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

Anonymous Author(s)
Affiliation
Address
email

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)** computing an optimal shift in the constructed space, and **(c)** reflecting the corresponding shift in the T2I latent space. Extensive evaluations demonstrate that LatentCompass effectively **(i)** mitigates generative stereotypes by $\mathbf{100}\%$, **(ii)** reduces unsafe concept generation by $\mathbf{58}\%$, **(iii)** enhances aesthetic quality by $\mathbf{27}\%$ on average, **(iv)** boosts red-teaming success rates against Deepfake detectors by up to $\mathbf{47}\%$, and **(v)** enables high-fidelity style and face attribute editing without attribute leakage. *CAUTION: This paper includes content that may be inappropriate or offensive.*

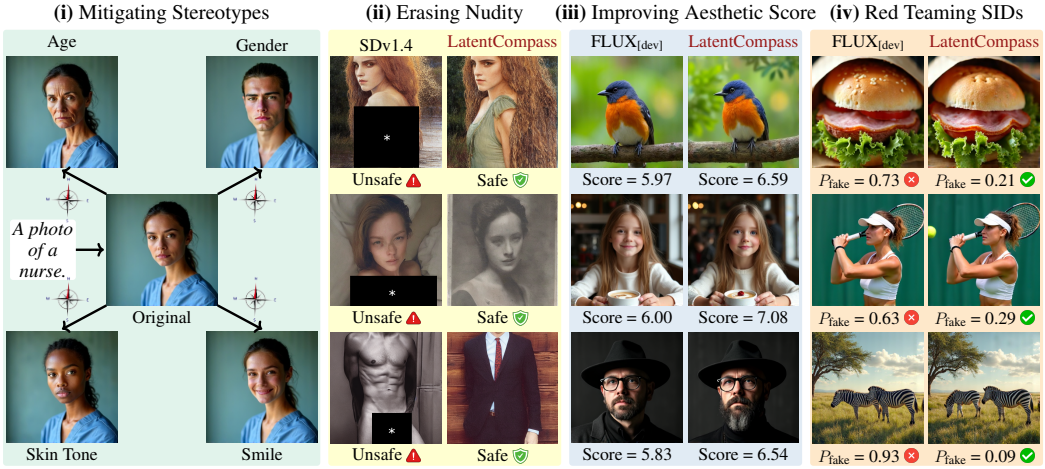


Figure 1: **Independent control of descriptive and evaluative concepts in T2I models**: **(i)** Facial attribute steering, which can be used to debias profession generation, **(ii)** improving content safety by nudity erasure, **(iii)** enhancing human aesthetic preferences of synthesized images, **(iv)** Red teaming, by finding examples that can deceive synthetic image detectors (SID).

1 Introduction

The case for controllable generation. While text-to-image (T2I) diffusion models excel at high-fidelity synthesis, their generative priors are entangled due to the correlations in their training data

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

[12, 21, 37, 59]. Consequently, T2I generation defaults to stereotypical attribute combinations, neglects rare joint conditions, and fails to correctly follow user intentions, perpetuating harmful societal biases [59, 15, 18, 54]. A *controllable* T2I model overcomes these limitations by overriding default priors to synthesize targeted distributions, ensuring prompt faithfulness, and suppressing 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. **(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** [26, 30, 32, 35, 22] directly optimize model weights to instill new concepts, but their computational cost and lack of modularity do not support R3. **(2) Prompt-based methods** [59, 47, 45, 38, 55] 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** [58, 1] 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** [16, 2] 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 and shifted toward a target semantic profile via likelihood maximization. **(P3)** Finally, this optimized state is integrated back into the T2I sampling trajectory by solving the pre-image, ensuring the final output adheres to the desired control attributes (satisfies R2).

Applications of LatentCompass includes 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, namely, achieving a $\mathbf{100}\%$ **mitigation** in generative stereotypes, **up to** $\mathbf{58}\%$ improvement in unsafe concept erasure, a $\mathbf{27}\%$ enhancement in human aesthetic preference, and **up to** $\mathbf{47}\%$ boost in red-teaming success against synthetic image detectors in a black-box manner.

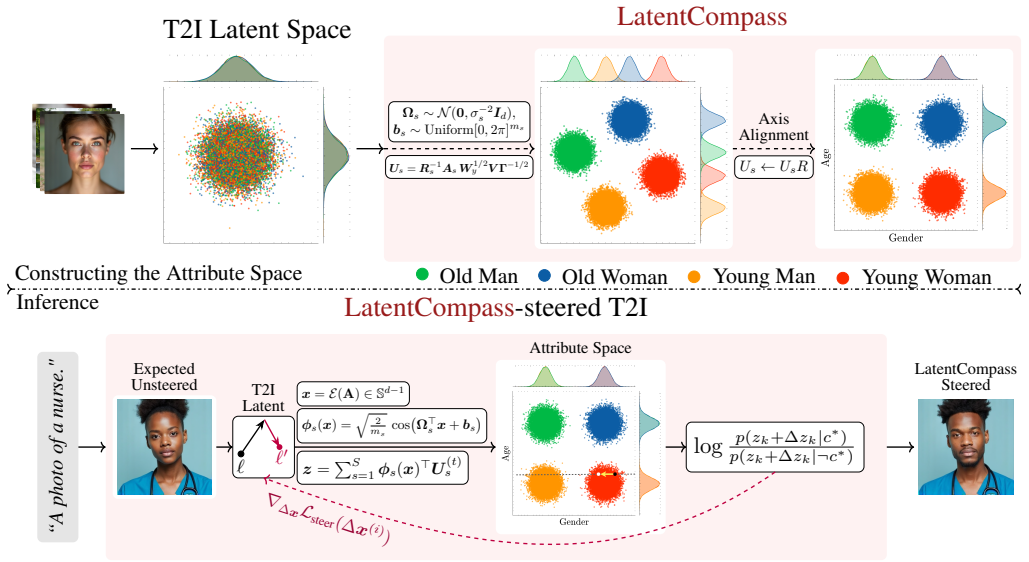


Figure 2: **Overview of LatentCompass: (Top)** LatentCompass uses labeled examples to construct an attribute space where each axis controls a target attribute. **(Bottom)** At inference, LatentCompass steers the T2I latent space to shift the image into the desired attribute region.

