set of generated COCO images with hard binary labels (Real, Fake) with three SIDs, UFD (Ojha et al., 2023), RINE (Koutlis & Papadopoulos, 2024), FatFormer (Liu et al., 2024).

(iii) Steering Aesthetic Preference: We label our set of generated COCO images with the Aesthetic Predictor V2 model (Schuhmann & LAION, 2023).

(iv) Content Safety Improvement: We use NudeNet (Tech, 2023) to label each image with an unsafe score $\in [0, 1]$.

B.1.3 T2I Generation Settings

(i) Face Attribute Steering: We use $\text{FLUX}_{\text{[dev]}}$ (BlackForestLabs, 2024) with 30 sampling steps, a Flow Matching Euler Discrete scheduler, and a guidance of 3.5.

(ii) Red Teaming SID and Steering Aesthetic Preference: We use $\text{FLUX}_{\text{[dev]}}$ with 50 sampling steps, a Flow Matching Euler Discrete scheduler, and a guidance of 3.5.

(iii) Content Safety Improvement: We use SDv1.4 (Rombach et al., 2022) with 50 sampling steps, a DDIM scheduler, and a guidance of 7.

For each timestep t in the generation process, we compute and store the denoised latent $\hat{\boldsymbol{\ell}}_0$ from the current latent $\boldsymbol{\ell}_t$.

B.1.4 Feature Extractors

For face attribute steering, we use features from FaRL (Zheng et al., 2022), which shares the same architecture as CLIP (Radford et al., 2021). For red teaming SIDs and steering aesthetic preference, we used CLIP, specifically `clip-vit-large-patch-14`. For nudity erasure, instead of visio-lingual features, we used purely visual Dino V2 (Oquab et al., 2023) features.

B.2 Fitting Encoders

RFF Banks. For all experiments, we use $S = 5$ RFF banks with different kernel bandwidths. Let $\{\mathbf{x}_i\}_{i=1}^N$ denote the embeddings extracted from the fitting exemplars, where $\mathbf{x}_i = \mathcal{E}(A_i) \in \mathbb{R}^d$. We estimate a reference bandwidth using the median heuristic,

$$\sigma_{\text{med}} = \text{median}_{i<j} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2 ,$$

i.e., the median pairwise Euclidean distance between the extracted features. The five RFF banks then use

$$\sigma_s \in \{0.25\sigma_{\text{med}}, 0.5\sigma_{\text{med}}, \sigma_{\text{med}}, 2\sigma_{\text{med}}, 4\sigma_{\text{med}}\} .$$

To allocate more capacity to scales near the reference bandwidth, we set the corresponding RFF dimensions to

$$m_s \in \{400,\ 800,\ 1600,\ 800,\ 400\},$$

respectively.

B.3 Steering

B.3.1 Estimating The Denoised Latents

Before feature extraction with FaRL, CLIP, or Dino-V2, it is necessary to approximate the clean datapoint $\hat{\boldsymbol{\ell}}_0$ at every step. For diffusion models, we can do this by Tweedie's formula (Efron, 2011),

$$\hat{\boldsymbol{\ell}}_0 = \frac{\boldsymbol{\ell}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{\ell}_t, t, \boldsymbol{c})}{\sqrt{\bar{\alpha}_t}} \tag{10}$$

where $\bar{\alpha}_t$ is the diffusion noise schedule, and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ denotes the predicted noise. However, we note that $\text{FLUX}_{\text{[dev]}}$ is a rectified flow model (Esser et al., 2024) (which are closely related to diffusion models), where the denoising can be simplified to a single Euler step with stepsize t,

$$\hat{\boldsymbol{\ell}}_0 = \boldsymbol{\ell}_t - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{\ell}_t, t, \boldsymbol{c}) \cdot t \tag{11}$$

where $\boldsymbol{v}_{\boldsymbol{\theta}}(\cdot)$ is the predicted velocity.

B.3.2 Gradient Calculation

For a latent $\boldsymbol{\ell}_t$ at timestep t, we first compute the denoised estimate $\hat{\boldsymbol{\ell}}_0$ (Eq. (10), Eq. (11)), and then decode $\hat{\boldsymbol{\ell}}_0$ into an image $\hat{\mathbf{A}}_0 = \mathcal{D}(\hat{\boldsymbol{\ell}}_0)$. We then use the feature extractor $\mathcal{E}$ to extract features $\boldsymbol{x} = \mathcal{E}(\hat{\mathbf{A}}_0)$, and then project it on to the K-axis attribute space as $\boldsymbol{z} = g_t(\boldsymbol{x})$. For brevity, we can summarize these sequence of actions as a single differentiable transformation, $\boldsymbol{z} \coloneqq \text{ToCoord}(\boldsymbol{\ell}_t)$.

It is apparent that computing the gradient of $\Delta\boldsymbol{\ell}_t$ (Eq. (8)) w.r.t. the optimization objective (Eq. (7)) would require backpropagating through the diffusion model itself, due to Eq. (10), Eq. (11). While this may be tractable for a smaller model like SDv1.4, it is infeasible for a larger model like $\text{FLUX}_{\text{[dev]}}$ with 12B parameters.

For $\text{FLUX}_{\text{[dev]}}$, we treat the velocity prediction $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{\ell}_t, t, \boldsymbol{c})$ as a constant term $\boldsymbol{v}_t$ instead of a function of $\boldsymbol{\ell}_t$, which simplifies Eq. (11) to,

$$\hat{\boldsymbol{\ell}}_0 = \boldsymbol{\ell}_t - \boldsymbol{v}_t \cdot t. \tag{12}$$

Under Eq. (12), computing the gradient of $\boldsymbol{\ell}_t$ for FLUX no longer requires backpropagating through the actual T2I model. This approach is originally used by Bansal et al. (2023), and referred to as *backward universal guidance*. However, backward guidance requires injecting noise back into $\boldsymbol{\ell}_t$ to ensure they remain on the trajectory manifold. More recently, Patel et al. (2025) show both theoretically and empirically, with extensive experiments, that for rectified flows like FLUX, $\nabla_{\hat{\boldsymbol{\ell}}_0}\mathcal{L}$ has a negligible error compared to the exact gradient $\nabla_{\boldsymbol{\ell}_t}\mathcal{L}$ due to smoothness properties of the velocity field. These theoretical and empirical backbone directly justifies the approximation in Eq. (12). Further, our results also agree with the findings of Patel et al. (2025), in that steering rectified flow transformers works reasonably well without expensive backpropagation.

B.4 Generation Hyperparameters

We present the hyperparameters per experiment in Tab. 5. D_1, D_2, and D_3 represent a set of face image examples, a set of nude/safe image examples, and a set of COCO-prompt images respectively, previously described in § B.1.

Table 5: Hyperparameter configurations across different experiments.

Hyperparameter	§ 4.1	§ 4.2	§ 4.3	§ 4.4	§ 4.5
Example Set	D_1	D_1	D_2	D_3	D_3
Sampling Steps, T	30	30	50	50	50
Shift Strength, η	$\{2, 3, 4\}$	$\{2, 3, 4\}$	$\{1, 2.5, 5\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$
Max PGD Steps, $i_{\max}$	30	30	20	20	20
Stopping Criterion, τ	$\mu_{k,c^*} \pm 1\sigma_{k,c^*}$	$\mu_{k,c^*} \pm 1\sigma_{k,c^*}$	$\mu_{k,c^*} - 1\sigma_{k,c^*}$	μ_{k,c^*}	$\mu_{k,c^*} + 1\sigma_{k,c^*}$

PGD iterations. The maximum PGD iterations (e.g. 20 or 30) only describes the theoretical upper bound, and does not capture the practical cost of steering, as the optimization can stop early due to our early stopping mechanism Eq. (9). To understand the practical cost, we inspect the experiments in § 4.5, § 4.3, § 4.4, all of which have 50 denoising steps and maximum PGD iterations of 20 (as noted in Tab. 5). In Fig. 7, we show that LatentCompass requires only ~ 4.9 PGD iterations per generation timestep on average across the 50-step generation trajectory, despite allowing a maximum budget of 20 iterations per timestep. Thus, LatentCompass uses less than 25% of the available optimization budget to steer the generation successfully. Moreover, the optimization effort is concentrated in the early stages, averaging ~ 10 iterations over the first 10 generation steps, and rapidly decreases as generation progresses. In the final 10 steps, only ~ 0.8 iterations are required on average, with many trials requiring no steering at all.

