CONSTRUCT BEFORE COMPOSE: PRESERVING CONDITIONAL STRUCTURE IN REPRESENTATION-BASED BIAS MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Social biases in large language models can vary across combinations of demographic attributes, yet existing representation-based interventions often construct directions from isolated attributes and attempt to recover such variation only during composition. However, marginal direction construction can discard cross-attribute information before intervention is applied. Hence, we propose Construct Before Compose (CBC), a condition-aware framework that constructs attribute directions while preserving cross-attribute conditions and then applies composition mechanisms. We evaluate CBC on race–gender, age–gender, and age–race settings across four instruction-tuned models from three model families. Compared with single-attribute construction under the same composition rule, CBC improves the mean conditional bias reduction from 7.70% to 12.13% across the three attribute pairs. Further analyses show that mitigation on marginal attributes does not reliably transfer to conditional targets, and that conditional evaluation is necessary to reveal residual slice-level disparities and cross-attribute interaction effects. Across six general-language benchmarks, CBC maintains comparable utility, with a four-model macro-average accuracy change of -0.30 percentage points relative to no intervention. These results identify condition-aware direction construction as a promising design choice for multi-attribute representation-based intervention..

1 INTRODUCTION

Large language models (LLMs) can reproduce stereotypical associations related to demographic attributes such as race, gender, and age (Weidinger et al., 2022). Such biases, however, need not be expressed independently along each demographic axis. In the same occupational context, for example, a model's preference for *he* versus *she* may change when the person is described with different racial or age attributes (Guo & Caliskan, 2021; Charlesworth et al., 2024). Consequently, improvements measured separately for race, gender, or age need not carry over to their combinations. This concern is closely related to intersectionality, which emphasizes that social categories and associated inequalities cannot always be understood in isolation or as simple additive effects (Crenshaw, 1989; 1991; Hancock, 2007). Motivated by this perspective, we study **whether representation-based interventions can reduce bias associated with one attribute across conditions defined by another**. Notably, we operationalize this question using controlled lexical preference disparities across attribute combinations, while recognizing that this captures only a model-level aspect of the broader social notion of intersectionality.

This conditional view of bias raises a corresponding intervention problem: **where should cross-attribute information enter the steering pipeline?** Inference-time activation steering modifies model behavior through internal representations without updating model parameters (Zou et al., 2023; Rimsky et al., 2024; Li et al., 2023). Recent multi-attribute methods extend this paradigm through direction composition, joint control, or input-dependent steering (Scalena et al., 2024; Nguyen et al., 2025; Oozeer et al., 2025; Jiang et al., 2025). Most of these studies focus on how multiple controls are combined or applied after their construction. We instead identify an earlier failure mode: **when steering directions are estimated from isolated single-attribute preferences, cross-attribute variation can be averaged away before composition begins**. A downstream composer cannot recover information that is absent from the direction bank.

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.

Therefore, we introduce *Construct Before Compose* (CBC), a condition-aware framework that moves cross-attribute conditioning into direction construction. For a pair of binary attributes, conventional single-attribute construction estimates one direction per attribute, whereas CBC estimates attribute directions separately under each value of the other attribute. We instantiate CBC with weighted mean difference (Cyberey et al., 2025) and then apply different downstream composition strategies, including input-independent static weights (*RepE Mean* and *TrainBias*), weights derived from external semantic relevance (*DynEmb*) and weights conditioned on the target model's internal activations (*APC*). This design lets us ask whether preserving conditional information during construction can improve the bias mitigation performance. We additionally vary the composition rule while holding the CBC direction bank fixed to test whether the effect depends on a particular composer.

Across four instruction-tuned checkpoints from three model families and race–gender, age–gender, and age–race settings, *CBC+APC* increases the equal-pair mean conditional bias reduction from 7.70% to 12.13% relative to single-attribute construction under the same APC pipeline. The gain is substantial for race–gender and age–gender, while small and uncertain for age–race, indicating that the benefit of conditional construction varies across attribute pairs. Multiple alternative composition rules also yield positive conditional attenuation with the same CBC direction bank, suggesting that the robustness of the identified effect.

