GGSS: Geodesic-Gated Spherical Steering for Inference-Time Debiasing of Generative Vision–Language Models

Yiqun Sun[1], Junyu Chen[2], Pengfei Wei[1]*, Lawrence B. Hsieh[1]
[1]Magellan Technology Research Institute (MTRI), [2]National University of Singapore
{duke.sun, pengfei.wei, lawrence.hsieh}@mtri.co.jp, chenjunyu@u.nus.edu

Abstract

Generative vision–language models (VLMs) are increasingly used in human-centered settings, yet they can produce demographically biased outputs even when images differ only in controlled attributes such as perceived race or gender. However, existing inference-time debiasers were largely designed for static embeddings or CLIP-like models rather than generative VLMs. We propose **GGSS**—**G**eodesic-**G**ated **S**pherical **S**teering—a norm-preserving intervention that discovers a counterfactual bias subspace on the unit hypersphere, steers visual tokens along geodesic arcs, and uses an adaptive gate to focus correction on tokens that carry stronger demographic signal. We evaluate four generative VLMs against ten adapted inference-time debiasing baselines and prompt-based mitigation under a single operating-point protocol across categorical, pairwise, and occupation-gender bias tests, while also measuring general visual-language capability. **GGSS** achieves the lowest average bias on all four models, significant on three of four backbones under paired permutation tests, while preserving MMStar accuracy within ± 0.6 p.p. of the unsteered baseline. Code is available at https://github.com/dukesun99/GGSS.

1 Introduction

Vision–language models (VLMs) now appear in human-centered, and sometimes high-stakes, decision-making settings, including recruitment and decision support (Liu et al., 2023; Wu et al., 2024; Banerjee et al., 2024). As they become more capable and more widely deployed, a growing body of work has documented that modern VLMs produce systematically different outputs across demographic groups even when the underlying visual content is controlled (Chen et al., 2026; Howard et al., 2024; Huang et al., 2025). Debiasing such

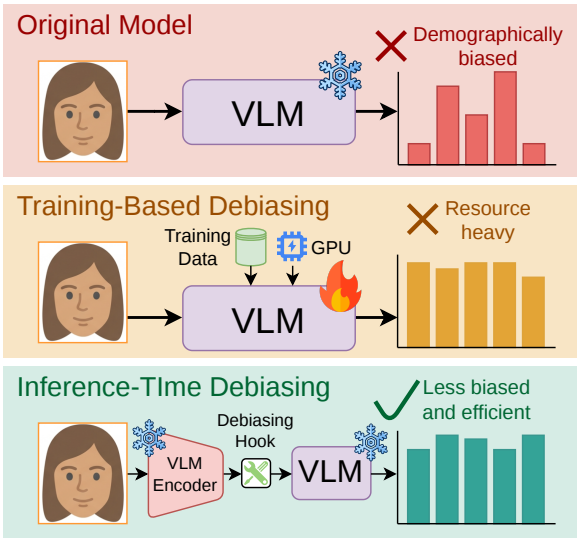

Figure 1: **Motivation for inference-time debiasing.** Instead of retraining a VLM with additional data and heavy GPU resources, GGSS inserts a lightweight debiasing hook during inference to reduce demographic bias efficiently while keeping the original model frozen.

models at *inference time* (Figure 1) is attractive because it avoids retraining, applies to frozen checkpoints, and can be reconfigured as the target notion of bias changes (Chuang et al., 2023; Gerych et al., 2024, 2026). However, many existing inference-time debiasing methods were developed for CLIP-like (Radford et al., 2021) settings where an input is summarized by a single global embedding. INLP, R-LACE, and LEACE erase concepts through linear subspace operations (Ravfogel et al., 2020, 2022; Belrose et al., 2023). Classifier-guided methods apply mean-shift corrections. BendVLM performs retrieval-based equalization in CLIP-style embedding spaces (Gerych et al., 2024; Radford et al., 2021). Modern *generative* VLMs (Bai et al., 2025; Liu et al., 2024) instead encode an image as many visual tokens that are fused with a decoder-only language model. Reusing CLIP-era methods in this setting can be overly destructive and, as our experiments show, can even be dominated by the

* Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 5875–5894
October 24-29, 2026 ©2026 Association for Computational Linguistics

unsteered baseline.

We identify three concrete challenges that a successful inference-time debiaser for generative VLMs must address. First, **geometry preservation**: visual activations in recent VLMs enter the language model through a small set of projection layers whose outputs are later consumed by attention and decoding. Euclidean subtraction of a bias direction distorts both directions and norms; empirically this matters increasingly with model scale, causing sharp drops in general quality and over-steering failure modes that on-sphere steering avoids (§4.3). Second, **token-level bias heterogeneity**: demographic signal in a generative VLM is concentrated in a small fraction of visual tokens (e.g., face or clothing tokens) while many tokens encode pose, background, or textual cues and carry essentially none. Hard subspace projection applied uniformly over-corrects the latter and silently degrades task performance. Third, **single-direction rigidity**: many existing methods assume a binary protected attribute and a single "bias direction," but race-like attributes in practice have multi-class structure and must be handled as a low-dimensional subspace (Manzini et al., 2019).

We propose **GGSS** (**G**eodesic-**G**ated **S**pherical **S**teering), an inference-time debiasing framework that matches these three challenges with three coordinated design choices. First, we discover a *bias subspace* $\mathbf{V}_{\text{bias}}$ from counterfactual image sets that vary demographic attributes while holding other visual factors as fixed as possible. We collect pooled visual activations and apply singular value decomposition to tangent-space shifts around per-group Fréchet means on the unit hypersphere, yielding a multi-dimensional and norm-aware estimate of demographic variation. Second, at inference time we replace hard null-space projection with *spherical linear interpolation* (Shoemake, 1985) between the original token direction and a debiased target, producing a smooth geodesic rotation that preserves norms exactly. Third, we add an *adaptive confidence gate* g_i per token, calibrated from the distribution of bias-projection magnitudes observed during discovery. This gate applies little or no steering to low-bias tokens and stronger steering to tokens with unusually large bias components, which makes the intervention more selective than a uniform projection.

We evaluate **GGSS** on four generative VLMs against ten adapted inference-time debiasing baselines, a prompt-based mitigation family, and four structural ablations. The evaluation covers race- and gender-related bias tests together with MMStar capability preservation, and all methods are compared under a fair protocol, with bootstrap confidence intervals and paired permutation tests for every reported comparison. Experiments show **GGSS** attains the lowest average bias on all four models, with statistically significant reductions on three of the four backbones and per-task reductions up to 96%, 84%, and 61%, while keeping MMStar accuracy within ± 0.6 p.p. of baseline. These results support our central claim: for generative VLMs, *adaptive* geodesic steering is a better primitive for inference-time bias mitigation than hard, uniform projection. Matched-gate ablations attribute robustness at scale to on-sphere steering and selectivity to the gate.

Our contributions are fourfold:

- We identify why hard, token-uniform, single-direction debiasing operators under-perform on generative VLMs, and propose **GGSS** as a norm-preserving geodesic alternative.
- We introduce an adaptive gate that uses the discovery-set bias-norm distribution to assign per-token steering strengths without additional training.
- We port representative CLIP- and LM-era debiasers to the same generative-VLM intervention layer, producing a unified comparison suite.
- We evaluate four VLMs under one best-avg-α protocol; **GGSS** achieves the lowest average bias on all four while preserving MMStar capability.

2 Related Work

Social biases in vision–language models. Prior work shows that VLMs encode and express social biases across representations, benchmarks, and downstream behaviors. Grounded vision-language embeddings inherit demographic associations from earlier representation models (Ross et al., 2021; Srinivasan and Bisk, 2022), and later studies measure disparities across gender, race, age, nationality, religion, and other attributes (Zhou et al., 2022; Janghorbani and De Melo, 2023; Ruggeri and Nozza, 2023). These biases also surface in captioning, visual question answering, retrieval, open-ended prompting, counterfactual evaluation, and real-image settings (Hirota et al., 2022; Hall et al., 2023; Ghate et al., 2025; Howard et al., 2024; Wang et al., 2024; Huang et al., 2025; Narnaware

et al., 2025). This motivates inference-time mitigation methods that can be applied to frozen models without retraining.

VLM debiasing: training-based vs. inference-time. Most VLM debiasing methods target CLIP-style dual encoders, not generative VLMs. Training-based approaches learn lightweight transformations, pruning or imputation modules, causal adjustments, bias-corpus disentanglement, LLM-guided projections, or joint image–text corrections (Seth et al., 2023; Jung et al., 2024; Pang et al., 2025; Jang et al., 2025; Molahasani et al., 2025; Zhang et al., 2025). Inference-time methods instead modify prompts or embedding geometry at test time, including calibrated prompt-defined projection, query-specific local debiasing, and rotation-based correction (Chuang et al., 2023; Gerych et al., 2024, 2026). Post-hoc transformations of embedding geometry also serve instruction following (Feng et al., 2025) and concept suppression in text-to-image generation (Chen et al., 2025; Li et al., 2025). Debiasing for generative VLMs is less explored, with recent work moving toward internal activation intervention and monitoring (Ratzlaff et al., 2024; An et al., 2026; Huben et al., 2024; Cheng et al., 2025; Singh et al., 2026; Wang et al., 2026). **GGSS** follows this activation-level direction, but focuses on generative VLM on token activations.

Concept erasure and activation steering. Our baselines draw on linear concept erasure and inference-time activation steering. INLP, R-LACE, and LEACE remove protected information by learning or solving for linear subspace projections (Ravfogel et al., 2020, 2022; Belrose et al., 2023). These classical erasers are useful reference points, but they were designed for static embeddings or encoder representations and typically act as hard, token-uniform projections. Activation steering methods such as ActAdd, Representation Engineering, and Inference-Time Intervention instead manipulate hidden states directly at test time (Turner et al., 2023; Zou et al., 2023; Li et al., 2023). **GGSS** shares the inference-time spirit of this line, but replaces additive Euclidean steering with spherical geometry. The unit hypersphere and spherical interpolation are standard tools in directional statistics and computer graphics (Mardia and Jupp, 2000; Pennec, 2006; Shoemake, 1985), and related geometric views of concept directions have appeared in representation analysis and null-space steering (Park et al., 2023; Sheng et al., 2026; Sun et al., 2025a). We instantiate these ideas for image-side activations in generative VLMs and add a bias-norm-calibrated gate that makes steering token-adaptive.

3 The GGSS Framework

We propose **GGSS**—**G**eodesic-**G**ated **S**pherical **S**teering—a two-stage framework for debiasing generative VLMs, summarized in Figure 2. Offline discovery passes counterfactual images through the frozen VLM to estimate the steering quantities: a protected-attribute basis, a global spherical reference point $\boldsymbol{\mu}$, and gate calibration statistics $(\tilde{b}, \sigma_b)$. Online inference is separate: a new test input produces an activation tensor, and **GGSS** steers its visual tokens using the discovered quantities. The learned quantities are reused for new inputs at inference time.

Notation. Let

$$\mathbb{S}^{D-1} = \{\mathbf{x} \in \mathbb{R}^D : \|\mathbf{x}\|_2 = 1\}$$

be the unit sphere in $\mathbb{R}^D$, and let $\nu(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ be the normalization operator. For $\mathbf{p}, \mathbf{q} \in \mathbb{S}^{D-1}$, define $d(\mathbf{p}, \mathbf{q}) = \arccos\langle \mathbf{p}, \mathbf{q}\rangle$ and

$$\log_{\mathbf{p}}(\mathbf{q}) = d(\mathbf{p}, \mathbf{q}) \frac{\mathbf{q} - \langle \mathbf{p}, \mathbf{q}\rangle \mathbf{p}}{\|\mathbf{q} - \langle \mathbf{p}, \mathbf{q}\rangle \mathbf{p}\|_2},$$
$$\exp_{\mathbf{p}}(\mathbf{t}) = \cos(\|\mathbf{t}\|_2)\mathbf{p} + \sin(\|\mathbf{t}\|_2) \frac{\mathbf{t}}{\|\mathbf{t}\|_2},$$

with the standard conventions $\log_{\mathbf{p}}(\mathbf{p}) = \mathbf{0}$ and $\exp_{\mathbf{p}}(\mathbf{0}) = \mathbf{p}$. For $\mathbf{p}, \mathbf{q} \in \mathbb{S}^{D-1}$ with $\mathbf{p} \neq \mathbf{q}$ and $\mathbf{q} \neq -\mathbf{p}$, define spherical linear interpolation by

$$\mathrm{Slerp}(\mathbf{p}, \mathbf{q}; \beta) = \frac{\sin((1-\beta)\theta)}{\sin\theta}\mathbf{p} + \frac{\sin(\beta\theta)}{\sin\theta}\mathbf{q},$$
$$\theta = \arccos\langle \mathbf{p}, \mathbf{q}\rangle.$$

When $\mathbf{p} = \mathbf{q}$, we use the continuous extension $\mathrm{Slerp}(\mathbf{p}, \mathbf{p}; \beta) = \mathbf{p}$. For $0 \leq \beta \leq 1$, Slerp interpolates between $\mathbf{p}$ and $\mathbf{q}$; for $\beta > 1$, it continues along the same geodesic beyond the target.

