# KEEP IDENTITY, HIDE ATTRIBUTES: SPECTRAL SUBSPACES FOR SELECTIVE FACIAL PRIVACY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Face images used for identity verification inevitably expose soft biometric attributes such as gender, age, and facial expression, enabling downstream models to infer information that users never intended to reveal. Selectively protecting these attributes remains difficult. Generative editing methods often modify identity-relevant facial structure, whereas adversarial perturbations distribute changes across the entire image and unintentionally affect attributes that users wish to preserve. We ask a simple question: *if an attribute classifier relies on only a few directions in image space, why perturb the whole face?* We propose *GAUSS (Geometric Attribute Suppression via Spectral Subspaces)* answers this by constructing perturbations within low-dimensional, attribute-specific spectral subspaces learned from classifier input gradients. An instance-adaptive coefficient network then generates image-specific perturbations, allowing a single model to protect arbitrary combinations of user-selected attributes without a generative backbone. Beyond the method itself, we show that attribute subspaces are largely classifier-specific and that language-defined semantic supervision substantially improves transfer across architecture families. On CelebA-HQ, GAUSS suppresses each target attribute in at least 92% of test images, achieves the highest non-target retention across all attribute pairs among eight baselines, preserves face verification performance with $71\times$ fewer trainable parameters, and transfers to unseen classifiers at more than twice the rate of prior generative approaches while maintaining 35.79 dB PSNR.

## 1 INTRODUCTION

Modern face recognition systems are designed to answer one question: *Who is this person?* Yet the same face image also answers many other questions that users never agreed to disclose: *Is this person male or female? Young or old? Smiling or serious?* Beyond identity, a single face exposes *soft biometric attributes* such as gender, age, ethnicity, and expression, which downstream models can infer automatically without the user's awareness or consent (Dhar et al., 2021; Gong et al., 2020). As face recognition becomes embedded in everyday systems (Gururaj et al., 2024; Liu et al., 2017; Meng et al., 2021), such unauthorized inference enables demographic profiling, targeted advertising, and algorithmic discrimination (Meden et al., 2021; Kumar et al., 2025). These concerns have motivated *Soft Biometric Privacy Enhancement (SBPE)* (Ji et al., 2022; Mirjalili & Ross, 2017), which aims to suppress sensitive attributes while preserving the image's utility for legitimate purposes. A practical SBPE system must satisfy four competing requirements, which are missing in the majority of the existing works are: (1) *attribute suppression*, (2) *retention of non-target attributes and identity*, (3) *visual imperceptibility*, and (4) *transferability* to unseen attribute classifiers.

Existing SBPE methods satisfy only subsets of these requirements. *Generative approaches* (Mirjalili et al., 2018; 2020; Wang et al., 2026) suppress attributes by synthesizing new facial content, but they often alter identity-relevant structure and require expensive training pipelines. As Figure 1 illustrates, the recent PrivAD (Wang et al., 2026) flips the *Young* attribute by adding a beard, visibly altering the face in the process. *Adversarial privacy methods* (Goodfellow et al., 2015; Shan et al., 2020; Cherepanova et al., 2021; Sun et al., 2024; Zhang et al., 2025) preserve the original facial content by introducing imperceptible pixel-level perturbations, but these perturbations are unstructured, may lie anywhere in $\mathbb{R}^{3HW}$, and frequently interfere with non-target attributes. A recent alternative



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

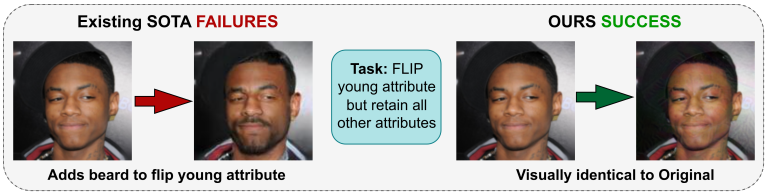


Figure 1: Motivating example. The recent PrivAD method (Wang et al., 2026) flips the *Young* attribute by adding a beard, visibly altering the face. Our proposed GAUSS flips the same attribute while the protected image stays visually close to the original.

restricts the modification to a structured subspace: MIRAGE (Kumar & Agarwal, 2026) confines identity perturbation to the low-frequency wavelet subband, though the structure there is fixed by the transform rather than derived from the model being protected. This raises a fundamental question: *if a classifier's decision about an attribute depends on only a few directions in image space, why perturb the whole face?* Prior work shows that semantic attributes correspond to approximately linear directions in deep representation spaces (Upchurch et al., 2017; Shen et al., 2020; Park et al., 2023). We therefore ask whether a comparable low-dimensional structure exists directly in the input space of an attribute classifier. Rather than assuming such a structure, we investigate it empirically through spectral analysis and a dimensionality study. The results show that the dominant attribute-sensitive directions are consistently recovered from independent subsets of the training data, and that a compact subspace of eight directions captures most of the suppression performance. These findings suggest that optimizing perturbations within a low-dimensional subspace is sufficient in practice while reducing interference with attributes the user intends to preserve.

Motivated by this observation, we propose **GAUSS** (**G**eometric **A**ttribute S**u**ppression via **S**pectral **S**ubspaces), a training-efficient and interpretable framework for selective soft biometric privacy. GAUSS identifies attribute-sensitive low-dimensional subspaces through spectral decomposition of classifier gradients, refines them through orthogonality-constrained learning, and generates instance-adaptive perturbations through dynamic coefficient prediction and spatial masking, so that a single forward pass handles any combination of user-requested attributes with no GAN inversion and no latent optimization. A gradient subspace is, however, largely specific to the classifier that produced it: estimated independently from six architectures on identical data, two subspaces for the same attribute share on average only $0.108$ of their span. Perturbations confined to such a subspace consequently transfer poorly across architecture families, and averaging further classifiers into the objective does not address this, since each additional model contributes one more boundary normal drawn from the same near-orthogonal pool. GAUSS therefore supervises the perturbation not only with trained classifiers but also with a *text-defined* attribute direction obtained from a pair of CLIP text embeddings. This direction specifies the attribute semantically rather than through a fitted decision boundary, enabling supervision that transfers across encoder architectures despite their differing decision boundaries. Our contributions are summarized as follows:

- We formulate selective soft biometric privacy enhancement as generating perturbations from low-dimensional, attribute-specific subspaces of a classifier's input-gradient space, instead of optimizing freely over the full image space.
- We propose a *Spectral Attribute Subspace (SAS)* module that initializes attribute-specific bases from the singular value decomposition of classifier input gradients and refines them through orthogonal parameterization and cross-attribute disentanglement.
- We develop an *Instance-Adaptive Perturbation (IAP)* module that predicts image-specific subspace coefficients and spatial masks, enabling a single model to satisfy arbitrary user-specified combinations of attributes in a single forward pass.
- We demonstrate the effectiveness of GAUSS on CelebA-HQ through comprehensive evaluations against eight baselines, including white-box and black-box suppression, non-target retention, identity verification, robustness, and ablation studies.

## 2 Related Work

Prior work on facial attribute privacy differs in where privacy is enforced: in synthesized image content, in facial templates, or in bounded perturbations applied to the original image.

**Soft Biometric Privacy Enhancement:** Early approaches rely on signal-level transformations such as blurring, masking, and pixelization, which Lander et al. (2001) and Pulfer (2019) show suppress sensitive facial information but significantly degrade identity utility. To address this limitation, Mirjalili & Ross (2017) formalizes SBPE and later develops Semi-Adversarial Networks (Mirjalili et al., 2018) to jointly optimize privacy and recognition utility, while PrivacyNet (Mirjalili et al., 2020) extends the idea to selective suppression through binary control masks, Li et al. (2023) improves localization using attribution maps, and Tang et al. (2022) propose modular architectures for better attribute disentanglement. More recently, PrivAD (Wang et al., 2026) and PFNet (Jin et al., 2026) achieve multi-attribute privacy through disentangled latent editing, the latter adding spatial attention and dual-frequency compensation. A parallel line protects face *templates* rather than images: SensitiveNets (Morales et al., 2020) removes selected attributes from embeddings through an adversarial regularizer, PFRNet (Bortolato et al., 2020) disentangles identity from attribute information through an autoencoder, and Face-CPFNet (Chen et al., 2025) protects both levels by reconstructing images in which the sensitive attribute is replaced by a random guess. Template-level methods protect features exchanged within a recognition system and are therefore complementary to GAUSS, which targets the image itself. In contrast, image-level methods often introduce identity degradation, visual artifacts, and high computational cost.

**Adversarial Perturbations for Privacy:** Adversarial perturbations provide a lightweight alternative that modifies existing images instead of synthesizing new ones. Goodfellow et al. (2015) introduce FGSM, demonstrating the vulnerability of deep networks to imperceptible perturbations, while Fawkes (Shan et al., 2020) and LowKey (Cherepanova et al., 2021) protect facial identities from recognition systems. Closer to our setting, Chhabra et al. (2018) anonymize a subset of facial attributes via adversarial perturbations, and Rezgui et al. (2024) craft mutual information-based image attacks that hide gender from face templates. These methods perturb the entire image space and often affect non-target attributes, and their transferability is conventionally improved through input diversity (Xie et al., 2019) or by averaging several white-box models. Unlike these methods, GAUSS explicitly learns attribute-specific perturbation subspaces and confines the perturbation to them.