Our analyses further show why conditional evaluation matters. Methods that perform well on single-attribute targets do not reliably retain those gains on conditional targets, and stronger average conditional attenuation does not necessarily imply reduced cross-attribute interaction. Importantly, CBC does not introduce substantial general-language utility degradation: across six benchmarks, its four-model macro-average accuracy change is -0.30 percentage points relative to no intervention. Together, these results suggest that both intervention and evaluation should preserve the conditional structure that marginal attribute-wise analysis can obscure.

Our contributions are:

- We introduce *Construct Before Compose* (CBC), a condition-aware direction construction framework that preserves cross-attribute information before multi-direction composition.
- Across three attribute pairs and four model checkpoints, we show that CBC yields stronger average conditional bias mitigation than the evaluated single-attribute construction strategy, remains effective across multiple composition strategies, and preserves comparable general-language utility, providing evidence that preserving conditional structure during direction construction can benefit representation-based bias mitigation.

2 Related Work

Intersectionality and computational bias analysis. Intersectionality research emphasizes that demographic categories are mutually constitutive and that additive analyses may miss variation across attribute combinations (Hancock, 2007; Collins, 2015; Bauer et al., 2021). Computational studies have similarly documented cross-category stereotypes and subgroup disparities in language models (Guo & Caliskan, 2021; Charlesworth et al., 2024; Khan et al., 2025). However, subgroup evaluation alone does not explain how such conditional variation is represented or affected by intervention (Ovalle et al., 2023). Our work therefore focuses on preserving cross-attribute conditions during both direction construction and post-intervention evaluation.

Representation-based and multi-attribute steering. Representation-based interventions identify directions in activation space and modify them to control model behavior (Cyberey et al., 2025; Siddique et al., 2026; Park et al., 2024). Recent methods further combine, select, or adapt multiple steering signals through dynamic composition, joint subspaces, sparse interventions, classifier gradients, or conditional gating (Scalena et al., 2024; Nguyen et al., 2025; Jiang et al., 2025; Bayat et al., 2025; Wang et al., 2025; Oozeer et al., 2025; Lee et al., 2025; Li et al., 2025). These methods largely focus on how representation-level controls are combined, selected, or applied after construction. In contrast, our study moves conditioning to the construction stage, preserving cross-attribute information before composition.

3 FROM MARGINAL TO CONDITIONAL BIAS MEASUREMENT

3.1 FROM INTERSECTIONALITY TO CONDITIONAL BIAS TARGETS

Intersectionality motivates evaluating bias within specific attribute combinations rather than treating success along each attribute in isolation as sufficient (Crenshaw, 1989; Hancock, 2007). Accordingly, we preserve cross-attribute conditions both when constructing the preference scores used to estimate intervention directions and when evaluating model behavior after intervention. Although our operationalization does not capture the broader social, structural, or experiential dimensions of intersectionality, it provides a controlled model-level measure of how lexical preferences vary across attribute combinations.

Consider two binary attributes S and T with candidate values s_0, s_1 and t_0, t_1 which represents the lexical contrasts evaluated in our experiments. For a fixed prompt x, let $\ell_{jk}(x) \in \mathbb{R}$ denote the log-probability of the complete candidate sequence corresponding to the joint condition $S = s_j$ and $T = t_k$. Under a fixed intervention setting, we define four conditional preference contrasts:

$$\begin{aligned} \delta_{T|s_0}(x) &= \ell_{00}(x) - \ell_{01}(x), & \delta_{T|s_1}(x) &= \ell_{10}(x) - \ell_{11}(x), \\ \delta_{S|t_0}(x) &= \ell_{00}(x) - \ell_{10}(x), & \delta_{S|t_1}(x) &= \ell_{01}(x) - \ell_{11}(x). \end{aligned} \tag{1}$$

The first two contrasts measure preference over T while holding S fixed, whereas the latter two measure preference over S while holding T fixed. We collect them as

$$\boldsymbol{\delta}(x) = \left(\delta_{T|s_0}(x), \delta_{T|s_1}(x), \delta_{S|t_0}(x), \delta_{S|t_1}(x)\right)^\top. \tag{2}$$

The preference change across conditioning values is identical whether computed along S or T:

$$\delta_{T|s_0}(x) - \delta_{T|s_1}(x) = \delta_{S|t_0}(x) - \delta_{S|t_1}(x) = \ell_{00}(x) - \ell_{01}(x) - \ell_{10}(x) + \ell_{11}(x). \tag{3}$$

This quantity is the log cross-product ratio of the 2×2 candidate preference table and captures cross-attribute conditional variation.

