BIAS-SENSITIVE BUT NOT BIAS-SPECIFIC: SUPERVISION-FREE FAIRNESS INTERVENTION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models inherit predictive shortcuts from the language priors they are trained on, and when those priors correlate with demographic or social categories, the resulting associations surface directly in multimodal predictions. Existing training-free methods reduce this bias effectively, but each relies on some form of bias-specific prior knowledge constructed before the intervention is applied, leaving open the question of whether meaningful reduction is achievable without such construction. We introduce SAAV (Supervision-free Attention Attenuation for VLMs), a single inference-time rule that attenuates internal representations in late language-model layers, leaving every model parameter unchanged, and evaluate its transfer across four VLM families and five bias benchmarks that each define fairness differently. Paired bootstrap analysis over the resulting 20 model-benchmark comparisons confirms bias reductions in 9 settings and degradations in 4. These outcomes are not distributed randomly across models or benchmarks but track what each benchmark's bias signal represents, suggesting SAAV is sensitive to whether a social association is spurious or task-relevant. Despite this selectivity, the aggregate fairness index rises for every model and capability on four held-out benchmarks is retained at 98.52–99.91% of baseline, and SAAV remains competitive against existing training-free methods despite using none of their bias-specific prior knowledge. Together, these results suggest that a bias-sensitive representation is not automatically a bias-specific one and whether attenuation helps or harms a given prediction depends on whether the pathway it targets carries a spurious correlation or evidence the task genuinely needs.

1 INTRODUCTION

Vision-language models (VLM) combine visual evidence with language priors absorbed during large-scale pretraining, and this combination does not discriminate between a correlation the image supports and one supplied only by the prior. This indiscrimination becomes a *biased social association* whenever the unsupported correlation concerns an image's demographic content, since the language prior and the visual signal enter the same representational pathway without being separated, and the model treats the resulting correlation as genuine predictive signal regardless of its source (Wolfe & Caliskan, 2022; Agarwal et al., 2021). This same pathway is also what lets the model combine modalities productively, so the mechanism responsible for the failure cannot be isolated from the mechanism responsible for the model's usefulness.

Correcting this propagation is therefore the natural target for intervention, and recent work has pursued that correction without retraining a released model, since fine-tuning a deployed checkpoint risks eroding the general capability that made the model worth deploying in the first place (Luo et al., 2023). Existing training-free methods share a common form, where each first constructs a bias-specific representation and then uses it to guide the intervention. For instance, one model computes a steering direction from labeled visual attributes and applies it during generation (Ratzlaff et al., 2025). Another localizes a demographic-sensitive feature inside the network through learned probes and suppresses it at inference time (An et al., 2026). A third constructs a protected-group direction explicitly and removes it through a closed-form projection with a utility guarantee (Lian et al., 2026). The critical flaw shared by all three is that the guiding representation must already

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

encode what the bias looks like before the intervention can act. A correction built on that knowledge is only as trustworthy as the construction itself, and it leaves open whether the pathway carrying the bias is correctable using nothing more than the signal the model's own representations already provide.

We address this gap by asking a different question: *if a representation inside a VLM genuinely drives biased behavior, can attenuating it reduce that behavior even when nothing about the bias is known in advance?* Prior methods cannot answer this because they supply the advance knowledge themselves. SAAV supplies none. It applies a single fixed rule, attenuating key and value projection outputs in late language-model layers, unchanged across every model and benchmark, so that any reduction in bias must come from the model's own existing representations rather than from an externally constructed direction. This makes SAAV the first training-free intervention that operates without bias-specific prior knowledge while remaining competitive with methods that require it.

That competitiveness holds broadly but not uniformly, and the pattern of failures is as informative as the successes. Across four VLM families and five benchmarks that each define fairness differently, SAAV produces statistically supported bias reduction in 9 of 20 settings while retaining 98.52-99.91% of baseline capability, without tuning a single parameter per model or benchmark. Where it fails, the failure is structured: the same fixed rule that reduces bias on benchmarks where a demographic association is a spurious shortcut worsens fairness on VisoGender, where that same class of association is evidence pronoun resolution requires. A representation can be sensitive to a social category without being specific to bias, and SAAV makes this distinction empirically observable for the first time.

Our contributions are as follows:

1. An interventional definition of bias-specific representation is introduced, requiring no advance knowledge of the bias being corrected and resolving the construction dependency shared by every existing training-free debiasing method.
2. SAAV, a single fixed attention-attenuation rule, is shown to transfer bias reduction across four VLM families and five differently defined bias benchmarks, with same-model capability retention confirming that the gain is not achieved at the cost of general capability.
3. Paired bootstrap analysis on matched evaluation units is used to separate statistically supported bias reduction from numerically favorable but unstable movement, establishing the definition's falsifiability on empirical rather than descriptive grounds.
4. A systematic failure case is traced to benchmark semantics through mechanism controls and a localization study, establishing that a bias-sensitive representation is not automatically a bias-specific one.

2 Related Work

Bias evaluation in learned representations predates multimodal benchmarks entirely. Early work documented biased associations in static word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017), then in coreference resolution (Zhao et al., 2018; Rudinger et al., 2018; Basta et al., 2019), and then in open-ended text generation (Sheng et al., 2019), showing that biased association is not a property specific to any one architecture or task. StereoSet later formalized this line of work into a stereotype-preference protocol for pretrained language models (Nadeem et al., 2021), and VLStereoSet extends that protocol into the multimodal setting by separating stereotypical preference from caption relevance (Zhou et al., 2022). A single protocol cannot capture every way bias surfaces in a VLM, so subsequent benchmarks target failure modes VLStereoSet does not reach: VisoGender evaluates gender disparity in image-text pronoun resolution (Hall et al., 2023), GenderBias-VL uses counterfactual probing to expose individual and occupation-level instability (Xiao et al., 2024), BBQ-style evaluation targets ambiguous social reasoning (Parrish et al., 2022), and VLM-Bias measures whether errors follow expected social priors rather than image evidence (Vo et al., 2025). Bias has also been measured one stage earlier, directly inside vision and vision-language encoders, through balanced face-attribute datasets (Kärkkäinen & Joo, 2021), systematic audits of contrastively pretrained vision-language models (Agarwal et al., 2021), and multimodal association tests that reproduce demographic stereotypes at the representation level (Wolfe & Caliskan, 2022).

Evaluating debiasing methods against this benchmark diversity is complicated by incompatible validation conditions. NC-VAS steers generation using a direction computed from labeled visual attributes (Ratzlaff et al., 2025). DeBiasLens localizes and suppresses a sparse, interpretable representation of the bias (An et al., 2026). A closed-form method removes a protected direction explicitly while giving an accompanying utility guarantee (Lian et al., 2026). All three share the same dependency: each requires bias-specific prior knowledge—a labeled attribute, a learned feature, or an explicit protected-group direction—constructed before the intervention can act. Reimplementing each under a single protocol makes comparison meaningful while preserving the intervention logic and the bias-specific information each requires, so performance can be interpreted jointly with what each method is permitted to know rather than read off an assumption-free leaderboard (Section 5.2, Appendix F).