**Directional Representations and Language Supervision:** Although primarily designed for facial editing rather than privacy, attribute manipulation methods offer important insights into selective attribute control. VecGAN (Dalva et al., 2022) performs controllable semantic editing through latent directions, StyleFeatureEditor (Bobkov et al., 2024) targets detail-preserving editing, and BS-GAN (Ren et al., 2025) addresses fine-grained attribute editing, demonstrating that directional representations improve controllability and reduce attribute interference. More broadly, Upchurch et al. (2017) and Shen et al. (2020) show that semantic attributes correspond to approximately linear directions in representation space, and Chen et al. (2016) and Karras et al. (2019) explore disentangled representations for controllable generation. These directions, however, reside within a generator and require GAN inversion or iterative latent optimization to apply, making them computationally expensive and less suitable for privacy-preserving deployment, whereas GAUSS operates directly in the classifier's input space.

## 3 Proposed Privacy-Preserving Methodology

We propose GAUSS, which generates selective privacy perturbations from low-dimensional, attribute-specific subspaces. The Spectral Attribute Subspace (SAS) module learns an orthonormal basis for each attribute from the spectrum of classifier input gradients (Section 3.2), and the Instance-Adaptive Perturbation (IAP) module predicts per-image coefficients over these bases together with a spatial mask (Section 3.3). Both are trained jointly under a multi-objective loss in which the attribute is defined by several observers, one of whom specifies it in language rather than via a trained classifier (Section 3.4). An overview of the proposed framework is shown in Figure 2.

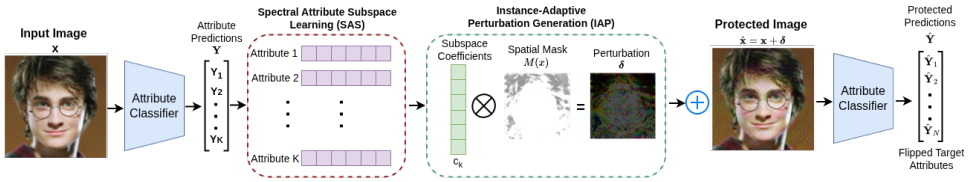


Figure 2: Overview of GAUSS. The Spectral Attribute Subspace (SAS) module learns an orthonormal perturbation basis for each attribute from the input gradients of the frozen attribute classifiers. The Instance-Adaptive Perturbation (IAP) module predicts image-specific coefficients over the bases selected by the privacy prompt, together with a spatial mask that allocates the resulting perturbation. The suppression objective is evaluated on several observers, one of which defines the attribute through a text direction rather than a fitted decision boundary.

### 3.1 Problem Formulation and Threat Model

Let $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ be an input face image normalized to $[-1, 1]$, with $d = 3HW$. We write $\mathrm{vec}(\cdot)$ for flattening and $\mathrm{res}(\cdot)$ for its inverse, $\sigma(\cdot)$ for the sigmoid, $\odot$ for the element-wise product, and $[\,u\,]_+ = \max(0, u)$. Let $\{f_a\}_{a=1}^{A}$ be frozen binary classifiers for the sensitive attributes $a \in \mathcal{A}$ (here *Male*, *Young*, *Smiling*), where $f_a(\mathbf{x}) \in \mathbb{R}$ is the scalar logit and $y_a \in \{0, 1\}$ the ground-truth label. Following PrivAD (Wang et al., 2026), each $f_a$ is a separate ResNet-18 (He et al., 2016) with a single-logit head, sharing no weights. A user supplies a binary privacy prompt $\mathbf{p} \in \{0, 1\}^A$, where $p_a = 1$ requests that attribute $a$ be suppressed and $p_a = 0$ that it remain unchanged.

The task is a protected image $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$ that satisfies the prompt while remaining visually close to $\mathbf{x}$. For a single image this is

$$\min_{\boldsymbol{\delta}} \sum_{a=1}^{A} \Big( p_a\, \mathcal{L}^{(a)}_{flip}(\mathbf{x}+\boldsymbol{\delta}) + (1-p_a)\, \mathcal{L}^{(a)}_{preserve}(\mathbf{x}+\boldsymbol{\delta}) \Big) \quad \text{s.t. } \|\boldsymbol{\delta}\|_\infty \le \varepsilon, \tag{1}$$

where $\mathcal{L}^{(a)}_{flip}$ drives the prediction of $f_a$ away from the true value $y_a$ for each suppressed attribute, $\mathcal{L}^{(a)}_{preserve}$ constrains it to remain at $f_a(\mathbf{x})$ for each retained attribute, and $\varepsilon$ is measured on the $[-1, 1]$ scale, corresponding to $\varepsilon/2$ on the conventional $[0, 1]$ pixel scale. Solving Eq. 1 per image is costly and, with $\boldsymbol{\delta}$ unconstrained in $\mathbb{R}^d$, admits solutions that disturb non-target attributes. GAUSS instead learns a generator $\boldsymbol{\delta} = g_\theta(\mathbf{x}, \mathbf{p})$ that amortizes Eq. 1 over the training set and restricts $\boldsymbol{\delta}$ to attribute-specific subspaces of dimension $k \ll d$.

**Threat model:** The adversary applies pre-trained attribute inference models to protected images: the white-box classifiers $f_a$, independently trained classifiers of unseen architecture, or linear probes on frozen foundation-model features.

### 3.2 Spectral Attribute Subspace Learning (SAS)

**Initialization from classifier gradients:** To first order, the input gradient of $f_a$ identifies the direction of steepest change in the attribute logit. For a training image $\mathbf{x}_n$ we take

$$\mathbf{g}_n = \frac{\mathrm{vec}\big(\nabla_{\mathbf{x}} f_a(\mathbf{x}_n)\big)}{\|\nabla_{\mathbf{x}} f_a(\mathbf{x}_n)\|_2} \in \mathbb{R}^d, \tag{2}$$

and stack $N$ such gradients, drawn in equal numbers from both classes, into $\mathbf{G}_a \in \mathbb{R}^{N\times d}$. Each row is the sensitive direction for one face, and the directions shared across faces follow from the decomposition $\mathbf{G}_a = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, whose top $k$ right singular vectors span the subspace carrying the largest share of $\|\mathbf{G}_a\|_F^2$. These raw directions carry high-frequency components that are visible in the image but contribute little to the classifier response, so we reshape each to $3 \times H \times W$, convolve it with a Gaussian kernel $\mathcal{G}_s$ of bandwidth $s$, and renormalize, giving $\tilde{\mathbf{V}}_a \in \mathbb{R}^{k\times d}$.

**Learnable orthonormal refinement:** Smoothing destroys mutual orthogonality, and directions selected by the gradient spectrum need not be optimal for the privacy objective. We therefore treat the basis as a learnable parameter $\mathbf{B}_a \in \mathbb{R}^{k\times d}$ initialized to $\tilde{\mathbf{V}}_a$ and orthonormalize it at every forward

pass. Given the thin decomposition $\mathbf{B}_a = \mathbf{U}_a \mathbf{\Sigma}_a \mathbf{V}_{h,a}$,

$$\mathbf{V}_a = \mathbf{U}_a \mathbf{V}_{h,a}, \qquad \mathbf{V}_a \mathbf{V}_a^\top = \mathbf{I}_k. \tag{3}$$

$\mathbf{V}_a$ is the polar factor of $\mathbf{B}_a$, the orthonormal-row matrix nearest to it in Frobenius norm (Appendix A.2, Proposition 2), so the projection enforces orthonormality while minimally perturbing the learned parameter. Unlike the individual factors, which are non-unique when singular values coincide, their product is invariant to that ambiguity, and its derivative remains bounded provided $\mathbf{B}_a$ has full row rank.

### 3.3 Instance-Adaptive Perturbation Generation (IAP)

**Coefficient prediction:** The basis $\mathbf{V}_a$ is shared by all images, but attributes vary across individuals in how strongly they are expressed and where they are visible, so its coefficients are predicted per image. Let $\mathbf{z}(\mathbf{x}) = \frac{1}{A}\sum_a \boldsymbol{\phi}_a(\mathbf{x}) \in \mathbb{R}^{512}$ average the penultimate-layer representations $\boldsymbol{\phi}_a$ of the frozen classifiers. A two-layer network predicts

$$\tilde{\boldsymbol{\alpha}}_a = \tanh\Big(\mathbf{P}_2\,\mathrm{GELU}\big(\mathrm{LN}(\mathbf{P}_1\mathbf{z}(\mathbf{x}))\big)\Big) \in [-1,1]^k, \tag{4}$$

where $\mathbf{P}_1, \mathbf{P}_2$ are learnable projections and LN is layer normalization. Since the map from coefficients to image space is an isometry (Appendix A.2, Proposition 3), bounding $\tilde{\boldsymbol{\alpha}}_a$ bounds the $\ell_2$ magnitude of the perturbation. Suppression additionally requires that positive and negative instances be displaced in opposite directions along the same basis, and we therefore condition the sign of the coefficients on the attribute class $c_a \in \{0,1\}$,

$$\boldsymbol{\alpha}_a = (1 - 2c_a)\,\tilde{\boldsymbol{\alpha}}_a, \tag{5}$$

with $c_a = y_a$ during training. One basis and one model therefore serve both classes, at the cost that the orientation of $\boldsymbol{\delta}$ is determined by the attribute value.