The rest of this section follows the two stages of GGSS. Section 3.1 presents the offline discovery stage, where counterfactual activations are used to estimate a protected-attribute subspace, a global reference point, and gate calibration statistics. Section 3.2 presents the inference-time steering stage, where each visual token is projected away from the learned protected-attribute subspace, updated along the sphere, and rescaled to its original norm.

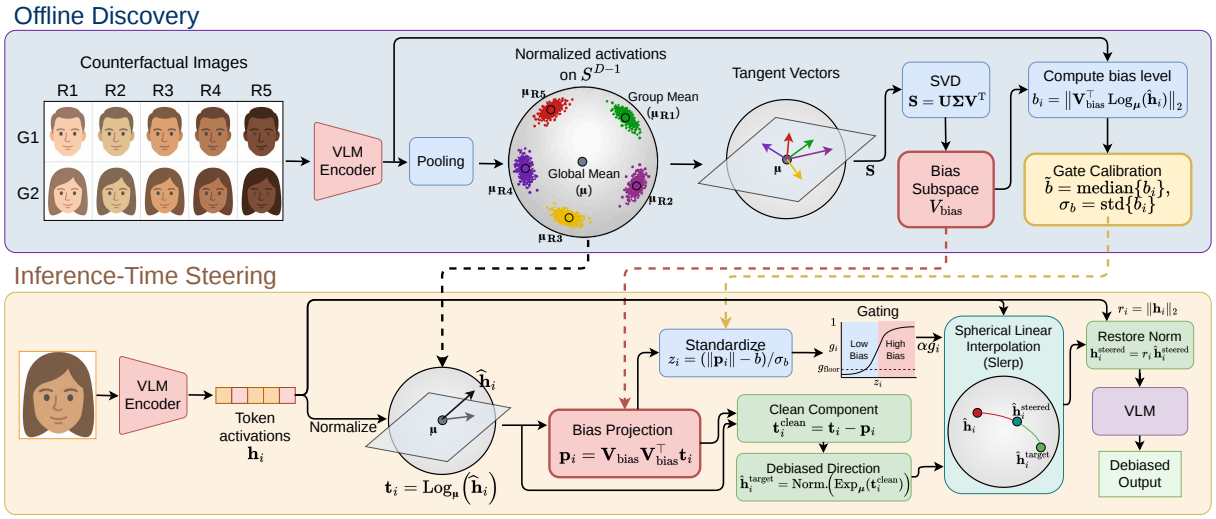

Figure 2: **Overview of the GGSS framework.** The offline stage estimates a spherical bias subspace from counterfactual image activations and calibrates a bias-sensitive token gate. At inference time, each visual token is mapped to tangent space, decomposed into bias and clean components, gated according to its bias magnitude, geodesically rotated toward a debiased direction via Slerp, and finally norm-restored before being passed to the LLM decoder of the VLM.

3.1 Counterfactual Bias Subspace Discovery

Counterfactual representatives. We write the discovery images as

$$\mathbf{x}_{u,a}, \qquad u \in \mathcal{U}, \quad a \in \mathcal{A}.$$

Here a indexes the protected attribute value, such as a race category, while u collects the fixed factors, such as identity, occupation, gender, pose, clothing, background, and lighting. This abstracts the concrete occupation/identity/race/gender indexing specified in Appendix A.1. Running the frozen model on $\mathbf{x}_{u,a}$ with a fixed *discovery prompt*, the text instruction paired with each discovery image ("*Describe this image in detail.*"), gives the target-layer activation

$$\mathbf{H}_{u,a} \in \mathbb{R}^{T \times D}.$$

The choice of discovery prompt is immaterial at our intervention layer: the vision-to-language projection computes its visual-token activations before any interaction with the text prompt, so the prompt architecturally cannot influence them, and re-running discovery with four alternative prompts yields subspaces identical to numerical precision (principal-angle cosines ≥ 0.9997) and identical end-to-end results (Appendix B.13). We summarize each image by the normalized pooled representative

$$\mathbf{y}_{u,a} = \nu\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_{u,a,t}\right) \in \mathbb{S}^{D-1}.$$

Within a fixed context u, varying a is designed to emphasize protected-attribute variation while keeping other visual factors fixed. Pooling yields one stable spherical representative per counterfactual image.

Subspace estimation. For each fixed context u, let the spherical Fréchet mean be

$$\mathbf{c}_u \in \arg\min_{\mathbf{z} \in \mathbb{S}^{D-1}} \sum_{a \in \mathcal{A}} d(\mathbf{z}, \mathbf{y}_{u,a})^2.$$

The protected-attribute shift for value a is

$$\mathbf{s}_{u,a} = \log_{\mathbf{c}_u}(\mathbf{y}_{u,a}).$$

Stacking the shifts $\{\mathbf{s}_{u,a} : u \in \mathcal{U}, a \in \mathcal{A}\}$ as rows gives the shift matrix

$$\mathbf{S} \in \mathbb{R}^{N_{\text{shifts}} \times D}, \qquad N_{\text{shifts}} = |\mathcal{U}||\mathcal{A}|.$$

Each row of $\mathbf{S}$ is one tangent shift induced by changing the protected attribute within a fixed context. We then apply SVD:

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \qquad \mathbf{V}_{\text{bias}} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{D \times k}. \tag{1}$$

Here $\boldsymbol{v}_j$ is the j-th column of $\mathbf{V}$. Thus $\mathbf{V}_{\text{bias}}$ is the matrix of the top-k right singular vectors of $\mathbf{S}$, and $\mathrm{span}(\mathbf{V}_{\text{bias}})$ is the learned protected-attribute subspace. Since the columns of $\mathbf{V}$ are orthonormal, $\mathbf{V}_{\text{bias}}^{\top}\mathbf{V}_{\text{bias}} = I_k$. We set $k = |\mathcal{A}| - 1$ by default. The global reference point is the Fréchet mean of all discovery representatives:

$$\boldsymbol{\mu} \in \arg\min_{\mathbf{z} \in \mathbb{S}^{D-1}} \sum_{u \in \mathcal{U}} \sum_{a \in \mathcal{A}} d(\mathbf{z}, \mathbf{y}_{u,a})^2.$$

Together, $\mathbf{V}_{\text{bias}}$ and $\boldsymbol{\mu}$ parameterize the steering geometry. Unlike classifier-based discovery methods, the subspace is estimated directly from counterfactual activation differences and requires no supervised probe.

Gate calibration statistics. For each discovery representative $\mathbf{c} \in \{\mathbf{y}_{u,a} : u \in \mathcal{U}, a \in \mathcal{A}\}$, compute

$$b(\mathbf{c}) = \|\mathbf{V}_{\text{bias}}^{\top} \log_{\boldsymbol{\mu}}(\mathbf{c})\|_2,$$

where $\mathbf{V}_{\text{bias}}^{\top} \log_{\boldsymbol{\mu}}(\mathbf{c})$ gives the coordinates of $\mathbf{c}$, after mapping to the reference point $\boldsymbol{\mu}$, along the learned protected-attribute basis. Thus $b(\mathbf{c})$ is the protected-coordinate magnitude of the discovery representative. We store

$$\tilde{b} = \text{median}\{b(\mathbf{c})\}, \qquad \sigma_b = \text{std}\{b(\mathbf{c})\}.$$

These scalars calibrate the token gate at inference time.

3.2 Token-Level Geodesic Steering

At inference time, the new input produces a target-layer activation tensor

$$\mathbf{Z} \in \mathbb{R}^{n_{\text{batch}} \times T' \times D}.$$

We flatten it to token vectors $\{\mathbf{h}_i\}_{i=1}^{N}$, with $N = n_{\text{batch}} T'$, and process each nonzero token independently. Let

$$r_i = \|\mathbf{h}_i\|_2, \qquad \widehat{\mathbf{h}}_i = \nu(\mathbf{h}_i), \qquad \mathbf{t}_i = \log_{\boldsymbol{\mu}}(\widehat{\mathbf{h}}_i).$$

The radius r_i is kept outside the debiasing step and restored at the end.

Projection and target direction. Because $\mathbf{V}_{\text{bias}}$ has orthonormal columns, $\mathbf{V}_{\text{bias}}\mathbf{V}_{\text{bias}}^{\top}$ is the orthogonal projector onto $\text{span}(\mathbf{V}_{\text{bias}})$. The protected-coordinate component and its orthogonal complement are

$$\begin{aligned} \mathbf{p}_i &= \mathbf{V}_{\text{bias}}\mathbf{V}_{\text{bias}}^{\top}\mathbf{t}_i, \\ \mathbf{t}_i^{\text{clean}} &= (I - \mathbf{V}_{\text{bias}}\mathbf{V}_{\text{bias}}^{\top})\mathbf{t}_i. \end{aligned} \tag{2}$$

Here $\mathbf{p}_i$ is the component removed by **GGSS**, while $\mathbf{t}_i^{\text{clean}}$ is the steering coordinate after zeroing the learned protected-attribute coordinates. Mapping the projected tangent vector back to the sphere gives the target direction

$$\widehat{\mathbf{h}}_i^{\text{target}} = \exp_{\boldsymbol{\mu}}(\mathbf{t}_i^{\text{clean}}).$$

This defines the target direction by first projecting in the steering coordinates and then mapping the projected coordinate back to the sphere.

Calibrated token gate. The token-specific gate is

$$\begin{aligned} z_i &= \frac{\|\mathbf{V}_{\text{bias}}^{\top}\mathbf{t}_i\|_2 - \tilde{b}}{\sigma_b + \varepsilon}, \\ g_i &= g_{\text{floor}} + (1 - g_{\text{floor}})\,\text{sigmoid}(\kappa z_i), \end{aligned} \tag{3}$$

where $\varepsilon > 0$ avoids division by zero, κ controls gate sharpness, and g_{floor} sets the minimum steering strength. Since $\mathbf{V}_{\text{bias}}$ has orthonormal columns, $\|\mathbf{V}_{\text{bias}}^{\top}\mathbf{t}_i\|_2 = \|\mathbf{p}_i\|_2$. Thus the gate compares the token's protected-coordinate magnitude with the discovery statistics $(\tilde{b}, \sigma_b)$. Tokens with larger protected-coordinate magnitude receive stronger steering, while low-bias tokens stay closer to their original directions.

Geodesic update. Set $\beta_i = \alpha g_i$ and update the token direction by

$$\widehat{\mathbf{h}}_i^{\text{steered}} = \text{Slerp}(\widehat{\mathbf{h}}_i, \widehat{\mathbf{h}}_i^{\text{target}}; \beta_i), \tag{4}$$

then restore the original radius:

$$\mathbf{h}_i^{\text{steered}} = r_i \widehat{\mathbf{h}}_i^{\text{steered}}. \tag{5}$$

When $\beta_i = 0$ the direction is unchanged, while $\beta_i = 1$ reaches the projected target. Larger values continue along the same geodesic, allowing the steering strength α to control how aggressively the protected-coordinate component is reduced.