2 Problem Setup: Controllable Generation in Text-to-Image Diffusion 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 $\hat{\mathbf{A}} = \mathcal{S}(G, \boldsymbol{c}, \mathcal{K})$ whose attributes on $\mathcal{K}$ match the targets. Formally:

Problem 1. Concept-Guided Generation

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

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,

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

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 itself. However, prompt-based control is fundamentally limited. **(L1)** Evaluative attributes cannot be expressed in natural language, as these concepts are external to the generative process (e.g., a T2I model has no inherent 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 [6, 31]. For instance,

prompting 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)** computing an optimal shift within this manifold; and **(P3)** propagating the shift back into 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 under the data distribution to prevent attribute leakage (Eq. (1)), 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) [44], which provide an explicit 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 independent RFF banks. In the following, we describe the encoder fitting at a single 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

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

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 [25, 3] 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) from the population HSIC in § G.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

$$\boldsymbol{U}_s \;=\; \boldsymbol{R}_s^{-1} \, \boldsymbol{A}_s \, \boldsymbol{W}_y^{1/2} \, \boldsymbol{V} \, \boldsymbol{\Gamma}^{-1/2} \;\in\; \mathbb{R}^{m_s \times K}. \tag{5}$$

The derivation of this closed-form solution by Lagrangian analysis is given in § G.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 [48]. 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 § G.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:

$$\boldsymbol{z} = g_t(\boldsymbol{x}) = \textstyle\sum_{s=1}^{S} \boldsymbol{\phi}_s(\boldsymbol{x})^\top \boldsymbol{U}_s^{(t)} \in \mathbb{R}^K. \tag{6}$$

3.2 P2: Computing the Optimal Coordinate Shift

Given the disentangled attribute space, we seek an optimal intervention $\Delta \boldsymbol{z}$ to achieve a target attribute profile $\boldsymbol{y}^*$. By design (E2, E3), each axis independently controls its corresponding attribute, reducing the manipulation to a set of independent 1D optimizations.

Gaussian Density Estimation. For each attribute $k \in \{1, \ldots, K\}$ and class $c \in \{-1, +1\}$, we model the class-conditional density $p(z_k \mid c) \sim \mathcal{N}(\mu_{k,c}, \sigma_{k,c}^2)$ under the assumption that the distribution of a low-dimensional projection of high-dimensional data is approximately normal [17]. For continuous labels, we set a threshold at $\mathbb{E}[y_k]$ to obtain binary c. 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:

$$\mathcal{L}_{\mathrm{LLR}}^{(c^*)}(\tilde{z}_k) = \log \frac{p(\tilde{z}_k \mid c^*)}{p(\tilde{z}_k \mid \neg c^*)} = \frac{(\tilde{z}_k - \mu_{k,\neg c^*})^2}{2\sigma_{k,\neg c^*}^2} - \frac{(\tilde{z}_k - \mu_{k,c^*})^2}{2\sigma_{k,c^*}^2} + \log \frac{\sigma_{k,\neg c^*}}{\sigma_{k,c^*}}. \tag{7}$$

Optimizing the Shift. The optimal multi-attribute shift $\Delta \boldsymbol{z}^*$ maximizes this objective across the targeted attribute subset $\mathcal{K}$, regularized by $\beta \geq 0$ to penalize extreme deviations:

$$\Delta \boldsymbol{z}^* = \arg\max_{\Delta \boldsymbol{z}} \sum_{k \in \mathcal{K}} \mathcal{L}_{\mathrm{LLR}}^{(c_k^*)}(z_k + \Delta z_k) - \beta \|\Delta \boldsymbol{z}\|_2^2. \tag{8}$$

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 evidence c_k^* while penalizing non-target deviations c_j^{orig}:

$$\mathcal{L}_{\text{steer}}(\Delta\boldsymbol{\ell}) = \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}\Big(\tilde{z}_k(\Delta\boldsymbol{\ell})\Big) + \frac{1}{K-|\mathcal{K}|}\sum_{j\notin\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_j^{\text{orig}})}\Big(\tilde{z}_j(\Delta\boldsymbol{\ell})\Big), \tag{9}$$

where $\tilde{z}_m(\Delta\boldsymbol{\ell}) = \big[g_t(\boldsymbol{x}')\big]_m$ 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 $\|\boldsymbol{\ell}\|_2$:

$$\widetilde{\Delta\boldsymbol{\ell}} = \Delta\boldsymbol{\ell}^{(i)} + \eta\nabla_{\Delta\boldsymbol{\ell}}\mathcal{L}_{\text{steer}}(\Delta\boldsymbol{\ell}^{(i)}), \qquad \Delta\boldsymbol{\ell}^{(i+1)} = \left(\frac{\boldsymbol{\ell}+\widetilde{\Delta\boldsymbol{\ell}}}{\|\boldsymbol{\ell}+\widetilde{\Delta\boldsymbol{\ell}}\|_2}\|\boldsymbol{\ell}\|_2\right) - \boldsymbol{\ell}, \tag{10}$$

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:

$$\min_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}\Big(\tilde{z}_k(\Delta\boldsymbol{\ell}^{(i+1)})\Big) \geq \tau_k(\mu_{k,c^*}, \sigma_{k,c^*}). \tag{11}$$

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 medical profession generation (§ 4.1), disentangled face attribute editing (§ 4.2), and improving content safety by erasing nudity (§ 4.3). 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 synthesize images that better match the aesthetic taste of observers (§ 4.5). We use FLUX$_{\text{[dev]}}$[7] and SDv1.4 [47] as the T2I generator.