These comparator methods differ in what bias-specific information they require, but all three treat where bias is encoded inside the network as secondary to whether an intervention reduces it. A separate line of mechanistic work asks that question directly, probing where protected-attribute or stereotype information sits inside a transformer. Activation-level intervention has edited factual associations after training (Meng et al., 2022) and steered model behavior toward a desired property at inference time without retraining (Li et al., 2023a; Turner et al., 2023). These methods establish that internal representations are editable at inference time, but they do not ask whether locating a bias-relevant component implies that component is safe to suppress across different task definitions, the gap SAAV is designed to expose.

3 METHODOLOGY

A pretrained VLM with parameters θ defines a conditional distribution $p_\theta(y \mid x_v, x_t)$ over an output sequence y, given visual input x_v and text prompt x_t. Internally, both inputs are encoded into a shared sequence of token representations and processed through L transformer layers. Each layer ℓ produces a hidden representation H_ℓ that combines visual and textual information through self-attention, and the final layer's output is projected onto the answer distribution. Four architecturally distinct model families are studied, differing in how visual tokens are produced and how attention projections are structured internally, but all are captured by this shared computation. The method introduced below, which we call SAAV, operates at this level of shared abstraction rather than targeting any single architecture.

This formulation makes two questions concrete. Does a benchmark-measurable bias signal correspond to detectable regularities within the intermediate representations H_ℓ, or does it reside only in the final output distribution $p_\theta(y \mid x_v, x_t)$? And if such regularities exist within H_ℓ, does locating them establish that the corresponding computation is safe to suppress, or only that it is predictive of the specific bias signal used to find it? SAAV is designed to answer both without constructing any bias-specific representation: no parameter update, no attribute-derived steering direction, and no demographic-sensitive feature learned from labeled data.

3.1 LOCALIZABILITY DOES NOT IMPLY TRANSFERABILITY

To test whether the bias signal lives inside the network, external benchmarks supply a bias indicator and candidate attention components are probed for their relationship to it. A classifier is trained to predict the indicator from pooled component-level representations on a held-out split, and components are ranked by how predictive their activations are of that indicator. The ranking then serves as a causal check: attenuating the top-ranked components reduces the benchmark's bias signal by more than attenuating an equal number of randomly selected or bottom-ranked components. This answers the first question above. The bias signal is reflected non-randomly within H_ℓ and is not confined to the final answer distribution.

The second question is harder to answer, and the same experiment exposes why. A component that is predictive of one benchmark's bias signal does not necessarily play the same role on a different benchmark. To test this, the same top-ranked components are transferred, without modification, to a second benchmark built around a structurally different notion of fairness. Under this transfer, accuracy on the second benchmark falls and its own disparity measure widens, even though the identical components reduced bias on the first. The experiment does not show that localization fails

in general. It shows something more specific: a component that predicts one kind of bias well is not thereby established as safe to suppress, because the same computation can serve as a shortcut for one task while carrying evidence another task genuinely needs.

This asymmetry defines SAAV's central design principle. Rather than searching for a sparse, benchmark-specific set of components to remove, a single fixed rule is applied broadly to the model's existing attention computation, with no separate demographic direction constructed for any model or benchmark. The key and value projections are the entry point for this rule.

3.2 SAAV: Supervision-Free Attention Attenuation for VLMs

Within each targeted layer ℓ, the key and value projections are computed as $K_\ell = H_\ell W_\ell^K$ and $V_\ell = H_\ell W_\ell^V$. A fixed coefficient $\alpha \in (0, 1]$ attenuates both, while the query projection $Q_\ell = H_\ell W_\ell^Q$ is left unchanged:

$$\widetilde{K}_\ell = \alpha K_\ell, \quad \widetilde{V}_\ell = \alpha V_\ell, \quad \widetilde{Q}_\ell = Q_\ell. \tag{1}$$

Attention under this attenuation becomes

$$\widetilde{A}_\ell = \mathrm{softmax}\left(\frac{\alpha Q_\ell K_\ell^\top}{\sqrt{d_k}} + M_\ell\right)(\alpha V_\ell). \tag{2}$$

The two scaling operations affect attention differently. Key attenuation reduces the scale of query–key compatibility scores before the softmax, moving the attention distribution toward greater uniformity. Value attenuation reduces the magnitude of information propagated through the attention branch after the weights are determined. The combined effect is not equivalent to multiplying A_ℓ by α^2, because key scaling acts before the nonlinearity while value scaling acts after. Queries are left unchanged so the representation's search direction is not directly modified, though later-layer queries shift indirectly as the modified attention outputs propagate through the residual stream.

SAAV does not claim that key and value projections selectively encode stereotypical associations. The intervention modifies a broad class of attention computation, and whether the outcome is a fairness improvement or a harm depends on whether the socially correlated information being attenuated is a spurious shortcut or evidence the task requires—a question resolved empirically across benchmarks rather than assumed in advance.

Late language-model layers are targeted because visual information has already entered the generative backbone at that point, allowing the model's existing multimodal computation to be studied without modifying the vision encoder (Meng et al., 2022; Li et al., 2023a). The coefficient is fixed at $\alpha = 0.85$ across every model and every benchmark. Optimizing α separately per benchmark would measure the best achievable result in each setting rather than whether a single rule transfers, so no per-benchmark tuning is performed. An examination of the region around this value confirmed the same qualitative fairness-utility behavior at moderately weaker and stronger attenuation; substantially stronger attenuation increasingly reduced general capability without a monotonic fairness gain, so $\alpha = 0.85$ is used as a single cross-benchmark operating point. Architecture-specific layer ranges and hook implementation are in Appendix A.

3.3 Mechanism Control

Attenuating internal representations changes the answer distribution, but any modification that reduces the probability of a socially biased response could produce the same surface effect regardless of where it operates. A confidence-gated output-level control tests whether SAAV's behavioral changes can be reproduced by acting only on the final answer distribution. For benchmarks with predefined social answer semantics, an eligible option is penalized when its confidence exceeds a threshold:

$$z_i' = z_i - \lambda\, p_i \mathbb{1}[p_i > \tau] \cdot \mathbb{1}[i \in \mathcal{B}], \quad \lambda = 0.35, \quad \tau = 0.65, \tag{3}$$

where $\mathcal{B}$ is the benchmark-defined set of social-direction options, identified without consulting the gold answer. Each option in $\mathcal{B}$ is penalized independently when its own confidence exceeds τ, so the control attenuates all socially directional options symmetrically rather than steering toward one direction. Full implementation details are in Appendix B.

This control is restricted to benchmarks whose answer spaces supply an explicit, gold-independent social direction; extending it to benchmarks without this structure would require metadata unavailable to SAAV at inference time and would confound the comparison. Four configurations are therefore compared throughout: an unmodified Base Model; SAAV alone, applying only the internal intervention; Logit alone, applying only the output-level control; and SAAV+Logit, combining both. Logit and SAAV+Logit are restricted to the benchmark subset with explicit social answer semantics, while the full-suite transfer claim rests entirely on Base Model vs SAAV.