The following proposition records the mechanical guarantees of the update: GGSS removes the learned protected-attribute coordinates by an orthogonal projection in the steering coordinates, and the final spherical update preserves the original token norm exactly. Whether the steering reduces bias is an empirical question, addressed in §4.

Proposition 1. *For the basis $\mathbf{V}_{\text{bias}}$ defined in Eq. (1) and the token update defined in Eqs. (2)–(5), the projected coordinate $\mathbf{t}_i^{\text{clean}}$ is the Euclidean projection of $\mathbf{t}_i$ onto the set of vectors whose protected-attribute coordinates are zero:*

$$\mathbf{t}_i^{\text{clean}} = \arg\min_{\mathbf{z} \in \mathbb{R}^D} \left\{ \|\mathbf{z} - \mathbf{t}_i\|_2^2 : \mathbf{V}_{\text{bias}}^{\top}\mathbf{z} = \mathbf{0} \right\}.$$

Moreover, if $\widehat{\mathbf{h}}_i^{\text{target}} \neq -\widehat{\mathbf{h}}_i$, with the convention $\text{Slerp}(\mathbf{p}, \mathbf{p}; \beta) = \mathbf{p}$, then

$$\|\mathbf{h}_i^{\text{steered}}\|_2 = \|\mathbf{h}_i\|_2.$$

The proof is given in Appendix A.3.

The proposition shows that GGSS removes the learned protected-attribute coordinates by an orthogonal projection and preserves the token norm after the spherical update.

The steered tokens are reshaped back into the activation tensor and passed to the remaining layers of the frozen VLM. Because the gate is token-dependent, low-bias tokens are changed less, while tokens with larger protected-coordinate magnitude are steered more strongly.

4 Experiments

4.1 Setup

Models. We evaluate **GGSS** on four generative VLMs spanning three model families and a range of parameter scales: Pixtral-12B (12B) (Agrawal et al., 2024), LLaVA-1.6-Vicuna-7B (7B), LLaVA-1.6-Mistral-7B (7B) (Liu et al., 2024), and Qwen3-VL-4B-Instruct (4B) (Bai et al., 2025). For every method, we intervene at the same late vision-to-language projection layer so that differences reflect the steering rule rather than the hook location. Checkpoint identifiers, exact hook paths, activations, and results for additional hook sites appear in Appendix B.3.

Counterfactual images. The counterfactual image sets are not generated by us: we use the 480 real-photograph, face-only counterfactual images released with the REFLECT/FOCUS dataset (Chen et al., 2026), covering 6 occupations $\times$ 8 source identities $\times$ 5 perceived races $\times$ 2 genders. Each identity's variants are pixel-aligned, so activation differences isolate the protected attribute; discovery occupations are disjoint from the evaluated occupation (*Protocol* below).

Tasks and metrics. We evaluate with three bias protocols on categorical and pairwise attributes. For each counterfactual image, the probes ask:

- **MCQ** (race, multiple choice; REFLECT (Chen et al., 2026)): a salary/education multiple-choice question about the pictured person, scored by the Jensen–Shannon divergence of answer distributions across races. The metric is `mean_race_jsd`$\times 10^3$.
- **2AFC** (race, pairwise; REFLECT (Chen et al., 2026)): a two-alternative forced choice ("which of two versions of the same person has higher income") on race-paired images. The metric is `race_bias_std`.
- **Nurse/Doctor** (gender, classification): a single-word occupation query on nurse/doctor images. The metric is the gender gap in correct classification, $|P(\text{nurse}|\text{man}) - P(\text{nurse}|\text{woman})|$.

Two of the four models (LLaVA-1.6-Vicuna-7B and LLaVA-1.6-Mistral-7B) exhibit zero baseline 2AFC bias. Steered values are reported for completeness but do not contribute to the average. We also report accuracy on MMStar (Chen et al., 2024), a 1,500-question multiple-choice benchmark of general multimodal capability spanning six dimensions from coarse/fine-grained perception to reasoning, math, and science, scored by exact matching with VLMEvalKit (Duan et al., 2024); per-dimension breakdown in Appendix B.6.

Baselines. This work *re-implements and adapts* existing inference-time debiasers from CLIP- and LM-era settings to generative VLMs. These methods were originally developed for static word embeddings, BERT-style encoders, or CLIP dual-encoders; here, we make them target the same late vision-to-language projection layer for a controlled comparison. Concretely, we compare **GGSS** against ten external steering baselines spanning four families:

- **INLP** (Ravfogel et al., 2020): iterative linear-probe null-space stacking, ported from word-embedding debiasing, with Euclidean and spherical variants.
- **MeanDiff**: classifier-guided mean-shift correction in the style of Zou et al. (2023) probing. We use both RBF-SVM and multinomial logistic-regression probes, each with Euclidean and spherical variants trained on pooled visual activations.
- **BendVLM** (Gerych et al., 2024): we adapt BendVLM's CLIP-embedding retrieval and Lagrangian equalization from the shared text–image embedding space to generative-VLM hidden activations, evaluating both Euclidean and spherical variants.
- **LEACE** (Belrose et al., 2023): the closed-form least-squares concept eraser, fit with the authors' official implementation, as a direct spherical port and combined with our calibrated gate (Appendix A.5).

We additionally evaluate a *prompt-space* family: the identical probes with a fairness instruction appended to every prompt, with no steering (§4.2). All adapted implementations are released together with this paper as a benchmark suite for inference-time debiasing of generative VLMs.

Protocol. For every (method, model) pair we sweep $\alpha \in \{0.25, 0.5, 0.75, 1.0, 1.5\}$ across the bias tasks and pick a *single* α that minimizes the unweighted mean per-task % change relative to the unsteered baseline ("best-avg-α" protocol). Only

Method	Pixtral-12B				LLaVA-1.6-Vicuna-7B				LLaVA-1.6-Mistral-7B				Qwen3-VL-4B			
	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$
Baseline	25.75	0.243	0.425	0%	8.70	0.000	0.600	0%	26.87	0.000	0.625	0%	14.65	0.368	0.400	0%
INLP (Ravfogel et al., 2020)																
INLP (Eucl.)	15.05 −42%	0.131 −46%	0.350 −18%	−35%	1.87 −79%	0.000 –	0.625 +4%	−37%	10.52 −61%	0.000 –	0.375 −40%	−50%	5.71 −61%	0.279 −24%	0.375 −6%	−31%
INLP (sph.)	**0.47** −98%	0.297 +22%	0.350 −18%	−31%	2.05 −76%	0.000 –	**0.025** −96%	−86%	9.87 −63%	0.000 –	**0.050** −92%	−78%	11.00 −25%	0.238 −35%	**0.050** −88%	−49%
MeanDiff																
MeanDiff-SVM (Eucl.)	25.75 +0%	0.247 +2%	0.425 +0%	+1%	0.86 −90%	0.000 –	0.600 +0%	−45%	11.05 −59%	0.000 –	0.625 +0%	−29%	5.24 −64%	0.368 +0%	0.375 −6%	−23%
MeanDiff-SVM (sph.)	17.73 −31%	0.248 +2%	0.425 +0%	−10%	2.76 −68%	0.000 –	0.625 +4%	−32%	13.67 −49%	0.000 –	0.625 +0%	−25%	14.02 −4%	0.412 +12%	0.375 −6%	+0%
MeanDiff-LR (Eucl.)	25.75 +0%	0.247 +2%	0.425 +0%	+1%	3.09 −64%	0.000 –	0.600 +0%	−32%	12.14 −55%	0.000 –	0.625 +0%	−27%	5.77 −61%	0.368 +0%	0.375 −6%	−22%
MeanDiff-LR (sph.)	25.75 +0%	0.239 −2%	0.425 +0%	−1%	6.15 −29%	0.000 –	0.625 +4%	−13%	21.21 −21%	0.000 –	0.625 +0%	−11%	13.56 −7%	0.400 +9%	0.400 +0%	+0%
BendVLM (Gerych et al., 2024)																
BendVLM (Eucl.)	13.03 −49%	0.267 +10%	0.125 −71%	−37%	2.71 −69%	0.000 –	0.625 +4%	−32%	10.81 −60%	0.000 –	0.625 +0%	−30%	5.31 −64%	0.374 +2%	0.350 −13%	−25%
BendVLM (sph.)	112.03 +335%	**0.000** −100%	**0.000** −100%	+45%	**0.00** −100%	0.452 –	*	−100%†	**2.83** −89%	0.000 –	0.550 −12%	−51%	**2.86** −80%	**0.152** −59%	0.425 +6%	−44%
GGSS (Ours)	16.85 −35%	0.096 −61%	0.125 −71%	**−55%**	1.36 −84%	0.000 –	**0.025** −96%	**−90%**	8.48 −68%	0.000 –	**0.050** −92%	**−80%**	5.10 −65%	0.154 −58%	0.175 −56%	**−60%**

Table 1: **Main results against external baselines.** Each (method, model) row uses a single best-avg-α selected from $\{0.25, 0.5, 0.75, 1.0, 1.5\}$. A gray "–" on the second line means the baseline is zero, so percent change is undefined. * marks runs where the steered model produced no parseable answers (BendVLM (sph.) collapses generation there, consistent with Table 2; likely a difficulty of porting a CLIP-space method to generative activations, not a flaw of the original method). † average over a single task, excluded from cross-method comparison. Avg $\Delta\%$ averages over non-zero-baseline tasks: three on Pixtral-12B and Qwen3-VL-4B, two on the LLaVA models, so it is not strictly comparable across models.

Method	Race-task steering				Gender-task steering				Avg.
	Pixtral	Vicuna	Mistral	Qwen3	Pixtral	Vicuna	Mistral	Qwen3	
Baseline	53.6 0.0%	37.3 0.0%	38.5 0.0%	61.5 0.0%	53.6 0.0%	37.3 0.0%	38.5 0.0%	61.5 0.0%	47.7 0.0%
INLP (Eucl.)	52.1 −1.5%	37.2 −0.1%	36.4 −2.1%	62.1 +0.6%	53.3 −0.3%	37.6 +0.3%	38.0 −0.5%	62.0 +0.5%	47.3 −0.4%
INLP (sph.)	47.5 −6.1%	36.8 −0.5%	38.4 −0.1%	**62.4** +0.9%	50.3 −3.3%	37.1 −0.2%	38.6 +0.1%	61.3 −0.2%	46.5 −1.2%
MeanDiff-SVM (Eucl.)	53.3 −0.3%	37.1 −0.2%	39.0 +0.5%	61.4 −0.1%	53.1 −0.5%	37.4 +0.1%	38.2 −0.3%	61.9 +0.4%	47.7 −0.1%
MeanDiff-SVM (sph.)	53.2 −0.4%	37.3 +0.0%	39.2 +0.7%	61.8 +0.3%	53.1 −0.5%	37.2 −0.1%	39.0 +0.5%	61.8 +0.3%	**47.8** +0.1%
MeanDiff-LR (Eucl.)	**53.6** +0.0%	36.9 −0.4%	38.7 +0.2%	61.4 −0.1%	**53.7** +0.1%	37.2 −0.1%	38.8 +0.3%	61.8 +0.3%	**47.8** +0.0%
MeanDiff-LR (sph.)	53.0 −0.6%	37.1 −0.2%	39.4 +0.9%	61.7 +0.2%	53.2 −0.4%	36.6 −0.7%	38.8 +0.3%	61.9 +0.4%	47.7 +0.0%
BendVLM (Eucl.)	53.0 −0.6%	37.1 −0.2%	39.0 +0.5%	61.6 +0.1%	45.1 −8.5%	**37.8** +0.5%	38.6 +0.1%	61.7 +0.2%	46.7 −1.0%
BendVLM (sph.)	6.9 −46.7%	2.9 −34.4%	**39.5** +1.0%	59.0 −2.5%	5.3 −48.3%	2.6 −34.7%	**39.8** +1.3%	59.3 −2.2%	26.9 −20.8%
Pooled SVD (Eucl.)	47.3 −6.3%	35.0 −2.3%	38.9 +0.4%	60.8 −0.7%	46.5 −7.1%	37.5 +0.2%	39.0 +0.5%	61.8 +0.3%	45.9 −1.9%
Pooled SVD (sph.)	53.5 −0.1%	**38.2** +0.9%	38.5 +0.0%	61.7 +0.2%	53.3 −0.3%	37.2 −0.1%	38.4 −0.1%	61.8 +0.3%	**47.8** +0.1%
Per-token SVD (Eucl.)	42.0 −11.6%	37.2 −0.1%	37.7 −0.8%	60.2 −1.3%	22.1 −31.5%	36.9 −0.4%	38.5 +0.0%	61.0 −0.5%	42.0 −5.8%
Per-token SVD (sph.)	50.1 −3.5%	36.2 −1.1%	38.1 −0.4%	61.1 −0.4%	50.9 −2.7%	37.4 +0.1%	38.2 −0.3%	61.2 −0.3%	46.6 −1.1%
GGSS (ours)	**53.2** −0.4%	**37.8** +0.5%	**38.9** +0.4%	**61.6** +0.1%	**53.3** −0.3%	**37.1** −0.2%	**38.4** −0.1%	**62.1** +0.6%	**47.8** +0.1%

Table 2: **MMStar capability preservation at best-avg-α.** Best-avg-α choices are listed in Appendix B.2.

tasks with non-zero baselines contribute to the average. All bias metrics, the Avg $\Delta\%$ column, and MMStar for that method and model are reported at this selected operating point. For **GGSS** we fix $\kappa = 5$ and $g_{\text{floor}} = 0.3$ throughout, with sensitivity analyses in §4.3 and Appendix B.2.