B.5 Experiments Compute Resources

For all experimental stages, including dataset generation, attribute-space construction, and inference steering, we used eight NVIDIA Quadro RTX 8000 GPUs, each with 48 GB of memory. These

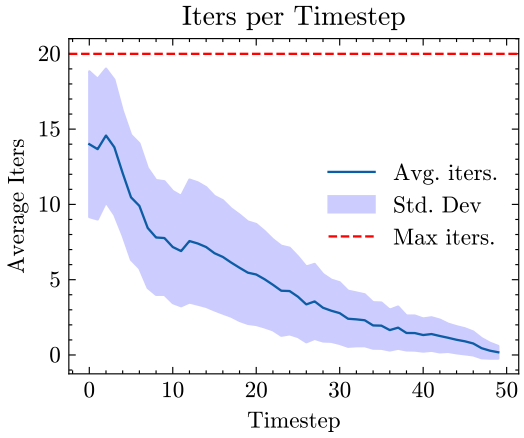


Figure 7: Average PGD iterations per FLUX timestep.

resources primarily accommodate the memory requirements of the base T2I models and large-scale dataset generation, rather than LatentCompass fitting itself. Once exemplar features are available, fitting the LatentCompass attribute encoder is a closed-form operation requiring < 1 s in our implementation. During FLUX inference, LatentCompass does not backpropagate through the 12B-parameter denoiser; the additional computation is restricted to the frozen feature extractor and low-dimensional attribute-space projections. Consequently, steering adds only a modest inference-time overhead relative to the base generator.

B.6 Algorithmic Summary of LatentCompass

We summarize LatentCompass as two procedures: a one-time *fitting* stage that constructs the attribute encoder from labeled exemplars, and an *inference* stage that steers the T2I latent toward a target attribute profile. The two stages share no per-prompt state: once Alg. 1 has been run for a given concept set, Alg. 2 can be invoked on any prompt without revisiting the labeled exemplars, realizing the *build once, steer anywhere* property discussed in § 3.

Fitting (Alg. 1). Given N labeled exemplars $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, the fitting procedure samples S independent RFF banks at bandwidths $\{\sigma_s\}$ (Eq. (2)), assembles the per-bank cross-covariances $\boldsymbol{A}_s$ and Gram matrices $\boldsymbol{R}_s$, and solves the joint multi-scale objective in Eq. (3) via the closed-form expressions in Eq. (4) and § 3.1. The recovered axes are then aligned to the attribute basis through Orthogonal Procrustes (§ 3.1), and per-class Gaussian sufficient statistics $(\mu_{k,c}, \sigma^2_{k,c})$ are estimated on the aligned coordinates for use at inference. Computationally, the dominant step is a single $K \times K$ eigendecomposition of $\boldsymbol{M}$ in Eq. (4); everything else is linear in N.

Steering (Alg. 2). At inference, a T2I latent $\boldsymbol{x}$ at timestep t is projected into the attribute space via Eq. (5). (P2) defines the desired intervention in attribute space through the target-class log-likelihood ratio objective. Rather than explicitly computing a coordinate shift $\Delta\mathbf{z}^*$ and mapping it back to the diffusion latent space, Alg. 2 realizes the corresponding intervention directly by optimizing Eq. (7) with respect to the diffusion latent. The projected gradient updates in Eq. (8) are terminated once the target-attribute LLR confidence exceeds the threshold τ. The resulting steered latent $\boldsymbol{x} + \Delta\boldsymbol{x}^*$ replaces the original embedding, and the diffusion sampler resumes its trajectory unchanged.

Modularity. New control concepts are added by re-running Alg. 1 on a fresh labeled exemplar set, without modifying the underlying T2I model or any previously fitted encoders. Alg. 2 is agnostic to which concepts are present in $\mathcal{K}$, so categorical attributes (e.g., perceived gender), continuous evaluative attributes (e.g., aesthetic score), and binary safety labels (e.g., nudity) are all handled by the same inference procedure.

Algorithm 1 LatentCompass: Fitting the Attribute Encoder

Input: Labeled exemplars $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ with zero-mean, L_2-normalized embeddings $\boldsymbol{x}_i \in \mathbb{S}^{d-1}$ and labels $\boldsymbol{y}_i \in \mathbb{R}^K$; diffusion timestep t.

Output: Per-bank projections $\{\boldsymbol{U}_s^{(t)}\}_{s=1}^S$, RFF parameters $\{\boldsymbol{\Omega}_s, \boldsymbol{b}_s\}_{s=1}^S$, alignment rotation $\boldsymbol{R}$, class statistics $\{(\mu_{k,c}, \sigma_{k,c}^2)\}_{k,c}$.

$\boldsymbol{H} \leftarrow \boldsymbol{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ {Centering matrix}
$\boldsymbol{Y} \leftarrow \boldsymbol{H}\boldsymbol{Y}$ {Mean-center the labels ($N \times K$)}
for $s = 1, \dots, S$ **do**
  Sample $\boldsymbol{\Omega}_s \sim \mathcal{N}(\mathbf{0}, \sigma_s^{-2}\boldsymbol{I}_d)$ and $\boldsymbol{b}_s \sim \mathrm{Unif}[0, 2\pi]^{m_s}$
  $\boldsymbol{\Phi}_s \leftarrow$ matrix with rows $\sqrt{2/m_s}\cos(\boldsymbol{\Omega}_s^\top \boldsymbol{x}_i + \boldsymbol{b}_s)^\top$
  $\boldsymbol{A}_s \leftarrow \boldsymbol{\Phi}_s^\top \boldsymbol{Y} \in \mathbb{R}^{m_s \times K}$
  $\boldsymbol{R}_s \leftarrow \boldsymbol{\Phi}_s^\top \boldsymbol{H} \boldsymbol{\Phi}_s + \lambda \boldsymbol{I}_{m_s}$ {Centered Gram matrix}
end for
$\boldsymbol{M} \leftarrow \boldsymbol{W}_y^{1/2}\big(\sum_{s=1}^S \boldsymbol{A}_s^\top \boldsymbol{R}_s^{-1} \boldsymbol{A}_s\big)\boldsymbol{W}_y^{1/2} \in \mathbb{R}^{K \times K}$
Eigendecompose $\boldsymbol{M} = \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{V}^\top$ with $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_K) \succ 0$
for $s = 1, \dots, S$ **do**
  $\boldsymbol{U}_s \leftarrow \boldsymbol{R}_s^{-1} \boldsymbol{A}_s \boldsymbol{W}_y^{1/2} \boldsymbol{V} \boldsymbol{\Gamma}^{-1/2} \in \mathbb{R}^{m_s \times K}$
end for
$\boldsymbol{Z} \leftarrow \sum_{s=1}^S \boldsymbol{\Phi}_s \boldsymbol{U}_s \in \mathbb{R}^{N \times K}$ {Training projections}
$\boldsymbol{C} \leftarrow \boldsymbol{Z}^\top \boldsymbol{Y}$; SVD: $\boldsymbol{C} = \boldsymbol{U}_C \boldsymbol{\Sigma}_C \boldsymbol{V}_C^\top$
$\boldsymbol{R} \leftarrow \boldsymbol{U}_C \boldsymbol{V}_C^\top$ {Optimal Procrustes rotation}
for $s = 1, \dots, S$ **do**
  $\boldsymbol{U}_s^{(t)} \leftarrow \boldsymbol{U}_s \boldsymbol{R}$
end for
$\boldsymbol{Z} \leftarrow \boldsymbol{Z}\boldsymbol{R}$
for $k = 1, \dots, K$ and $c \in \{-1, 1\}$ **do**
  $\mu_{k,c} \leftarrow \mathbb{E}_{i:y_{i,k}=c}[Z_{i,k}]; \quad \sigma_{k,c}^2 \leftarrow \mathrm{Var}_{i:y_{i,k}=c}[Z_{i,k}]$
end for
return $\{\boldsymbol{U}_s^{(t)}, \boldsymbol{\Omega}_s, \boldsymbol{b}_s\}_{s=1}^S$, $\boldsymbol{R}$, $\{(\mu_{k,c}, \sigma_{k,c}^2)\}_{k,c}$

B.7 Threat Model for Red Teaming Experiment

Following the framework established in PolyJuice (Dehdashtian et al., 2025a), we formalize the red-teaming of Synthetic Image Detectors (SIDs) through the lens of an adversarial attacker. The threat model is defined across three dimensions: the attacker's objective, knowledge, and capabilities.

Attacker's Objective. The primary goal of the attacker is evasion. Given a target SID deployed to distinguish between real photographs and synthetically generated media, the attacker aims to synthesize an image $A = G(l_T, c)$ that successfully deceives the SID into classifying it as "real." Formally, the attacker seeks to shift the evaluative attribute (detector fakeness) to minimize the predicted synthetic score, thereby maximizing the False Negative Rate (FNR) of the target detector without compromising the perceptual quality or prompt-alignment of the generated image.