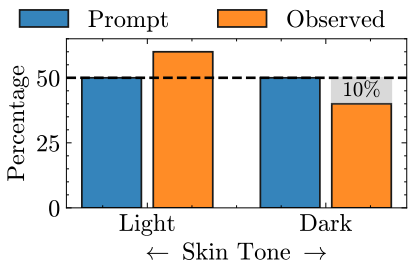


Figure 3: Prompt non-adherence.

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 [4, 15, 37]. Crucially, explicit prompting often fails to override these priors, occasionally yielding images that contradict the requested demographic attributes (see Fig. 3). We evaluate LatentCompass's ability to mitigate these demographic predispositions *without* altering the prompt that elicits the concept.

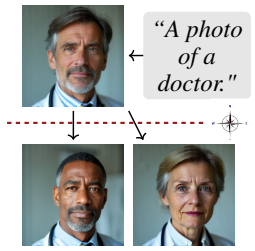


Figure 4: **(Top)** Unsteered **(Bottom)** LatentCompass.

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** [15], $\Psi = \max\big(0,\ P_{D,C}(y) - P^*_{D,C}(y)\big)$, where $P^*_C(y)$ is the real-world occupational density obtained from U.S. census data. By construction, the Stereotype Score penalizes only *directional* over-representation of the stereotypical attribute, separating genuine stereotype reinforcement from harmless under-representation. **2) Demographic Parity (DP)** as 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.

Table 1: Evaluation of gender and skin tone bias and stereotypes in Doctor and Nurse images generated by unsteered versus LatentCompass-steered FLUX$_{[dev]}$.

	Method	Gender Stereotype			Gender Bias		Skin Tone Stereotype			Skin Tone Bias	
		$P_C(y = male)$	$P_C^*(y = male)$	$\Psi \downarrow$	$P_C(y = male)$	DP $\downarrow$	$P_C(y = light)$	$P_C^*(y = light)$	$\Psi \downarrow$	$P_C(y = light)$	DP $\downarrow$
Doctor	Unsteered	82.6	61.0	21.6	82.6	65.2	99.0	94.0	5.0	99.0	98.0
	LatentCompass	61.0	61.0	**0.0**	49.1	**1.8**	94.0	94.0	**0.0**	50.3	**0.6**
Nurse	Unsteered	89.3	88.0	1.3	89.3	68.6	98.0	89.0	9.0	98.0	96.0
	LatentCompass	88.0	88.0	**0.0**	50.8	**1.6**	89.0	89.0	**0.0**	50.0	**0.0**

Results. Tab. 1 summarizes the impact of LatentCompass on neutralizing systemic biases regarding perceived gender and skin tone. Our framework significantly attenuates gender and skin tone stereotypes, yielding average reductions of Ψ by $\mathbf{100}\%$. Further, we reduce DP by $\mathbf{98.57}\%$ on average, demonstrating robust performance in mitigating historical biases [5] 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.

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.

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 $\mathbf{99.7}\%$.

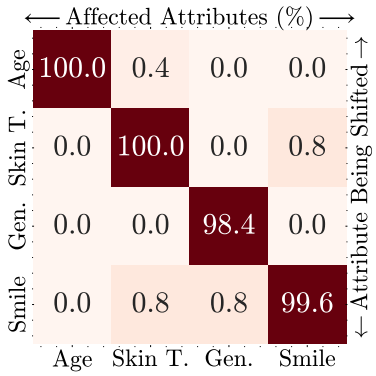


Figure 5: Edit Confusion Matrix.

Remark. LatentCompass precisely edits the target attribute without affecting 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 [58] suppress NSFW concepts mostly through prompt embedding manipulation, but these methods remain susceptible to adversarial and implicit prompts as they are *input-level* interventions [58]. Instead, we explore an *output-level* erasure approach where the T2I latents containing nudity are explicitly 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 [58], we evaluate LatentCompass on removing nudity in SDv1.4 on three adversarial prompt benchmarks: P4D [11], Ring-A-Bell [53], and MMA-Diffusion [57]. 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)** [27] and **CLIP score** on COCO [33].

Tab. 2 compares the effectiveness of LatentCompass and baselines in erasing nudity. LatentCompass is able to achieve state-of-the-art erasure by nearly eliminating adversarial vulnerabilities across all benchmarks. By explicitly steering latents at the output level, our method yields substantial absolute reductions in Attack Success Rate (ASR) over the strongest fine-tuning-free baselines, and even surpasses fine-tuning approaches. Crucially, this robust safety mechanism avoids the severe utility trade-offs typical of concept erasure. LatentCompass attains the highest image fidelity (FID) among all evaluated methods while maintaining text-alignment (CLIP) on par with the foundational model. Ultimately, these results demonstrate that our output-level intervention provides a fundamentally stronger, more resilient defense against implicit and adversarial prompts than existing input-level manipulations, successfully decoupling NSFW concept erasure from generation degradation.

Table 2: Comparison of Attack Success Rate (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 [46]	-	98.7	83.1	95.7	-	31.3
ESD [22]	×	75.0	52.8	87.3	-	30.7
SA [26]	×	62.3	32.9	20.5	54.98	30.6
CA [30]	×	92.7	77.3	85.5	40.99	31.2
MACE [35]	×	14.6	7.6	18.3	52.24	29.4
SDID [32]	×	93.3	69.6	90.7	22.99	30.5
UCE [23]	✓	66.7	33.1	86.7	31.25	**31.3**
RECE [24]	✓	38.1	13.4	67.5	37.60	30.9
SLD-Medium [49]	✓	93.4	64.6	94.2	31.47	31.0
SLD-Strong [49]	✓	86.1	62.0	92.0	40.88	29.6
SLD-Max [49]	✓	74.2	57.0	83.7	50.51	28.5
SAFREE [58]	✓	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

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 [51, 39, 29, 14]. 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.