3.2 AVERAGED PREFERENCES DO NOT DETERMINE CONDITIONAL PREFERENCES

To characterize what is lost when the conditional preference contrasts are summarized by their averages, we decompose the four contrasts into average single-attribute components (m_T and m_S) and a cross-attribute interaction component (q).

$$m_T = \frac{\delta_{T|s_0} + \delta_{T|s_1}}{2}, \qquad m_S = \frac{\delta_{S|t_0} + \delta_{S|t_1}}{2}, \qquad q = \frac{\delta_{T|s_0} - \delta_{T|s_1}}{2} = \frac{\delta_{S|t_0} - \delta_{S|t_1}}{2}. \tag{4}$$

Where, m_T and m_S summarize the average preference contrast for each attribute across the two conditions of the other attribute, whereas q captures how that contrast changes across conditions. By Eq. 3, q can equivalently be computed along either attribute dimension, and $2q$ is the log cross-product ratio of the 2×2 candidate preference table.

Proposition 1 (Average preference contrasts do not identify conditional variation) *For any binary two-factor preference table, the conditional contrast vector admits the decomposition*

$$\boldsymbol{\delta} = (m_T + q,\, m_T - q,\, m_S + q,\, m_S - q)^\top. \tag{5}$$

Moreover, its total contrast energy satisfies

$$\frac{1}{4}\|\boldsymbol{\delta}\|_2^2 = \frac{1}{2}\left(m_T^2 + m_S^2\right) + q^2. \tag{6}$$

Consequently, the average components (m_T, m_S) alone do not determine the cross-attribute interaction component q, and therefore do not determine the four conditional preference contrasts.

This decomposition shows that zero average preference does not imply the absence of conditional variation. In particular, $(m_T, m_S) = (0, 0)$ can coexist with a nonzero interaction component q, and thus with nonzero conditional preference contrasts. Moreover, a reduction in total contrast energy does not necessarily imply a reduction in interaction energy q^2. This distinction motivates preserving cross-attribute conditions during direction construction rather than relying on averaged single-attribute representations alone.

4 CONSTRUCT BEFORE COMPOSE

Building on the distinction established in Section 3.2, Construct Before Compose (CBC) preserves cross-attribute conditions when intervention directions are estimated, before any directions are combined. We first define the conditional targets used for direction construction, then describe how these targets are converted into direction banks, composed, and applied during intervention.

4.1 CONDITIONAL TARGET CONSTRUCTION ON A CONTROLLED LEXICAL PROBE

To make CBC clearer, we use WinoGenerated occupation templates (Perez et al., 2023), which extend Winogender (Rudinger et al., 2018). The controlled evaluation considers three binary attribute pairs: race–gender, age–gender, and age–race, using `white`/`black`, `young`/`old`, and `he`/`she` as their respective lexical realizations. For each attribute pair, the single-attribute bank and the CBC conditional bank are constructed from the same template source and data split, but use different preference targets. Consider race–gender as an example. Let x_i denote the ith template, and let $y_{i,r,g}$ denote the complete candidate answer associated with race value r and gender value g, written in the order of their slots in the template. CBC retains four conditional targets:

$$\mathcal{T}_{\mathrm{RG}} = \{G \mid W,\ G \mid B,\ R \mid M,\ R \mid F\}, \tag{7}$$

where, for example, $G \mid W$ denotes gender preference under the `white` condition, whereas $R \mid F$ denotes race preference under the `she` condition.

For template x_i and target $t \in \mathcal{T}_{\mathrm{RG}}$, we write

$$\Delta_{i,t} \equiv \delta_t(x_i), \tag{8}$$

where $\delta_t(x_i)$ is the corresponding conditional preference contrast defined in Equation 1. We score the log-probability of the complete candidate answer under the template prompt, rather than scoring a fully filled sentence. The age–gender and age–race settings follow the same construction, replacing the race–gender targets with the corresponding four conditional targets for each attribute pair.