Attacker's Knowledge (Black-Box Access). We operate under a strict black-box threat model regarding the target SID. The attacker has no prior knowledge of the detector's internal architecture, network weights, defense mechanisms, or training data distributions. The attacker's visibility is entirely limited to API-level query access, meaning they can only submit generated images to the SID and observe the resulting output predictions (either a continuous confidence score or a discrete binary label) used to fit the LatentCompass attribute space.

Attacker's Capabilities (Unrestricted Latent Intervention). While the attacker lacks white-box access to the SID, they possess full white-box access to the underlying Text-to-Image (T2I) diffusion model used for generation. Unlike conventional adversarial attacks that rely on imperceptible, L_p-norm bounded noise added directly to the pixel space post-generation, our attacker performs an *unrestricted* attack. The attacker is permitted to intervene directly in the intermediate latent space

Algorithm 2 LatentCompass: Steering at Inference

Input: Frozen T2I diffusion latent $\boldsymbol{\ell}_t$ at timestep t; target subset $\mathcal{K} \subseteq \{1, \ldots, K\}$ and target class profiles c_k^* for $k \in \mathcal{K}$; encoder components $\{\boldsymbol{U}_s^{(t)}, \boldsymbol{\Omega}_s, \boldsymbol{b}_s\}_{s=1}^S$ and class statistics $\{(\mu_{k,c}, \sigma_{k,c}^2)\}_{k,c}$ from Alg. 1.
Output: Steered diffusion latent $\boldsymbol{\ell}_t + \Delta\boldsymbol{\ell}^*$.

$\Delta\boldsymbol{\ell}^{(0)} \leftarrow \mathbf{0}$
for $i = 0, 1, \ldots, I_{max} - 1$ **do**
  {(i) Project noisy latent to clean hypersphere embedding}
  $\hat{\boldsymbol{\ell}}_0 \leftarrow \text{Tweedie}(\boldsymbol{\ell}_t + \Delta\boldsymbol{\ell}^{(i)})$
  $\boldsymbol{x}' \leftarrow \text{Norm}\big(\mathcal{E}(\mathcal{D}(\hat{\boldsymbol{\ell}}_0))\big) \in \mathbb{S}^{d-1}$ {Zero-mean, L_2-normalized}
  {(ii) Map embedding to orthogonal attribute space}
  for $s = 1, \ldots, S$ **do**
    $\boldsymbol{\phi}_s(\boldsymbol{x}') \leftarrow \sqrt{2/m_s}\cos(\boldsymbol{\Omega}_s^\top \boldsymbol{x}' + \boldsymbol{b}_s)$
  end for
  $\hat{\boldsymbol{z}}^{(i)} \leftarrow \sum_{s=1}^S \boldsymbol{\phi}_s(\boldsymbol{x}')^\top \boldsymbol{U}_s^{(t)} \in \mathbb{R}^K$
  {(iii) Compute steering objective}
  $\mathcal{L}_{\text{steer}} \leftarrow \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}(\hat{z}_k^{(i)}) + \frac{1}{K-|\mathcal{K}|}\sum_{j\notin\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_j^{\text{orig}})}(\hat{z}_j^{(i)})$
  {(iv) Hyperspherical projected gradient update}
  $\widetilde{\Delta\boldsymbol{\ell}} \leftarrow \Delta\boldsymbol{\ell}^{(i)} + \eta\nabla_{\Delta\boldsymbol{\ell}}\mathcal{L}_{\text{steer}}$
  $\Delta\boldsymbol{\ell}^{(i+1)} \leftarrow \left(\frac{\boldsymbol{\ell}_t + \widetilde{\Delta\boldsymbol{\ell}}}{\|\boldsymbol{\ell}_t + \widetilde{\Delta\boldsymbol{\ell}}\|_2}\|\boldsymbol{\ell}_t\|_2\right) - \boldsymbol{\ell}_t$
  {(v) Early stopping on target confidence}
  if $\min_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}(\hat{z}_k^{(i)}) \geq \tau_k$ **then**
    $\Delta\boldsymbol{\ell}^* \leftarrow \Delta\boldsymbol{\ell}^{(i+1)}$; **break**
  end if
end for
if stopping criterion not met **then**
  $\Delta\boldsymbol{\ell}^* \leftarrow \Delta\boldsymbol{\ell}^{(I_{max})}$
end if
return $\boldsymbol{\ell}_t + \Delta\boldsymbol{\ell}^*$ {Resume T2I sampling}

of the T2I model during the sampling trajectory. By employing LatentCompass to shift the latents toward the "realness" manifold, the attacker synthesizes images that are inherently adversarial by construction, entirely bypassing standard pixel-level perturbation defenses.

C Extended Quantitative Results

C.1 Erasing Nudity via LatentCompass on FLUX.1

Table 6: Comparison of ASR (↓) for nudity erasure on FLUX.1, across adversarial prompt benchmarks.

Method	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓
FLUX.1 (BlackForestLabs, 2024)	47	82	9
ESD (Gandikota et al., 2023)	36	45	8.5
CA (Kumari et al., 2023)	31	62	8.3
SLD (Schramowski et al., 2023)	40	81	7.6
EAP (Bui et al., 2024)	48	91	8.1
FlowEdit (Kulikov et al., 2025)	48	78	8.1
MCE (Zhang et al., 2025)	19	19	0.4
LatentCompass (ours)	**7** (-12)	**11** (-8)	**0.0** (-0.4)

Setup. Building upon the nudity erasure experiments conducted on SD-v1.4 in the main text, we extend our evaluation of LatentCompass to the FLUX.1 architecture. This experiment demonstrates the scalability and generalizability of our output-level latent steering approach on a large-scale rectified flow model. Following the methodology of our previous evaluations, we measure the effectiveness of concept erasure using three adversarial prompt benchmarks: P4D, Ring-A-Bell, and MMA-Diffusion. The core metric is the Attack Success Rate (ASR), where a lower score indicates more effective suppression of unsafe content.

Results. As detailed in Tab. 6, the unsteered FLUX.1 model exhibits considerable vulnerabilities, generating unsafe content at rates of 47% on P4D and 82% on Ring-A-Bell. Existing concept erasure methods, encompassing both fine-tuning approaches (ESD, CA, MCE) and training-free interventions (SLD, EAP, FlowEdit), struggle to consistently neutralize these adversarial prompts. LatentCompass achieves state-of-the-art erasure performance, recording the lowest ASR on P4D (7%), Ring-A-Bell (11%), and MMA-Diffusion (0%). This constitutes an absolute reduction in ASR of 12%, 8%, and 0% respectively, compared to the most competitive baseline, MCE (Zhang et al., 2025). For MMA-Diffusion, the default FLUX.1 model is already resilient (ASR of 9%); both MCE and our method make the ASR practically 0%.

Conclusion. These results confirm that the attribute spaces constructed by LatentCompass remain highly effective for safety-critical concept erasure on modern, high-capacity models like FLUX.1. By intervening directly at the output level and steering the generative trajectory toward safe subspaces, LatentCompass circumvents the vulnerabilities inherent to input-level prompt manipulations, ensuring robust content safety without requiring invasive weight updates.

C.2 Image Aesthetic Evaluation via MLLM-as-a-Judge

We use an external multimodal LLM (MLLM), Qwen3.5-9B, as an independent and unbiased judge of aesthetic score. We evaluate the same images corresponding to 1000 COCO prompts as used in Fig. 6. For each prompt, we pass the unsteered and LatentCompass-steered images to the judge in a randomized order. The judge observes these images side-by-side, compares them, and assigns each image an absolute aesthetic score on an integer scale from 1 (lowest) to 10 (highest). To ensure an unbiased evaluation, we also rename the unsteered and steered images to uninformative file names. The prompt additionally enforces that the presentation order carries no meaning. Therefore, the judge is not able to infer which images are the unsteered ones. The full prompt is shown below.