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

Method	UFD [39]	RINE [29]	FatFormer [34]	AVG ± STD	Time (s) ↓
No Steering	67.6	52.4	55.1	58.4 ± 6.6	∼ 5.0
PolyJuice [14]	96.3	81.2	83.2	86.8± 6.6	∼ 198.6
LatentCompass	**100.0** (+32)	**99.5** (+47)	**96.4** (+41)	98.6 ± 1.6	∼ 8.2

Setup. Following [14], 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 [39], RINE [29], and FatFormer [34]. 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 [14] 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 [56], that performs white-box attacks on a proxy detector trained on RINE labels.

Table 4: LatentCompassASR (%) against transfer-based baseline.

Method	ASR (↑)
Unsteered	52.4
DiffPGD^t (x^n) [56]	57.2
DiffPGD^t (x_0^n) [56]	70.1
LatentCompass (ours)	**99.5** (+46)

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

4.5 Aligning T2I Generation with Human Aesthetic Preferences using 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 LAION-Aesthetics_Predictor V2 [50] 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-

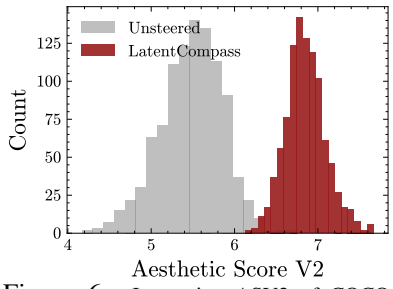


Figure 6: Improving ASV2 of COCO-prompted images via LatentCompass.

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 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.

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 [1, 55, 38] or CLIP-based text guidance [59, 47, 45]. 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 [1, 16, 59] steers latent trajectories using gradients from external classifiers. While operating directly on latents, this approach struggles with attribute leakage [59].

Generative stereotypes and bias. T2I models inherently default to skewed demographic distributions, an issue often exacerbated rather than resolved by prompt engineering [49, 59]. Although open-set auditing has improved [18, 15, 4, 37, 36, 8, 9], current surface-level text mitigations fail to explicitly disentangle correlated attributes within generative priors, leaving the visual latent space structurally biased and resistant to prompt-agnostic correction. The ability of LatentCompass in disentangled attribute editing allows us to mitigate visual bias and stereotypes in T2I models.

Red-teaming of synthetic image detectors. Although diffusion-based adversarial attacks [56, 13, 10] are effective on SIDs, they demand white-box access, which limits their utility for commercial black-box detectors. Alternatively, while black-box methods like PolyJuice [14] 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 permanently alter model weights to suppress concepts, but suffer from prohibitive computational costs and catastrophic forgetting [26, 30, 32, 35, 22]. Alternatively, training-free inference interventions bypass retraining by manipulating classifier-free guidance (SLD [49]) or steering text embeddings away from toxic subspaces (SAFREE [58]). However, since these models are vision-unaware, they struggle with safeguarding against subversive prompts that induce harmful results. 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.

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 § F.

References

[1] 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.

[2] 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*, pages 843–852, 2023.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016.

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

[10] 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.

[11] 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*, pages 8468–8486, 2024.

[12] 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.

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

[14] 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*, 2025.

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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*, pages 5111–5120, 2024.

[24] 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*, pages 73–88. Springer, 2024.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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*, pages 22691–22702, 2023.

[31] 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.

[32] 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*, pages 12006–12016, 2024.

[33] 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*, pages 740–755. Springer, 2014.

[34] 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*, pages 10770–10780, 2024.

[35] 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.

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

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

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

[39] 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*, pages 24480–24489, 2023.

[40] 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.

[41] 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)*, pages 15308–15318, October 2025.

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

[43] 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*, pages 8748–8763. PmLR, 2021.

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

[45] 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.

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

[47] 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*, pages 10684–10695, 2022.

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

[49] 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.

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

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

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

[53] 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.

[54] 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.

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

[56] 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.

[57] 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*, pages 7737–7746, 2024.

[58] 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.

[59] 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*, pages 3969–3980, 2023.

[60] 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 Implementation Details

A.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. 11, we perform an ablation on the effect of the number of examples used to construct the LatentCompass attribute space.

A.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 prompts from the COCO dataset [33]. No additional quality prompts are appended.

(iii) Content Safety Improvement: We use 1K prompts from the MMA-Diffusion [57] benchmark, specifically the 'target_prompt' column that contains non-adversarial 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.

A.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 [42]. We refer to the attributes specified before generation, and the result attributes labeled by Qwen, as 'prompt' and 'observed' labels respectively in Fig. 3. 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 [39], RINE [29], FatFormer [34].

(iii) Steering Aesthetic Preference: We label our set of generated COCO images with the Aesthetic Predictor V2 model [50].

(iv) Content Safety Improvement: We use NudeNet [52] to label each image with an unsafe score $\in [0, 1]$.

A.1.3 T2I Generation Settings.

(i) Face Attribute Steering: We use $\text{FLUX}_{\text{[dev]}}$ [7] 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 [47] 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 $\boldsymbol{\ell}_t$.

A.1.4 Feature Extractors

For face attribute steering, we use features from FaRL [60], which shares the same architecture as CLIP [43]. 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 [40] features.