**Composition and spatial allocation:** The prompt selects which attributes contribute, weighted by learnable scalars $\beta_a$,

$$\boldsymbol{\delta}_{base} = \sum_{a=1}^{A} \beta_a\, p_a\, \mathrm{res}\big(\mathbf{V}_a^\top \boldsymbol{\alpha}_a\big), \tag{6}$$

so $\boldsymbol{\delta}_{base}$ lies in the span of the selected subspaces by construction. A mask network predicts $\mathbf{m} = \sigma\big(\mathrm{MaskNet}([\mathbf{x}, \boldsymbol{\delta}_{base}])\big) \in [0,1]^{H\times W}$ from the image and the base perturbation, and

$$\boldsymbol{\delta} = \varepsilon\, \mathbf{m} \odot \tanh\left(\frac{\boldsymbol{\delta}_{base}}{\varepsilon}\right), \qquad \mathbf{x}' = \mathrm{clip}_{[-1,1]}(\mathbf{x} + \boldsymbol{\delta}), \tag{7}$$

with $\mathbf{m}$ broadcast over color channels. Since $|\tanh(\cdot)| < 1$ and $\mathbf{m} \le 1$, the constraint of Eq. 1 is satisfied without clipping (Appendix A.2, Proposition 1) and the mapping remains differentiable. Both operations act element-wise, so the released $\boldsymbol{\delta}$ departs slightly from that span; we measure rather than assume this, finding 87–91% of $\|\boldsymbol{\delta}\|^2$ inside it against 0.05% for a random perturbation (Appendix A.3). During training $\mathbf{p}$ is sampled uniformly over all $2^A$ combinations, so a single model with 1.38M trainable parameters serves arbitrary privacy requests at inference.

### 3.4 Training Objective

The protected image should suppress the requested attributes, preserve the remaining ones, stay visually faithful, and be built from well-separated subspaces:

$$\mathcal{L} = \lambda_{flip}\mathcal{L}_{flip} + \lambda_{preserve}\mathcal{L}_{preserve} + \lambda_{norm}\mathcal{L}_{norm} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{disent}\mathcal{L}_{disent}, \tag{8}$$

with weights given in Appendix A.1.

**Attribute suppression:** For each attribute marked for suppression we penalize the margin by which the classifier still supports the ground-truth label. Let $\mathrm{ID}(\cdot)$ denote the input-diversity transform of Xie et al. (2019), which randomly rescales and pads the protected image, and $\mathcal{O}_a$ a set of *observers* of attribute $a$, each a scalar-valued function of the image. Then

$$\mathcal{L}_{flip} = \sum_{a=1}^{A} p_a \cdot \frac{1}{|\mathcal{O}_a|} \sum_{h\in\mathcal{O}_a} \Big[\big(2y_a - 1\big)\, h\big(\mathrm{ID}(\mathbf{x}')\big) + \gamma\Big]_+, \tag{9}$$

with margin $\gamma$. The term vanishes once $h$ classifies the image beyond its decision boundary by $\gamma$, so the perturbation budget is allocated only to instances not yet suppressed. Resampling $\mathrm{ID}(\cdot)$ at each step discourages the perturbation from overfitting the spatial sampling of any single network, while averaging over $\mathcal{O}_a$ discourages it from overfitting a single decision boundary.

**Observers:** $\mathcal{O}_a$ comprises the white-box classifier $f_a$, two further CNNs, and one observer that defines the attribute without fitting a decision boundary. Writing $E_{\text{img}}$ and $E_{\text{txt}}$ for the image and text encoders of CLIP (Radford et al., 2021) and $\tau_a^{\pm}$ for prompts naming the two poles of attribute $a$, such as *"a photo of a man"* and *"a photo of a woman"*, this observer scores the attribute as

$$\Delta \mathbf{t}_a = \frac{E_{\text{txt}}(\tau_a^+)}{\|E_{\text{txt}}(\tau_a^+)\|} - \frac{E_{\text{txt}}(\tau_a^-)}{\|E_{\text{txt}}(\tau_a^-)\|}, \qquad h_a^{\text{txt}}(\mathbf{x}) = \left\langle \tfrac{E_{\text{img}}(\mathbf{x})}{\|E_{\text{img}}(\mathbf{x})\|}, \Delta \mathbf{t}_a \right\rangle, \tag{10}$$

thresholded at zero, so that no attribute labels enter its construction. Section 5.3 shows that the presence of this observer, rather than their number, accounts for the improvement in cross-architecture transfer.

**Retention and fidelity:** For retained attributes the reference is the classifier's output on the clean image rather than the ground-truth label, $\mathcal{L}_{preserve} = \sum_a (1 - p_a)\,\mathrm{BCE}\big(f_a(\mathbf{x}'),\, \sigma(f_a(\mathbf{x}))\big)$, where $\mathrm{BCE}(u, t)$ denotes the binary cross-entropy between $\sigma(u)$ and target $t$. The perturbation is therefore constrained to preserve the classifier's original predictions, including those that are incorrect, rather than to revise them. Since $\varepsilon$ bounds each pixel change but permits that bound to be attained everywhere, we additionally penalize the perturbation magnitude, $\mathcal{L}_{norm} = \|\boldsymbol{\delta}\|_2$, and the structural difference between the two images, $\mathcal{L}_{ssim} = 1 - \mathrm{SSIM}(\mathbf{x}, \mathbf{x}')$.

**Subspace separation:** Retention is easier when the directions used for one attribute do not move another, so we penalize overlap between subspaces,

$$\mathcal{L}_{disent} = \sum_{a \neq b} \|\mathbf{V}_a \mathbf{V}_b^\top\|_F^2, \tag{11}$$

whose entries are inner products between basis vectors of the two attributes, so the penalty vanishes exactly when their row spaces are mutually orthogonal (Appendix A.2, Proposition 4). It is effective at negligible cost: raising $\lambda_{disent}$ from 0 to 2 reduces the overlap from $0.201$ to $0.006$ with no loss of flip rate or fidelity, leaving a mean principal angle of $89.8°$ between attribute subspaces (Appendix A.7).

## 4 Experimental Setup

**Dataset and protocol:** All experiments are conducted on CelebA-HQ (Lee et al., 2020), comprising 30,000 face images downsampled to $128 \times 128$, with three binary soft biometric attributes: *Male*, *Young*, and *Smiling*. We use a fixed split of 25,000 training and 5,000 test images, identical to the PrivAD protocol (Wang et al., 2026), enabling fair comparison with baselines. Suppression and retention are evaluated under single-attribute prompts, whereas the identity evaluation and the ablations use the hardest all-attribute setting $\mathbf{p} = [1, 1, 1]$.

**Evaluation metrics:** We evaluate GAUSS across four dimensions. *(1) Attribute suppression* is measured using *Flip Rate*, the percentage of test images whose predicted attribute differs from the ground truth after anonymization; for unseen evaluators, we instead measure flips relative to each evaluator's prediction on the clean image so that its own classification errors are not credited to the anonymizer. *(2) Attribute specificity* is measured using *Non-Target Preservation (NTP)*, the percentage of non-target attributes whose predictions remain unchanged. *(3) Identity preservation* is evaluated with ArcFace (Deng et al., 2019), FaceNet (Schroff et al., 2015), VGG-Face (Parkhi et al., 2015), and FaceNet512 (Schroff et al., 2015) over genuine and impostor pairs, and is always reported alongside the unprotected reference because absolute AUC primarily reflects the verifier rather than the anonymizer. *(4) Perceptual quality* is measured using PSNR (Hore & Ziou, 2010) and SSIM (Hore & Ziou, 2010). To evaluate transferability, we consider two tiers of unseen models: independently trained ResNet-50 (He et al., 2016), EfficientNet-B0 (Tan & Le, 2019), and MobileNetV3 (Howard et al., 2019), together with linear probes on frozen DINOv2-L (Oquab et al., 2023), SigLIP-B (Zhai et al., 2023), and CLIP-H (Cherti et al., 2023) features. None of these evaluators is used during training; the text observer in Eq. 10 instead employs CLIP ViT-B/16, which is distinct from the CLIP-H evaluator.

Table 1: White-box evaluation on CelebA-HQ with single-attribute prompts. **(a)** Attribute suppression accuracy (%, ↑). **(b)** Non-target attribute retention (%, ↑), measured relative to the classifier's prediction on the clean image. **Bold** indicates the best result in each row. BSGAN does not evaluate *Smiling* ("–"), and † denotes baseline results reproduced under our evaluation protocol.

| Attribute | | HiSD | RAPP | PrivacyNet | VecGAN | GenderAN | SFE | BSGAN | PrivAD | GAUSS (OURS) |
|---|---|---|---|---|---|---|---|---|---|---|
| *(a) Target attribute suppression* | | | | | | | | | | |
| Male | | 93.93† | 93.87† | 72.13† | 81.05† | 93.13† | 90.12† | 87.81† | 94.66† | **96.82** |
| Young | | 63.28† | 36.30† | 87.03† | 77.15† | 48.47† | 84.97† | 47.78† | 82.78† | **99.56** |
| Smiling | | 72.65† | 25.60† | 88.37† | 91.13† | 68.47† | 91.42† | – | 91.73† | **92.88** |
| *(b) Non-target attribute retention (target → retained)* | | | | | | | | | | |
| Male | Young | 77.18† | 72.83† | 74.13† | 72.38† | 80.23† | 72.16† | 56.98† | 79.65† | **97.30** |
| Male | Smile | 91.49† | 94.17† | 86.10† | 83.96† | 92.40† | 59.46† | 58.14† | 91.86† | **98.18** |
| Young | Male | 91.63† | 92.67† | 89.67† | 83.86† | 98.30† | 61.23† | 68.96† | 90.12† | **99.66** |
| Young | Smile | 91.53† | 90.67† | 91.53† | 90.79† | 91.50† | 65.26† | 57.05† | 92.70† | **98.36** |
| Smiling | Young | 88.19† | 79.53† | 83.03† | 84.99† | 87.00† | 73.27† | – | 88.89† | **97.36** |
| Smiling | Male | 96.07† | 95.83† | 90.77† | 95.46† | 95.67† | 67.91† | – | 96.76† | **99.58** |
| **Retention mean** | | 89.35 | 87.62 | 85.87 | 85.24 | 90.85 | 66.55 | 60.28 | 90.00 | **98.41** |




Figure 3: Qualitative comparison. **(a)** Original images, PrivAD, and GAUSS. GAUSS preserves facial structure and identity consistency while performing targeted attribute suppression, whereas PrivAD introduces visible changes and artifacts. **(b)** Original image, proposed protected image, perturbation $\boldsymbol{\delta}$, and spatial mask $M$. The perturbation is amplified $10\times$ and offset to mid-grey for display, since at $\varepsilon = 16/255$ it is not visible at its true magnitude.