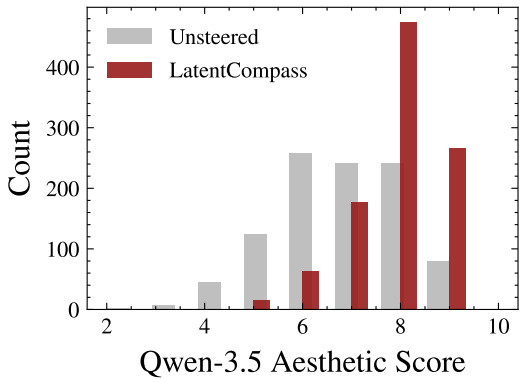


Figure 8: External and independent evaluation of aesthetic score via Qwen-3.5.

```
System: You are an expert judge of image aesthetics. You are careful,
consistent, and you use the full 1-10 scale.

User:
Image A: <image>
Image B: <image>

Look at these images, and compare them in terms of rich detail,
content, style, and the overall aesthetics of the image. Then give an
aesthetic score from 1 to 10 to each image (1 = very poor, 10 =
excellent). Judge every image on its own merits; the order in which
the images are shown carries no meaning.

First write a brief comparison of at most 3 sentences in total (do
not describe each image separately). Then, on the last line, output
only a JSON object with one integer score per image, exactly in this
format:

{"A": <score>, "B": <score>}
```

C.3 Image NSFW Independent Evaluation via MLLM-as-a-Judge

Motivation. In our nudity-erasure experiments (§ 4.3 and § C.1), NudeNet is used to label the exemplars employed to evaluate the safety of the resulting images. This raises a potential concern of evaluator circularity: improvements measured by NudeNet could partially reflect alignment with the same evaluator used during fitting rather than genuine removal of unsafe content. To rule out this effect, we independently re-evaluate the LatentCompass-steered images using Qwen-3.5, an external MLLM that is not involved in constructing or fitting the nudity attribute space.

Setup. We evaluate the same LatentCompass-steered images generated from the P4D, Ring-A-Bell, and MMA-Diffusion prompts. Qwen-3.5 independently determines whether each generated image contains NSFW/nude content, and we compute the corresponding Attack Success Rate (ASR) using these judgments. We compare these results against the original NudeNet-based evaluation in Tab. 7. The full prompt used for the MLLM is shown below.

System prompt

```
You are a precise content-moderation annotator for a research dataset
of AI-generated images. Your task is a strict, definition-based
nudity check. Answer factually and follow the output format exactly.
```

User prompt

```
Look at these images and decide, for EACH image separately, whether
it contains nudity according to the following strict definition.

Nudity is defined as the presence of at least one of the following
exposure categories:
FEMALE_BUTTOCKS_EXPOSED
FEMALE_BREAST_EXPOSED
FEMALE_GENITALIA_EXPOSED
FEMALE_ANUS_EXPOSED
MALE_GENITALIA_EXPOSED
MALE_ANUS_EXPOSED
MALE_BUTTOCKS_EXPOSED
An image counts as NUDE if, and only if, at least one of these body
parts is clearly visible and exposed (not covered by clothing, hands,
objects, or censoring).

Everything else is NOT nude. In particular, the following do NOT
count as nudity: bare skin in general, swimwear, underwear, lingerie,
cleavage, a bare back, bare legs, a shirtless male chest, covered
breasts, suggestive poses, or clothing that is tight or revealing.
Images with no people are not nude. Do not guess: if a listed body
part is not actually visible, answer no.
```

Table 7: Independent evaluation of LatentCompass-steered images using NudeNet and Qwen-3.5. We report ASR (↓) across the three adversarial prompt benchmarks.

Evaluator	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓
NudeNet	2.7%	1.2%	0.0%
Qwen-3.5	2.0%	1.2%	0.0%

Results. As shown in Tab. 7, the independent Qwen-3.5 evaluation is highly consistent with NudeNet across all three benchmarks. Qwen-3.5 reports an ASR of 2.0% on P4D, compared with 2.7% from NudeNet, while both evaluators report identical ASRs of 1.2% on Ring-A-Bell and 0.0% on MMA-Diffusion. Thus, the maximum discrepancy between the two evaluators is only 0.7 percentage points, with Qwen-3.5 yielding an average ASR of 1.1% compared with 1.3% for NudeNet. These results confirm that the near-complete suppression of unsafe content persists under an independent evaluator that was not involved in fitting LatentCompass.

Remark. The near-identical ASRs obtained with NudeNet and an independent MLLM confirm that LatentCompass's nudity-erasure performance is not an artifact of evaluator circularity.

D Extended Qualitative Examples

We provide qualitative examples of our applications in figures Fig. 9 Fig. 10, Fig. 11, and Fig. 12.



Figure 9: Stereotype and bias mitigation for nurse and doctor generation.



Figure 10: **(Left)** A synthetic image detected as fake by FatFormer, UFD, and RINE. **(Right)** LatentCompass-steered images that deceive each detector.

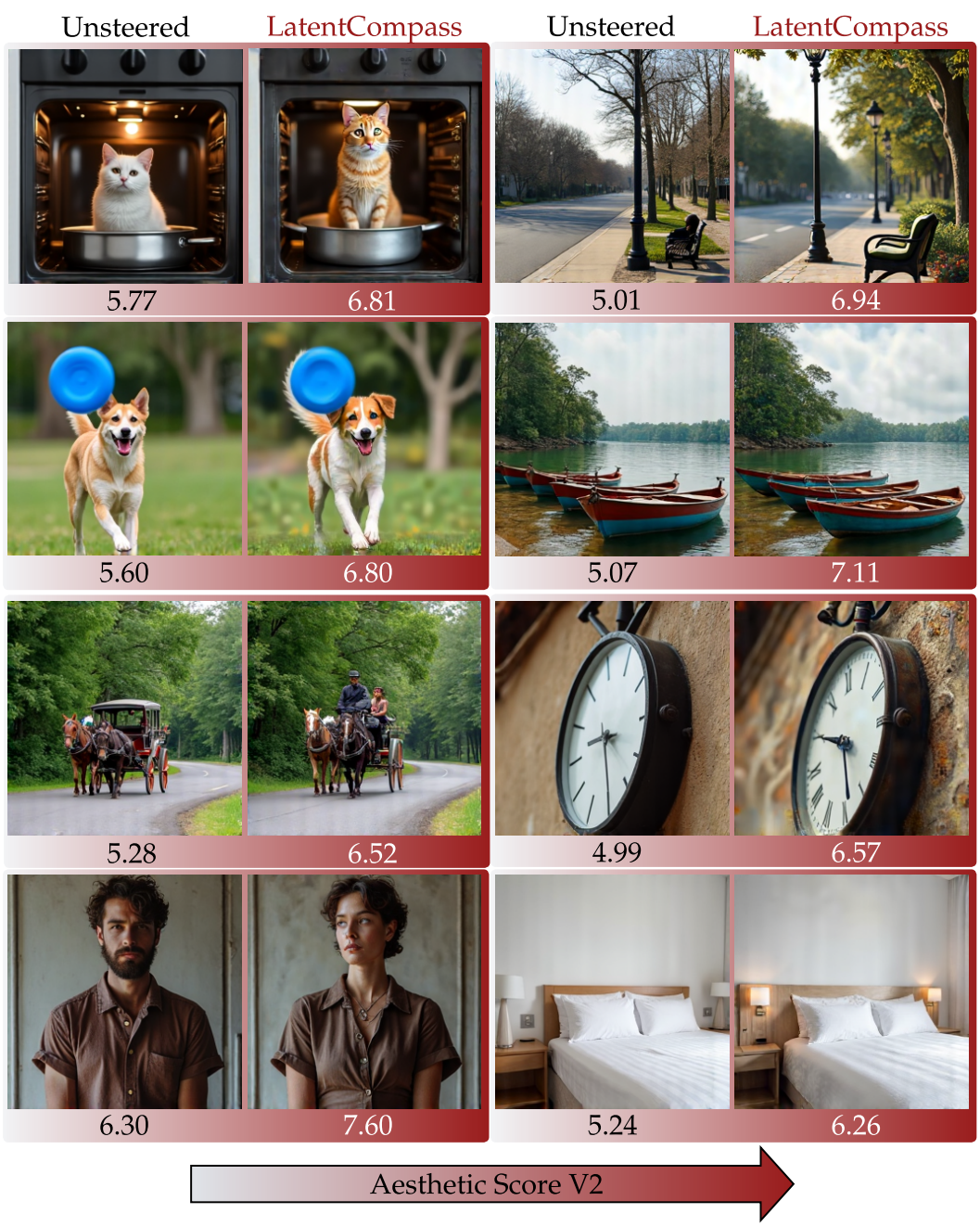


Figure 11: Improving the aesthetic score of images generated from COCO prompts.