A.2 Steering

A.2.1 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 [19],

$$\hat{\boldsymbol{\ell}}_0 = \frac{\boldsymbol{\ell}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{\ell}_t, t, \boldsymbol{c})}{\sqrt{\bar{\alpha}_t}} \tag{12}$$

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 [20] (which are closely related to diffusion models), where the denoising can be simplified to a single Euler step with stepsize t,

$$\hat{\boldsymbol{\ell}}_0 = \boldsymbol{\ell}_t - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{\ell}_t, t, \boldsymbol{c}) \cdot t \tag{13}$$

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

A.2.2 Gradient Calculation

For a latent $\boldsymbol{\ell}_t$ at timestep t, we first compute the denoised estimate $\hat{\boldsymbol{\ell}}_0$ (Eq. (12), Eq. (13)), 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. (10)) w.r.t. the optimization objective (Eq. (9)) would require backpropagating through the diffusion model itself, due to Eq. (12), Eq. (13). 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. (13) to,

$$\hat{\boldsymbol{\ell}}_0 = \boldsymbol{\ell}_t - \boldsymbol{v}_t \cdot t. \tag{14}$$

Under Eq. (14), computing the gradient of $\boldsymbol{\ell}_t$ no longer requires backpropagating through the actual T2I model. This approach is originally used by Bansal et al. [2], 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. [41] theoretically show that for rectified flows, due to smoothness properties, $\nabla_{\hat{\boldsymbol{\ell}}_0}\mathcal{L}$ has a negligible error compared to the exact gradient $\nabla_{\boldsymbol{\ell}_t}\mathcal{L}$. As such, we use the approximation in Eq. (14), and **we do not compute the gradient through the T2I model for $\text{FLUX}_{\text{[dev]}}$.**

A.3 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 § A.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^*}	$\mu_{k,c^*} + 1\sigma_{k,c^*}$

A.4 Experiments Compute Resources

For all steps of experiments, including dataset generation, attribute space construction, and inference steering, we used eight NVIDIA Quadro RTX 8000 GPUs, each with 48 GB of memory. The primary computational bottleneck arises from the memory requirements of the T2I models during image generation; LatentCompass itself adds negligible overhead.

A.5 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

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 $s = 1, \ldots, S$ **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 $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_K) \succ 0$
for $s = 1, \ldots, S$ **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 $s = 1, \ldots, S$ **do**
  $\boldsymbol{U}_s^{(t)} \leftarrow \boldsymbol{U}_s \boldsymbol{R}$
end for
$\boldsymbol{Z} \leftarrow \boldsymbol{Z}\boldsymbol{R}$
for $k = 1, \ldots, K$ and $c \in \{-1, 1\}$ **do**
  $\mu_{k,c} \leftarrow \mathbb{E}_{i:y_{i,k}=c}[Z_{i,k}]$; $\sigma_{k,c}^2 \leftarrow \mathrm{Var}_{i:y_{i,k}=c}[Z_{i,k}]$
end for
return $\{\boldsymbol{U}_s^{(t)}, \boldsymbol{\Omega}_s, \boldsymbol{b}_s\}_{s=1}^S$, $\boldsymbol{R}$, $\{(\mu_{k,c}, \sigma_{k,c}^2)\}_{k,c}$

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 Eqs. (4) and (5). The recovered axes are then aligned to the attribute basis through Orthogonal Procrustes (§ 3.1), and per-class Gaussian sufficient statistics $(\mu_{k,c}, \sigma_{k,c}^2)$ 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. (6). The optimal per-axis shift is obtained in closed form by maximizing the regularized log-likelihood ratio of the target class against its complement (P2). This shift is then propagated back to the d-dimensional latent space through the differentiable encoder g_t by gradient ascent on Eq. (9), with the iteration in Eq. (10) terminated as soon as the per-axis 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 2 LatentCompass: Steering at Inference

Input: Frozen T2I diffusion latent $\boldsymbol{\ell}_t$ at timestep t; target subset $\mathcal{K} \subseteq \{1, \dots, K\}$ 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 $i = 0, 1, \dots, I_{max} - 1$ **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 $s = 1, \dots, S$ **do**
    $\boldsymbol{\phi}_s(\boldsymbol{x}') \leftarrow \sqrt{2/m_s}\cos(\boldsymbol{\Omega}_s^\top \boldsymbol{x}' + \boldsymbol{b}_s)$
  end for
  $\tilde{\boldsymbol{z}}^{(i)} \leftarrow \sum_{s=1}^S \boldsymbol{\phi}_s(\boldsymbol{x}')^\top \boldsymbol{U}_s^{(t)} \in \mathbb{R}^K$
  {(iii) Compute steering objective}
  $\mathcal{L}_{\text{steer}} \leftarrow \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}(\tilde{z}_k^{(i)}) + \lambda_{\text{pres}} \sum_{j\notin\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_j^{\text{orig}})}(\tilde{z}_j^{(i)})$
  {(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 $\min_{k\in\mathcal{K}} \mathcal{L}_{\text{LLR}}^{(c_k^*)}(\tilde{z}_k^{(i)}) \geq \tau_k$ **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}

A.6 Threat Model for Red Teaming Experiment

Following the framework established in PolyJuice [14], 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 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.

B Extended Qualitative Examples

We provide qualitative examples of our applications in figures Fig. 7 Fig. 8, Fig. 9, and Fig. 10.

Figure 7: Stereotype and bias mitigation for nurse and doctor generation.

Figure 8: **(Left)** A synthetic image detected as fake by FatFormer, UFD, and RINE. **(Right)** LatentCompass-steered images that deceive each detector.