**Baselines:** For a fair comparison, we evaluate GAUSS against eight methods spanning soft biometric privacy enhancement and facial attribute editing from 2020 to 2026: PrivacyNet (Mirjalili et al., 2020), HiSD (Li et al., 2021), VecGAN (Dalva et al., 2022), RAPP (Zhang et al., 2023), GenderAN (Tang et al., 2022), SFE (Bobkov et al., 2024), BSGAN (Ren et al., 2025), and PrivAD (Wang et al., 2026). Values marked † are obtained by re-running under our evaluation protocol.

## 5 Experimental Results and Analysis

### 5.1 Attribute Suppression and Non-Target Retention

We first evaluate whether GAUSS suppresses user-specified attributes while preserving the facial attributes that users choose to retain. Table 1(a) reports target attribute suppression, and Table 1(b) reports the corresponding non-target retention. GAUSS consistently outperforms all baselines in target attribute suppression. The largest improvement is observed for *Young*, where existing methods range from $36.30\%$ to $87.03\%$, while GAUSS reaches $99.56\%$. It also improves upon the best-performing baseline on *Male* ($96.82\%$ vs. $94.66\%$) and *Smiling* ($92.88\%$ vs. $91.73\%$). A paired comparison against PrivAD is reported in (Appendix A.9).

GAUSS likewise outperforms all competing methods in non-target attribute retention. It achieves the highest retention across all six target–retained attribute pairs, with a mean of $98.41\%$, compared with $90.85\%$ for the closest baseline and $90.00\%$ for PrivAD. It is also the only method to exceed $97\%$ retention for every attribute pair, demonstrating that the learned spectral subspaces effectively disentangle attribute-specific perturbations. These gains are achieved with substantially higher visual fidelity. GAUSS preserves image quality at $35.79$ dB PSNR and $0.95$ SSIM, compared with

Table 2: Identity verification under all-attribute protection ($\mathbf{p} = [1, 1, 1]$), comparing the proposed GAUSS with the strongest prior baseline, PrivAD. Original denotes the unprotected reference, and the identity degradation introduced by each protection method is measured relative to it.

| Verifier | AUC ↑ | | | EER ↓ | | | TAR@1% FAR ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Original | PrivAD | Ours | Original | PrivAD | Ours | Original | PrivAD | Ours |
| ArcFace | 0.7546 | 0.7015 | **0.7256** | 31.84 | 35.91 | **33.43** | 8.73 | 5.30 | **7.49** |
| FaceNet | 0.9682 | 0.8986 | **0.9624** | 9.57 | 18.44 | **10.29** | 68.65 | 33.72 | **64.73** |
| VGG-Face | 0.9834 | 0.9467 | **0.9815** | 5.68 | 12.25 | **6.28** | 86.05 | 54.12 | **84.52** |
| FaceNet512 | 0.9880 | 0.9435 | **0.9868** | 4.84 | 12.85 | **5.27** | 88.04 | 56.28 | **86.40** |

Table 3: Generalization to unseen attribute inference models, comparing GAUSS with the strongest prior baseline, PrivAD. Flip rate (%, ↑) is measured relative to each evaluator's prediction on the corresponding clean image. The Clean row reports the attribute prediction accuracy of each unseen evaluator on unprotected images.

| Method | Held-out CNNs | | | Foundation | | PSNR |
|---|---|---|---|---|---|---|
| | RN-50 | EN-B0 | MNv3 | SigLIP-B | CLIP-H | |
| Clean | 93.7 | 94.2 | 93.5 | 93.5 | 92.4 | |
| PrivAD | 29.4 | 31.0 | 28.5 | 28.6 | **28.4** | 19.40 |
| Ours | **68.5** | **74.5** | **48.5** | **41.0** | 26.7 | **35.79** |

$19.40$ dB and $0.72$ for PrivAD. As illustrated in Figure 3, PrivAD suppresses attributes through visible facial edits that introduce structural changes and identity drift, whereas GAUSS produces visually consistent images while preserving facial structure. Additional qualitative comparisons are provided in Appendix A.12.

### 5.2 Does Soft Privacy Come at the Cost of Identity?

Selective privacy is meaningful only if the protected image remains recognizable as its owner. Table 2 therefore evaluates identity verification under the all-attribute setting, always using the unprotected image as the reference because absolute verification accuracy primarily reflects the verifier rather than the anonymizer. Across all four face verifiers, GAUSS introduces only a negligible identity gap. Relative to the original images, AUC decreases by at most $0.0058$, while EER increases by less than $0.8$ percentage points. In contrast, PrivAD reduces AUC by up to $0.0696$ and more than doubles the verification error on the strongest verifiers. The difference is even more pronounced at deployment-relevant operating points: on VGG-Face, GAUSS retains $84.52\%$ TAR at $1\%$ FAR, compared with $54.12\%$ for PrivAD. Notably, these results are obtained without using any face recognition model during training. Identity preservation arises solely from the bounded-perturbation formulation and the fidelity objectives, indicating that selective attribute suppression does not require sacrificing verification utility. The same perturbation also transfers to unaligned in-the-wild faces, changing LFW verification AUC by only $-0.0009$ (Appendix A.10).

### 5.3 Generalization to Unseen Classifiers and Foundation Models

The previous experiments evaluate GAUSS against the same attribute classifiers used for supervision. In practice, however, the inference model applied to a shared image is unknown. We therefore evaluate transfer on three independently trained CNNs and two frozen foundation-model probes, none of which participates in training (Table 3). GAUSS generalizes substantially better to unseen CNN architectures, achieving flip rates of $68.5\%$, $74.5\%$, and $48.5\%$ on ResNet-50, EfficientNet-B0, and MobileNetV3, compared with $29.4\%$, $31.0\%$, and $28.5\%$ for PrivAD. At the same time, it preserves considerably higher visual fidelity, reaching $35.79$ dB PSNR versus $19.40$ dB. On the foundation-model probes, the gap narrows: GAUSS improves suppression on SigLIP-B ($41.0\%$ vs. $28.6\%$), while CLIP-H exhibits comparable performance between the two methods with a minor performance improvement of $1.7\%$ of PrivAD against our proposed approach.

This behavior reflects the geometry of the learned perturbation. Content-editing methods such as PrivAD modify facial appearance and therefore exhibit relatively uniform transfer across evaluators. In contrast, bounded perturbations depend on the representational similarity between the supervised

Table 4: **(a)** Component ablation study. **(b)** Robustness under JPEG compression and downsampling. Both under all-attribute suppression.

(a) Component ablation

| Variant | Male | Young | Smiling | PSNR |
|---|---|---|---|---|
| w/o basis learning | 32.92 | 65.86 | 24.12 | 37.19 |
| w/o IAP | 31.92 | 65.32 | 41.34 | **40.70** |
| GAUSS | **96.82** | **99.56** | **92.88** | 35.79 |

(b) Robustness under JPEG compression and resizing.

| Attr. | JPEG Compression | | Resizing | |
|---|---|---|---|---|
| | PrivAD | Ours | PrivAD | Ours |
| Male | 59.20 | **72.80** | 59.00 | **70.60** |
| Young | 28.40 | **90.40** | 18.80 | **66.60** |
| Smiling | 13.80 | **60.60** | 19.00 | **33.60** |

classifiers and the target model, making transfer naturally stronger to held-out CNNs than to image-text foundation models. The same observation explains why simply supervising with additional classifiers provides limited benefit. When the *Male* subspace is estimated independently from six architectures, two classifiers share only $0.108$ of their subspace on average, decreasing to $0.037$ between the white-box ResNet-18 and ViT-B/16, against a chance overlap of $0.0002$ (Appendix A.5). Rather than averaging increasingly distinct decision boundaries, GAUSS introduces a text-defined semantic direction that provides supervision beyond any single classifier family.