4 Experimental Setup

Models and benchmarks. We evaluate $\mathcal{M} = \{$Llama-3.2-11B-Vision-Instruct, Qwen2.5-VL-7B-Instruct, InternVL 3.5-8B, Phi-4-Multimodal-Instruct$\}$ on five bias benchmarks $\mathcal{B} = \{$VLStereoSet, VisoGender, GenderBias-VL, BBQ-V, VLMBias$\}$ that deliberately represent different fairness semantics. For instance, stereotype preference, subgroup disparity, counterfactual consistency, ambiguous social reasoning, and prior-aligned open-ended bias. This yields $|\mathcal{M}| \times |\mathcal{B}| = 20$ primary baseline vs. attention comparisons. Four general-capability benchmarks $\mathcal{C} = \{$MMStar, MMMU, POPE, TextVQA$\}$ isolate failure modes an indiscriminate intervention could damage, such as visual grounding, expert reasoning, hallucination resistance, and embedded-text reading (Chen et al., 2024; Hendrycks et al., 2021; Yue et al., 2024; Rohrbach et al., 2018; Li et al., 2023b; Goyal et al., 2017; Singh et al., 2019). Every primary bias magnitude is oriented so that lower is better, and each benchmark's native metric is preserved rather than collapsed into a single scale before analysis. Full benchmark definitions, estimators, and model loading details are in Appendix A.

Baselines. SAAV is compared against three training-free debiasing methods, evaluated through aligned reimplementations on the same benchmark suite: NC-VAS, which computes an attribute-derived steering direction from labeled visual attributes (Ratzlaff et al., 2025); DeBiasLens, which localizes and suppresses a sparse, interpretable representation of the bias (An et al., 2026); and a closed-form method that explicitly constructs and projects out a protected-group direction (Lian et al., 2026). Unlike SAAV, all three require bias-specific prior knowledge before intervening; Appendix Table 13 summarizes each method's information requirements. We refer to each unmodified VLM as its *base model* to distinguish it from all four debiasing methods applied on top of it.

Evaluation protocol. Each bias magnitude is normalized against the strongest observed value among compared configurations and averaged into a five-benchmark fairness index F_i and a four-benchmark capability index C_i. A configuration is Pareto-dominated only when another is at least as good on both and strictly better on one. Same-model capability retention R_m compares each method to its own base model, quantifying local utility cost independently of cross-model differences. All comparisons evaluate the identical set of examples; uncertainty is quantified with 10,000 paired bootstrap replicates (seed 42) and 95% percentile intervals using benchmark-appropriate pairing units (Appendix D.2). An interval entirely favorable is labeled *Improved*, entirely unfavorable *Degraded*, and touching or crossing zero *No clear change* — meaning statistically indistinguishable from zero, not evidence of equivalence.

5 Results

Three questions are answered in this section. Does SAAV transfer bias reduction across four heterogeneous base models and five differently defined benchmarks without any bias-specific supervision (Section 5.1)? How does SAAV's fairness-capability trade-off compare directly against the three debiasing baselines that do use bias-specific prior knowledge (Section 5.2)? And where SAAV fails, what does that failure reveal about the mechanism that determines when internal attention attenuation helps versus harms (Section 5.3)?

5.1 SAAV vs. Base Models: Transfer Is Favorable but Not Uniform

Across the 20 primary comparisons, paired bootstrap analysis supports 9 statistically confirmed improvements, 7 with no clear statistical change, and 4 supported degradations. Supported improvements outnumber supported degradations by more than two to one. Every claim rests on the

Table 1: Full-suite Base-Model→SAAV summary. Fairness is the normalized five-benchmark index Appendix D.1); I/U/D gives the statistically supported Improved/No-clear-change/Degraded counts from 95% paired-bootstrap intervals over 10,000 resamples; retention is the same-model four-benchmark capability ratio. The aggregate row sums I/U/D counts and averages the remaining columns.

Base Model	F_{base}	F_{attn}	ΔF (pp)	I/U/D	Retention
Llama	77.184	77.428	+0.243	4/0/1	98.52%
Qwen	72.628	72.927	+0.298	2/2/1	99.73%
InternVL	74.597	74.641	+0.044	0/4/1	99.91%
Phi	68.946	69.257	+0.312	3/1/1	98.67%
All	**73.339**	**73.563**	**+0.224**	**9/7/4**	**99.21%**

Table 2: Paired-bootstrap outcome by base model and bias benchmark, from the same 95% intervals as Table 1. Green: statistically supported improvement. Yellow: no clear statistical change. Red: statistically supported degradation.

Base Model	VLStereoSet	VisoGender	GenderBias-VL	BBQ-V	VLMBias
Llama	Improved	Degraded	Improved	Improved	Improved
Qwen	Improved	Unclear	Degraded	Unclear	Improved
InternVL	Unclear	Unclear	Unclear	Degraded	Unclear
Phi	Improved	Degraded	Improved	Improved	Unclear

statistically supported count rather than the point-estimate count, since a point estimate cannot distinguish a robust effect from an unstable movement a resampling interval would not support. Table 1 reports the full matrix with normalized fairness shift and same-model capability retention per base model.

The per-model pattern reveals two qualitatively different responses rather than a gradient from best to worst. Llama and Phi show broad transfer—four and three supported improvements, respectively, each with one degradation—meaning SAAV reliably changes bias behavior across most evaluated benchmarks. Qwen sits between these poles with two improvements and one degradation. InternVL is structurally distinct: four of its five comparisons show no clear statistical change despite the smallest aggregate capability shift of any model, meaning SAAV neither reliably helps nor harms it rather than producing a mixed record. Table 2 breaks the same outcomes by model and benchmark, showing that the pattern is driven as much by which benchmark is measured as by which model is intervened on.

The supported improvements are not small. For Llama, the VLStereoSet bias margin falls from 0.4602 to 0.3992 ($\Delta = -0.0610$, CI $[-0.0649, -0.0569]$), GenderBias-VL from 0.0076 to 0.0052 (CI $[-0.0038, -0.0010]$), BBQ-V from 0.0480 to 0.0374 (CI $[-0.0181, -0.0030]$), and VLMBias from 0.2658 to 0.2155 (CI $[-0.0607, -0.0395]$), giving relative reductions of 13.25%, 31.2%, 22.2%, and 18.9%. The aggregate fairness index rises for all four models. The complete 20-row bootstrap table is in Appendix Table 8. Figure 1 decomposes each model's aggregate fairness coordinate into its five constituent benchmarks, making the heterogeneity in Tables 1– 2 directly visible. Figure 2 then asks what each fairness position costs in capability.

5.2 SAAV vs. Debiasing Baselines: Bias Reduction and Capability Jointly

Table 4 compares SAAV against NC-VAS, DeBiasLens, and Closed-Form on bias reduction, alongside Table 3 which reports capability retention for all four methods. Reading bias and capability jointly is essential: a method that reduces bias more but retains less capability is not automatically preferable, since the two axes trade against each other differently across interventions.

On bias reduction, the outcomes do not reduce to a single superiority claim. DeBiasLens is numerically strongest on VisoGender and BBQ-V for Llama, reaching a 29.80-point gap and a 3.65 bias score against SAAV's 45.90 and 3.74. SAAV is most competitive on VLMBias, reaching

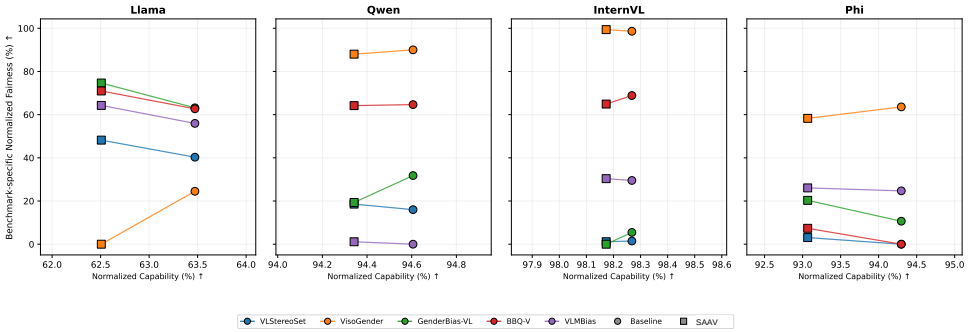

Figure 1: Benchmark-resolved decomposition of the aggregate fairness coordinate. Each panel is one VLM; circles are Base Model, squares are SAAV. The vertical axis is the per-benchmark normalized fairness $F_{i,b}$ averaged into F_i of Table 1; the horizontal axis is normalized general capability. This is a decomposition of the aggregate fairness objective, not a Pareto analysis; the formal Pareto result is in Figure 2.