Component ablation on Qwen3-VL-4B, MCQ (race)		
Variant	JSD×1k ↓	Bias Red. % ↑
Baseline (unsteered)	14.646	–
Pooled SVD (Eucl.; hard proj., no gate)	4.261	70.9
Pooled SVD (sph.; hard proj., no gate)	5.612	61.7
GGSS w/o gate ($g_i \equiv 1$)	5.612	61.7
GGSS w/o Slerp (hard proj. + gate)	4.533	69.1
GGSS (full)	**2.414**	**83.5**

Table 3: **Component ablation of GGSS** on Qwen3-VL-4B, MCQ salary (race) at $\alpha = 1.0$.

Sensitivity to $(\kappa, g_{\text{floor}})$ on Qwen3-VL-4B, MCQ				
$g_{\text{floor}} \backslash \kappa$	1	2	5	10
0.0	4.851	4.578	10.971	9.196
0.3	4.578	5.781	**2.414**	4.533
0.5	6.128	6.714	4.578	4.069

Table 4: **Sensitivity of GGSS** to gate sharpness κ and floor g_{floor} on Qwen3-VL-4B, MCQ salary (race) at $\alpha = 1.0$. Entries are JSD×1k; lower is better.

Discovery uses counterfactual image sets whose identities are held out from the evaluation occupation. For the race-based tasks we discover on {`cook`, `doctor`, `lawyer`, `nurse`, `teacher`}, and for the gender-based Nurse/Doctor task we discover on {`cook`, `lawyer`, `teacher`} while holding out both `nurse` and `doctor`. MCQ and 2AFC are scored on the SocialCounterfactuals probe suite (Howard et al., 2024), which does not share identities with the discovery pool.

Reading the tables. Tables 1 and 2 should be read as operating-point comparisons: for each method and model, the same selected α is used for all bias metrics and for MMStar. They separate two questions that are often conflated: whether a steering rule reduces demographic sensitivity, and whether the same operating point preserves general multimodal reasoning. A useful method must do both.

4.2 Main Results

Table 1 reports best-avg-α metrics for every method across all models and bias tasks; **GGSS** is the *single most reliable* method in the comparison.

Ranking. **GGSS** attains the lowest Avg $\Delta\%$ among all external baselines on all 4 models. The gaps are -55% vs. -37% on Pixtral-12B, where BendVLM (Eucl.) is the strongest external baseline, then -90% vs. -86% on LLaVA-1.6-Vicuna-7B, -80% vs. -78% on LLaVA-1.6-Mistral-7B, and -60% vs. -49% on Qwen3-VL-4B, all three against INLP (sph.), the strongest competitor overall.

Magnitude. **GGSS** produces substantial bias reductions: up to -96% on Nurse/Doctor (LLaVA-1.6-Vicuna-7B: $0.600 \rightarrow 0.025$), -84% on MCQ (LLaVA-1.6-Vicuna-7B: $8.70 \rightarrow 1.36$), and -61% on 2AFC (Pixtral-12B: $0.243 \rightarrow 0.096$).

MCQ and Nurse/Doctor measure different output formats—categorical answer distributions and binary occupation classification—yet **GGSS** reduces both without changing the selected operating point for a given model; on the two LLaVA models the zero-baseline 2AFC entries serve as sanity checks.

Capability preservation (MMStar). Table 2 reports MMStar results at each method's best-avg-α (the same α used to report its bias metrics in Table 1) across all steering settings. **GGSS** stays within ± 0.6 p.p. of baseline across these evaluations: adaptive geodesic steering preserves general capability. A steering method that reduces benchmark bias by collapsing general VLM capability is not useful as a deployment primitive, and several baselines suffer exactly such drops; BendVLM (sph.), for example, collapses on Pixtral and LLaVA-Vicuna. **GGSS**'s MMStar changes remain small under both race and gender steering: the gate confines the intervention to directions that carry the targeted bias and leaves the remaining visual tokens largely untouched.

Model	**GGSS** vs. unsteered (p)	MMStar Δpp (race / gender)	McNemar p (race / gender)
Pixtral-12B	$< 10^{-4}$	-0.40 / -0.33	0.58 / 0.49
LLaVA-Vicuna-7B	0.068^{a}	$+0.47$ / -0.20	0.58 / 0.77
LLaVA-Mistral-7B	0.007	$+0.40$ / -0.07	0.51 / 1.00
Qwen3-VL-4B	0.002	$+0.13$ / $+0.60$	0.86 / 0.09

Table 5: **Significance summary at the paper's operating points.** Column 2: paired sign-flip permutation test of **GGSS** vs. the unsteered model, combined across non-zero-baseline bias tasks ([a]on LLaVA-Vicuna the Nurse/Doctor task alone has $p < 10^{-4}$). Columns 3–4: MMStar paired per-question statistics under race / gender steering. Protocol details in Appendix B.4.

Statistical significance. Decoding is greedy and deterministic, so the relevant uncertainty is sampling over evaluation items. For every reported cell we compute 95% bootstrap CIs (10,000 resamples, over items and identities), paired sign-flip permutation tests, and exact McNemar tests on MMStar (Appendix B.4). Table 5 summarizes: **GGSS**'s reductions are significant on three of four backbones (on LLaVA-Vicuna the Nurse/Doctor reduction alone has $p < 10^{-4}$), and its MMStar changes are

Method	Pix-tral	Qwen3-VL	Mis-tral	Vi-cuna	mean	worst
GGSS (paper)	-55.3	-59.9	-80.2	-90.1	-71.4	-55.3
LEACE + our gate	-72.2	-30.8	-66.9	-83.9	-63.5	-30.8
INLP (sph.)	-31.2	-49.3	-77.6	-86.1	-61.1	-31.2
LEACE (sph.)	-70.6	-26.4	-51.5	-45.4	-48.5	-26.4

Table 6: **Added baselines.** Avg $\Delta\%$ at each method's best-avg-α (bias tasks with non-zero baseline; lower is better), with the four-model mean and worst case. INLP (sph.) is the strongest baseline from Table 1, shown for reference. Full details in Appendix B.10.

statistically indistinguishable from the unsteered model in all eight runs, whereas the strongest competitor, INLP (sph.), significantly damages Pixtral's MMStar (-6.1 pp race, $p < 10^{-4}$).

Held-out α selection. Table 1 selects α on the probe suite it reports, identically for every method, so no method is advantaged, but absolute reductions may be optimistic. Under a cross-fitted protocol selecting α on one identity fold and evaluating on the other (Appendix B.5), **GGSS** stays near the top on every model, matches the paper's α on six of eight folds, and reduces bias in every held-out fold.

Additional baseline families. Table 6 adds LEACE (Belrose et al., 2023) at the same layer under the same protocol. Ported this way, LEACE is a serious baseline (it wins the Pixtral column while preserving MMStar there) but is less consistent across backbones (-26% to -31% on Qwen3-VL), while **GGSS** keeps the best four-model mean and worst case. Prompt-space mitigation (a fairness instruction appended to every prompt) is inconsistent and can even amplify bias (e.g., Vicuna MCQ $+85\%$); full results in Appendix B.9.

Debiasing, not blanket removal. A steering rule could reduce measured bias by deleting demographic perception outright; **GGSS** does not. Steering one attribute leaves recognition of the *other* attribute intact on all four backbones, all six MMStar dimensions stay within binomial noise, and MME (Qwen3-VL) moves only -0.4% / -1.8% with no subcategory collapsing. The trade-off with the *targeted* attribute is controlled by α: at $\alpha = 0.5$ on Qwen3-VL, race recognition stays within 4 pp of baseline while 55% of the MCQ reduction is already realized, so tasks that need the attribute can run at moderate α or gate steering off (Appendix B.8).

4.3 Ablations and Sensitivity

Component and sensitivity. Table 3 isolates the contribution of each **GGSS** ingredient on a representative setting (Qwen3-VL-4B, MCQ salary, $\alpha = 1.0$). Disabling the gate ($g_i \equiv 1$) gives the same JSD as the ungated spherical hard-projection variant (5.61). Adding the gate *without* Slerp improves this value from 5.61 to 4.53 (-69% from baseline), showing that selective gating alone already helps. The full method, combining gating *and* Slerp interpolation, achieves 2.41 (-84%), a further 14 pp improvement in this setting.

Table 4 sweeps gate sharpness $\kappa \in \{1, 2, 5, 10\}$ and floor $g_{\text{floor}} \in \{0.0, 0.3, 0.5\}$. All twelve settings reduce bias well below the unsteered baseline (14.65); the best reduction (-84%) occurs at $\kappa = 5$, $g_{\text{floor}} = 0.3$. High κ with $g_{\text{floor}} = 0$ can overselect, but a positive floor recovers performance.

Geometry isolated at matched gate. Toggling Euclidean versus spherical steering at matched gate condition on all four backbones, with paired sign-flip tests (Appendix B.11), shows the toggle is within sampling noise at the 4B–7B scales but decisive at 12B: on Pixtral the spherical form is significantly better under both toggles ($p = 0.031$ / 0.015), both Euclidean forms significantly damage MMStar there (-6.3 pp, $p < 10^{-8}$) while the spherical forms preserve it everywhere, and over-steering exposes failure modes the geodesic never showed ($9\times$ overshoot; a degenerate cell with $72/80$ unparseable). The ablations thus attribute *robustness that grows with model scale* to on-sphere steering, and *selectivity and controllability* to the gate.

5 Conclusion

We introduced **GGSS**, an inference-time debiasing framework for generative VLMs that discovers counterfactual spherical bias subspaces, applies a calibrated token gate, and steers with norm-preserving Slerp. Across four generative VLMs and three bias protocols, **GGSS** attains the lowest Avg $\Delta\%$ on all 4 models against ten external steering baselines and prompt-based mitigation, significant on three of four backbones, while keeping MMStar indistinguishable from the unsteered model. Matched-gate ablations attribute robustness at scale to on-sphere steering and controllable selectivity to the gate, supporting adaptive geodesic steering as a practical primitive for inference-time bias mitigation.

Limitations

Scope. Our evaluation covers four generative VLMs, three demographic-bias protocols, and MMStar, but does not establish that **GGSS** generalizes to all architectures, languages, domains, or bias types. We focus on perceived race and gender in image-conditioned settings; other protected attributes, intersectional groups, multilingual prompts, and open-ended uses remain future work. The subspace construction may also extend to non-demographic bias axes such as political leaning, whose embedding-space structure has been studied in text models (Sun et al., 2025b).

Calibration. **GGSS** still requires selecting a steering strength. We choose one best-avg-α per (method, model) from $\{0.25, 0.5, 0.75, 1.0, 1.5\}$ using held-out bias measurements, and the optimal value is model-dependent. A tuning-free rule, potentially using the bias-norm statistics $\tilde{b}$ and σ_b to set token-level strengths, remains an open direction.