### 5.4 Ablation and Robustness Studies

We next examine the contribution of the two learned components of GAUSS. As shown in Table 4(a), both the learned subspace basis and the Instance-Adaptive Perturbation (IAP) module are essential. Freezing the basis at its gradient-derived initialization, without the orthonormal refinement of Eq. 3, reduces suppression from $96.82\%$ to $32.92\%$ for *Male* and from $92.88\%$ to $24.12\%$ for *Smiling*, demonstrating that the spectral initialization alone is insufficient and that most of the suppression is obtained during refinement. Removing IAP similarly reduces suppression to $31.92\%$ and $41.34\%$, while producing the highest PSNR among all variants. This indicates that fixed perturbation coefficients preserve visual quality by underutilizing the perturbation budget, whereas instance-adaptive coefficient prediction is necessary for effective attribute suppression. The accompanying subspace-dimension study further shows that performance saturates at $k = 8$, with larger subspaces providing negligible gains while reducing reproducibility across independent training subsets (Appendix A.4).

We further evaluate robustness under common post-processing operations. Table 4(b) shows that GAUSS consistently retains higher suppression than PrivAD under both JPEG compression and resizing. The largest margin occurs for *Young*, where GAUSS preserves $90.40\%$ suppression under JPEG compression compared with $28.40\%$ for PrivAD. Resizing is more challenging for both methods because spatial resampling attenuates the high-frequency components of the perturbation, with *Smiling* exhibiting the greatest degradation. A comprehensive evaluation of sixteen image transformations is provided in Appendix A.8. Finally, GAUSS remains substantially more efficient, requiring only $1.38$M trainable parameters, $4.5$ MB of protection state, and $9.33$ ms inference time, compared with $98.0$M parameters, $749.8$ MB, and $24.86$ ms for PrivAD (Appendix A.11).

## 6 Conclusion

Selective soft biometric privacy requires suppressing only the attributes a user chooses while preserving identity, visual fidelity, and the remaining facial characteristics. This work shows that such selective protection does not require unconstrained pixel-space optimization or generative image editing. Instead, the dominant attribute-sensitive directions of a classifier form compact spectral subspaces that provide an effective geometric prior for privacy-preserving perturbations. By combining spectral attribute subspace learning, instance-adaptive perturbation generation, and language-guided supervision, GAUSS achieves highly selective attribute suppression while preserving non-target attributes and identity, and generalizes substantially better to unseen classifier architectures than existing generative approaches. These results suggest that the geometry of classifier sensitivity, rather than image synthesis, offers a principled foundation for selective facial privacy. A promising direction is to learn language-defined semantic subspaces, enabling open-vocabulary protection of unseen soft biometric attributes.

### AI USE STATEMENT

AI tools are used solely to assist with language refinement, improving clarity and readability, and LaTeX formatting during manuscript preparation. Generative AI tools are not used to formulate research ideas, design methodologies, conduct experiments, implement proposed models, analyze results, or draw scientific conclusions. All AI-assisted edits are reviewed and verified by the authors, who take full responsibility for the final content of this manuscript, including all text, figures, tables, claims, and experimental results.

### ETHICS STATEMENT

This work develops a privacy-enhancing method to selectively suppress soft biometric attributes while preserving the utility of identity verification. The intended application is to reduce unauthorized demographic inference from facial images shared for authentication and other legitimate purposes. Our evaluation uses the publicly available CelebA-HQ dataset and does not involve human subjects or the collection of new personal data. The proposed threat model is limited to pretrained attribute inference models and does not claim protection against adversaries who retrain on protected images. As with other privacy-preserving technologies, responsible deployment requires transparency about the intended threat model and application context.

### REPRODUCIBILITY STATEMENT

We make reproducibility a central design goal of this work. The main paper provides the complete problem formulation, model architecture, training objective, dataset split, threat model, and evaluation protocol, while the appendix reports implementation details, hyperparameters, optimization settings, statistical significance tests, robustness evaluations, efficiency analysis, and additional ablation studies. The source code, trained models, and scripts required to reproduce all reported experiments will be released upon publication.

## REFERENCES

Denis Bobkov, Vadim Titov, Aibek Alanov, and Dmitry Vetrov. The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9337–9346, 2024.

Blaž Bortolato, Marija Ivanovska, Peter Rot, Janez Križaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Štruc. Learning privacy-enhancing face representations through feature disentanglement. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 495–502. IEEE, 2020.

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

Zheyu Chen, Zhiqiang Yao, Biao Jin, Jianting Ning, and Mingwei Lin. Face-cpfnet: Leveraging disentangled representations for dual-level soft-biometric privacy-enhancement. *IEEE Transactions on Dependable and Secure Computing*, 2025.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k facial attributes via adversarial perturbations. pp. 656–662, 07 2018. doi: 10.24963/ijcai.2018/91.

Yusuf Dalva, Said Fahri Altındiş, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. In *European conference on computer vision*, pp. 153–169. Springer, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, and Rama Chellappa. Pass: protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15087–15096, 2021.

Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European conference on computer vision*, pp. 330–347. Springer, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *stat*, 1050:20, 2015.

HL Gururaj, BC Soundarya, S Priya, J Shreyas, and Francesco Flammini. A comprehensive review of face recognition techniques, trends, and challenges. *IEEE Access*, 12:107903–107926, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, and Rongrong Ji. Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *European Conference on Computer Vision*, pp. 475–491. Springer, 2022.

Biao Jin, Haowei Huang, Zhiqiang Yao, Jinbo Xiong, Xuan Li, Xing Wang, and Li Lin. Pfnet: A face soft biometric privacy enhancement method based on attribute disentanglement and frequency compensation. *Expert Systems with Applications*, pp. 133230, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Atul Kumar and Akshay Agarwal. Visible yet unrecognizable: Frequency-selective facial privacy via attention. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2026.

Atul Kumar, Akshay Agarwal, and Nalini Ratha. Your face, your privacy: Combating unauthorized usage. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10. IEEE, 2025.

Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 15(1):101–116, 2001.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5549–5558, 2020.

Jingzhi Li, Hua Zhang, Siyuan Liang, Pengwen Dai, and Xiaochun Cao. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Transactions on Information Forensics and Security*, 18:3632–3646, 2023.

Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8639–8648, 2021.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14225–14234, 2021.

Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International joint conference on biometrics (IJCB)*, pp. 564–573. IEEE, 2017.

Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, pp. 82–89. IEEE, 2018.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020.

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*. British Machine Vision Association, 2015.

Erich-Matthew Pulfer. Different approaches to blurring digital images and their effect on facial detection. 2019.

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

Fanghui Ren, Wenpeng Liu, Fasheng Wang, Bo Wang, and Fuming Sun. Facial attribute editing via a balanced simple attention generative adversarial network. *Expert Systems with Applications*, 277:127245, 2025.

Zohra Rezgui, Nicola Strisciuglio, and Raymond Veldhuis. Enhancing soft biometric face template privacy with mutual information-based image attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1141–1149, 2024.

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1589–1604, 2020.

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

Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24584–24594. IEEE, 2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Deyan Tang, Siwang Zhou, Hongbo Jiang, Haowen Chen, and Yonghe Liu. Gender-adversarial networks for face privacy preserving. *IEEE Internet of Things Journal*, 9(18):17568–17576, 2022.

Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7064–7073, 2017.

Yue Wang, Biao Jin, Zheyu Chen, Jinsen Lin, and Zhiqiang Yao. Privacy preservation in face soft biometrics via attribute disentanglement. *Expert Systems with Applications*, pp. 131520, 2026.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2725–2734. IEEE, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Yushu Zhang, Tao Wang, Ruoyu Zhao, Wenying Wen, and Youwen Zhu. Rapp: Reversible privacy preservation for various face attributes. *IEEE Transactions on Information Forensics and Security*, 18:3074–3087, 2023.

Yushu Zhang, Zixuan Yang, Tao Wang, Zhongyun Hua, and Jian Weng. Tailor-made face privacy protection via class-wise targeted universal adversarial perturbations. *IEEE Transactions on Dependable and Secure Computing*, 22(5):5108–5120, 2025.

# A Appendix

The appendix follows the organization of the main text. Appendices A.1–A.3 support the method of Section 3, giving implementation details, the properties underlying the construction, and a measurement of how closely the released perturbation adheres to the learned subspaces. Appendices A.4-A.9 support the analysis of Section 5, covering the subspace dimension study, cross-architecture overlap, evaluator accuracies, hyperparameter sweeps, extended robustness, and statistical testing. Appendices A.10-A.13 report supplementary results and the broader impact statement.

## A.1 Implementation Details

GAUSS is implemented in PyTorch and trained on a single NVIDIA H200 GPU with a fixed random seed of 42.

**SAS:** The Spectral Attribute Subspace module is initialized from $N = 4{,}000$ gradient samples per attribute (2,000 per class) drawn from the training set. Each basis is smoothed with bandwidth

$s = 2.0$ over a $9 \times 9$ kernel and truncated to $k = 8$ basis vectors. The orthonormalization of Eq. 3 is computed with `torch.linalg.svd` on the thin factorization at every forward pass; no non-finite gradients were observed in any training run. A norm-scaled diagonal jitter ($10^{-6}$) and a QR fallback are available as safeguards against degenerate singular values, but neither was triggered and the reported runs do not use them.

**IAP:** The coefficient prediction MLP of the Instance-Adaptive Perturbation module uses a hidden dimension of 128. The mask network takes the channel-wise concatenation of the image and the unmasked perturbation (six channels) and applies a $3 \times 3$ convolution to 16 channels with batch normalization and ReLU, a second $3 \times 3$ convolution at 16 channels with batch normalization and ReLU, and a $1 \times 1$ convolution to a single channel followed by a sigmoid, all at stride 1 with padding 1 so that spatial resolution is preserved. The final convolution is zero-initialized with bias 0.5, so training begins from a near-uniform mask and any spatial selectivity must be learned. The mask network accounts for 3,281 of the 1.38M trainable parameters. The perturbation budget is $\varepsilon = 16/255$ in $\ell_\infty$ norm on the $[-1, 1]$ scale, equivalently $8/255$ on the $[0, 1]$ scale.