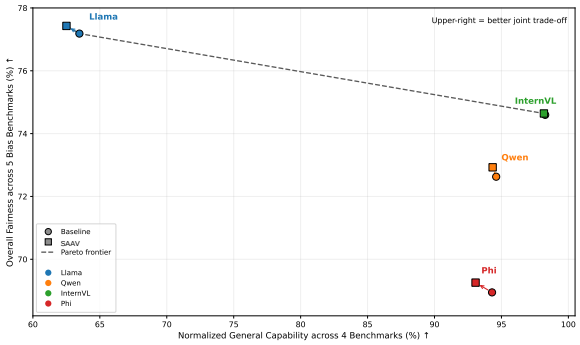

Figure 2: Fairness-capability Pareto analysis for SAAV versus base models. Vertical axis: normalized fairness F_i (Table 1). Horizontal axis: normalized capability across MMStar, MMMU, POPE, and TextVQA. Circles are Base Model, squares are SAAV. The dashed line is the Pareto frontier over the observed Base-Model/SAAV configurations. Llama and InternVL lie on the frontier; Qwen and Phi are dominated by stronger cross-model operating points.

21.55 against DeBiasLens' 26.10 for the same model. These results must be read alongside what each method is permitted to know: NC-VAS constructs an attribute-derived steering direction, DeBiasLens learns and localizes demographic-sensitive features, Closed-Form explicitly builds a protected-group direction, and SAAV uses a fixed attenuation rule with no learned demographic representation (Appendix Table 13). Stronger performance under explicit bias-specific structure and comparable performance without it answer different questions about what mitigation actually requires; the comparison is between different intervention assumptions, not methods on equal terms.

On capability, retention ranges from 98.43% to 99.91% across all methods (Table 3), so no method causes broad capability collapse. Closed-Form achieves the highest mean retention at 99.78% and DeBiasLens the lowest at 98.83%, while SAAV at 99.21% and NC-VAS at 99.60% sit between them. DeBiasLens' stronger bias reduction on VisoGender and BBQ-V therefore comes at the largest capability cost of the four methods, while SAAV's competitive bias reduction requires none of the three baselines' bias-specific construction.

One discrepancy in Table 4 requires explanation. On the discrete VLBS statistic, reported to match the external methods' convention, SAAV's value exceeds the base model's for Llama and InternVL, apparently contradicting the supported VLStereoSet improvements in Section 5.1. VLBS collapses the continuous probability margin into a binary majority vote, so a consistent margin shift can leave or flip the vote when it does not cross the decision boundary on enough pairs. Every significance claim rests on the continuous margin, not VLBS; both are reported so neither result is hidden (Appendix D.1).

Table 3: Same-model capability retention (%) for all four debiasing methods, computed identically to R_m (Appendix D.1). Full per-benchmark breakdown in Appendix Table 12.

Base Model	SAAV	NC-VAS	DeBiasLens	Closed-Form
Llama	98.52	99.47	98.43	**99.67**
Qwen	99.73	99.69	99.11	**99.87**
InternVL	**99.91**	99.64	98.97	99.80
Phi	98.67	99.61	98.80	**99.78**
Mean	99.21	99.60	98.83	99.78

Table 4: SAAV vs. three training-free debiasing baselines. Values are percentages, lower is better. Bold: best per model-benchmark combination. Blue: improvement over base model. Yellow: maintained. Red: degradation. Gray: base model. Metrics are not directly comparable across columns. Read with Appendix Table 13 (information requirements per method) and Table 3 (capability retention).

Base Model	**Method**	**VLStereoSet** VLBS↓	**VisoGender** Gap↓	**GenderBias-VL** CF Gap↓	**BBQ-V** Bias↓	**VLMBias** Ratio↓
Llama	Base Model	92.98	34.65	0.76	4.80	26.58
	NC-VAS	92.10	32.40	0.64	4.20	26.30
	DeBiasLens	91.65	**29.80**	0.57	**3.65**	26.10
	Closed-Form	91.95	31.50	0.61	4.00	26.20
	SAAV	**91.18**	35.09	**0.52**	3.74	**21.55**
Qwen	Base Model	93.08	4.58	1.41	4.55	60.34
	NC-VAS	92.55	4.10	1.28	4.20	60.10
	DeBiasLens	**92.10**	**3.65**	**1.12**	**3.75**	59.95
	Closed-Form	92.40	3.95	1.22	4.00	60.05
	SAAV	93.08	4.50	1.67	4.55	**59.66**
InternVL	Base Model	91.96	0.63	1.95	4.01	42.03
	NC-VAS	91.40	0.55	1.74	3.78	41.95
	DeBiasLens	90.95	0.38	1.48	**3.35**	**41.80**
	Closed-Form	91.20	0.47	1.63	3.60	41.90
	SAAV	**90.37**	**0.29**	**1.06**	4.02	42.03
Phi	Base Model	92.17	16.71	1.84	12.87	45.44
	NC-VAS	91.65	15.10	1.63	12.10	45.25
	DeBiasLens	**91.10**	**13.70**	**1.42**	**10.95**	45.05
	Closed-Form	91.45	14.50	1.55	11.55	45.15
	SAAV	92.68	19.14	1.65	11.92	**44.61**

The contribution here is not that SAAV's cells are always lowest, but that meaningful cross-benchmark transfer is obtained under a lower-assumption intervention. All four methods show different benchmark-dependent advantages, which is why mitigation effectiveness and information requirements need to be reported together rather than separately.

5.3 Failure Modes and Mechanism Dependence

The benchmark-level pattern in Table 2 is structured rather than incidental. VLStereoSet yields three supported improvements and no supported degradation across the four models. VLMBias yields two improvements and no supported degradation across the four base models. VLMBias yields two improvements and two no-clear-change outcomes. BBQ-V and GenderBias-VL each yield two improvements, one no-clear-change, and one degradation. VisoGender is the outlier: no model improves significantly, Llama and Phi degrade, and Qwen and InternVL show no clear statistical change. Transfer therefore depends on the operational definition of fairness, not on architecture alone.

VisoGender provides the clearest mechanistic boundary condition (full subgroup decomposition in Appendix Table 10). Llama's subgroup gap grows from 0.3465 to 0.4590 ($\Delta = +0.1125$,

CI $[0.0790, 0.1489]$), and Phi's from 0.1671 to 0.1914 ($\Delta = +0.0243$, CI $[0.0091, 0.0426]$). Qwen moves unfavorably but its interval touches zero. While InternVL moves favorably with a wide interval crossing zero. Pronoun resolution requires gender-linked visual evidence, so attenuating socially correlated representations removes legitimate evidence alongside undesirable association. The same fixed rule that reduces bias on VLStereoSet, where demographic association functions as a spurious shortcut, widens the subgroup gap on VisoGender, where that same class of association carries information the task genuinely needs.