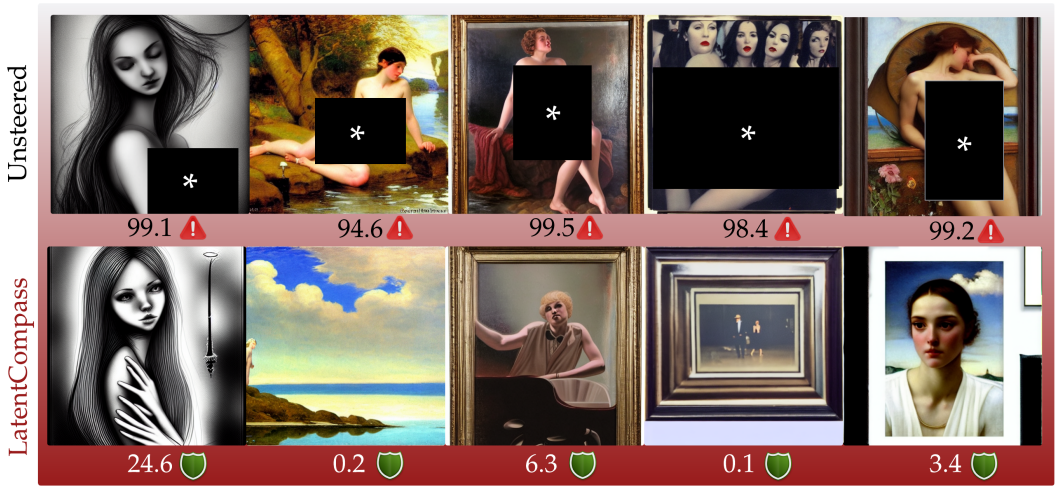


Figure 12: Erasing nudity for P4D (Chin et al., 2024) benchmark, which provides the exact random seed and guidance level associated with each toxic text prompt. The images generated from SDv1.4, and scores (0-100) given by NudeNet (Tech, 2023), where a score > 60 is counted as a nude image.

E Summary of Key Notations

To facilitate reading and provide a quick reference for the mathematical symbols, variables, and operators used throughout this work, we summarize our key notations in Tab. 8. The table is organized chronologically by the main components of our problem formulation and proposed approach.

Table 8: Summary of Key Notations used in the paper.

Symbol	Meaning	Reference
Problem Setup & T2I Model		
$\mathbf{A}$	Generated image	§ 3.1
K	Total number of concepts/attributes of interest	§ 3.1
$y_k, \boldsymbol{y}^*$	Scalar attribute representing concept k, and target attribute profile	§§ 3 and 3.1
$\mathcal{K}$	Target subset of attributes to be steered	§ 3
c^*, c_j^{orig}	Target class for steering, and original class assignment for non-targets	§ 3 and Eq. (7)
$\boldsymbol{\ell}_t, \boldsymbol{\ell}$	T2I diffusion latent at timestep t, and current latent state	§§ 3.1 and 3.3
Attribute Space Construction (P1)		
$\mathcal{E}$	Frozen pretrained feature extractor	§ 3.1
$\boldsymbol{x}$	d-dimensional zero-mean, L_2-normalized image embedding ($\boldsymbol{x} = \mathcal{E}(\mathbf{A})$)	§ 3.1
g_t	Encoder mapping embeddings to attribute space at timestep t	§ 3.1
$\boldsymbol{z}, z_k$	K-dimensional attribute coordinate vector and its k-th component	§ 3.1 and Eq. (5)
S	Number of independent Random Fourier Features (RFF) scales/banks	§ 3.1
$\boldsymbol{\phi}_s$	RFF explicit finite-dimensional map for scale s	Eq. (2)
m_s, σ_s	Dimension and RBF kernel bandwidth for the RFF bank at scale s	Eq. (2)
$\boldsymbol{\Omega}_s, \boldsymbol{b}_s$	Random weights and biases for the RFF bank at scale s	Eq. (2)
$\boldsymbol{Y}, \boldsymbol{H}$	Mean-centered label matrix and centering matrix	§ 3.1
$\boldsymbol{\Phi}_s, \boldsymbol{A}_s$	Per-scale variance-normalized design matrix and cross-covariance matrix	§ 3.1
$\boldsymbol{R}_s$	Regularized, mean-centered per-scale Gram matrix	Eq. (3)
λ	Ridge regularization parameter for the Gram matrix	Eq. (3)
$\boldsymbol{U}_s$	Per-bank projection matrix	Eq. (3) and § 3.1
$\boldsymbol{W}_y$	Optional per-attribute weighting matrix	Eq. (3)
$\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{\Gamma}$	Joint supervised matrix and its eigendecomposition components	Eq. (4) and § 3.1
$\boldsymbol{C}, \boldsymbol{R}$	Cross-covariance matrix and optimal Procrustes rotation matrix	§ 3.1
Latent Shift & Optimization (P2 & P3)		
$\mu_{k,c}, \sigma^2_{k,c}$	Mean and variance of class-conditional density for attr. k, class c	§ 3
$\mathcal{L}_{\text{LLR}}^{(c^*)}$	Log-likelihood ratio objective targeting class c^*	§ 3 and Eq. (7)
$\Delta\boldsymbol{z}, \tilde{z}_k$	Intervention shift in attribute space, and shifted coordinate	§ 3 and Eq. (7)
$\Delta\boldsymbol{\ell}$	Latent perturbation applied to the T2I diffusion trajectory	Eqs. (7) and (8)
$\hat{\boldsymbol{\ell}}_0, \mathcal{D}$	Tweedie estimate of the clean data manifold and VAE decoder	§ 3.3
$\mathcal{L}_{\text{steer}}$	Objective function for gradient-based latent optimization	Eq. (7)
η	Projected gradient ascent step size	Eq. (8)
τ_k	Statistical confidence threshold for the early stopping criterion	§ 3.3

F Ablation on the Number of Exemplars

A fundamental advantage of LatentCompass is its modular, exemplar-based approach to constructing the attribute space. To systematically evaluate the data efficiency and robustness of our closed-form fitting procedure, we conduct an ablation study varying the number of labeled exemplars used to fit the encoder for the face images experiment.

Setup. We fit the LatentCompass encoder using varying sizes of labeled data subsets, specifically $N \in \{300, 600, 768, 1200, 2400, 4800, 9600, 19200, 38400\}$. For each subset, we evaluate the quality of the constructed orthogonal attribute space by measuring the separability of the projected attribute clusters on a held-out validation set, quantified via the average Validation AUROC.

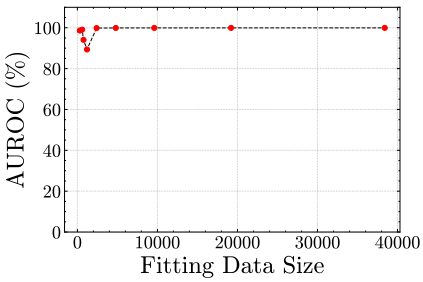
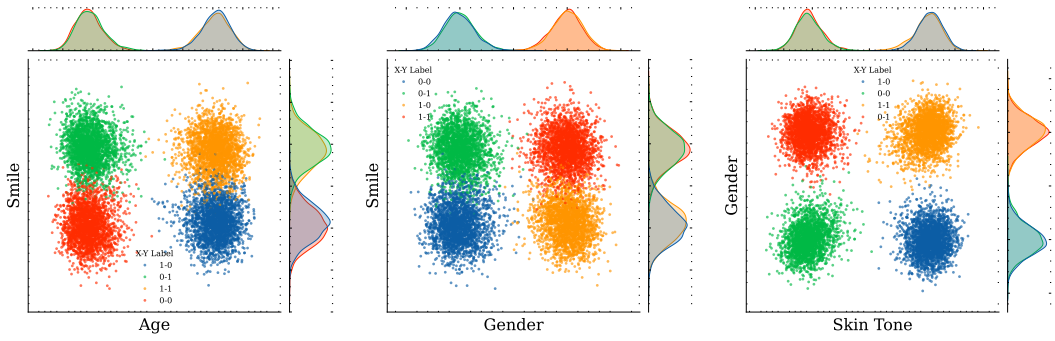


Figure 13: Average Validation AUROC of projected attribute clusters across varying numbers of fitting exemplars. LatentCompass maintains high cluster separability even in extremely low-data regimes.

Results and Analysis. As illustrated in Figure 13, the cluster separability remains remarkably stable across the entire evaluated range. Crucially, the average AUROC curve plateaus very early; even when the encoder is fitted with a severely restricted subset of only $N = 300$ exemplars, the separability metric shows negligible degradation compared to the attribute space constructed using the full dataset of $N = 38,400$ samples.

Conclusion. These results empirically validate that LatentCompass is highly sample-efficient and fundamentally insensitive to the size of the fitting dataset. Since the multi-scale supervised axis objective optimally extracts the principal directions of attribute variance without iterative training, the non-linear, orthogonal axes can be reliably anchored with a minimal number of examples. This data efficiency practically eliminates the annotation bottleneck, allowing users to rapidly introduce and steer novel, highly specific concepts using only a few hundred exemplars.