Deployment. Lower benchmark bias should not be interpreted as complete fairness or safety. **GGSS** changes activations at inference time, but does not remove biased knowledge from model parameters or guarantee robustness under distribution shift. Misconfigured steering may also suppress useful demographic information or affect tasks beyond our capability checks: at the bias-minimizing steering strength, the targeted attribute's reportability on face-only close-ups is strongly reduced (Appendix B.8), so tasks that legitimately require the attribute, such as requested demographic descriptions or clinical documentation, should operate at a moderate α chosen from the trade-off curve, or gate steering off. Deployment should include broader auditing, human review where appropriate, and monitoring for residual or newly introduced harms.

Ethics Statement

Intended use and dual use. **GGSS** is designed to reduce demographic bias in deployed VLMs without retraining. The same mechanism, however, is dual-use: a steering subspace can be aimed at amplifying demographic sensitivity as easily as at reducing it, and inference-time manipulation of model behavior is demonstrably exploitable (Liu et al., 2026). At high steering strength the intervention also shades from *debiasing* into *attribute removal* (Appendix B.8). We release the trade-off measurements needed to choose an operating point deliberately; deployments should measure both bias and attribute retention on the target task before fixing a steering strength.

Demographic labels. The counterfactual datasets we use annotate *perceived* race and binary gender as assigned by dataset curators. These labels are operational constructs for measuring output disparities; they do not capture self-identification, intersectional identity, or the full diversity of human appearance, and results should not be read as claims about any individual's identity.

Data and licensing. All images come from publicly released research datasets (REFLECT/FOCUS, SocialCounterfactuals) used within their stated licenses; we generate no new images of people and release no new personal data. Artifact documentation and intended-use notes accompany the released benchmark suite (Appendix B.14).

Residual risk. Reduced benchmark bias is not fairness. Steering does not remove biased knowledge from model parameters, may behave differently under distribution shift, and covers only the attributes and tasks we measure. Systems using **GGSS** in consequential settings should retain human oversight and independent auditing.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Na Min An, Yoonna Jang, Yusuke Hirota, Ryo Hachiuma, Isabelle Augenstein, and Hyunjung Shim. 2026. Interpretable debiasing of vision-language models for social fairness. *arXiv preprint arXiv:2602.24014*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Debodeep Banerjee, Stefano Teso, Burcu Sayin, and Andrea Passerini. 2024. Learning to guide human decision makers with vision-language models. *arXiv preprint arXiv:2403.16501*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063.

Die Chen, Zhiwen Li, Mingyuan Fan, Cen Chen, Wenmeng Zhou, Yanhao Wang, and Yaliang Li. 2025. Growth inhibitors for suppressing inappropriate image concepts in diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 79164–79184.

Haodong Chen, Qiang Huang, Jiaqi Zhao, Qiuping Jiang, Xiaojun Chang, and Jun Yu. 2026. Measuring social bias in vision-language models with face-only counterfactuals from real photos. *arXiv preprint arXiv:2601.06931*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Harry Cheng, Yangyang Guo, Qingpei Guo, Ming Yang, Tian Gan, Weili Guan, and Liqiang Nie. 2025. Social debiasing for fair multi-modal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1740–1750.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11198–11201.

Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, and Wei Chen. 2025. Don't reinvent the wheel: Efficient instruction-following text embedding based on guided space transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24511–24525.

Walter Gerych, Cassandra Parent, Quinn Perian, Rafiya Javed, Justin Solomon, and Marzyeh Ghassemi. 2026. Wring out the bias: A rotation-based alternative to projection debiasing. In *The Fourteenth International Conference on Learning Representations*.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. 2024. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502.

Kshitish Ghate, Tessa Charlesworth, Mona Diab, and Aylin Caliskan. 2025. Biases propagate in encoder-based vision-language models: A systematic analysis from intrinsic measures to zero-shot retrieval outcomes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18562–18580.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-language models performing zero-shot tasks exhibit disparities between gender groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2778–2785.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985.

Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. 2025. Visbias: Measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17981–18004.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Taeuk Jang, Hoin Jung, and Xiaoqian Wang. 2025. Target bias is all you need: Zero-shot debiasing of vision-language models with bias corpus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1935–1946.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735.

Hoin Jung, Taeuk Jang, and Xiaoqian Wang. 2024. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Zhiwen Li, Die Chen, Mingyuan Fan, Cen Chen, Yaliang Li, Yanhao Wang, and Wenmeng Zhou. 2025. Responsible diffusion models via constraining text embeddings within safe regions. In *Proceedings of the ACM on Web Conference 2025*, pages 1588–1601.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuansen Liu, Yixuan Tang, and Anthony Kum Hoe Tung. 2026. Reasoning hijacking: The fragility of reasoning alignment in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36646–36665.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Kanti V. Mardia and Peter E. Jupp. 2000. Directional statistics. In *Wiley Series in Probability and Statistics*. Wiley.

Mahdiyar Molahasani, Azadeh Motamedi, Michael Greenspan, Il-Min Kim, and Ali Etemad. 2025. Prism: Reducing spurious implicit biases in vision-language models with llm-guided embedding projection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 688–697.

Vishal Narnaware, Ashmal Vayani, Rohit Gupta, Sirnam Swetha, and Mubarak Shah. 2025. Bbq-v: Benchmarking visual stereotype bias in large multimodal models. *Preprint*, arXiv:2502.08779.

Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. Cabin: Debiasing vision-language models using backdoor adjustments. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 484–492.

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

Xavier Pennec. 2006. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Neale Ratzlaff, Matthew Lyle Olson, Musashi Hinck, Shao-Yen Tseng, Vasudev Lal, and Phillip Howard. 2024. Debiasing large vision-language models by ablating protected attribute representations. *arXiv preprint arXiv:2410.13976*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008.

Gabriele Ruggeri and Debora Nozza. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. 2026. AlphaSteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 245–254.

Himanshu Singh, Ziwei Xu, AV Subramanyam, and Mohan Kankanhalli. 2026. Do prompts guarantee safety? mitigating toxicity from llm generations through subspace intervention. *arXiv preprint arXiv:2602.06623*.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85.

Yandong Sun, Qiang Huang, Ziwei Xu, Yiqun Sun, Yixuan Tang, and Anthony KH Tung. 2025a. One swallow does not make a summer: Understanding semantic structures in embedding spaces. *arXiv preprint arXiv:2512.00852*.

Yiqun Sun, Qiang Huang, Anthony Kum Hoe Tung, and Jun Yu. 2025b. Prism: A framework for producing interpretable political bias embeddings with political-aware cross-encoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27719–27733.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Hao Wang, Yiqun Sun, Pengfei Wei, Lawrence B Hsieh, and Daisuke Kawahara. 2026. Sparse autoencoders as plug-and-play firewalls for adversarial attack detection in vlms. *arXiv preprint arXiv:2605.07447*.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*.

Xing Wu, Kehong Liu, Jianjia Wang, Junfeng Yao, Bin Deng, Rongqi Lv, and Jun Song. 2024. Candidate evaluation with multimodal data-driven for recruitment. In *International Conference on Pattern Recognition*, pages 81–96. Springer.

Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. 2025. Joint vision-language social bias removal for clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4246–4255.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. Vl-stereoset: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Implementation Details

A.1 Notation

The main text uses the compact notation $\mathbf{x}_{u,a}$, where a is the protected attribute value and u collects the visual factors held fixed. In the race-debiasing implementation, $a = r \in \mathcal{R}$ and $u = (o, b, g)$, where $o \in \mathcal{O}$ is the occupation, $b \in \mathcal{B}_o$ is the base identity, and $g \in \mathcal{G}$ is the gender label. Thus

$$\mathbf{x}_{u,a} = I_{o,b}^{(r,g)}.$$

The corresponding activation matrix is

$$\mathbf{H}_{u,a} = \mathbf{H}_{o,b}^{(r,g)} \in \mathbb{R}^{T \times D},$$

where D is the hidden dimension, T is the number of discovery visual tokens, and the t-th token activation is $\mathbf{h}_{o,b,t}^{(r,g)} \in \mathbb{R}^D$. The normalized pooled representative in the main text is

$$\mathbf{y}_{u,a} = \nu\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_{u,a,t}\right) = \widehat{\mathbf{h}}_{o,b}^{(r,g)}.$$

The within-context Fréchet mean and shift are similarly related by

$$\begin{aligned} \mathbf{c}_u &= \boldsymbol{c}_{o,b,g}, \\ \mathbf{s}_{u,a} &= \mathbf{s}_{o,b}^{(r,g)} \quad \text{when } u = (o,b,g),\ a = r. \end{aligned}$$

For gender debiasing, the roles of r and g are exchanged: $a = g$ and $u = (o, b, r)$. The learned basis is denoted by $\mathbf{V}_{\text{bias}}$ throughout the paper. At inference time, the hooked activation tensor has shape

$$\mathbf{Z} \in \mathbb{R}^{n_{\text{batch}} \times T' \times D},$$

where T' is the sequence length at the intervention layer. We use n_{batch}, rather than B, to denote batch size. The steering strength is denoted by $\alpha \in \mathbb{R}$.

A.2 Implementation Details of GGSS

A.2.1 Discovery data and activation extraction

Discovery occupations are held out on a per-task basis so that the occupation used to score the downstream bias metric never appears in the discovery pool. For the race-based tasks (MCQ, 2AFC) we discover on {`cook`, `doctor`, `lawyer`, `nurse`, `teacher`}. For the gender-based Nurse/Doctor task we discover on {`cook`, `lawyer`, `teacher`} and hold out both `nurse` and `doctor`. For each base identity in the active

discovery pool, we use all $K = 5$ race variants and both gender variants, yielding 10 counterfactual images per identity. All images within the same counterfactual group are required to have identical pixel dimensions. This is enforced by a runtime check before activation extraction to guarantee a consistent number of visual tokens.

For each image, we run a single forward pass with the fixed discovery prompt

`"Describe this image in detail."`

and capture the output of the target module using `register_forward_hook`. If the hooked module returns a tuple, we cache the first element. Otherwise, we cache the output tensor directly. The captured tensor is reshaped into

$$\mathbf{H}_{o,b}^{(r,g)} \in \mathbb{R}^{T \times D}.$$

When T exceeds a configurable cap `max_tokens` (default: 2048), we draw a deterministic random subset of token indices using seed 42. The same token indices are reused for all (r, g) variants of the same base identity b, so token positions remain aligned across counterfactual images. **GGSS** then mean-pools the retained tokens to obtain one vector per image.

A.2.2 Numerical spherical primitives

Section 3 defines the ideal spherical operations ν, log, exp, and Slerp. In Appendix A, we use $\mathrm{Normalize}_\varepsilon$, Log, and Exp to denote their numerical implementations. Specifically,

$$\mathrm{Normalize}_\varepsilon(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2 + \varepsilon}, \qquad \varepsilon = 10^{-10}.$$

Inner products passed to arccos are clamped to $[-1 + 10^{-7}, 1 - 10^{-7}]$. Degenerate cases with geodesic distance below $\varepsilon_{\text{geo}} = 10^{-8}$ return the zero tangent vector for the logarithmic map, and tangent vectors with norm below ε_{geo} return the base point under the exponential map. For Slerp, when the angle is below ε_{geo}, we use the continuous small-angle limit. All spherical operations are implemented in vectorized form over tokens.