4.2 CONDITION-INDEXED DIRECTION EXTRACTION WITH MAGNITUDE-STRATIFIED WMD

For each conditional target t, CBC requires a separate intervention direction. CBC itself does not prescribe how this direction is estimated; in our experiments, we instantiate the direction extractor using an adaptation of weighted mean difference (WMD) (Cyberey et al., 2025). The extractor contrasts residual-stream activations associated with weaker and stronger preference magnitudes for a conditional target. We convert the signed log-probability difference to a bounded preference score:

$$s_{i,t} = \tanh(\Delta_{i,t}). \tag{9}$$

Because $\tanh(\cdot)$ is monotonic, $|s_{i,t}|$ preserves the ordering of preference magnitude while bounding the influence of extreme values. For each target t, we sort the n training examples by $|s_{i,t}|$, assigning the lower half to L_t and the upper half to H_t. Repeating this for the four conditional targets gives

$$\mathcal{V}_4 = \{v_{G|W}, v_{G|B}, v_{R|M}, v_{R|F}\}. \tag{10}$$

The single-attribute bank instead estimates v_G, v_R from separate single-slot prompts. Prompt view, scoring target, direction count, and parameter sharing therefore differ together in the construction comparison. The age–gender and age–race instances repeat this procedure with their own four conditional targets and independently estimated no-intervention baselines.

4.3 DIRECTION COMPOSITION OVER A FIXED BANK

After constructing the direction bank in Section 4.2, we next specify how these directions are combined during intervention. Importantly, the composer operates on an already constructed bank and does not modify the directions themselves. Let

$$V = [v_1, \ldots, v_K] \in \mathbb{R}^{d \times K} \tag{11}$$

denote a bank of K constructed directions. For token position k in scored sequence i, a composer assigns coefficients $\boldsymbol{\alpha}_{i,k} \in \mathbb{R}^K$. When $V\boldsymbol{\alpha}_{i,k} \neq 0$, the resulting composed intervention direction is

$$u_{i,k} = \mathrm{norm}(V\boldsymbol{\alpha}_{i,k})\,. \tag{12}$$

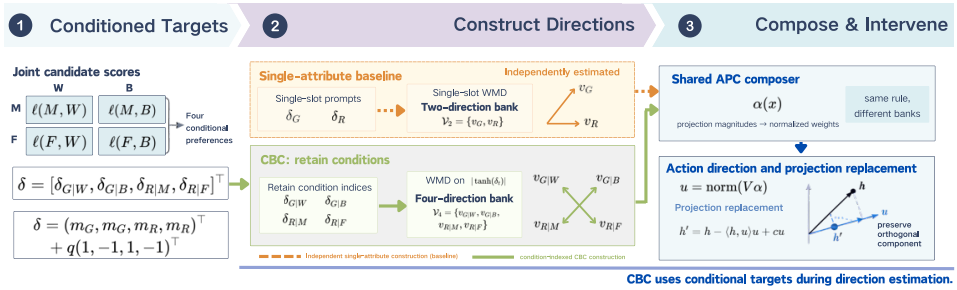


Figure 1: From Marginal Direction Construction to Conditional Representation Construction.

We consider three families of composition rules that span distinct sources of coefficient information: input-independent static weights, weights conditioned on the target model's internal activations, and weights derived from external semantic relevance.

Static Composition. **RepE Mean** assigns equal weights to all directions in the bank. **TrainBias** further derives fixed weights from the training-set RMS magnitudes of the preference scores $s_{i,t}$. Because these coefficients do not depend on the current input, both methods produce a single composed direction for each checkpoint and direction bank.

External Semantic based Composition. **DynEmb** provides a complementary input-dependent composer whose coefficients are derived from external semantic relevance rather than from the target model's internal activations. We use a frozen `bert-base-uncased` encoder to compute signed relevance scores and evaluate concentration powers $p \in \{0.5, 1, 2\}$. The resulting coefficients are shared across token positions within a prompt, while the intervention itself remains on the target model's residual stream.