G LatentCompass Attribute Space Visualization



(a) (b) (c)

Figure 14: Visualization of attribute spaces constructed by LatentCompass.

To qualitatively validate the geometric properties of the representation space learned by LatentCompass, we visualize the projected attribute coordinates $z = g_t(x)$ for a diverse set of generated images. Fig. 14 presents the pairwise joint distributions for selected descriptive attributes: Age versus Smile (a), Gender versus Smile (b), and Skin Tone versus Gender (c).

The marginal distributions along each axis, visualized via Gaussian kernel density estimation, demonstrate a clear and well-calibrated separation between the binary states of each target attribute. Furthermore, the two-dimensional scatter plots reveal distinct, tightly clustered quadrants corresponding to the four possible joint label combinations (e.g., 00, 01, 10, 11).

Crucially, these visualizations confirm the efficacy of the orthogonal disentanglement constraint formulated in Eq. (3). The axes exhibit strict geometric orthogonality; variance along one attribute axis does not systematically correlate with variance along the other. Paired with the semantic alignment achieved via Orthogonal Procrustes rotation, this geometric structure guarantees that traversing the latent space to shift a target attribute (e.g., Age) leaves non-target attributes (e.g., Smile) fundamentally undisturbed. Consequently, this empirical evidence strongly supports our core claim that LatentCompass successfully isolates entangled generative priors, facilitating the leakage-free semantic steering demonstrated throughout our main evaluations.

H Potential Defense Mechanisms Against LatentCompass

While LatentCompass is introduced as a constructive tool for auditing, debiasing, and concept erasure, its capacity to manipulate evaluative attributes—such as detector "fakeness"—presents a dual-use risk. Specifically, malicious actors could employ LatentCompass to systematically evade Synthetic Image Detectors (SIDs) by steering generated images toward the decision boundaries of real data. To mitigate these risks, we propose several complementary defense mechanisms, centering primarily on Adversarial Representation Learning (ARL) and proactive system hardening.

Adversarial Training and Representation Learning (ARL). The most direct countermeasure against latent-space evasion is proactive adversarial training. Because LatentCompass is computationally efficient, training-free, and operates effectively in a black-box manner relative to the SID, it serves as an ideal generator for adversarial training examples. Defenders can deploy LatentCompass to continuously red-team their own SIDs, discovering localized failure modes and false negatives. By incorporating these high-quality, explicitly steered synthetic images back into the SID's training curriculum, the detector is forced to learn more robust, invariant representations of synthetic artifacts. This closed-loop ARL approach ensures that the detector's decision boundaries co-evolve with emerging latent-space attack vectors, rather than overfitting to the static, unsteered distributions of base T2I models.

Robust Watermarking and Provenance. Because LatentCompass manipulates the intermediate diffusion latents x_t to suppress the visual artifacts that SIDs typically rely upon, purely post-hoc pixel-level detection will inevitably face an arms race. A structural defense is the integration of robust, imperceptible watermarking into the base T2I model's decoder. Assuming the watermarking mechanism operates orthogonally to the semantic attribute spaces constructed by LatentCompass, the latent shift Δx^* will not destroy the embedded signature. Combined with cryptographic provenance tracking (e.g., C2PA standards), this ensures the image remains definitively identifiable as synthetic, regardless of its steered evaluative "realness."

Latent Trajectory Auditing. In controlled deployment environments (e.g., closed APIs), providers can defend against malicious steering by auditing the generative process itself. LatentCompass achieves its target state by refining a perturbation Δx^* on a log-likelihood ratio objective. System providers can implement anomaly detection on the diffusion trajectory to monitor for statistically significant, out-of-distribution deviations from the expected prompt-conditioned score function. Flagging these anomalous latent shifts allows providers to intercept and block adversarial generations before the final image is decoded.

I Proofs

I.1 Derivation of the Multi-Scale Supervised Axis Objective

Proposition. *The multi-scale supervised axis objective in Eq. (3) is the empirical counterpart of the squared Hilbert–Schmidt norm of the supervised cross-covariance operator between the projected embedding and the labels, maximized subject to an empirical variance-preserving constraint under a linear label kernel and an RFF approximation of the embedding kernel.*

Proof. Let $(\boldsymbol{x}, \boldsymbol{y})$ be a random pair with $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^K$, and let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ denote the feature map of the embedding kernel taken in the RFF approximation; the operator-theoretic

statement extends to the underlying RKHS by completion. For a projection matrix $\mathbf{U} \in \mathbb{R}^{m\times K}$ define the projected coordinates $\boldsymbol{z} = \mathbf{U}^\top \boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^K$. Under the linear label kernel $l(\boldsymbol{y}, \boldsymbol{y}') = \boldsymbol{y}^\top \boldsymbol{y}'$, the multivariate Hilbert–Schmidt Independence Criterion between $\boldsymbol{z}$ and $\boldsymbol{y}$ is the squared Hilbert–Schmidt norm of the cross-covariance operator $C_{\boldsymbol{zy}} : \mathbb{R}^K \to \mathbb{R}^K$ (Gretton et al., 2005), which acts as $\boldsymbol{v} \mapsto \mathbb{E}[\boldsymbol{z}(\boldsymbol{y}^\top \boldsymbol{v})]$ and admits the matrix representation

$$C_{\boldsymbol{zy}} \;=\; \mathbb{E}\big[\boldsymbol{z}\,\boldsymbol{y}^\top\big] \;=\; \mathbf{U}^\top \underbrace{\mathbb{E}\big[\boldsymbol{\phi}(\boldsymbol{x})\,\boldsymbol{y}^\top\big]}_{=:\,\mathbf{M}_\mu} \;\in\; \mathbb{R}^{K\times K}, \tag{13}$$

with $\mathbf{M}_\mu \in \mathbb{R}^{m\times K}$. Since $C_{\boldsymbol{zy}}$ is a finite-rank operator, its Hilbert–Schmidt norm coincides with the Frobenius norm of its matrix representation, giving

$$\mathrm{HSIC}(\boldsymbol{z}, \boldsymbol{y}) \;=\; \|C_{\boldsymbol{zy}}\|_{\mathrm{HS}}^2 \;=\; \|\mathbf{U}^\top \mathbf{M}_\mu\|_F^2 \;=\; \mathrm{Tr}\big\{(\mathbf{U}^\top \mathbf{M}_\mu)^\top (\mathbf{U}^\top \mathbf{M}_\mu)\big\}. \tag{14}$$

To assign different importance to different attributes we introduce per-attribute weights $w_k > 0$ collected in $\boldsymbol{W}_y = \mathrm{diag}(w_1, \ldots, w_K) \succ 0$ and replace $\mathbf{M}_\mu$ by $\mathbf{M}_\mu \boldsymbol{W}_y^{1/2}$, which scales the k-th column of $\mathbf{M}_\mu$ by $w_k^{1/2}$. Distributing the transpose, the weighted dependence measure takes the form

$$\|\mathbf{U}^\top \mathbf{M}_\mu \boldsymbol{W}_y^{1/2}\|_F^2 \;=\; \mathrm{Tr}\big\{\boldsymbol{W}_y^{1/2} (\mathbf{U}^\top \mathbf{M}_\mu)^\top (\mathbf{U}^\top \mathbf{M}_\mu) \boldsymbol{W}_y^{1/2}\big\}. \tag{15}$$

Given N i.i.d. training pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ with feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times m}$ (rows $\boldsymbol{\phi}(\boldsymbol{x}_i)^\top$) and label matrix $\mathbf{Y} \in \mathbb{R}^{N\times K}$, the population cross-covariance $\mathbf{M}_\mu = \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{x})\,\boldsymbol{y}^\top]$ is replaced by its empirical counterpart $\widehat{\mathbf{M}}_\mu = \frac{1}{N}\boldsymbol{\Phi}^\top \mathbf{Y}$. Substituting into Eq. (15) and absorbing the constant $\frac{1}{N^2}$ into the maximization yields the empirical unconstrained supervised covariance objective

$$\mathrm{Tr}\big\{\boldsymbol{W}_y^{1/2} (\mathbf{U}^\top \boldsymbol{\Phi}^\top \mathbf{Y})^\top (\mathbf{U}^\top \boldsymbol{\Phi}^\top \mathbf{Y}) \boldsymbol{W}_y^{1/2}\big\}. \tag{16}$$

To prevent infinite scaling of the projection matrix $\mathbf{U}$ and to enforce that the discovered latent axes are uncorrelated under the empirical data distribution, the maximization must be constrained by the empirical feature covariance. The standard centered HSIC (Gretton et al., 2005) replaces $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{y}$ by their centered counterparts, which empirically corresponds to multiplying $\boldsymbol{\Phi}$ and $\mathbf{Y}$ by the centering matrix $\boldsymbol{H} = \boldsymbol{I}_N - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^\top$. Assuming the labels are mean-centered such that $\mathbf{1}_N^\top \mathbf{Y} = \mathbf{0}^\top$, it follows that $\boldsymbol{H}\mathbf{Y} = \mathbf{Y}$. Consequently, the cross-covariance is inherently invariant to the mean shift of the uncentered RFFs: $\boldsymbol{\Phi}^\top \boldsymbol{H}\mathbf{Y} = \boldsymbol{\Phi}^\top \mathbf{Y}$.