A.2.3 Pooled counterfactual bias subspace discovery

For each image, we first mean-pool the token activations and project onto the unit sphere:

$$\widehat{\bar{\mathbf{h}}}_{o,b}^{(r,g)} = \mathrm{Normalize}_\varepsilon\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_{o,b,t}^{(r,g)}\right). \tag{6}$$

For race debiasing, $u = (o, b, g)$ and $a = r$. Thus the main-body center $\mathbf{c}_u$ is written concretely as $\boldsymbol{c}_{o,b,g}$. For each counterfactual group (o, b, g), this center is computed by iterative Fréchet (Karcher) mean. Starting from the normalized Euclidean mean $\boldsymbol{\eta}^{(0)}$, the update at iteration m is

$$\begin{aligned} \bar{\mathbf{t}}^{(m)} &= \frac{1}{K}\sum_r \mathrm{Log}_{\boldsymbol{\eta}^{(m)}}\left(\widehat{\bar{\mathbf{h}}}_{o,b}^{(r,g)}\right), \\ \boldsymbol{\eta}^{(m+1)} &= \mathrm{Normalize}_\varepsilon\left(\mathrm{Exp}_{\boldsymbol{\eta}^{(m)}}(\bar{\mathbf{t}}^{(m)})\right). \end{aligned} \tag{7}$$

The iteration terminates when $\|\bar{\mathbf{t}}^{(m)}\|_2 < 10^{-7}$ or after 100 iterations. We denote the converged center by $\boldsymbol{c}_{o,b,g}$.

The main-body shift $\mathbf{s}_{u,a}$ is written concretely as $\mathbf{s}_{o,b}^{(r,g)}$:

$$\mathbf{s}_{u,a} = \mathbf{s}_{o,b}^{(r,g)} = \log_{\mathbf{c}_u}(\mathbf{y}_{u,a}) = \mathrm{Log}_{\boldsymbol{c}_{o,b,g}}\left(\widehat{\bar{\mathbf{h}}}_{o,b}^{(r,g)}\right). \tag{8}$$

Concatenating across occupations, identities, genders, and races yields a shift matrix

$$\mathbf{S} \in \mathbb{R}^{N_{\text{shifts}} \times D}, \qquad N_{\text{shifts}} = \sum_{o \in \mathcal{O}} |\mathcal{B}_o|\,|\mathcal{G}|\,|\mathcal{R}|. \tag{9}$$

The SVD $\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ yields the bias subspace

$$\mathbf{V}_{\text{bias}} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_k] \in \mathbb{R}^{D \times k}, \tag{10}$$

with $k = K - 1 = 4$ by default. The global reference point is the Fréchet mean of all pooled unit vectors, computed with the same iteration as Eq. (7).

A.2.4 Gate calibration statistics

For each normalized pooled discovery representative $\mathbf{c} = \mathbf{y}_{u,a}$, we compute the protected-coordinate magnitude

$$b(\mathbf{c}) = \|\mathbf{V}_{\text{bias}}^\top \log_{\boldsymbol{\mu}}(\mathbf{c})\|_2.$$

Equivalently, in the concrete implementation notation, $\mathbf{c} = \widehat{\bar{\mathbf{h}}}_{o,b}^{(r,g)}$. We cache

$$\tilde{b} = \mathrm{median}\{b(\mathbf{c})\}, \qquad \sigma_b = \mathrm{std}\{b(\mathbf{c})\}.$$

These scalars are loaded at inference time to compute the calibrated token gate.

A.2.5 Inference-time geodesic-gated steering

At inference time, a forward hook intercepts the output activation tensor $\mathbf{Z} \in \mathbb{R}^{n_{\text{batch}} \times T' \times D}$, flattens it into $N = n_{\text{batch}} T'$ token vectors, and processes each nonzero token independently. For token $\mathbf{h}_i$, we record its norm $r_i = \|\mathbf{h}_i\|_2$ and normalized direction $\widehat{\mathbf{h}}_i = \text{Normalize}_\varepsilon(\mathbf{h}_i)$. The implemented token update is

$$\begin{aligned} \mathbf{t}_i &= \text{Log}_{\boldsymbol{\mu}}(\widehat{\mathbf{h}}_i), \\ \mathbf{p}_i &= \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top \mathbf{t}_i, \\ \mathbf{t}_i^{\text{clean}} &= (I - \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top) \mathbf{t}_i. \end{aligned} \tag{11}$$

The target direction is

$$\widehat{\mathbf{h}}_i^{\text{target}} = \text{Normalize}_\varepsilon\left(\text{Exp}_{\boldsymbol{\mu}}(\mathbf{t}_i^{\text{clean}})\right). \tag{12}$$

The per-token gate uses the discovery statistics:

$$\begin{aligned} z_i &= \frac{\|\mathbf{V}_{\text{bias}}^\top \mathbf{t}_i\|_2 - \tilde{b}}{\sigma_b + \varepsilon}, \\ g_i &= g_{\text{floor}} + (1 - g_{\text{floor}}) \, \text{sigmoid}(\kappa z_i), \end{aligned} \tag{13}$$

Since $\mathbf{V}_{\text{bias}}$ has orthonormal columns, $\|\mathbf{V}_{\text{bias}}^\top \mathbf{t}_i\|_2 = \|\mathbf{p}_i\|_2$. For the main experiments, we use $\kappa = 5$ and $g_{\text{floor}} = 0.3$; other g_{floor} values are used only in the sensitivity analysis. Slerp then rotates toward the target:

$$\widehat{\mathbf{h}}_i^{\text{steered}} = \text{Slerp}\big(\widehat{\mathbf{h}}_i, \widehat{\mathbf{h}}_i^{\text{target}}; \alpha g_i\big), \tag{14}$$

and norm restoration produces the final output:

$$\mathbf{h}_i^{\text{steered}} = r_i \, \widehat{\mathbf{h}}_i^{\text{steered}}. \tag{15}$$

The steered tensor is reshaped to its original layout and cast to the input dtype before being returned by the hook.

A.2.6 Saved artifacts

The discovery stage saves a `.pt` checkpoint containing $(\mathbf{V}_{\text{bias}}, \boldsymbol{\mu}, k, \tilde{b}, \sigma_b)$ together with metadata: singular values σ_j, explained-variance ratios $\rho_j = \sigma_j^2 / \sum_\ell \sigma_\ell^2$, the full protected-coordinate magnitude distribution $\{b(\mathbf{c})\}$ (for diagnostic histograms, see Appendix B.2), and discovery-set sizes. This checkpoint is sufficient for inference-time loading.

A.3 Proof of Proposition 1

Proof. By Eq. (1), the columns of $\mathbf{V}_{\text{bias}}$ are right singular vectors of $\mathbf{S}$. Hence $\mathbf{V}_{\text{bias}}^\top \mathbf{V}_{\text{bias}} = I_k$. Therefore $\mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top$ is symmetric and idempotent, so it is the orthogonal projector onto $\text{span}(\mathbf{V}_{\text{bias}})$. Consequently,

$$I - \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top$$

is the orthogonal projector onto

$$\text{span}(\mathbf{V}_{\text{bias}})^\perp = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{V}_{\text{bias}}^\top \mathbf{z} = \mathbf{0}\}.$$

Using Eq. (2), we have

$$\mathbf{t}_i^{\text{clean}} = (I - \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top) \mathbf{t}_i,$$

which proves the projection claim.

It remains to prove norm preservation. Let $\mathbf{p} = \widehat{\mathbf{h}}_i$, $\mathbf{q} = \widehat{\mathbf{h}}_i^{\text{target}}$, and $\theta = \arccos\langle \mathbf{p}, \mathbf{q} \rangle$. If $\mathbf{p} \neq \mathbf{q}$ and $\mathbf{q} \neq -\mathbf{p}$, define

$$\mathbf{u} = \frac{\mathbf{q} - \cos(\theta)\mathbf{p}}{\sin(\theta)}.$$

Then $\mathbf{u}$ has unit norm and is orthogonal to $\mathbf{p}$. The Slerp formula can be written as

$$\text{Slerp}(\mathbf{p}, \mathbf{q}; \beta_i) = \cos(\beta_i \theta)\mathbf{p} + \sin(\beta_i \theta)\mathbf{u},$$

which has unit norm. If $\mathbf{p} = \mathbf{q}$, the convention $\text{Slerp}(\mathbf{p}, \mathbf{p}; \beta_i) = \mathbf{p}$ also gives a unit vector. Therefore $\widehat{\mathbf{h}}_i^{\text{steered}}$ has unit norm under the stated cases. By Eq. (5) and $r_i = \|\mathbf{h}_i\|_2$,

$$\|\mathbf{h}_i^{\text{steered}}\|_2 = \|r_i \widehat{\mathbf{h}}_i^{\text{steered}}\|_2 = r_i = \|\mathbf{h}_i\|_2.$$

This proves the result. □

A.4 Ablation Variants

We evaluate four SVD ablations that isolate individual design choices of **GGSS**.

Pooled SVD (Eucl.). This removes both spherical geometry and the Slerp/gate. Images are Euclidean-pooled, per-group Euclidean means are subtracted, and SVD on the resulting shift matrix yields a Euclidean bias basis. Inference subtracts $\alpha \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top (\mathbf{h}_i - \boldsymbol{\mu})$ from every token. Because it is Euclidean, this variant does not preserve token norms.

Pooled SVD (sph.). Retains spherical geometry (Fréchet means, tangent shifts, exp-map projection, norm restoration) but uses hard null-space projection in tangent space:

$$\mathbf{t}_i^{\text{steered}} = \mathbf{t}_i - \alpha \mathbf{V}_{\text{bias}} \mathbf{V}_{\text{bias}}^\top \mathbf{t}_i.$$

This is identical to **GGSS** with the Slerp step replaced by hard projection and the gate disabled ($g_i \equiv 1$).

Per-token SVD (Eucl.) and Per-token SVD (sph.). These variants compute per-token centers $\boldsymbol{c}_{o,b,g,t}$ and per-token shifts before SVD, preserving token position information in discovery. The inference step then applies standard null-space projection. Per-token SVD is more expensive and, on most (model, task) cells, dominated by the pooled variant.

Method	Pixtral-12B				LLaVA-1.6-Vicuna-7B				LLaVA-1.6-Mistral-7B				Qwen3-VL-4B			
	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$	MCQ $\times 10^3 \downarrow$	2AFC $\downarrow$	N/D $\downarrow$	Avg $\Delta\% \downarrow$
Baseline	25.75	0.243	0.425	0%	8.70	0.000	0.600	0%	26.87	0.000	0.625	0%	14.65	0.368	0.400	0%
Hard projection, no Slerp, no gate																
Pooled SVD (Eucl.)	**10.30** −60%	0.173 −29%	0.425 +0%	−30%	9.91 +14%	0.000 –	**0.025** −96%	−41%	5.81 −78%	0.000 –	0.300 −52%	−65%	6.49 −56%	0.190 −48%	**0.100** −75%	**−60%**
Pooled SVD (sph.)	15.54 −40%	0.098 −60%	**0.125** −71%	**−57%**	**1.36** −84%	0.000 –	**0.025** −96%	**−90%**	11.20 −58%	0.000 –	**0.050** −92%	−75%	8.37 −43%	0.173 −53%	0.150 −62%	−53%
Per-token hard projection																
Per-token SVD (Eucl.)	24.59 −4%	0.172 −29%	0.522 +23%	−4%	4.26 −51%	0.000 –	0.600 +0%	−25%	**2.08** −92%	0.000 –	0.625 +0%	−46%	4.35 −70%	0.245 −33%	0.475 +19%	−28%
Per-token SVD (sph.)	32.84 +28%	0.227 −7%	0.400 −6%	+5%	2.05 −76%	0.000 –	0.625 +4%	−36%	2.74 −90%	0.000 –	0.625 +0%	−45%	**4.19** −71%	0.286 −22%	0.450 +12%	−27%
GGSS (Ours)	**16.85** **−35%**	**0.096** **−61%**	**0.125** **−71%**	**−55%**	**1.36** **−84%**	**0.000** –	**0.025** **−96%**	**−90%**	**8.48** **−68%**	**0.000** –	**0.050** **−92%**	**−80%**	**5.10** **−65%**	**0.154** **−58%**	**0.175** **−56%**	**−60%**

Table 7: **SVD ablations of GGSS across all models and bias tasks.** Rows remove the gate and Slerp, and per-token rows additionally replace pooled SVD with per-token SVD. Results use the same best-avg-α protocol as Table 1; per-column best values within this ablation set are bolded.