Internal Activation based Composition. **Activation projection composition (APC)**, adapts the composition coefficients conditioned on the target model's internal activations:

$$\alpha_{i,k,j} = \frac{\exp\Big(|\langle h^{\ell}_{i,k}, v_j\rangle|/\tau\Big)}{\sum_{j'=1}^{K} \exp\Big(|\langle h^{\ell}_{i,k}, v_{j'}\rangle|/\tau\Big)}, \qquad \tau = 1. \tag{13}$$

Here, $h^{\ell}_{i,k}$ denotes the residual-stream activation at block ℓ for token position k of scored sequence i. Hence, APC assigns larger weights to directions with stronger absolute alignment to the current activation. Full coefficient definitions are provided in Appendix B.2.

4.4 Intervention by Projection Replacement

After constructing and composing the direction bank, we apply the resulting unit direction through the same residual-stream intervention for all WMD-based variants. Given an activation h, a unit intervention direction u, and intervention strength c, projection replacement updates the activation as

$$\begin{aligned} h' &= \underbrace{(h - \langle h, u\rangle u)}_{h_\perp} + \underbrace{cu}_{\text{new component along } u} \\ &= h_\perp + cu. \end{aligned} \tag{14}$$

Consequently,$\langle h', u\rangle = c$.This update leaves the component of h orthogonal ($h_\perp$) to u unchanged, while replacing its original projection onto u with the fixed value c. Thus, projection replacement sets the activation along u to a prescribed value rather than adding a multiple of u to it.

5 Research Questions and Experimental Setup

5.1 Research questions

We organize our experiments around three questions:

Q1: Does CBC improve conditional mitigation over existing construction strategies? We compare two direction-bank construction strategies under the same APC composer and intervention procedure. *Single+APC* constructs one direction per attribute, whereas *CBC+APC* constructs directions conditioned on cross-attribute targets. Both methods use the same intervention layer, strength, and data split. *Pair4* is averaged across four checkpoints within each attribute pair, followed by an equal-weight mean across the three attribute pairs. We further test CBC robustness by varying the composer while fixing the CBC direction bank. Notably, this comparison evaluates complete construction strategies rather than the isolated effect as prompt view, scoring target, direction count, and parameter sharing differ jointly.

Q2: Why is condition-aware construction needed? Using race–gender as the representative setting, we examine whether single-attribute interventions transfer to conditional targets without retraining. We adapt MAT-Steer, FairSteer, K-Steering, CAST-inspired, and DSAS-inspired methods to a shared residual-stream interface (Nguyen et al., 2025; Li et al., 2025; Oozeer et al., 2025; Lee et al., 2025; Ferrando et al., 2025) and evaluate them on held-out Marginal2 and Cross4 targets. We further analyze whether average conditional attenuation corresponds to cross-attribute interaction reduction using the interaction component q.

Q3: Does CBC preserve general-language utility? To assess whether improved conditional mitigation introduces utility costs, we evaluate No-Intervention, Single+APC, and CBC+APC on six general-language benchmarks: MMLU, HellaSwag, ARC-Challenge, GSM8K, TruthfulQA MC1, and WinoGrande. TruthfulQA uses all 817 examples, while the remaining tasks use fixed 1,000-example subsets (see Appendix A.4 for details). We report accuracy changes relative to No-Intervention and compute an equal-weight six-task macro average for each checkpoint. We then average these four checkpoint-level macro averages equally across the four evaluated models, which we refer to as the *four-model mean*.

5.2 Models, data, and intervention configuration

We evaluate four instruction-tuned checkpoints: Llama-3.2-3B-Instruct (Meta, 2024), Llama-3-8B-Instruct (Llama Team, AI @ Meta, 2024), Mistral-7B-Instruct-v0.2 (Mistral AI, 2023), and Qwen2.5-3B-Instruct (Yang et al., 2024). The suite spans the Llama, Mistral, and Qwen families with public weights and no parameter updates.