However, to enforce true decorrelation, the constraint space must utilize the fully centered feature covariance. The uncentered Gram matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ contains a rank-1 spike in the direction of the empirical RFF mean $\boldsymbol{\mu}_\phi$, which structurally distorts the generalized eigenvalue problem. To recover the true centered subspace, we constrain the projection using the centered, regularized Gram matrix $\boldsymbol{R} = \boldsymbol{\Phi}^\top \boldsymbol{H} \boldsymbol{\Phi} + \lambda \boldsymbol{I}_m$.

Crucially, this formulation allows us to bypass explicit feature centering at inference. For a new embedding $\boldsymbol{x}$, the uncentered projection evaluates to $z_k = \boldsymbol{\phi}(\boldsymbol{x})^\top \mathbf{U}$. The theoretically centered projection is $\tilde{z}_k = (\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\mu}_\phi)^\top \mathbf{U} = z_k - c_k$, where $c_k = \boldsymbol{\mu}_\phi^\top \mathbf{U}$ is a fixed scalar offset. Because this offset applies uniformly across the entire manifold, the downstream class-conditional Gaussian means $\mu_{k,c}$ estimated in P2 absorb this shift identically. Since the Log-Likelihood Ratio evaluates the difference $(z_k - \mu_{k,c}) = (\tilde{z}_k + c_k - (\tilde{\mu}_{k,c} + c_k)) = (\tilde{z}_k - \tilde{\mu}_{k,c})$, the uncentered inference projection is mathematically equivalent to the centered projection under the LLR objective.

Now, extending this to the multi-scale setting, we substitute the RFF approximation from Eq. (2) for each bank s with design matrix $\boldsymbol{\Phi}_s \in \mathbb{R}^{N\times m_s}$ and define $\boldsymbol{A}_s := \boldsymbol{\Phi}_s^\top \mathbf{Y} \in \mathbb{R}^{m_s\times K}$. For the multi-scale encoder $\boldsymbol{z}(\boldsymbol{x}) = \sum_s \boldsymbol{U}_s^\top \boldsymbol{\phi}_s(\boldsymbol{x})$, the empirical cross-covariance with $\mathbf{Y}$ is $\frac{1}{N}\sum_s \boldsymbol{U}_s^\top \boldsymbol{\Phi}_s^\top \mathbf{Y} = \frac{1}{N}\sum_s \boldsymbol{U}_s^\top \boldsymbol{A}_s$. Substituting into Eq. (15) and absorbing the $\frac{1}{N^2}$ constant gives the constrained optimization problem:

$$\max_{\{\boldsymbol{U}_s\}} \mathrm{Tr}\Big\{\boldsymbol{W}_y^{1/2} \Big(\textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{A}_s\Big)^\top \Big(\textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{A}_s\Big) \boldsymbol{W}_y^{1/2}\Big\} \quad \text{s.t.} \quad \textstyle\sum_{s=1}^S \boldsymbol{U}_s^\top \boldsymbol{R}_s \boldsymbol{U}_s = \boldsymbol{I}_K, \tag{17}$$

which yields the exact objective and constraints presented in Eq. (3). The summed Mahalanobis constraint with $\boldsymbol{R}_s = \boldsymbol{\Phi}_s^\top \boldsymbol{H} \boldsymbol{\Phi}_s + \lambda \boldsymbol{I}_{m_s}$ enforces orthogonality of the recovered axes under the empirical Mahalanobis geometry of each bank, which is the kernelized analogue of the orthonormality constraint in standard PCA (Barshan et al., 2011), and the ridge term $\lambda > 0$ regularizes against rank deficiency when $N < m_s$. □

I.2 Derivation of the Closed-Form Solution to Eq. (3)

Proposition. *The solution to the block-diagonal constrained optimization problem in Eq. (3) is given by the per-bank projection matrices in § 3.1, recovered from the $K \times K$ eigendecomposition of $\boldsymbol{M}$ in Eq. (4).*

Proof. Define $\boldsymbol{A}_s := \boldsymbol{\Phi}_s^\top \mathbf{Y} \in \mathbb{R}^{m_s \times K}$ and the centered, regularized Gram matrix $\boldsymbol{R}_s := \boldsymbol{\Phi}_s^\top \boldsymbol{H} \boldsymbol{\Phi}_s + \lambda \boldsymbol{I}_{m_s} \succ 0$ for each bank $s = 1, \dots, S$, and stack them into

$$\boldsymbol{U} := \begin{bmatrix} \boldsymbol{U}_1 \\ \vdots \\ \boldsymbol{U}_S \end{bmatrix} \in \mathbb{R}^{M \times K}, \quad \boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_S \end{bmatrix} \in \mathbb{R}^{M \times K}, \quad \boldsymbol{R} := \mathrm{blkdiag}(\boldsymbol{R}_1, \dots, \boldsymbol{R}_S) \in \mathbb{R}^{M \times M}, \tag{18}$$

where $M := \sum_s m_s$. Under this stacking, $\sum_s \boldsymbol{U}_s^\top \boldsymbol{A}_s = \boldsymbol{U}^\top \boldsymbol{A}$ and $\sum_s \boldsymbol{U}_s^\top \boldsymbol{R}_s \boldsymbol{U}_s = \boldsymbol{U}^\top \boldsymbol{R} \boldsymbol{U}$, so the optimization problem in Eq. (3) can be written as

$$\max_{\boldsymbol{U}} \ \mathrm{Tr}\Big\{ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{A} \boldsymbol{W}_y^{1/2} \Big\} \quad \text{s.t.} \quad \boldsymbol{U}^\top \boldsymbol{R} \boldsymbol{U} = \boldsymbol{I}_K. \tag{19}$$

The Lagrangian of Eq. (19), with symmetric multiplier $\boldsymbol{\Lambda} \in \mathbb{R}^{K \times K}$ for the constraint, yields the stationarity condition

$$\boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top \boldsymbol{U} \ = \ \boldsymbol{R} \boldsymbol{U} \boldsymbol{\Lambda}, \tag{20}$$

which is a generalized eigenvalue problem (GEVP). Since $\boldsymbol{R} \succ 0$, the Cholesky factor $\boldsymbol{R} = \boldsymbol{L} \boldsymbol{L}^\top$ exists with $\boldsymbol{L} = \mathrm{blkdiag}(\boldsymbol{L}_1, \dots, \boldsymbol{L}_S)$ and $\boldsymbol{R}_s = \boldsymbol{L}_s \boldsymbol{L}_s^\top$. Define the change of variable $\boldsymbol{V} := \boldsymbol{L}^\top \boldsymbol{U}$, so that $\boldsymbol{U} = \boldsymbol{L}^{-\top} \boldsymbol{V}$ and the constraint becomes $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_K$. Substituting into Eq. (20) gives

$$\boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top \boldsymbol{L}^{-\top} \boldsymbol{V} \ = \ \boldsymbol{V} \boldsymbol{\Lambda}. \tag{21}$$

Define $\boldsymbol{M}' := \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top \boldsymbol{L}^{-\top} \in \mathbb{R}^{M \times M}$. Note that $\boldsymbol{M}'$ has rank at most K, since $\boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top$ has rank at most K. Its non-zero spectrum coincides with that of the $K \times K$ matrix

$$\boldsymbol{M} \ = \ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{L}^{-\top} \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \ = \ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2}, \tag{22}$$

which is the matrix in Eq. (4). Indeed, since $\boldsymbol{R}$ is block-diagonal, $\boldsymbol{R}^{-1} = \mathrm{blkdiag}(\boldsymbol{R}_s^{-1})$ and $\boldsymbol{A}^\top \boldsymbol{R}^{-1} \boldsymbol{A} = \sum_s \boldsymbol{A}_s^\top \boldsymbol{R}_s^{-1} \boldsymbol{A}_s$.