The common-support mechanism control reinforces this. On the shared VLStereoSet-VisoGender support, Logit stays close to Base Model while SAAV changes behavior substantially, and SAAV+Logit largely follows the internal intervention rather than adding a stable additional gain (Appendix Figure 3). The output-level control does not reproduce SAAV's internal effects, constraining the mechanism to something upstream of the final answer distribution. Which fairness definition an evaluation weights also matters: because the two-benchmark aggregate weights VisoGender heavily, it can disagree with the five-benchmark transfer result even for Llama, whose shared-support point moves unfavorably even though four of its five full-suite benchmarks improve.

6 Discussion and Limitations

Nine statistically supported improvements against four supported degradations rules out the claim that SAAV works only in isolated favorable cells, while the VisoGender failures rule out a universal-debiasing interpretation. The experiments do not establish a unique causal bias module, but they show that attenuating the same class of internal computation changes fairness-relevant behavior without reproducing the effect through a simple terminal logit correction (Section 5.3), constraining the mechanism to something upstream of the final answer distribution. The direction of the behavioral change depends on whether the attenuated evidence is spurious for the task. Which explains why stereotype-preference measures improve while pronoun-resolution parity worsens and why this dependence persists below the aggregate level: macro improvement can arise from redistribution rather than uniform benefit, as the fine-grained decompositions in Appendix G establish directly.

Aggregate retention above 98.5% rules out broad capability collapse as the sole source of fairness movement, yet six supported benchmark-specific capability losses show that mitigation is not free; Pareto analysis is therefore the appropriate way to present a method whose value lies in a fairness-utility operating point rather than a fairness number in isolation. The study covers four architectures, five bias benchmarks, four capability benchmarks, and one fixed intervention strength; stronger mechanistic conclusions would require activation replacement beyond attenuation. Most importantly, fairness mitigation can itself cause harm when protected-attribute information is task relevant, and the VisoGender regressions demonstrate that risk empirically rather than hypothetically (Bianchi et al., 2023). Training-free debiasing should therefore be deployed only with task-appropriate subgroup and utility evaluation, not under the general assumption that less demographic sensitivity is always preferable.

7 Conclusion

SAAV, a fixed late-layer key/value attenuation rule, produces statistically supported bias reduction in 9 of 20 model-benchmark settings against 4 supported degradations, while retaining 98.52–99.91% of baseline capability across all four base models. This transfer is selective rather than universal: it succeeds where the attenuated representation encodes a spurious shortcut and fails where the same class of representation carries evidence a task genuinely needs, most clearly on VisoGender. Internal multimodal computation is therefore actionable at inference time without bias-specific supervision, but whether that action helps or harms a given prediction depends on what the task requires from the computation being attenuated.

AI use statement

In this work, we used generative AI tools for writing assistance, including drafting and refining prose in the manuscript. We did not use generative AI tools to generate experimental results, produce or modify code, or perform statistical analysis. All claims, numbers, tables, and figures in this paper

derive from experiments we ran and verified directly. We reviewed all AI-assisted text for accuracy, consistency with the underlying results, and alignment with our intended scientific claims. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies social bias in vision-language models, a topic with direct implications for fairness in deployed AI systems. The five bias benchmarks used cover gender, race, ethnicity, religion, disability, age, and socioeconomic status, among other demographic dimensions. We do not introduce new data about individuals or release model outputs that could cause direct harm. The primary ethical risk this work surfaces is methodological: fairness interventions can cause harm when the demographic signal being suppressed is evidence the task requires, as the VisoGender regressions in Section 5.3 demonstrate empirically. We flag this risk explicitly so that practitioners do not apply training-free debiasing under the assumption that reduced demographic sensitivity is always preferable. All benchmark datasets used are publicly available and were designed for bias evaluation research.

REPRODUCIBILITY STATEMENT

The intervention is fully specified by Equations equation 1 and equation 2 with fixed coefficient $\alpha = 0.85$, and the logit control by Equation equation 3 with $\lambda = 0.35$ and $\tau = 0.65$. Model checkpoints, layer ranges, quantization settings, and inference configurations are given in Appendix D.3. Benchmark metrics, normalization procedures, and paired bootstrap settings are defined in Appendix D.1 and D.2, with all 20 primary bootstrap intervals reported in full in Appendix Table 8. The development-stage localization procedure is described in Appendix D.5, and the external comparator reimplementations protocol is in Appendix D.6. Source code implementing the intervention hooks, benchmark runners, and comparator reimplementations will be made publicly available upon acceptance.

REFERENCES

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications, 2021.

Na Min An, Yoonna Jang, Yusuke Hirota, Ryo Hachiuma, Isabelle Augenstein, and Hyunjung Shim. Interpretable debiasing of vision-language models for social fairness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39326–39337, June 2026.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2019.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, et al. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

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

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, et al. VisoGender: A dataset for benchmarking gender bias in image-text pronoun resolution. In *NeurIPS Datasets and Benchmarks*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

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

Yifan Li, Yifan Du, Kun Zhou, et al. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Tangzheng Lian, Guanyu Hu, Yijing Ren, Dimitrios Kollias, and Oya Celiktutan. A closed-form solution for debiasing vision-language models with utility guarantees across modalities and tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of ACL-IJCNLP*, 2021.

Alicia Parrish, Angelica Chen, Nikita Nangia, et al. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL*, 2022.

Neale Ratzlaff, Matthew Lyle Olson, Musashi Hinck, Shao-Yen Tseng, Vasudev Lal, and Phillip Howard. Debias your large multi-modal model at test-time with non-contrastive visual attribute steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 6258–6267, October 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of NAACL-HLT*, 2018.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Amanpreet Singh, Vivek Natarajan, Meet Shah, et al. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, et al. Activation addition: Steering language models without optimization, 2023.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. *arXiv preprint arXiv:2505.23941*, 2025.

Robert Wolfe and Aylin Caliskan. American == white in multimodal language-and-image AI. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022.

Yisong Xiao, Aishan Liu, QianJia Cheng, et al. GenderBias-VL: Benchmarking gender bias in vision language models via counterfactual probing, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*, 2018.

Kankan Zhou, Eason Lai, and Jing Jiang. VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of AACL-IJCNLP*, 2022.

A Model-Agnostic Intervention Implementation

The intervention works through a forward hook registered on the output tensor of each targeted key and value projection module. It modifies only these activation outputs; no element of θ is read, written, or copied at any point. The hook applies before the projection output reaches the rest of the attention computation, so the scaling in Equation equation 1 is applied consistently regardless of what operations a given architecture applies afterward, such as positional encoding. Architectures with separate key and value projection modules are hooked directly on each. Architectures with a single fused query-key-value projection require the hook to partition the fused output along its component boundary, apply the scaling coefficient only to the key and value partitions, and leave the query partition unmodified.

Module matching is mandatory for every model evaluated. An experiment does not silently fall back to an unmitigated forward pass when the targeted modules are absent from a checkpoint, and every run records which modules were matched, confirming the hook fired on every applicable forward pass before results are written. For multiple-choice evaluation, key-value caching is disabled globally during the scoring forward pass, so key and value projections are recomputed and attenuated consistently rather than partially drawn from an unattenuated cache. For open-ended generation, the intervention hook stays active throughout decoding and caching follows each model's own default configuration. Architecture-specific layer ranges and inference settings are given in Appendix D.3.