**Objective:** The hinge margin is $\gamma = 0.5$, and the input-diversity transform (Xie et al., 2019) applies a random rescale with rate 0.85 followed by random padding, with probability 0.5. The observer set $\mathcal{O}_a$ contains the white-box ResNet-18, ResNet-101, DenseNet-121, and the text-defined head of Eq. 10 built on CLIP ViT-B/16. Loss weights are $\lambda_{\text{flip}} = 10.0$, $\lambda_{\text{preserve}} = 5.0$, $\lambda_{\text{ssim}} = 8.0$, $\lambda_{\text{norm}} = 1.0$ (adaptively scheduled to hold PSNR within $[28, 36]$ dB), and $\lambda_{\text{disent}} = 1.0$.

**Prompts:** For each attribute we average the text embeddings of three templates per polarity before taking the difference, which reduces the sensitivity of $\Delta\mathbf{t}_a$ to any single phrasing:

> **Male**: *"a photo of a man"* / *"a close-up photo of a man's face"* / *"a portrait of a male person"* versus *"a photo of a woman"* / *"a close-up photo of a woman's face"* / *"a portrait of a female person"*.
>
> **Young**: *"a photo of a young person"* / *"a portrait of a young adult"* / *"a close-up photo of a youthful face"* versus *"a photo of an old person"* / *"a portrait of an elderly person"* / *"a close-up photo of an older adult's face"*.
>
> **Smiling**: *"a photo of a smiling person"* / *"a portrait of a person smiling"* / *"a close-up photo of a face with a big smile"* versus *"a photo of a person with a neutral expression"* / *"a portrait of a person who is not smiling"* / *"a close-up photo of a serious face"*.

**Optimization:** The model is trained for 40 epochs with batch size 32 using AdamW (Kingma & Ba, 2014) with weight decay $10^{-4}$ and a cosine annealing schedule, with separate learning rates for the SAS parameters ($\eta_{\text{SAS}} = 5 \times 10^{-5}$) and all remaining parameters ($\eta = 2 \times 10^{-4}$). Binary privacy prompts are sampled uniformly across all attribute combinations during training.

**Baseline evaluation:** All baseline results marked $^\dagger$ are obtained by re-running under our evaluation protocol, using the same test split, the same attribute classifiers, and the same metric definitions applied to GAUSS.

### A.2 Properties of the Construction

The following properties justify design choices that are otherwise only asserted in Section 3: why the budget requires no clipping, why the polar factor is the appropriate orthonormalization, why bounding the coefficients bounds the distortion, and why the disentanglement penalty is an exact measure of overlap. Throughout, $\mathbf{V}_a \in \mathbb{R}^{k \times d}$ has orthonormal rows, $\mathbf{V}_a \mathbf{V}_a^\top = \mathbf{I}_k$.

**Proposition 1 (the budget holds by construction):** For any $\boldsymbol{\delta}_{base}$ and any mask $\mathbf{m} \in [0, 1]^{H \times W}$, the perturbation of Eq. 7 satisfies $\|\boldsymbol{\delta}\|_\infty < \varepsilon$.

*Proof:* Element-wise, $|\delta_i| = \varepsilon\, m_i\, |\tanh(\delta_{base,i}/\varepsilon)| \le \varepsilon \cdot 1 \cdot |\tanh(\cdot)| < \varepsilon$, since $m_i \le 1$ and $|\tanh(u)| < 1$ for all finite $u$. □

No clipping is applied, so the mapping remains differentiable everywhere and the gradient does not vanish at the boundary, unlike $\text{clip}(\cdot, \pm\varepsilon)$, whose gradient is zero once saturated.

**Proposition 2 (the polar factor is the nearest orthonormal basis):** Let $\mathbf{B}_a = \mathbf{U}_a \boldsymbol{\Sigma}_a \mathbf{V}_{h,a}$ be a thin SVD with $\mathbf{B}_a$ of full row rank. Then $\mathbf{V}_a = \mathbf{U}_a \mathbf{V}_{h,a}$ solves $\arg\min_{\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_k} \|\mathbf{Q} - \mathbf{B}_a\|_F$.

*Proof:* For any $\mathbf{Q}$ with orthonormal rows, $\|\mathbf{Q} - \mathbf{B}_a\|_F^2 = \|\mathbf{B}_a\|_F^2 + k - 2\,\mathrm{tr}(\mathbf{Q}\mathbf{B}_a^\top)$, so minimizing the distance maximizes $\mathrm{tr}(\mathbf{Q}\mathbf{B}_a^\top)$. Writing $\mathbf{Z} = \mathbf{V}_{h,a}\mathbf{Q}^\top\mathbf{U}_a$, whose entries are bounded by 1 in magnitude, $\mathrm{tr}(\mathbf{Q}\mathbf{B}_a^\top) = \sum_i \sigma_i Z_{ii} \le \sum_i \sigma_i$, with equality if and only if $Z_{ii} = 1$ for every $i$, that is, if and only if $\mathbf{Q} = \mathbf{U}_a\mathbf{V}_{h,a}$. □

This is the orthogonal Procrustes solution (Schönemann, 1966). Two consequences follow. Refinement displaces the basis as little as possible while restoring orthonormality, so the gradient-derived initialization is retained; and while the individual factors are non-unique when singular values coincide, the product $\mathbf{U}_a\mathbf{V}_{h,a}$ is invariant to their sign and rotation ambiguities, with derivative bounded while $\mathbf{B}_a$ keeps full row rank.

**Proposition 3 (coefficients control distortion):** $\|\boldsymbol{\alpha}_a^\top\mathbf{V}_a\|_2 = \|\boldsymbol{\alpha}_a\|_2$.

*Proof:* $\|\boldsymbol{\alpha}^\top\mathbf{V}\|_2^2 = \boldsymbol{\alpha}^\top\mathbf{V}\mathbf{V}^\top\boldsymbol{\alpha} = \boldsymbol{\alpha}^\top\mathbf{I}_k\boldsymbol{\alpha} = \|\boldsymbol{\alpha}\|_2^2$. □

The map from coefficients to perturbation is an isometry, so bounding $\boldsymbol{\alpha}$ bounds the $\ell_2$ distortion directly. The $\tanh$ on the coefficient network is therefore sufficient, and no separate norm constraint on $\boldsymbol{\delta}_{base}$ is required, which is why orthonormality is enforced rather than merely encouraged.

**Proposition 4 (the penalty measures subspace overlap exactly):** $\mathcal{L}_{disent} = \sum_{a\neq b}\|\mathbf{V}_a\mathbf{V}_b^\top\|_F^2 = 0$ if and only if the row spaces of $\mathbf{V}_a$ and $\mathbf{V}_b$ are mutually orthogonal, and $\|\mathbf{V}_a\mathbf{V}_b^\top\|_F \in [0, \sqrt{k}]$ with the upper bound attained if and only if the row spaces coincide.

*Proof:* $\|\mathbf{V}_a\mathbf{V}_b^\top\|_F = 0 \iff \mathbf{V}_a\mathbf{V}_b^\top = \mathbf{0} \iff \langle\mathbf{v}_i^{(a)}, \mathbf{v}_j^{(b)}\rangle = 0$ for all $i, j$. The singular values of $\mathbf{V}_a\mathbf{V}_b^\top$ are the cosines of the $k$ principal angles between the subspaces, each in $[0, 1]$, so $\|\mathbf{V}_a\mathbf{V}_b^\top\|_F^2 = \sum_i \cos^2\theta_i \le k$, with equality if and only if every $\theta_i = 0$. □

The penalty is therefore a normalized and interpretable measure rather than an arbitrary regularizer, and the principal angles it is built from are reported in Tables 6 and 7.

### A.3 Concentration of the Perturbation

Table 5: Fraction of released perturbation energy lying inside the active subspace, measured over the test set after the mask and $\tanh$ of Eq. 7.

| Prompt | Rank | Mean in-span | Std | Min |
|---|---|---|---|---|
| Male | 8 | 0.8742 | 0.0277 | 0.8047 |
| Young | 8 | 0.9085 | 0.0244 | 0.8382 |
| Smiling | 8 | 0.8803 | 0.0342 | 0.7869 |
| All-attribute | 24 | 0.8664 | 0.0179 | 0.8011 |
| *Random reference* | 8 | 0.0005 | – | – |

By construction $\boldsymbol{\delta}_{base}$ lies in the span of the selected subspaces, but the mask and the $\tanh$ of Eq. 7 act element-wise, so the released perturbation is not exactly confined to that span. Table 5 quantifies how closely it adheres. Between $87\%$ and $91\%$ of $\|\boldsymbol{\delta}\|^2$ projects onto the active subspace across all prompts, with a minimum of $0.787$ over the entire test set, against $0.05\%$ for a random perturbation of the same dimension. The geometric constraint therefore holds to within a small residual, and the selectivity reported in Section 5.1 follows from it.

### A.4 Subspace Dimension and Reproducibility

Section 5.4 states that suppression saturates at $k = 8$; Table 6 reports the measurements behind this choice. Suppression rises steeply to $k = 8$ and flattens thereafter, with the remaining twenty-four directions contributing at most $1.7$ points on any attribute while fidelity declines monotonically. Saturation holds for each attribute individually rather than only in the mean, so the trade-off is not an artifact of averaging.