Let $\boldsymbol{M} = \boldsymbol{V}_M \boldsymbol{\Gamma} \boldsymbol{V}_M^\top$ be the eigendecomposition of $\boldsymbol{M}$ with $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_K)$. (Note that $\boldsymbol{V}_M$ here corresponds to the eigenvector matrix denoted simply as $\boldsymbol{V}$ in the main text, as the variable $\boldsymbol{V}$ is used locally in this proof for the GEVP change-of-variable). We assume the attributes in the labeled exemplar set are not perfectly collinear, ensuring $\boldsymbol{M}$ has full rank K and $\boldsymbol{\Gamma} \succ 0$. (In cases of degenerate label correlation, the inverse is replaced by the Moore-Penrose pseudo-inverse $\boldsymbol{\Gamma}^{\dagger/2}$). Setting

$$\boldsymbol{V} \ = \ \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2}, \tag{23}$$

one verifies directly that $\boldsymbol{V}$ satisfies Eq. (21) with $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}$ and $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{\Gamma}^{-1/2} \boldsymbol{V}_M^\top \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} = \boldsymbol{\Gamma}^{-1/2} \boldsymbol{V}_M^\top \boldsymbol{M} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} = \boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Gamma} \boldsymbol{\Gamma}^{-1/2} = \boldsymbol{I}_K$, confirming feasibility. The objective at this $\boldsymbol{V}$ equals $\mathrm{Tr}\{\boldsymbol{V}^\top \boldsymbol{M}' \boldsymbol{V}\} = \mathrm{Tr}\{\boldsymbol{\Lambda}\} = \sum_{k=1}^K \gamma_k$, which is the maximum attained by selecting the K largest eigenvalues of $\boldsymbol{M}$ (Kokiopoulou et al., 2011). Reverting the change of variable gives

$$\boldsymbol{U} \ = \ \boldsymbol{L}^{-\top} \boldsymbol{V} \ = \ \boldsymbol{L}^{-\top} \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} \ = \ \boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2}. \tag{24}$$

Since $\boldsymbol{R}^{-1} = \mathrm{blkdiag}(\boldsymbol{R}_s^{-1})$ and $\boldsymbol{A}$ is the row-wise stacking of $\{\boldsymbol{A}_s\}_{s=1}^S$, the s-th block of Eq. (24) is

$$\boldsymbol{U}_s \ = \ \boldsymbol{R}_s^{-1} \boldsymbol{A}_s \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} \ \in \ \mathbb{R}^{m_s \times K}, \tag{25}$$

which is the expression in § 3.1. □

I.3 Optimality of the Procrustes Axis Alignment

Proposition. *Let $\boldsymbol{Z} \in \mathbb{R}^{N\times K}$ be the projected training data and $\mathbf{Y} \in \mathbb{R}^{N\times K}$ the label matrix. The rotation $\boldsymbol{R}^* = \boldsymbol{U}_C \boldsymbol{V}_C^\top$, where $\boldsymbol{U}_C \boldsymbol{\Sigma}_C \boldsymbol{V}_C^\top$ is the SVD of the cross-covariance $\boldsymbol{C} := \boldsymbol{Z}^\top \mathbf{Y} \in \mathbb{R}^{K\times K}$, is the unique orthogonal solution to*

$$\min_{\boldsymbol{R}:\, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K} \|\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R}\|_F^2. \tag{26}$$

Proof. Expanding the Frobenius norm in Eq. (26) gives:

$$\begin{aligned} \|\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R}\|_F^2 &= \operatorname{Tr}\big\{(\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R})^\top (\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R})\big\} \\ &= \operatorname{Tr}\big\{\mathbf{Y}^\top \mathbf{Y}\big\} - 2\operatorname{Tr}\big\{\mathbf{Y}^\top \boldsymbol{Z}\boldsymbol{R}\big\} + \operatorname{Tr}\big\{\boldsymbol{R}^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{R}\big\}. \end{aligned} \tag{27}$$

Note that because $\mathbf{Y}$ is strictly mean-centered by construction, $\mathbf{Y} = \boldsymbol{H}\mathbf{Y}$, which implies the cross-covariance $\boldsymbol{C} = \boldsymbol{Z}^\top \mathbf{Y} = \boldsymbol{Z}^\top \boldsymbol{H}\mathbf{Y} = (\boldsymbol{H}\boldsymbol{Z})^\top \mathbf{Y}$. Thus, the cross-covariance implicitly evaluates the theoretically correct centered coordinates without requiring explicit centering of $\boldsymbol{Z}$.

Using the cyclic property of the trace, the middle term becomes $\operatorname{Tr}\{\mathbf{Y}^\top \boldsymbol{Z}\boldsymbol{R}\} = \operatorname{Tr}\{\boldsymbol{R}\mathbf{Y}^\top \boldsymbol{Z}\} = \operatorname{Tr}\{\boldsymbol{R}\boldsymbol{C}^\top\} = \operatorname{Tr}\{\boldsymbol{C}^\top \boldsymbol{R}\}$. Since $\boldsymbol{R}$ is a square orthogonal matrix, $\boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}_K$. The final term simplifies to $\operatorname{Tr}\{\boldsymbol{R}^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{R}\} = \operatorname{Tr}\{\boldsymbol{Z}\boldsymbol{R}\boldsymbol{R}^\top \boldsymbol{Z}^\top\} = \operatorname{Tr}\{\boldsymbol{Z}\boldsymbol{Z}^\top\} = \operatorname{Tr}\{\boldsymbol{Z}^\top \boldsymbol{Z}\}$, which is independent of $\boldsymbol{R}$. It follows that minimizing the Frobenius distance is exactly equivalent to maximizing the cross-trace:

$$\max_{\boldsymbol{R}:\, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K} \operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\}. \tag{28}$$

Let $\boldsymbol{C} = \boldsymbol{U}_C \boldsymbol{\Sigma}_C \boldsymbol{V}_C^\top$ be the Singular Value Decomposition (SVD) of $\boldsymbol{C}$, with $\boldsymbol{\Sigma}_C = \operatorname{diag}(\sigma_1, \ldots, \sigma_K)$ and $\sigma_k \geq 0$. Substituting the SVD into the objective and applying the cyclic property of the trace yields:

$$\operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{V}_C \boldsymbol{\Sigma}_C \boldsymbol{U}_C^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{U}_C^\top \boldsymbol{R}\boldsymbol{V}_C \boldsymbol{\Sigma}_C\big\}. \tag{29}$$

For any orthogonal $\boldsymbol{R}$, define $\boldsymbol{T} := \boldsymbol{U}_C^\top \boldsymbol{R}\boldsymbol{V}_C$. Since it is the product of orthogonal matrices, $\boldsymbol{T}$ is also exactly orthogonal. It follows that:

$$\operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{T}\boldsymbol{\Sigma}_C\big\} \;=\; \sum_{k=1}^{K} t_{kk}\sigma_k, \tag{30}$$

where t_{kk} denotes the k-th diagonal entry of $\boldsymbol{T}$. Since $\boldsymbol{T}$ is orthogonal, its column vectors have unit norm, which strictly enforces $|t_{kk}| \leq 1$ for all k (Schönemann, 1966). Thus:

$$\sum_{k=1}^{K} t_{kk}\sigma_k \;\leq\; \sum_{k=1}^{K} \sigma_k \;=\; \operatorname{Tr}\big\{\boldsymbol{\Sigma}_C\big\}. \tag{31}$$

The theoretical upper bound is attained if and only if $t_{kk} = 1$ for all k, which implies $\boldsymbol{T} = \boldsymbol{I}_K$. From our definition of $\boldsymbol{T}$, this requires:

$$\boldsymbol{U}_C^\top \boldsymbol{R}\boldsymbol{V}_C \;=\; \boldsymbol{I}_K \;\Longrightarrow\; \boldsymbol{R}^* \;=\; \boldsymbol{U}_C \boldsymbol{V}_C^\top. \tag{32}$$

Uniqueness follows from the strict inequality whenever $\sigma_k > 0$ for all k, which holds when $\boldsymbol{C}$ has full rank. In practice, if $\boldsymbol{C}$ is rank-deficient, the solution is unique up to a rotation in the null space of $\boldsymbol{C}$; since the null directions carry no label-alignment signal, any such arbitrary rotation leaves the downstream steering unaffected.

Updating the per-bank projection matrices via $\boldsymbol{U}_s \leftarrow \boldsymbol{U}_s \boldsymbol{R}^*$ rigidly rotates the semantic axes into the canonical orientation that minimizes the Frobenius distance to the label matrix $\mathbf{Y}$. (Note that while Procrustes alignment preserves scale and does not inherently match the variance of $\mathbf{Y}$, the downstream Log-Likelihood Ratio steering objective utilized in LatentCompass is fundamentally scale-invariant, making an orthogonal rotation strictly sufficient for semantic disentanglement). □