A.5 Baseline Methods

A.5.1 Classifier-Guided Mean-Shift Correction (MeanDiff)

We adapt classifier-based debiasing to the generative VLM setting using a single image-level representation per image. For each discovery image, the activation matrix is mean-pooled to obtain $\bar{\mathbf{h}} \in \mathbb{R}^D$. A race classifier is then trained on these pooled activations. We consider two classifiers: an RBF-kernel SVM (`sklearn.svm.SVC(kernel="rbf")`, C=1, γ=`scale`) and a multinomial logistic regression (GPU-accelerated cuML where available, otherwise scikit-learn). For the Euclidean variant, the class mean $\boldsymbol{\mu}_r$ and global mean $\boldsymbol{\mu}_{\text{global}}$ yield the class-specific shift $\boldsymbol{\delta}_r = \boldsymbol{\mu}_r - \boldsymbol{\mu}_{\text{global}}$. At inference time the pooled activation is classified and $\alpha\boldsymbol{\delta}_{\hat{r}}$ is subtracted from every token. The spherical variant uses Fréchet means and tangent-space subtraction with norm restoration.

A.5.2 INLP

We implement Iterative Null-space Projection (Ravfogel et al., 2020) in both Euclidean and spherical variants. Let $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $y \in \{1,\dots,K\}^N$. The iteration is $\mathbf{X}^{(0)} = \mathbf{X}$. For $m = 0, 1, \dots$ we fit a multinomial logistic-regression classifier $f^{(m)}$ on $\mathbf{X}^{(m)}$. If accuracy falls to within 10^{-6} of the uniform-chance baseline $1/K$, the iteration halts. Otherwise, the orthonormalized rows of the weight matrix are appended to the bias basis and $\mathbf{X}^{(m+1)}$ is obtained by projecting $\mathbf{X}^{(m)}$ into their null space. We use at most 10 iterations and C=1.0. After convergence, stacking and orthonormalizing the extracted directions yields $\mathbf{V}_{\text{bias}}$. The Euclidean variant applies standard null-space steering. The spherical variant normalizes, computes a Fréchet global mean, and projects in the tangent space before exp-mapping and norm-restoring.

A.5.3 LEACE

We additionally evaluate LEACE (Belrose et al., 2023), which provides a closed-form least-squares concept eraser. Let $\boldsymbol{\Sigma}_{\mathbf{x}}$ denote the covariance of pooled discovery activations $\bar{\mathbf{h}}$ and let $\boldsymbol{\Sigma}_{\mathbf{xz}}$ denote the cross-covariance with the one-hot race labels. LEACE computes a whitening $\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}$, takes the top-$K-1$ left singular vectors of $\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}}$, and applies the affine projection $P(\mathbf{x}) = \mathbf{x} - \mathbf{W}^{-1}\mathbf{U}\mathbf{U}^\top\mathbf{W}(\mathbf{x} - \boldsymbol{\mu}_{\text{global}})$. We evaluate two variants: spherical LEACE, applied in tangent space, and gated geodesic LEACE, which uses LEACE's $\mathbf{U}$ as $\mathbf{V}_{\text{bias}}$ inside the **GGSS** pipeline. The latter serves as a drop-in test of whether LEACE's closed-form subspace transfers under the Slerp + gate framework.

A.5.4 BendVLM-style Retrieval Equalization

Following Gerych et al. (2024), we represent each discovery image by a single *global* activation $\mathbf{e} = \text{vec}(\mathbf{H}) \in \mathbb{R}^{TD}$ and store L2-normalized reference vectors with their race labels. We replace BendVLM's text-side augmentation with an SVD-based demographic projector computed from race-center differences. At inference time we retrieve the 10 nearest references per race via cosine similarity, average to per-race prototypes $\boldsymbol{\pi}_r$, and solve for the minimum-norm correction that equalizes cosine distance to all prototypes. The Euclidean variant uses a linear equidistance constraint. The spherical variant lifts everything to the tangent space at the global Fréchet mean, solves the equidistance system there, and maps back to the sphere with

norm restoration. The correction tensor is cached after the first decoding step of each image because visual tokens remain unchanged across autoregressive steps.

A.6 Model-Specific Hook Locations and Runtime Setup

All methods intervene at the layer denoted by `projection-mlp2`, which resolves to the final `nn.Linear` layer of the vision-to-language projection MLP. For Pixtral-12B, this is the second linear in the multi-modal projector. For LLaVA-1.6-Vicuna-7B and LLaVA-1.6-Mistral-7B, it is `model.model.mm_projector[2]`. For Qwen3-VL-4B-Instruct, it is `model.model.visual.merger.linear_fc2`. Models are loaded with `device_map="auto"` and run in `bfloat16`. All steering computations use `float32` for numerical stability, and outputs are cast back to the input dtype before being returned by the hook. All stochastic components use seed 42. Inner products passed to $\arccos$ are clamped to $[-1+10^{-7}, 1-10^{-7}]$, and geodesic distances below 10^{-8} are treated as zero.

B Additional Experimental Details

B.1 Protocols

Sweep. All methods are evaluated at $\alpha \in \{0.25, 0.5, 0.75, 1.0, 1.5\}$ on three bias tasks. We report the best-α metric per (method, model, task). Discovery occupations are held out on a per-task basis as described in §A.2: race-task discovery uses $\{$`cook`, `doctor`, `lawyer`, `nurse`, `teacher`$\}$, and gender-task discovery uses $\{$`cook`, `lawyer`, `teacher`$\}$ with `nurse` and `doctor` held out.

Decoding. Greedy decoding (`do_sample=False`). Default budget `max_new_tokens`$=2048$. For MCQ we use 8, since the output is constrained to a single letter.

CEO Salary. We sample $N_{\text{bios}} = 10$ biographies, replace name placeholders with "the candidate," and ask the model to output only an integer salary. Parsing first searches for a 4–9 digit integer with commas/dollar signs stripped, then falls back to any integer ≥ 1000. The baseline maximum salary is used as an outlier cap for steered methods. Total queries per (method,α): $B_{\text{ceo}} \cdot N_{\text{bios}} \cdot 5 \cdot 2$, where $B_{\text{ceo}} = 8$. *Evidential role.* With only 10 biographies this probe is small, and we treat it as directional, corroborating evidence, not a stand-alone result; no headline claim rests on it. On both backbones where the probe elicits varying salaries, the point estimates move toward parity under **GGSS** (Qwen3-VL gender gap -88k $\rightarrow -55$k; Pixtral 144k $\rightarrow$ 33k), the same direction as the statistically significant reductions on the larger probes, but the bio-cluster bootstrap CIs are wide and overlap (Qwen3-VL: $[-213\text{k}, 0]$ unsteered vs. $[-165\text{k}, 0]$ steered).

MMStar. MMStar is evaluated with VLMEvalKit using exact matching rather than LLM-based judging to avoid judge-dependent confounds. We report $\Delta_{\text{acc}} = \text{Acc}^{\text{method}}(\alpha) - \text{Acc}^{\text{baseline}}$.

B.2 Hyperparameters

Table 8 summarizes the three **GGSS**-specific hyperparameters and the ranges we explored. Best-α is selected once per (method, model) pair under the best-avg-α protocol. κ and g_{floor} are *fixed* across all main-table experiments.

Hyperparameter	Default	Range explored
α (strength)	best-α	$\{0.25, 0.5, 0.75, 1.0, 1.5\}$
κ (gate sharpness)	5	$\{1, 2, 5, 10\}$
g_{floor} (gate floor)	0.3	$\{0.0, 0.3, 0.5\}$
k (subspace dim)	$K-1$	fixed

Table 8: **GGSS** hyperparameters and explored ranges.

Method	Pixtral	Vicuna	Mistral	Qwen3
INLP (Eucl.)	1.0	0.25	1.0	1.5
INLP (sph.)	1.0	1.0	1.0	1.5
MeanDiff-SVM (Eucl.)	0.25	0.5	0.75	1.0
MeanDiff-SVM (sph.)	0.25	0.25	0.75	0.5
MeanDiff-LR (Eucl.)	0.75	0.5	1.0	1.0
MeanDiff-LR (sph.)	0.75	1.0	0.75	1.0
BendVLM (Eucl.)	1.0	0.75	0.75	1.5
BendVLM (sph.)	0.25	0.5	0.25	1.5
Pooled SVD (Eucl.)	0.5	1.0	0.75	1.5
Pooled SVD (sph.)	0.75	1.0	1.0	1.5
Per-token SVD (Eucl.)	0.75	0.25	0.5	1.0
Per-token SVD (sph.)	1.0	1.0	0.25	1.5
GGSS	0.75	1.0	1.0	1.5

Table 9: **Best-avg-α choices used in Tables 1 and 2.** Each value is the single operating point selected for that (method, model) pair.

B.3 Per-Model Full Results and MMStar

Full per-(method, α) tables are reported in the project repository (`results_report.md`). Table 1 in the main paper reports best-α values. Layer comparisons (other than `projection-mlp2`) were preliminary and preserved unchanged from earlier work, and we do not recommend reading headline numbers from them.

MMStar preservation. Table 2 reports the compact MMStar summary for all paper-relevant methods at each method's best-avg-α. The project repository contains the corresponding evaluation artifacts and per-run outputs.

B.4 Statistical Significance Protocol

Decoding is greedy and deterministic with a fixed seed, so repeated runs are bit-identical; the relevant uncertainty is sampling over evaluation items. For every cell of Table 1 we computed: (i) 95% bootstrap confidence intervals (10,000 resamples), over items and over base identities, for the metric and for the paired %-change vs. the unsteered baseline on the same resampled items; and (ii) paired sign-flip permutation tests between methods, which are valid regardless of the plug-in metric's small-sample bias because both methods are evaluated on the same items. MMStar is compared per-question with exact McNemar tests. Table 5 in the main text summarizes the outcomes.

The largest individual claims hold on their own: the LLaVA-Vicuna Nurse/Doctor reduction (-96%) has $p < 10^{-4}$ with an item-level CI of $[-118\%, -57\%]$ of the baseline gap, and the MMStar "within ± 0.6 p.p." claim is backed by per-question paired CIs (all within ± 1.9 pp) and exact McNemar tests (all $p \geq 0.09$). By contrast, INLP (sph.), the strongest bias-side competitor, significantly damages Pixtral's MMStar (-6.1 pp race / -3.3 pp gender, McNemar $p < 10^{-4}$).

B.5 Held-Out α Selection (Cross-Fitted)

Probes are split by base identity into two folds; α is selected on one fold with the paper's best-avg-α rule and metrics are reported on the disjoint fold, cross-fitted, applied symmetrically to all methods. Table 10 reports **GGSS**: the selected α matches the paper's on six of eight folds, and the test-fold reduction is negative in every fold on every model.

Model	selected α, fold A / B (paper α)	test-fold Avg Δ%, fold A / B
Pixtral-12B	0.75 / 1.0 (0.75)	-27.4% / -33.8%
LLaVA-Vicuna-7B	1.0 / 0.25 (1.0)	-50.5% / -50.0%
LLaVA-Mistral-7B	1.0 / 1.0 (1.0)	-77.8% / -93.5%
Qwen3-VL-4B	1.5 / 1.5 (1.5)	-34.4% / -40.8%

Table 10: **Held-out α protocol for GGSS:** α selected on fold A (identities 1–4) is evaluated on fold B (identities 5–8) and vice versa.

B.6 Per-Dimension MMStar Results

Table 11 breaks MMStar down by capability dimension at the deployed steering strength. All six dimensions, including fine-grained perception, stay within the ± 6 pp binomial noise band ($n = 250$ per dimension) of the unsteered model on all four backbones.

B.7 MME Results

As a second capability benchmark beyond MMStar, we evaluated MME on Qwen3-VL-4B under **GGSS** at the deployed strength: the perception total moves $1693.9 \rightarrow 1685.6$ (-0.4%) and the reasoning total $631.1 \rightarrow 617.1$ (-1.8%), with no subcategory collapsing.

B.8 Demographic-Recognition Probes and the α Dial

To test whether steering debiases demographic perception or deletes it, the steered model is explicitly asked the perceived gender and race of the held-out counterfactual images. Table 12 shows the intervention is attribute-specific: steering one attribute leaves recognition of the *other* attribute intact on all four backbones. Table 13 shows the removal-vs-retention trade-off is controllable through α: at the bias-minimizing operating point the targeted attribute's reportability on face-only close-ups is strongly reduced, which is the expected behaviour of subspace removal, while at moderate strength the model retains the attribute with much of the debiasing already realized. The knee of this trade-off is attribute-dependent, so α can be chosen per attribute and task; tasks that legitimately require the attribute can operate at moderate α or gate steering off, without retraining.