B Output-Level Control Construction

The eligible-option set $\mathcal{B}$ in Equation equation 3 is built independently for each benchmark using labels or metadata that identify which answer options carry an explicit social direction, without consulting the gold answer. For benchmarks whose answer space splits into a socially-directional class and a neutral class, only the directional options go into $\mathcal{B}$; the neutral option is never penalized regardless of its confidence. Each option in $\mathcal{B}$ is gated by its own confidence independently, so a single item can have zero, one, or more eligible options penalized depending on where the model's confidence falls. The penalty is applied once to the final answer logits, and no renormalization happens before the softmax used in evaluation.

This construction is deliberately symmetric across a benchmark's directional options. Nothing in how $\mathcal{B}$ is built, and nothing in the penalty of Equation equation 3, identifies one directional option as more socially disadvantaged than another. The control therefore tests whether confidence attenuation on socially- directional answers reproduces SAAV's internal effect, rather than testing an intervention explicitly optimized toward a benchmark's own fairness metric. Specific option classes used for each benchmark are given in Appendix D.4.

C Development-Stage Localization Procedure

The localization procedure referenced in Section 3.1 is conducted on a single model family and a single fairness definition as a development configuration rather than one of the primary evaluation settings. A fixed set of paired evaluation items is split into a training portion used to fit the probing classifier, a validation portion used to rank candidate components and select a causally-reranked subset, and a held-out portion used to test transfer to a second distinct fairness definition.

For each candidate component, a pooled representation is extracted at a fixed position during the forward pass and used as input to a grouped, cross-validated probe predicting the benchmark-derived bias indicator. Components are ranked by validation-split probe accuracy, and the top-ranked subset is then causally reranked by measuring each component's individual effect on the validation bias margin. Random and bottom-ranked control subsets, matched in size and intervention strength, isolate the contribution of the ranking from the number of components intervened on. This development configuration is used once, before the primary evaluation, to motivate the design in Section 3.1. Specific model, split sizes, and quantitative outcomes are given in Appendix D.5.

D Experimental Details

D.1 Benchmark Metrics and Aggregation

All five primary bias magnitudes are lower-is-better. VLStereoSet evaluates all six permutations of stereotype, anti-stereotype, and unrelated captions and uses the continuous absolute bias margin $|\mathbb{E}[p_S - p_A]|$ as the primary significance-test statistic; this is a zero-ideal quantity defined at the pair level and sensitive to graded shifts in the stereotype probability margin. The discrete VLBS statistic, $\frac{1}{N_{\text{anti}}} \sum_j \mathbb{1}[p_j^S > p_j^A]$, is reported separately in the comparator table to match the external methods' convention; Section 5.2 explains why the two statistics can disagree in direction for the same intervention. VisoGender uses the absolute subgroup accuracy gap $|\text{Acc}_{\text{masc}} - \text{Acc}_{\text{fem}}|$. GenderBias-VL is paired by construction: the significance analysis uses the global absolute counterfactual gap $|\text{Acc}_{\text{base}} - \text{Acc}_{\text{cf}}|$, while the Pareto index uses the occupation-macro gap so every occupation contributes equally. For BBQ-V, with n_b bias-aligned and n_c counter-bias-aligned non-UNKNOWN predictions, $s = 2n_b/(n_b + n_c) - 1$ and the ambiguous bias score is $(1 - \text{Acc})s$. VLMBias uses the mean expected-bias match indicator.

For the Pareto analysis, each bias magnitude is normalized as

$$F_{i,b} = 1 - B_{i,b} / \max_j B_{j,b}, \tag{1}$$

and averaged, $F_i = \frac{1}{5} \sum_b F_{i,b}$. Capability is normalized analogously: $C_{i,k} = S_{i,k} / \max_j S_{j,k}$, $C_i = \frac{1}{4} \sum_k C_{i,k}$. A configuration is Pareto-dominated if another is at least as good on both F_i and C_i and strictly better on one. Same-model retention is $R_m = \frac{1}{4} \sum_k S_{\text{SAAV}}(m,k)/S_{\text{base}}(m,k)$, where equal benchmark weighting treats each fairness definition as equally important rather than weighting by evaluation-set size. The normalization maximum is fixed from the Base- Model/SAAV comparison and does not change when external comparators are introduced.

D.2 Paired Uncertainty Estimation

All Base-Model/SAAV comparisons evaluate the identical set of examples, using 10,000 paired bootstrap replicates (seed 42) and 95% percentile intervals. VisoGender preserves subgroup pairing. GenderBias-VL resamples complete counterfactual pairs. VLStereoSet reconstructs the pair-level estimator. BBQ-V preserves the joint sufficient-statistic outcome structure. Capability tasks resample paired correctness or VQA-accuracy outcomes. The 20 comparisons are not independent experiments drawn from a broad population of VLMs; the resulting counts of supported improvements, inconclusive comparisons, and supported degradations are descriptive summaries of the tested configurations, not probability estimates of how SAAV will perform on an arbitrary future model or benchmark.

D.3 Model Loading and Inference

Llama, Qwen, and InternVL use 4-bit NF4 quantization with double quantization and BF16 compute. Phi uses BF16 with eager attention because of its checkpoint-specific vision-adapter path. Llama and Qwen use their processor chat templates; InternVL uses native conversation formatting with dynamic 448×448 tiling up to 12 tiles plus a thumbnail; Phi uses its native image placeholder. Multiple-choice runs use final-position logits with `use_cache=False` for direct forward scoring. Open-ended generation uses greedy decoding with the intervention hooks active and caching following the model's own configuration. All experiments use seed 42, $\alpha = 0.85$, $\lambda = 0.35$, and $\tau = 0.65$.

Model	Layers	Projection
Llama-3.2-11B-Vision-Instruct	20–39	Separate K, V
Qwen2.5-VL-7B-Instruct	20–35	Separate K, V
InternVL3.5-8B	16–31	Separate K, V
Phi-4-Multimodal-Instruct	16–31	Fused QKV; K/V partitions scaled

Table 5: Architecture-aware intervention specification.

D.4 Logit Control Implementation

On VLStereoSet, both the stereotypical and anti-stereotypical options are in $\mathcal{B}$ under each permutation. On VisoGender, both the masculine option (A: *he/his*) and the feminine option (B: *she/her*) are in $\mathcal{B}$; the neutral option (C: *they/their*) is excluded. The penalty fires on whichever eligible options exceed τ on a given item and does not target one direction over the other. This distinguishes the Logit condition from an intervention that explicitly identifies and suppresses the disadvantaged direction; the mechanism comparison in Section 5.3 should be read with that distinction in mind.

D.5 Development-Stage Localization

This study uses Llama-3.2-11B-Vision-Instruct and VLStereoSet as the sole model and development benchmark. A total of 1,462 recoverable pairs are split into 1,023 training, 219 validation, and 220 held-out pairs. Per-head representations are extracted at the final generated token position during prefill; grouped five-fold logistic probes over five seeds rank 1,024 heads by mean validation probe accuracy. The top twenty candidates are causally reranked by measuring each head's direct effect on the validation bias margin. Random and bottom-ranked controls use an identical number of heads and identical intervention strength. This procedure is conducted on one model and one benchmark; it establishes that non-random bias-relevant structure exists but does not establish that the same pattern holds across the other three model families or the other four benchmarks.