The overlap columns explain why additional directions contribute so little. Estimating two bases from disjoint halves of the training set, the halves recover essentially the same leading direction at $k = 1$ and still agree on 68 to $76\%$ of the subspace at $k = 8$, two orders of magnitude above the chance level. Agreement declines as $k$ grows, indicating that the leading directions are a stable

Table 6: Subspace dimension study. Each row is an independent training run in which only $k$ varies, with all other settings and the random seed held fixed, so values are comparable across rows. Overlap is measured between two bases estimated from *disjoint* halves of the training set; two random $k$-dimensional subspaces of $\mathbb{R}^{49152}$ overlap below $0.002$.

| $k$ | Suppression and fidelity | | | | Subspace overlap | | |
|---|---|---|---|---|---|---|---|
| | Male | Young | Smiling | PSNR | Male | Young | Smiling |
| 1 | – | – | – | – | 0.999 | 0.999 | 0.999 |
| 2 | 81.34 | 90.62 | 65.64 | 35.07 | 0.877 | 0.966 | 0.978 |
| 4 | 85.84 | 93.22 | 69.02 | 34.48 | 0.789 | 0.828 | 0.767 |
| **8** | 88.58 | 94.56 | 71.14 | 34.01 | 0.684 | 0.762 | 0.691 |
| 16 | 89.38 | 95.16 | 71.94 | 33.73 | 0.610 | 0.619 | 0.527 |
| 32 | 90.00 | 95.34 | 72.20 | 33.61 | 0.524 | 0.479 | 0.420 |

property of the classifier while later directions are increasingly determined by the particular sample. The choice $k = 8$ therefore captures the reproducible structure while excluding directions that would not be recovered from different data, which is what makes the subspace a well-posed constraint on where the perturbation may act.

### A.5 Cross-Architecture Subspace Overlap

Table 7: Subspace overlap for the *same* attribute (*Male*, $k = 8$) estimated from six architectures on identical data. Each entry is the mean squared cosine of the principal angles between two bases; two random $8$-dimensional subspaces of $\mathbb{R}^{49152}$ overlap at $0.0002$.

| | ResNet-18 (WB) | ResNet-50 | DenseNet-121 | ViT-B/16 | MobileNetV3 | Swin-T |
|---|---|---|---|---|---|---|
| ResNet-18 (WB) | **1.000** | 0.173 | 0.251 | 0.037 | 0.088 | 0.072 |
| ResNet-50 | 0.173 | **1.000** | 0.281 | 0.073 | 0.055 | 0.073 |
| DenseNet-121 | 0.251 | 0.281 | **1.000** | 0.113 | 0.097 | 0.130 |
| ViT-B/16 | 0.037 | 0.073 | 0.113 | **1.000** | 0.026 | 0.129 |
| MobileNetV3 | 0.088 | 0.055 | 0.097 | 0.026 | **1.000** | 0.027 |
| Swin-T | 0.072 | 0.073 | 0.130 | 0.129 | 0.027 | **1.000** |

Table 7 reports the measurement underlying the transfer analysis of Section 5.3. Every network is trained on the same attribute using the same data, and the bases are extracted with the same $k$ and the same procedure, so the only difference between two subspaces is the architecture whose gradients produced them. The mean off-diagonal overlap is $0.108$. Agreement is highest within the convolutional family, reaching $0.251$ between ResNet-18 and DenseNet-121, whereas the ViT shares $0.037$ of its span with the white-box model and $0.026$ with MobileNetV3. Two design consequences follow. Attribute-sensitive directions are largely specific to the classifier that produced them, so supervision from a single classifier cannot be expected to generalize across architecture families; and averaging additional classifiers supplies further near-orthogonal boundary normals rather than a more general representation of the attribute. Both motivate the text-defined observer of Eq. 10, which specifies the attribute semantically and is therefore not tied to any one decision boundary.

### A.6 Evaluators and the Zero-Shot Attribute Head

Table 8 reports the clean accuracy of every model used to evaluate or supervise attribute prediction. The three held-out foundation evaluators read all attributes well above the majority baseline, so a flip rate measured against them in Table 3 reflects the behavior of the perturbation rather than a weak evaluator. The text-defined head classifies the three attributes at $98.92$, $85.86$, and $88.72\%$ using prompts alone, with its label-free and label-calibrated variants agreeing to within one point on every attribute. This supervision signal is therefore available without attribute annotations of any kind, which is what makes the observer of Eq. 10 applicable to attributes for which no labelled classifier exists.

Table 8: Clean accuracy (%) on CelebA-HQ. The upper block reports the held-out foundation evaluators used in Table 3, as linear probes on frozen features. The lower block reports the text-defined head of Eq. 10 used as a classifier in its own right, with CLIP ViT-B/16, under a label-free threshold at zero and under a label-calibrated threshold.

| | Model | Male | Young | Smiling |
|---|---|---|---|---|
| Held-out evaluators | DINOv2-L | 98.86 | 87.34 | 90.96 |
| | SigLIP-B | 99.10 | 90.12 | 92.38 |
| | CLIP-H | 99.02 | 89.34 | 91.94 |
| Text-defined head | Label-free | **98.92** | **85.86** | **88.72** |
| | Calibrated | 99.22 | 86.68 | 88.64 |
| | *Majority class* | 63.30 | 77.90 | 52.12 |

Table 9: Hyperparameter sweeps, each conducted as a separate set of training runs. The upper block varies the perturbation budget $\varepsilon$; the lower block varies the disentanglement weight $\lambda_{disent}$, where overlap is the normalized cross-attribute subspace overlap. Selected values are in bold.

| | Flip | PSNR | Overlap |
|---|---|---|---|
| *Perturbation budget* $\varepsilon$ | | | |
| 4/255 | 6.24 | 43.49 | – |
| 8/255 | 30.79 | 38.08 | – |
| 12/255 | 77.26 | 35.10 | – |
| **16/255** | **96.23** | **33.12** | – |
| 20/255 | 99.33 | 31.68 | – |
| 24/255 | 99.79 | 30.59 | – |
| *Disentanglement weight* $\lambda_{disent}$ | | | |
| 0.0 | 84.97 | 33.84 | 0.2012 |
| 0.5 | 85.38 | 33.87 | 0.0197 |
| **1.0** | **85.62** | **33.87** | **0.0109** |
| 2.0 | 85.75 | 33.86 | 0.0060 |

### A.7 Hyperparameter Sweeps

Table 9 reports two sweeps. Suppression is governed principally by the perturbation budget, rising from $6.24\%$ at $\varepsilon = 4/255$ to $96.23\%$ at $16/255$ while fidelity declines monotonically, which is why a flip rate reported without its corresponding budget is not interpretable and why we state $\varepsilon$ alongside every result. We adopt $16/255$ as the point at which suppression saturates while fidelity remains high. The disentanglement weight behaves differently: raising $\lambda_{disent}$ from $0$ to $2$ reduces cross-attribute subspace overlap by a factor of roughly thirty-three, from $0.2012$ to $0.0060$, with flip rate and fidelity essentially unchanged. The penalty of Eq. 11 therefore improves subspace separation at no measurable cost, and we retain $\lambda_{disent} = 1$.

### A.8 Extended Robustness Analysis

Table 10 extends the two transformations of Table 4(b) to sixteen operations spanning compression, resampling, blurring, additive noise, and photometric adjustment. The transformations separate into two groups. Additive noise and photometric adjustment leave suppression largely intact, reducing the mean by at most ten points, since neither alters the spatial structure of the perturbation. Operations that resample or low-pass the image are more damaging, as they attenuate the frequency content in which part of the perturbation resides. Suppression nonetheless remains effective throughout, retaining a mean of $41.31\%$ under aggressive JPEG compression and $30.73\%$ under a combined compression-and-blur transformation. The ordering across attributes is consistent with the geome-

Table 10: Flip rate (%) under sixteen image transformations, all-attribute prompt, ordered by mean. Values are measured on protected images after the transformation is applied.

| Transformation | Male | Young | Smiling | Mean |
|---|---|---|---|---|
| None (baseline) | 95.80 | 97.60 | 94.20 | **95.87** |
| Gaussian noise $\sigma$=0.01 | 74.76 | 88.18 | 67.80 | 76.91 |
| Gaussian noise $\sigma$=0.03 | 72.20 | 87.42 | 63.90 | 74.51 |
| Brightness $-20\%$ | 68.60 | 84.10 | 68.26 | 73.65 |
| Brightness $+20\%$ | 64.08 | 82.60 | 54.52 | 67.07 |
| JPEG $Q$=90 | 52.02 | 77.78 | 46.48 | 58.76 |
| Rotation $-5°$ | 52.88 | 77.20 | 42.54 | 57.54 |
| Rotation $+5°$ | 53.42 | 70.12 | 46.80 | 56.78 |
| Gaussian blur $s$=3 | 58.24 | 43.86 | 44.00 | 48.70 |
| JPEG $Q$=70 | 39.92 | 68.48 | 35.80 | 48.07 |
| Resize 32 $\rightarrow$ 128 | 57.62 | 52.90 | 31.90 | 47.47 |
| Gaussian blur $s$=1 | 49.18 | 53.58 | 29.30 | 44.02 |
| Resize 64 $\rightarrow$ 128 | 47.44 | 53.36 | 27.30 | 42.70 |
| Gaussian blur $s$=2 | 53.24 | 44.30 | 30.16 | 42.57 |
| JPEG $Q$=50 | 32.16 | 62.50 | 29.28 | 41.31 |
| Combined (JPEG $Q$=70 + blur $s$=1) | 28.22 | 41.92 | 22.06 | 30.73 |

try of the learned subspaces: *Young* is the most resilient to compression but the most affected by blurring, indicating that the three attribute subspaces occupy distinct frequency bands.