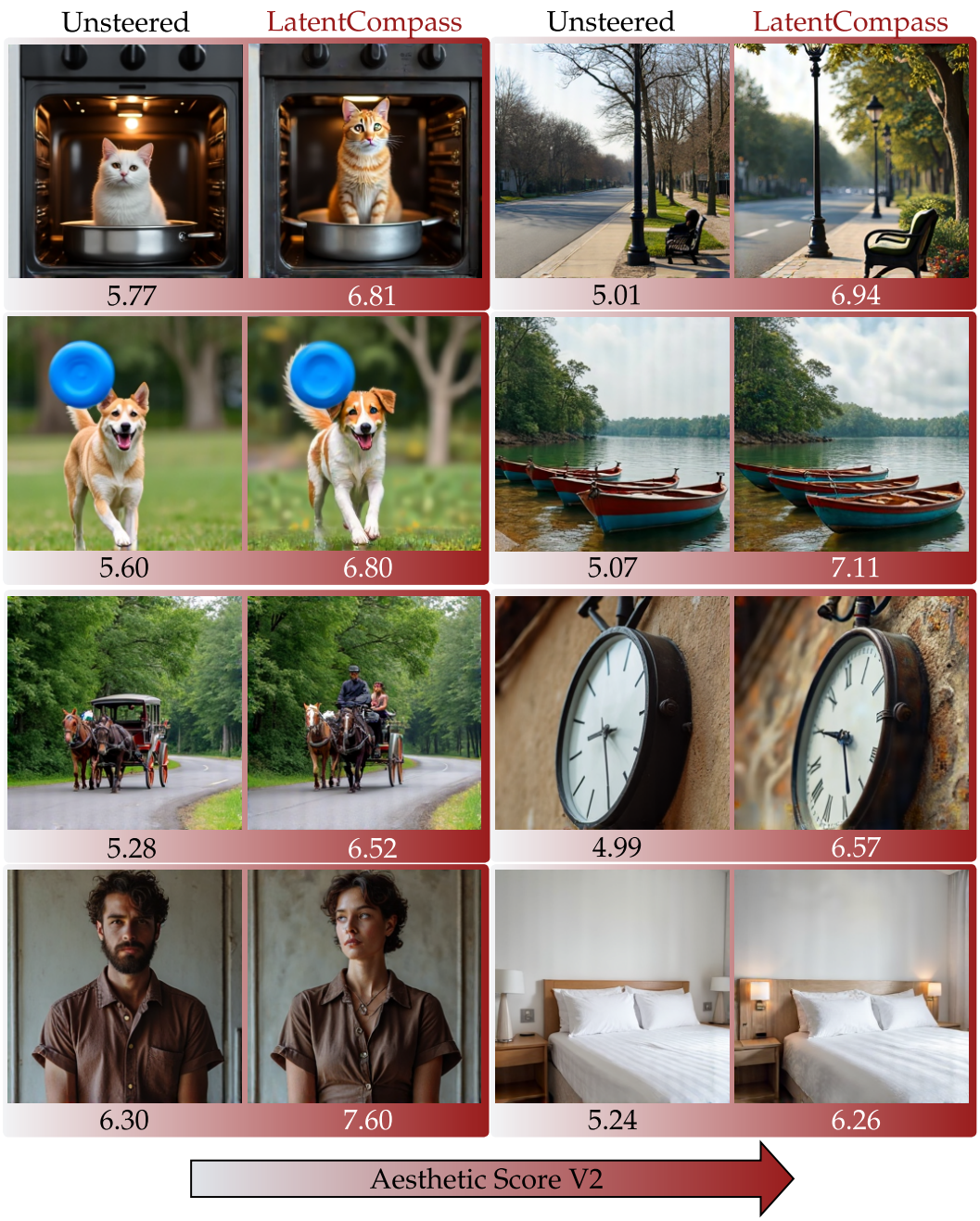


Figure 9: Improving the aesthetic score of images generated from COCO prompts.

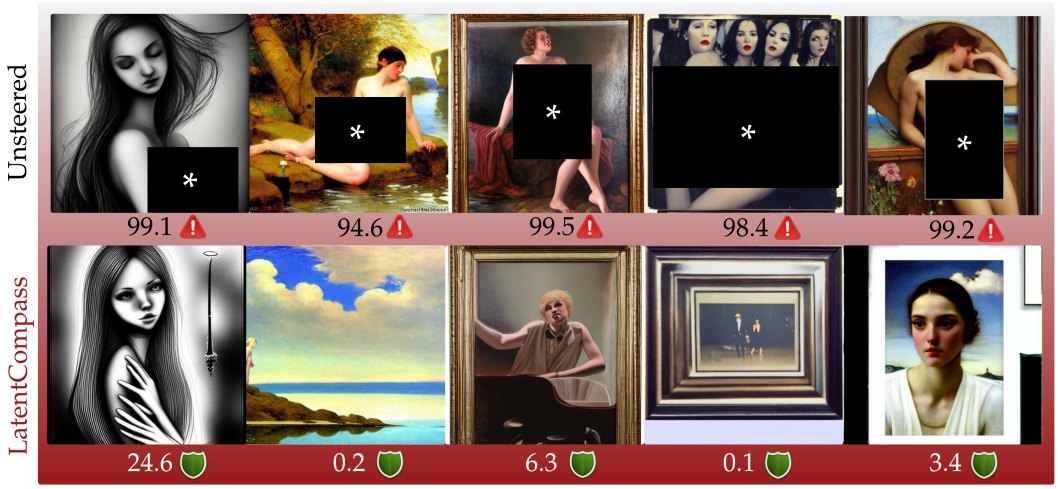


Figure 10: Erasing nudity for P4D [11] benchmark, which provides the exact random seed and guidance level associated with each toxic text prompt. The images generated from SDv1.4, and scores (0-100) given by NudeNet [52], where a score > 60 is counted as a nude image.

C 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. 6. The table is organized chronologically by the main components of our problem formulation and proposed approach.