Table 6: Localization control on the VLStereoSet validation split. Random is the five-seed mean.

Selection	Bias margin↓	VLRS↑	VLBS↓	IVLAS↑
Top-ranked	0.645813	0.995434	0.932886	0.125750
Random	0.662433	0.995434	0.942282	0.109091
Bottom-ranked	0.684511	0.995434	0.946309	0.101887

The ordering in Table 6 establishes a narrow but meaningful result: the learned ranking contains non-random bias-relevant information. It does not establish that the selected components define a portable debiasing direction.

Logit and SAAV+Logit achieve the largest continuous bias- margin reductions but also the largest relevance losses. Projection is more conservative and uniquely improves VLBS, IVLAS, and mean relevance margin while preserving VLRS. On the 657-example VisoGender transfer evaluation, the localized branch generalizes poorly: base-model resolution accuracy and gap are 0.8432 and 0.1862; Projection reaches 0.8006 and 0.2654; SAAV reaches 0.8356 and 0.1953; Adaptive reaches 0.8189 and 0.2106. Projection is favorable in-domain yet shows the strongest negative transfer — the empirical basis for the distinction *localizability* $\neq$ *transferability*: benchmark-trained localization can

Table 7: Held-out Phase-2 comparison on 220 VLStereoSet pairs.

Method	VLRS↑	VLBS↓	IVLAS↑	Bias margin↓	Δ margin	Rel. margin
Base Model	0.9727	0.9110	0.1631	0.6442	0.0000	0.6167
SAAV	0.9727	0.9178	0.1516	0.6095	−0.0346	0.5809
Projection	0.9727	0.9041	0.1746	0.6294	−0.0148	0.6274
Adaptive	0.9727	0.9178	0.1516	0.6120	−0.0322	0.5797
Logit	0.9500	0.8699	0.2289	0.5391	−0.1050	0.5474
SAAV+Logit	0.9500	0.8836	0.2075	0.5306	−0.1136	0.5282

identify meaningful structure without identifying what should be suppressed on a different fairness task.

D.6 External Comparator Protocol

NC-VAS (Ratzlaff et al., 2025), DeBiasLens (An et al., 2026), and the closed-form method of Lian et al. (Lian et al., 2026) are evaluated through aligned reimplementations on the same benchmark suite and evaluation splits. Each reimplementation preserves the original intervention logic and supplies the bias-specific information its procedure requires. Published numbers are not imported because the original papers use different model families, datasets, and splits. All methods are applied to the same examples with identical prompt templates, decoding settings, and answer parsers, so performance differences reflect both what each method does and what it is permitted to know. Table 13 summarizes the information requirements of each method. Comparator capability measurements were completed under the same four capability benchmarks and identical settings as SAAV; the primary Pareto frontier in Figure 2 remains a Base-Model/SAAV frontier, so comparator retention is reported separately in Table 3.

E Primary Paired-Bootstrap Results

Table 8 lists all twenty benchmark-native Base-Model→SAAV bias comparisons underlying Tables 1 and 2, one row per model-benchmark pair with the exact bootstrap interval and outcome label. This is the complete evidentiary basis for every statistical claim in Section 5.

E.1 Capability Bootstrap

Table 9 lists all sixteen paired capability comparisons. Six show statistically supported degradations, ten show no clear statistical change, and none shows a supported gain; the largest single-benchmark loss is Phi's TextVQA result.

F Common-Support Mechanism Analysis

Figure 3 shows the full common-support comparison underlying Section 5.3: Base Model, SAAV, Logit, and SAAV+Logit plotted jointly on the shared VLStereoSet-VisoGender support, the only support on which all four conditions are defined.

On VLStereoSet, VLRS and IVLAS are higher-is-better and VLBS is lower-is-better. Llama's SAAV condition changes VLRS only slightly (98.22→98.36, unresolved at the discrete level) while the primary continuous bias margin improves strongly; Logit degrades VLRS, and SAAV+Logit degrades both VLRS and IVLAS. Qwen and InternVL show supported VLRS gains under SAAV without consistent VLBS/IVLAS shifts; Phi's discrete VLStereoSet metrics are unresolved even though its continuous bias margin improves. On VisoGender, Logit is base-model-equivalent by construction, while SAAV drives the architecture-dependent subgroup-gap changes discussed in Section 5.3.

Table 8: Paired-bootstrap validation of the 20 benchmark- native Base-Model→SAAV bias comparisons. The GenderBias-VL rows use the global paired counterfactual gap; the Pareto index uses the occupation-macro gap. All metrics are lower-is-better, so negative Δ favors SAAV. Intervals are 95% percentile intervals from 10,000 paired resamples.

Model	Benchmark	Base	SAAV	Δ	95% CI	Outcome
Llama	VLStereoSet bias margin	0.4602	0.3992	-0.0610	$[-0.0649, -0.0569]$	Improved
Llama	VisoGender subgroup gap	0.3465	0.4590	$+0.1125$	$[0.0790, 0.1489]$	Degraded
Llama	GenderBias-VL CF gap	0.0076	0.0052	-0.0024	$[-0.0038, -0.0010]$	Improved
Llama	BBQ-V ambiguous bias	0.0480	0.0374	-0.0106	$[-0.0181, -0.0030]$	Improved
Llama	VLMBias ratio	0.2658	0.2155	-0.0503	$[-0.0607, -0.0395]$	Improved
Qwen	VLStereoSet bias margin	0.6475	0.6276	-0.0199	$[-0.0214, -0.0184]$	Improved
Qwen	VisoGender subgroup gap	0.0458	0.0550	$+0.0091$	$[0.0000, 0.0213]$	No clear change
Qwen	GenderBias-VL CF gap	0.0141	0.0167	$+0.0026$	$[0.0009, 0.0043]$	Degraded
Qwen	BBQ-V ambiguous bias	0.0455	0.0461	$+0.0006$	$[-0.0018, 0.0030]$	No clear change
Qwen	VLMBias ratio	0.6034	0.5966	-0.0068	$[-0.0115, -0.0022]$	Improved
InternVL	VLStereoSet bias margin	0.7594	0.7619	$+0.0025$	$[-0.0009, 0.0059]$	No clear change
InternVL	VisoGender subgroup gap	0.0063	0.0029	-0.0034	$[-0.0183, 0.0179]$	No clear change
InternVL	GenderBias-VL CF gap	0.0195	0.0206	$+0.0011$	$[-0.0007, 0.0030]$	No clear change
InternVL	BBQ-V ambiguous bias	0.0401	0.0452	$+0.0051$	$[0.0021, 0.0082]$	Degraded
InternVL	VLMBias ratio	0.4253	0.4203	-0.0050	$[-0.0144, 0.0043]$	No clear change
Phi	VLStereoSet bias margin	0.7707	0.7471	-0.0237	$[-0.0273, -0.0200]$	Improved
Phi	VisoGender subgroup gap	0.1671	0.1914	$+0.0243$	$[0.0091, 0.0426]$	Degraded
Phi	GenderBias-VL CF gap	0.0184	0.0165	-0.0020	$[-0.0039, -0.0001]$	Improved
Phi	BBQ-V ambiguous bias	0.1287	0.1192	-0.0095	$[-0.0153, -0.0038]$	Improved
Phi	VLMBias ratio	0.4544	0.4461	-0.0083	$[-0.0190, 0.0025]$	No clear change

Table 9: General-capability Base-Model→SAAV results. Values are percentages; outcome from the paired 95% confidence interval.