All three attribute-pair experiments use WinoGenerated occupation templates (Perez et al., 2023) and share the same seed-42 split, comprising 2,384 training templates and 596 validation templates. Direction extraction and any data-calibrated coefficients use only the training partition, while all primary bias metrics are computed on the validation partition, ensuring that intervention construction and evaluation remain strictly separated. Candidate answers are scored using teacher-forced log-probabilities of the complete candidate sequences. Each attribute pair constructs its own direction bank and no-intervention baseline; no controller is shared across attribute pairs.

To ensure a controlled comparison across models and construction strategies, we use the same intervention configuration for all WMD-based variants: zero-based Transformer block 20 and projection-replacement strength $c = 1.0$, without model- or method-specific tuning. Directions are extracted from the residual-stream activation at the final token of the prompt, while the resulting intervention is applied to the target block activations throughout candidate scoring. Baseline adapters are evaluated at the same block, with their perturbation magnitudes calibrated on the training partition according to their respective update rules. Because the evaluated checkpoints contain 28–36 Transformer blocks, block 20 represents a common absolute intervention location rather than a matched relative depth across architectures.

5.3 Evaluation metrics

We evaluate intervention effectiveness at two levels: conditional bias attenuation across attribute combinations and single-attribute bias attenuation.

Conditional bias attenuation across attribute combinations. For validation template i and conditional target t, we measure the magnitude of the preference contrast as $d_{i,t} = |\Delta_{i,t}|$. We then

aggregate these magnitudes over the validation set $\mathcal{D}_{\text{val}}$ using Root Mean Square (RMS):

$$\text{RMSBias}_t = \sqrt{\frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} d_{i,t}^2}. \tag{15}$$

Thus, RMSBias_t measures the overall magnitude of the conditional preference contrast for target t, with smaller values indicating weaker conditional preference disparities.

To quantify the effect of an intervention method m, we report its relative reduction from the corresponding no-intervention baseline:

$$R_t(m) = 100 * \frac{\text{RMSBias}_t(\text{none}) - \text{RMSBias}_t(m)}{\text{RMSBias}_t(\text{none})}. \tag{16}$$

Positive $R_t(m)$ indicates attenuation of the conditional preference contrast for slice t, whereas negative values indicate amplification. This quantity is defined whenever the corresponding no-intervention RMS is positive.

For an attribute pair (S, T), we summarize performance across its four conditional targets by

$$\text{Pair4}_{S,T}(m) = \frac{1}{4} \sum_{t \in \mathcal{T}_{S,T}} R_t(m). \tag{17}$$

We refer to this metric as **Pair4**; its race–gender instance is denoted **Cross4**. Pair4 therefore measures average relative attenuation across the four conditional slices. A positive Pair4 value does not imply that every slice improves, nor should it be interpreted as an absolute fairness score. Hence, we further separately evaluate single-attribute mitigation using the corresponding single-slot probes. Let $R_S(m)$ and $R_T(m)$ denote the relative RMS reductions for attributes S and T, respectively. We define

$$\text{Marginal2}_{S,T}(m) = \frac{R_S(m) + R_T(m)}{2}. \tag{18}$$

Marginal2 summarizes performance on the two single-attribute targets, whereas Pair4 evaluates performance on the four cross-attribute conditional targets.

Single-attribute bias attenuation. For single-attribute baseline methods, positive held-out Marginal2 is used as an eligibility criterion before assessing Pair4, so that Pair4 measures whether successful single-attribute mitigation transfers to conditional settings without retraining. For the direct comparison between single-attribute and CBC construction, Pair4 is the primary metric and no Marginal2 eligibility filter is applied. Because Marginal2 averages the two single-attribute reductions, a positive value does not necessarily imply improvement on both attributes individually.

5.4 Statistical reporting

For the fixed four-checkpoint suite, we report means together with checkpoint-level values and observed minima. For each attribute pair, we estimate a paired 95% bootstrap confidence interval using 10,000 replicates. Within each replicate, the same 596 validation-template indices are resampled across methods, checkpoints, and the four conditional targets, and all RMS-based quantities are recomputed before aggregation. These intervals quantify uncertainty arising from the sampled validation templates, conditional on the fitted controllers, the fixed checkpoint suite, and the evaluation distribution. These intervals reflect template-sampling uncertainty under the fixed fitted models and evaluation setting, rather than population-level uncertainty over models. The equal-weight mean across the three attribute pairs is reported descriptively and is not assigned a joint bootstrap confidence interval. Because relative-reduction metrics divide by the corresponding no-intervention RMS, their validity requires the baseline RMS to remain positive in every bootstrap replicate.