Table 6: 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. (9)
$\boldsymbol{\ell}_t, \boldsymbol{\ell}$	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. (6)
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	Eqs. (3) and (5)
$\boldsymbol{W}_y$	Optional per-attribute weighting matrix	Eq. (3)
$\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{\Gamma}$	Joint supervised matrix and its eigendecomposition components	Eqs. (4) and (5)
$\boldsymbol{C}, \boldsymbol{R}$	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. (9)
$\Delta\boldsymbol{z}, \tilde{z}_k$	Intervention shift in attribute space, and shifted coordinate	§ 3 and Eq. (9)
β	Regularization coefficient penalizing extreme attribute deviations	§ 3
$\Delta\boldsymbol{\ell}$	Latent perturbation applied to the T2I diffusion trajectory	Eqs. (9) and (10)
$\hat{\boldsymbol{\ell}}_0, \mathcal{D}$	Tweedie estimate of the clean data manifold and VAE decoder	§ 3.3
$\mathcal{L}_{\text{steer}}$	Objective function for gradient-based latent optimization	Eq. (9)
η	Projected gradient ascent step size	Eq. (10)
τ_k	Statistical confidence threshold for the early stopping criterion	§ 3.3

D 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.

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.

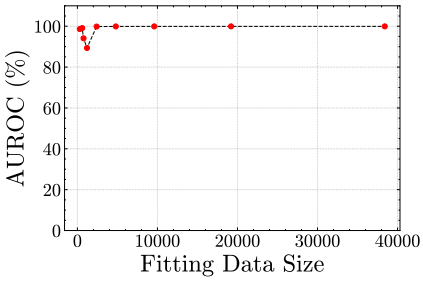
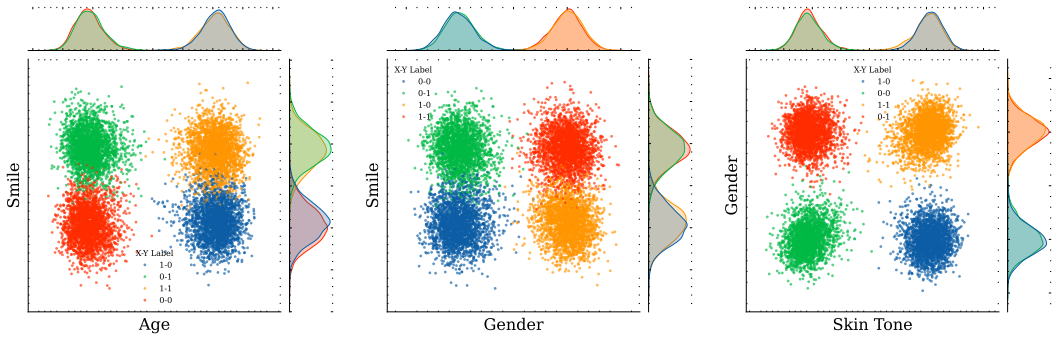


Figure 11: Average Validation AUROC of projected attribute clusters across varying numbers of fitting exemplars. LatentCompass maintains high cluster separability even in extremely low-data regimes.

Results and Analysis. As illustrated in Figure 11, 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.

E LatentCompass Attribute Space Visualization



(a) (b) (c)

Figure 12: 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. 12 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.

F 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.

G Proofs

G.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 $C_{\boldsymbol{zy}} : \mathbb{R}^K \to \mathbb{R}^K$ [25], which acts as $\boldsymbol{v} \mapsto \mathbb{E}[\boldsymbol{z}(\boldsymbol{y}^\top \boldsymbol{v})]$ and admits the matrix representation

$$C_{\boldsymbol{zy}} \;=\; \mathbb{E}\big[\boldsymbol{z}\,\boldsymbol{y}^\top\big] \;=\; \mathbf{U}^\top \underbrace{\mathbb{E}\big[\boldsymbol{\phi}(\boldsymbol{x})\,\boldsymbol{y}^\top\big]}_{=:\,\mathbf{M}_\mu} \;\in\; \mathbb{R}^{K\times K}, \tag{15}$$

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

$$\mathrm{HSIC}(\boldsymbol{z}, \boldsymbol{y}) \;=\; \|C_{\boldsymbol{zy}}\|_{\mathrm{HS}}^2 \;=\; \|\mathbf{U}^\top \mathbf{M}_\mu\|_F^2 \;=\; \mathrm{Tr}\big\{(\mathbf{U}^\top \mathbf{M}_\mu)^\top (\mathbf{U}^\top \mathbf{M}_\mu)\big\}. \tag{16}$$

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

$$\|\mathbf{U}^\top \mathbf{M}_\mu \boldsymbol{W}_y^{1/2}\|_F^2 \;=\; \mathrm{Tr}\big\{\boldsymbol{W}_y^{1/2} (\mathbf{U}^\top \mathbf{M}_\mu)^\top (\mathbf{U}^\top \mathbf{M}_\mu) \boldsymbol{W}_y^{1/2}\big\}. \tag{17}$$

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. (17) and absorbing the constant $\frac{1}{N^2}$ into the maximization yields the empirical unconstrained supervised covariance objective

$$\mathrm{Tr}\big\{\boldsymbol{W}_y^{1/2} (\mathbf{U}^\top \boldsymbol{\Phi}^\top \mathbf{Y})^\top (\mathbf{U}^\top \boldsymbol{\Phi}^\top \mathbf{Y}) \boldsymbol{W}_y^{1/2}\big\}. \tag{18}$$

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 [25] 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. (17) and absorbing the $\frac{1}{N^2}$ constant gives the constrained optimization problem:

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

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 [3], and the ridge term $\lambda > 0$ regularizes against rank deficiency when $N < m_s$. □

G.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 Eq. (5), 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

$$\boldsymbol{U} := \begin{bmatrix} \boldsymbol{U}_1 \\ \vdots \\ \boldsymbol{U}_S \end{bmatrix} \in \mathbb{R}^{M \times K}, \quad \boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_S \end{bmatrix} \in \mathbb{R}^{M \times K}, \quad \boldsymbol{R} := \mathrm{blkdiag}(\boldsymbol{R}_1, \dots, \boldsymbol{R}_S) \in \mathbb{R}^{M \times M}, \tag{20}$$