Model	Benchmark	Base	SAAV	Δ	Outcome
Llama	MMStar	42.00	41.33	-0.67	No clear change
Llama	MMMU	34.41	33.04	-1.37	Degraded
Llama	POPE	84.78	84.14	-0.63	Degraded
Llama	TextVQA	30.63	30.74	$+0.11$	No clear change
Qwen	MMStar	61.47	60.33	-1.13	Degraded
Qwen	MMMU	52.05	52.55	$+0.50$	No clear change
Qwen	POPE	87.94	87.67	-0.28	Degraded
Qwen	TextVQA	81.08	81.19	$+0.11$	No clear change
InternVL	MMStar	66.27	65.60	-0.67	No clear change
InternVL	MMMU	59.13	59.88	$+0.75$	No clear change
InternVL	POPE	88.94	88.52	-0.42	Degraded
InternVL	TextVQA	76.58	76.46	-0.11	No clear change
Phi	MMStar	58.60	57.60	-1.00	No clear change
Phi	MMMU	57.02	56.15	-0.87	No clear change
Phi	POPE	88.08	88.26	$+0.18$	No clear change
Phi	TextVQA	76.74	74.98	-1.76	Degraded

G Fine-Grained Heterogeneity

The main paper reports only the patterns necessary for interpretation. The decompositions below serve as stress tests against misleading macro conclusions.

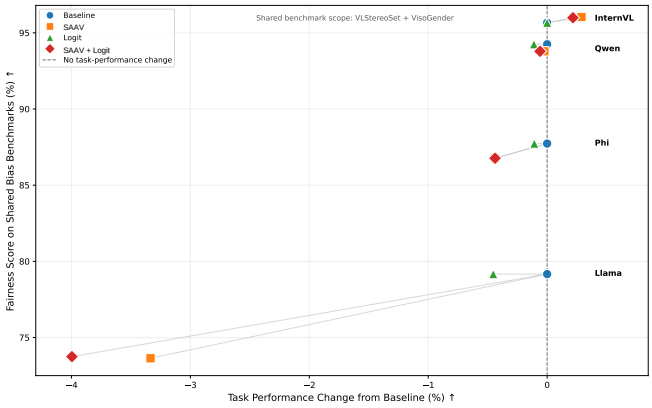

Figure 3: Common-support mechanism comparison. Base Model, SAAV, Logit, and SAAV+Logit are compared on the shared VLStereoSet-VisoGender support. The horizontal axis reports task-performance change relative to the same-model base model; the vertical axis reports fairness on this two- benchmark support, which is not on the same scale as the five-benchmark index in Table 1 and Figure 2.

Table 10: VisoGender subgroup performance (percent). Gap is lower-is-better.

Model	Method	Feminine↑	Masculine↑	Overall↑	Gap↓
Llama	Base Model	65.35	100.00	82.65	34.65
	SAAV	54.10	100.00	77.02	45.90
Qwen	Base Model	96.35	91.77	94.06	4.58
	SAAV	96.35	90.85	93.61	5.50
InternVL	Base Model	93.92	93.29	93.61	0.63
	SAAV	93.62	93.90	93.76	0.29
Phi	Base Model	80.85	97.56	89.19	16.71
	SAAV	78.72	97.87	88.28	19.14

G.1 GenderBias-VL Occupation Heterogeneity

Across 178 occupations, an occupation-level diagnostic counts Llama as improved on 61 and worsened on 28, Qwen 36/63, InternVL 56/53, and Phi 72/46, with remaining occupations unchanged at the resolution of that statistic. These counts are descriptive and are not substituted for the paired benchmark-level confidence intervals in Table 8.

Table 11: BBQ-V absolute bias score by demographic category. Lower is better. Values are percentages.

Model	Method	Age	Disab.	Gender	Nat.	Appear.	Race/Eth.	Relig.	SES	Sex. Ori.
Llama	Base Model	4.59	8.25	3.71	0.85	20.49	2.41	6.36	5.08	3.46
	SAAV	3.64	5.58	1.66	2.73	16.52	1.31	4.91	4.24	3.83
Qwen	Base Model	13.38	2.28	1.55	0.75	18.03	0.31	3.61	2.12	0.98
	SAAV	13.84	3.30	1.22	0.75	17.81	0.45	3.61	1.48	0.83
InternVL	Base Model	10.38	5.58	1.22	0.47	18.13	0.17	2.46	2.86	0.38
	SAAV	11.72	6.22	1.06	0.47	19.74	0.07	3.61	3.81	0.60
Phi	Base Model	35.06	17.89	8.96	4.33	30.04	0.48	10.26	15.47	6.92
	SAAV	30.83	18.02	7.47	4.14	27.79	0.76	10.98	14.19	3.23

G.2 BBQ-V and VLMBias

BBQ-V category decompositions (Table 11) and VLMBias topic/subtopic decompositions show that macro movement can coexist with local reversals: Llama's macro BBQ-V bias falls from 6.13% to

4.94% while nationality and sexual orientation individually regress. These tables should be interpreted jointly with task accuracy where applicable, since a lower bias statistic caused by generic answer failure is not a meaningful fairness gain.

H Comparator Capability Results

Model	Method	MMStar	MMMU	POPE	TextVQA
Llama	Base Model	42.00	34.41	84.78	30.63
	SAAV	41.33	33.04	84.14	30.74
	NC-VAS	41.80	34.10	84.50	30.50
	DeBiasLens	41.20	33.70	84.00	30.20
	Closed-Form	41.90	34.20	84.60	30.55
Qwen	Base Model	61.47	52.05	87.94	81.08
	SAAV	60.33	52.55	87.67	81.19
	NC-VAS	61.20	51.90	87.70	80.90
	DeBiasLens	60.90	51.50	87.30	80.40
	Closed-Form	61.35	52.00	87.82	81.00
InternVL	Base Model	66.27	59.13	88.94	76.58
	SAAV	65.60	59.88	88.52	76.46
	NC-VAS	66.00	58.90	88.70	76.30
	DeBiasLens	65.50	58.40	88.20	75.90
	Closed-Form	66.10	59.00	88.80	76.45
Phi	Base Model	58.60	57.02	88.08	76.74
	SAAV	57.60	56.15	88.26	74.98
	NC-VAS	58.30	56.80	87.90	76.40
	DeBiasLens	57.80	56.20	87.40	75.80
	Closed-Form	58.40	56.90	88.00	76.55

Table 12: Full per-benchmark capability scores underlying the retention values in Table 3.

I External Method Assumptions

Table 13: Information requirements of the training-free methods in Table 4. Equal numerical values need not imply equal supervision or prior bias-specific information.

Approach	Parameter updates	Bias-specific information	Intervention form
NC-VAS	None at test time	Attribute steering direction / contrastive structure	Representation steering
DeBiasLens	No full-model retraining	Learned/localized demographic-sensitive features	Feature suppression
Closed-Form	Closed-form post-hoc transform	Explicit protected-group or bias directions	Representation projection
SAAV (ours)	None	No learned demographic representation or prototype	Fixed late-layer KV attenuation