6 Results

CBC improves conditional mitigation across attribute pairs and composition strategies. Figure 2(a) compares Single+APC and CBC+APC under the same composition and intervention procedure. Across the three attribute pairs, CBC increases the equal-pair mean Pair4 reduction from 7.70%

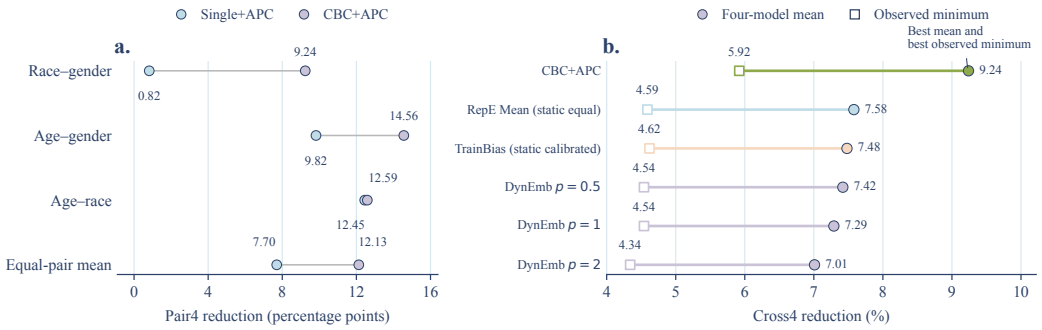


Figure 2: Unified construction and composer comparison. Panel (a) compares Pair4 reductions for Single+APC and CBC+APC across the three attribute pairs and their equal-pair mean. Panel (b) compares six coefficient rules over the fixed race–gender CBC direction bank, reporting the four-model mean and observed minimum Cross4 reduction for each rule.

to 12.13%. The gain is largest for race–gender (+8.42 percentage points; 95% CI: $[7.02, 9.92]$), followed by age–gender (+4.74 points; $[3.50, 5.96]$). For age–race, the difference is the smallest (+0.15 points; $[-1.06, 1.36]$).

We next assess whether CBC remains effective under alternative composition rules. Holding the race–gender CBC direction bank fixed, Figure 2(b) compares six prespecified composers. All six yield positive Cross4 reduction on every evaluated model, indicating that CBC's conditional attenuation is not tied to a single composer. APC achieves the largest four-model mean reduction (9.24%) and observed minimum (5.92%), although composer performance varies across individual models (see Appendix D.2).

Overall, CBC provides stronger conditional mitigation than single-attribute construction under the shared APC pipeline, and its effectiveness persists across multiple composition rules. Notably, as Single+APC and CBC+APC differ jointly in prompt view, scoring target, direction count, and parameter sharing, these observed improvements should not be attributed to condition indexing alone.

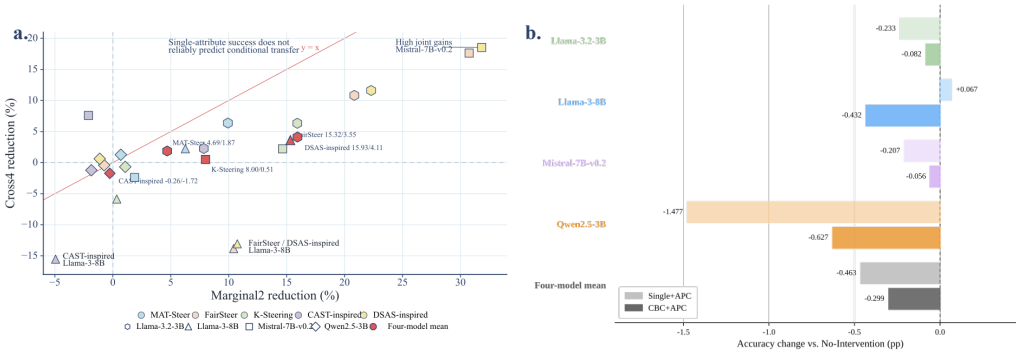


Figure 3: Conditional transfer and general-language utility across models. Panel (a) shows Marginal2 versus Cross4 reductions, with the red line denoting equal transfer. Panel (b) shows six-task macro-average accuracy changes relative to No-Intervention, in percentage points.