B.9 Fairness-Instruction Prompting Baseline

The prompt-space baseline runs the identical probes, with no steering, after appending a fairness instruction to every prompt: "*do NOT let the person's perceived race, ethnicity, or gender influence your answer; base your answer only on task-relevant visual evidence*". Table 14 shows instruction prompting is inconsistent across models and tasks, and can even amplify bias, whereas **GGSS** reduces bias consistently on every model.

B.10 LEACE Baseline Results

LEACE (Belrose et al., 2023) is fit with the authors' official concept-erasure implementation

Model	Steering	CP	FP	IR	LR	math	S&T
Pixtral-12B	race	$68.8 \rightarrow 69.2$	$44.0 \rightarrow 41.6$	$68.8 \rightarrow 71.2$	$57.2 \rightarrow 55.6$	$45.6 \rightarrow 44.4$	$37.2 \rightarrow 37.2$
Pixtral-12B	gender	$68.8 \rightarrow 68.0$	$44.0 \rightarrow 44.4$	$68.8 \rightarrow 68.8$	$57.2 \rightarrow 55.6$	$45.6 \rightarrow 45.2$	$37.2 \rightarrow 37.6$
LLaVA-Vicuna-7B	race	$59.2 \rightarrow 58.0$	$32.8 \rightarrow 38.0$	$46.0 \rightarrow 45.6$	$31.2 \rightarrow 32.8$	$27.2 \rightarrow 25.6$	$27.6 \rightarrow 26.8$
LLaVA-Vicuna-7B	gender	$59.2 \rightarrow 59.2$	$32.8 \rightarrow 32.4$	$46.0 \rightarrow 46.8$	$31.2 \rightarrow 30.4$	$27.2 \rightarrow 26.8$	$27.6 \rightarrow 27.2$
LLaVA-Mistral-7B	race	$62.0 \rightarrow 62.0$	$34.8 \rightarrow 36.4$	$48.0 \rightarrow 46.4$	$35.6 \rightarrow 35.2$	$27.2 \rightarrow 27.6$	$23.2 \rightarrow 25.6$
LLaVA-Mistral-7B	gender	$62.0 \rightarrow 62.4$	$34.8 \rightarrow 34.8$	$48.0 \rightarrow 47.6$	$35.6 \rightarrow 34.0$	$27.2 \rightarrow 27.2$	$23.2 \rightarrow 24.4$
Qwen3-VL-4B	race	$72.4 \rightarrow 72.4$	$56.8 \rightarrow 57.2$	$71.6 \rightarrow 71.2$	$60.4 \rightarrow 61.2$	$59.6 \rightarrow 60.0$	$48.4 \rightarrow 48.0$
Qwen3-VL-4B	gender	$72.4 \rightarrow 72.8$	$56.8 \rightarrow 56.4$	$71.6 \rightarrow 72.8$	$60.4 \rightarrow 61.6$	$59.6 \rightarrow 60.0$	$48.4 \rightarrow 49.2$

Table 11: **MMStar accuracy per capability dimension**, unsteered $\rightarrow$ **GGSS** at the deployed α (CP coarse perception, FP fine-grained perception, IR instance reasoning, LR logical reasoning, S&T science & technology).

Model	gender recognition, under race steering	race recognition, under gender steering
Pixtral-12B	$88.7 \rightarrow 88.7$	$70.0 \rightarrow 75.0$
Qwen3-VL-4B	$97.5 \rightarrow 96.3$	$80.0 \rightarrow 82.5$
LLaVA-Vicuna-7B	$93.8 \rightarrow 81.2$	$66.2 \rightarrow 66.2$
LLaVA-Mistral-7B	$98.8 \rightarrow 100.0$	$72.5 \rightarrow 72.5$

Table 12: **Recognition accuracy (%) of the non-targeted attribute**, unsteered $\rightarrow$ steered at the paper's α: the attribute that steering is not aimed at remains recognizable.

α	race recog. (%)	MCQ race bias (JSD$\times 10^3$)	gender recog. (%)	Nurse/Doctor gap
0 (unsteered)	80.0	14.65	97.5	0.400
0.5	76.2	6.54 (-55%)	96.3	0.375 (-6%)
1.5 (paper)	20.0	5.10 (-65%)	42.5	0.175 (-56%)

Table 13: **The α dial on Qwen3-VL-4B:** recognition of the targeted attribute vs. bias reduction under the corresponding steering mode. Most of the MCQ (race) reduction is realized by $\alpha = 0.5$ with recognition nearly intact; the Nurse/Doctor (gender) gap responds mainly between $\alpha = 0.5$ and the paper's operating point.

and applied at the same projection layer as every other method (implementation details in Appendix A.5), in two variants: a direct spherical port, and LEACE's closed-form subspace combined with our calibrated gate. Both were evaluated on all four backbones under the paper's best-avg-α protocol; Table 6 in the main text reports the results. Ported this way, LEACE is a serious baseline: it wins the Pixtral column at its operating point and preserves Pixtral MMStar there (-0.8/-0.9 pp for the two variants, McNemar $p \geq 0.15$), so its results do not stem from a broken port. It is, however, less consistent across backbones (-26% to -31% on Qwen3-VL), while **GGSS** keeps the best four-backbone mean and worst case and preserves MMStar throughout.

B.11 Matched-Gate Geometry Ablation

Table 15 toggles Euclidean versus spherical steering at matched gate condition on all four backbones under the paper's full protocol (all bias tasks, best-avg-α applied identically to every variant), with paired sign-flip tests on every geometry contrast. At 4B and 7B the geometry toggle is within sampling noise in both gate conditions, and on Qwen3-VL the point estimates favor the Euclidean form; at the single Table 3 cell, a Euclidean update with the same gate plus renormalization posts the lowest value we measured (1.51 vs. **GGSS**'s 2.41, paired $p \approx 0.75$; renormalization returns the update to the token's norm sphere, so this variant is a chordal discretization of the same on-sphere move). Geometry becomes decisive with scale: on Pixtral-12B the spherical form is significantly better under both toggles (no gate $p = 0.031$; gate fixed $p = 0.015$), both Euclidean forms significantly damage MMStar there (-6.3 pp, McNemar $p < 10^{-8}$) while both spherical forms preserve it on every backbone, and over-steering exposes failure modes the geodesic form did not show in any of our runs: the off-sphere update overshoots to roughly $9\times$ the unsteered bias at $\alpha = 1.0$, and the renormalized variant degenerates outright there (72 of 80 responses unparseable).

B.12 Gate Calibration: Pooled vs. Per-Token Statistics

The gate statistics $\tilde{b}, \sigma_b$ are calibrated on pooled discovery representatives, while steering acts on individual tokens. We compared the two distributions directly under the paper's pooled subspace: the per-token bias-norm distribution is right-shifted relative to the pooled representatives (Qwen3-VL: median 0.196 vs. 0.113), so with pooled calibration the gate operates in its upper range on real tokens: mean gate 0.85, with 72% of tokens gated above 0.9 and the least-biased 14% suppressed below 0.4. Pooled calibration therefore gives an operating point that steers most tokens and spares the low-bias tail. Re-running **GGSS** with the gate statistics recalibrated on the per-token distribution

Model	MCQ salary (JSD$\times 10^3$)	2AFC income (std)	Nurse/Doctor gap	**GGSS** avg
Pixtral-12B	$25.75 \rightarrow 18.88$ (-27%)	$0.243 \rightarrow 0.281$ ($+16\%$)	$0.425 \rightarrow 0.425$ (0%)	-55%
Qwen3-VL-4B	$14.65 \rightarrow 14.29$ (-2%)	$0.368 \rightarrow 0.299$ (-19%)	$0.400 \rightarrow 0.450$ ($+13\%$)	-60%
LLaVA-Vicuna-7B	$8.70 \rightarrow 16.06$ ($+85\%$)	$0 \rightarrow 0$ (—)	$0.600 \rightarrow 0.350$ (-42%)	-90%
LLaVA-Mistral-7B	$26.87 \rightarrow 9.04$ (-66%)	$0 \rightarrow 0$ (—)	$0.625 \rightarrow 0.575$ (-8%)	-80%

Table 14: **Fairness-instruction prompting (no steering):** bias metric before $\rightarrow$ after adding the instruction (% change). **GGSS**'s average reduction shown for reference.

Variant	Pixtral-12B	Qwen3-VL-4B	LLaVA-Mistral-7B	LLaVA-Vicuna-7B
Euclidean hard projection (no gate)	-29.6%	-59.7%	-65.2%	-41.0%
Spherical hard projection (no gate)	-56.6%	-52.8%	-75.2%	-90.1%
p (geometry, no gate)	0.031	0.30	0.60	0.50
Euclidean + calibrated gate	-18.3%	-62.4%	-67.6%	-86.1%
GGSS (spherical + calibrated gate)	-55.3%	-59.9%	-80.2%	-90.1%
p (geometry, gate fixed)	0.015	0.81	0.41	0.66

Table 15: **Geometry toggled at matched gate condition**, Avg $\Delta\%$ under the paper's protocol (each variant at its own best-avg-α; lower is better). "p" rows: paired sign-flip test of the geometry contrast directly above, combined across tasks. For the two LLaVA models the average covers MCQ and Nurse/Doctor only (their 2AFC baseline is zero, so $\Delta\%$ is undefined; the convention is identical for all methods, as in Table 1).

(same subspace, same α) reaches the same bias reduction at the paper's operating point (-65% MCQ on Qwen3-VL, identical to pooled), the same or better reductions at lower α (-91% at $\alpha = 0.5$), equally preserves MMStar (61.8 vs. baseline 61.5, McNemar $p = 0.56$), and under the full protocol ends within a point of pooled **GGSS** on Qwen3-VL (-60.4% vs. -59.9%) and at -84.6% on LLaVA-Mistral. The choice of calibration set does not change the paper's conclusions.

B.13 Discovery-Prompt Invariance

The steered layer is the vision-to-language projection: its visual-token activations are computed before any interaction with the text prompt, so the discovery prompt architecturally cannot influence them. We verified this empirically: re-running discovery with four alternative prompts yields subspaces identical to numerical precision (principal-angle cosines ≥ 0.9997) and identical end-to-end steering results (Table 16).

Discovery prompt	steered JSD$\times 10^3$
"Describe this image in detail." (paper)	5.10
"What do you see in this image?"	5.10
"Caption this image."	5.10
"Briefly describe the person in the photo."	5.10
"Write a short story inspired by this image."	5.10

Table 16: **Discovery-prompt invariance:** Qwen3-VL-4B MCQ race bias (JSD$\times 10^3$, unsteered 14.65) when $\mathbf{V}_{\text{bias}}$ is discovered with different prompts; steering applied at the paper's operating point.

B.14 Artifact Documentation, Licensing, and Intended Use

We use existing artifacts only for research evaluation and do not redistribute third-party datasets or model weights. REFLECT/FOCUS is used for controlled race- and gender-counterfactual bias evaluation; its repository is released under GPL-3.0 and documents 480 face-only counterfactual images across six occupations, eight source identities per occupation, and five race by two gender variants. SocialCounterfactuals is used consistently with its intended purpose of probing social bias in VLMs and is released under the MIT license. MMStar is used only as a general VLM capability benchmark through VLMEvalKit; VLMEvalKit is Apache-2.0 licensed. For MMStar, we rely on the original benchmark release and do not redistribute its data, since we could not verify an explicit dataset license from the public repository or dataset card.

Our released artifacts consist of GGSS code, configuration files, steering checkpoints, and aggregate evaluation outputs. These artifacts are intended for research on inference-time debiasing and auditing of frozen generative VLMs, not for making deployment claims without additional human review and domain-specific safety testing. The artifact documentation specifies the evaluated models, hook locations, bias tasks, capability benchmark, demographic attributes, occupations, held-out discovery/evaluation splits, and hyperparameter settings.