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

$$\max_{\boldsymbol{U}} \ \mathrm{Tr}\Big\{ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{U} \boldsymbol{U}^\top \boldsymbol{A} \boldsymbol{W}_y^{1/2} \Big\} \quad \text{s.t.} \quad \boldsymbol{U}^\top \boldsymbol{R} \boldsymbol{U} = \boldsymbol{I}_K. \tag{21}$$

The Lagrangian of Eq. (21), with symmetric multiplier $\boldsymbol{\Lambda} \in \mathbb{R}^{K \times K}$ for the constraint, yields the stationarity condition

$$\boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top \boldsymbol{U} \ = \ \boldsymbol{R} \boldsymbol{U} \boldsymbol{\Lambda}, \tag{22}$$

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. (22) gives

$$\boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y \boldsymbol{A}^\top \boldsymbol{L}^{-\top} \boldsymbol{V} \ = \ \boldsymbol{V} \boldsymbol{\Lambda}. \tag{23}$$

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

$$\boldsymbol{M} \ = \ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{L}^{-\top} \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \ = \ \boldsymbol{W}_y^{1/2} \boldsymbol{A}^\top \boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2}, \tag{24}$$

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

$$\boldsymbol{V} \ = \ \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2}, \tag{25}$$

one verifies directly that $\boldsymbol{V}$ satisfies Eq. (23) 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 $\boldsymbol{M}$ [28]. Reverting the change of variable gives

$$\boldsymbol{U} \ = \ \boldsymbol{L}^{-\top} \boldsymbol{V} \ = \ \boldsymbol{L}^{-\top} \boldsymbol{L}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} \ = \ \boldsymbol{R}^{-1} \boldsymbol{A} \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2}. \tag{26}$$

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. (26) is

$$\boldsymbol{U}_s \ = \ \boldsymbol{R}_s^{-1} \boldsymbol{A}_s \boldsymbol{W}_y^{1/2} \boldsymbol{V}_M \boldsymbol{\Gamma}^{-1/2} \ \in \ \mathbb{R}^{m_s \times K}, \tag{27}$$

which is the expression in Eq. (5). □

G.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*

$$\min_{\boldsymbol{R}:\, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K} \|\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R}\|_F^2. \tag{28}$$

Proof. Expanding the Frobenius norm in Eq. (28) gives:

$$\begin{aligned} \|\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R}\|_F^2 &= \operatorname{Tr}\big\{(\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R})^\top (\mathbf{Y} - \boldsymbol{Z}\boldsymbol{R})\big\} \\ &= \operatorname{Tr}\big\{\mathbf{Y}^\top \mathbf{Y}\big\} - 2\operatorname{Tr}\big\{\mathbf{Y}^\top \boldsymbol{Z}\boldsymbol{R}\big\} + \operatorname{Tr}\big\{\boldsymbol{R}^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{R}\big\}. \end{aligned} \tag{29}$$

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:

$$\max_{\boldsymbol{R}:\, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K} \operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\}. \tag{30}$$

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:

$$\operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{V}_C \boldsymbol{\Sigma}_C \boldsymbol{U}_C^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{U}_C^\top \boldsymbol{R}\boldsymbol{V}_C \boldsymbol{\Sigma}_C\big\}. \tag{31}$$

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:

$$\operatorname{Tr}\big\{\boldsymbol{C}^\top \boldsymbol{R}\big\} \;=\; \operatorname{Tr}\big\{\boldsymbol{T}\boldsymbol{\Sigma}_C\big\} \;=\; \sum_{k=1}^{K} t_{kk}\sigma_k, \tag{32}$$

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 [48]. Thus:

$$\sum_{k=1}^{K} t_{kk}\sigma_k \;\leq\; \sum_{k=1}^{K} \sigma_k \;=\; \operatorname{Tr}\big\{\boldsymbol{\Sigma}_C\big\}. \tag{33}$$

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:

$$\boldsymbol{U}_C^\top \boldsymbol{R}\boldsymbol{V}_C \;=\; \boldsymbol{I}_K \;\Longrightarrow\; \boldsymbol{R}^* \;=\; \boldsymbol{U}_C \boldsymbol{V}_C^\top. \tag{34}$$

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). □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We support our claims with our extensive experimental results in § 4

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our method in § 6

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide proofs for our used theories in § G

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Setup of each experiment in § 4 is mentioned in the corresponding subsections. Further, implementation details are provided in § A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Although the question is ambiguous to us, we interpret the question as "whether the paper will fully release all code and data in the future", as opposed to "whether the paper is submitting all code and data as part of supplemental material, at the time of submission". All the text-to-image generative models we use in this paper ($\text{FLUX}_{\text{[dev]}}$ and SDv1.4) are publicly available, and so are the text captions we use in image generation. We also include pseudo-code and detailed instructions to reproduce the main results in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Yes, please refer to "Setup" paragraph of the corresponding experiments and § A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide the average and standard deviation of our results in Tab. 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide compute information in § A.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: To our knowledge, we abide by all the guidelines presented in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss positive societal impacts throughout the paper, and also discuss potential negative societal impacts in the concluding remarks. Moreover, we provided some potential defenses against the malicious use of the LatentCompass in § F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not scrape any data from the internet, but use a well-established and audited dataset (COCO). All materials related to LatentCompass (model, data, code) will be extensively audited for safety prior to release. In particular, the LatentCompass-encoders for nudity erasure and SID red teaming will be gated models, and shared only upon request and verification.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We attribute original owners and respect their license, wherever applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: At the moment of submission we are not releasing any new asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.