### A.9 Statistical Significance Analysis

Table 11: Paired significance analysis (McNemar's test) of GAUSS against PrivAD on the CelebA-HQ test set ($n = 5{,}000$), with both methods evaluated on the same images and the same classifiers under the all-attribute prompt. $n_{01}$ counts images GAUSS suppresses and PrivAD does not; $n_{10}$ the reverse.

| Attribute | $n_{01}$ | $n_{10}$ | $\chi^2$ | $p$-value | Cohen's $h$ |
|---|---|---|---|---|---|
| Male | 1448 | 653 | 300.07 | $3.2 \times 10^{-67}$ | 0.78 |
| Young | 3575 | 157 | 3128.59 | $< 10^{-300}$ | 2.32 |
| Smiling | 1620 | 496 | 596.00 | $1.2 \times 10^{-131}$ | 1.12 |

Both methods are evaluated on the same 5,000 test images, so their per-image outcomes are paired rather than independent, and the two-proportion $Z$-test commonly used in this literature does not apply. Conditioning on the discordant pairs, McNemar's statistic is

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}, \tag{12}$$

distributed as $\chi^2$ with one degree of freedom, the $-1$ being the continuity correction. Table 11 reports the result. All three differences are significant far below the conventional threshold and the effect sizes are large throughout. The widest margin is on *Young*, where GAUSS succeeds on $3{,}575$ images that PrivAD does not while the reverse holds for only $157$, a ratio of more than twenty to one.

### A.10 In-the-Wild Transfer

The evaluation in the main paper uses aligned CelebA-HQ crops. Table 12 applies the same model, without fine-tuning or realignment, to two datasets of unaligned in-the-wild faces. Identity preservation transfers unchanged: on 400 LFW identities, verification AUC moves from $0.9954$ to $0.9944$ at $36.22$ dB. Suppression transfers as well, with *Male* suppressed in $86.45\%$ of 2,000 FairFace images at $42.11$ dB. Both properties therefore hold outside the alignment and capture conditions of the training distribution, without any adaptation of the learned subspaces.

Table 12: Transfer of the CelebA-HQ-trained model to unaligned in-the-wild faces, applied without fine-tuning or realignment. LFW measures identity preservation over 400 identities; FairFace measures *Male* suppression over 2,000 images against the classifier's own clean prediction.

| Dataset | Setting | AUC / Flip | PSNR |
|---|---|---|---|
| LFW (identity) | Unprotected | 0.9954 | – |
| | Protected | 0.9944 | 36.22 |
| FairFace (*Male*) | Protected | 86.45 | 42.11 |

Table 13: Computational cost, measured on a single NVIDIA H200 NVL. Trainable parameters count the protection model only; the three frozen attribute classifiers (33.5M) are the white-box target and are not trained.

| | PrivAD | GAUSS (OURS) |
|---|---|---|
| Trainable parameters (M) | 97.98 | **1.38** |
| Stored protection state (MB) | 749.8 | **4.5** |
| Latency, batch 1 (ms) | 24.86 | **9.33** |
| Latency, batch 32 (ms/image) | 0.792 | **0.322** |
| Peak GPU memory, batch 32 (MB) | 2021 | **738** |

### A.11 Computational Cost

Table 13 details the cost summarized in Section 5.4. The difference follows directly from the formulation: GAUSS learns a set of low-dimensional bases and two small networks, whereas PrivAD learns a generative model of faces. The gap is largest in stored state, at $167\times$, since the shipped protection model comprises only the bases, the coefficient network, the mask network, and the attribute weights. Latency is lower at both batch sizes and peak memory is reduced by a factor of nearly three, so protection can be applied on modest hardware.

### A.12 Extended Qualitative Comparison

Figure 4 extends the comparison of Figure 3 to all baselines. Several introduce visible distortions, including changes in facial structure and texture artifacts. PrivAD alters facial appearance to suppress attributes, for instance by adding a beard to flip the *Young* attribute, which produces identity inconsistencies, while SFE and GenderAN exhibit ghosting artifacts and misalignment of facial landmarks. The outputs of GAUSS remain visually consistent with the originals while suppressing the target attributes, which is the behavior the subspace constraint is designed to produce.

### A.13 Broader Impact

This work enables selective suppression of sensitive soft biometric attributes such as age, gender, and expression while preserving the utility of the remaining attributes. Such mechanisms can reduce unauthorized demographic profiling, targeted advertising, and attribute-based discrimination in surveillance, authentication, and social media pipelines, and offer individuals a degree of control over facial data collected without consent, while preserving legitimate applications such as authentication and verification. Two caveats apply. First, as stated in the threat model of Section 3.1, our evaluation measures whether pre-trained attribute inference models are misled, and deployments should state which adversaries they defend against rather than presenting such methods as guarantees. Second, as with most privacy-enhancing technologies, the same mechanism could be used to evade legitimate attribute-based systems; we judge this risk small relative to the benefit of giving individuals control over their own facial data, but it is not zero.

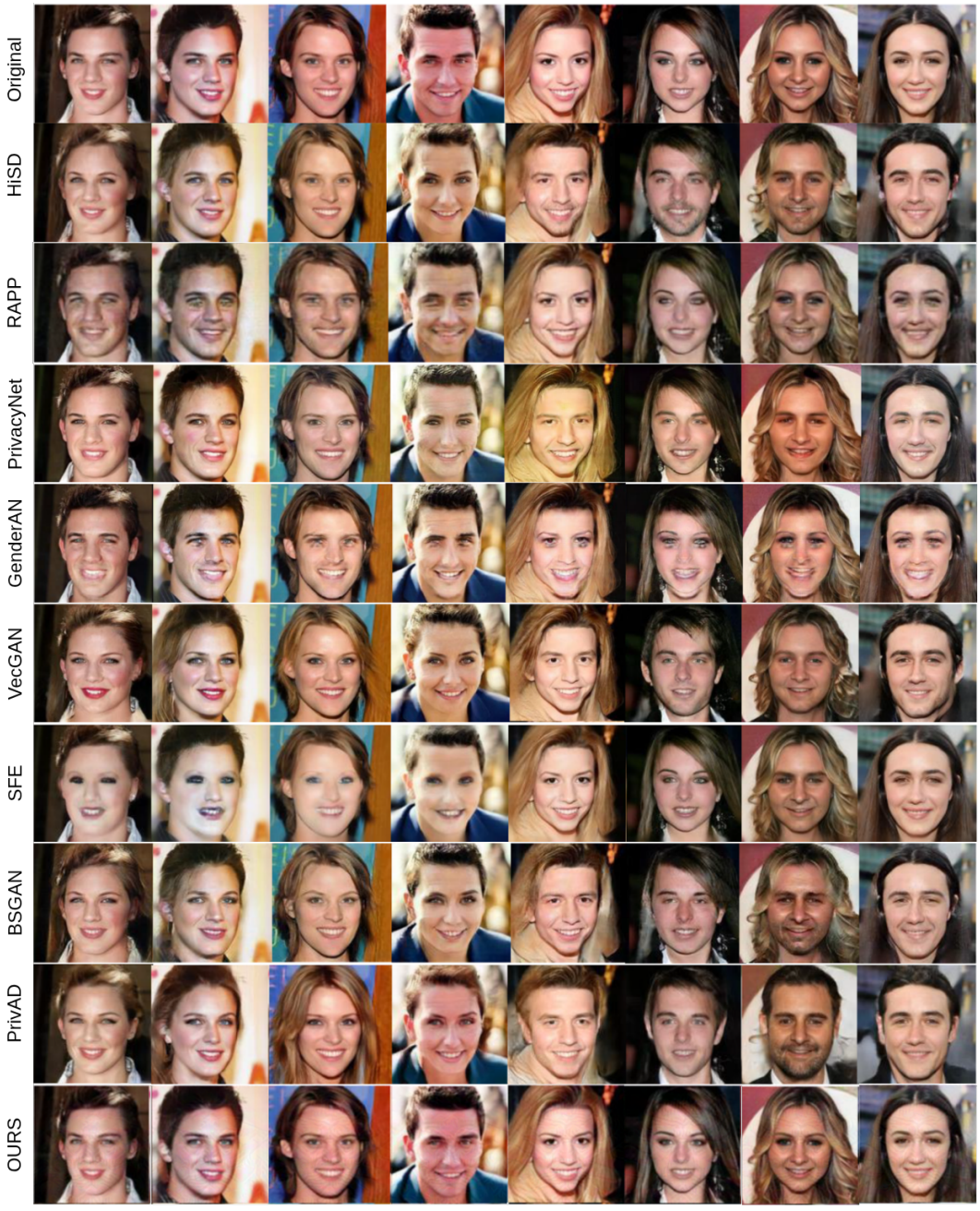


Figure 4: Qualitative comparison across baseline methods. The first row shows the original images and the last row the outputs of GAUSS; intermediate rows correspond to different baselines under the same attribute suppression setting.