Single-attribute mitigation transfers unevenly. We next evaluate whether mitigation learned from single-attribute targets transfers to conditional targets without retraining. For each method, we compare reduction on held-out Marginal2 probes with reduction on the Cross4 conditional targets.

Using race–gender as a representative setting, Figure 3 (a) summarizes this relationship at the model–method level. Points below the $y = x$ diagonal indicate weaker conditional than single-attribute mitigation, while negative Cross4 values indicate increased conditional bias relative to the no-intervention baseline. Most points fall below the diagonal, indicating a substantial transfer gap

for many model–method combinations. Averaged equally across the four models, FairSteer and DSAS-inspired reduce Marginal2 by 15.32% and 15.93%, respectively, but reduce Cross4 by only 3.55% and 4.11%. Transfer also varies markedly across models: on Llama-3-8B, both methods increase Cross4 bias (-13.83% and -13.06%), whereas on Mistral-7B-v0.2 they achieve Cross4 reductions of 17.63% and 18.53%. Similar model dependence appears for the other adapted baselines. MAT-Steer has a positive four-model Cross4 mean but a negative reduction on Mistral-7B-v0.2, while K-Steering yields negative Cross4 reductions on Llama-3-8B and Qwen2.5-3B.

A second transfer gap appears between average conditional attenuation and cross-attribute interaction reduction. Particularly, CBC+APC achieves a mean Cross4 reduction of 9.24% but an interaction reduction of only 3.26%, compared with 4.99% for Single+APC. On Llama-3.2-3B, CBC+APC instead increases interaction magnitude (-2.06%). Checkpoint-level results are reported in Appendix E.

Overall, mitigation does not transfer reliably across levels of conditional structure: strong single-attribute attenuation may weaken on conditional targets, and stronger average conditional attenuation need not imply greater reduction of cross-attribute interaction. These findings highlight the limitations of attribute-independent construction and motivate the need for condition-aware direction construction, as introduced by CBC.

Conditional mitigation preserves general-language utility. To examine whether conditional bias mitigation affects general model capability, Figure 3(b) reports accuracy changes on six general-language benchmarks relative to No-Intervention. Across the four checkpoints, CBC+APC produces a smaller average utility degradation, with a macro-average accuracy change of -0.299 percentage points, compared with -0.463 points for Single+APC. In other words, condition-aware direction construction does not introduce additional substantial utility loss under the shared intervention pipeline.

The aggregate result masks checkpoint-level variation. On Qwen2.5-3B, the accuracy loss decreases from 1.477 points with Single+APC to 0.627 points with CBC+APC, with similar reductions observed on Llama-3.2-3B and Mistral-7B-v0.2. In contrast, Llama-3-8B changes from a 0.067-point gain under Single+APC to a 0.432-point loss under CBC+APC. Therefore, the four-model average should not be interpreted as uniform improvement across checkpoints.

Overall, CBC-based intervention maintains comparable general-language utility, with macro-average accuracy changes remaining within 0.5 percentage points of the No-Intervention baseline.

7 Conclusion

We show that intersectional bias mitigation can be limited by how representation directions are constructed: estimating interventions from single-attribute preferences may discard cross-attribute information before composition. To address this limitation, we propose Construct Before Compose (CBC), a condition-aware framework that preserves attribute conditions during direction construction and organizes condition-indexed directions before intervention.

Across four model checkpoints and three attribute-pair settings, *CBC+APC* achieves stronger conditional bias attenuation than single-attribute construction, increasing the mean bias reduction from 7.70% to 12.13%. Our analyses further show that marginal mitigation is insufficient for characterizing conditional bias: improvements on single-attribute targets do not reliably transfer to attribute combinations, and reductions in average conditional bias do not necessarily correspond to reduced cross-attribute interaction effects. Importantly, these gains are achieved while maintaining general-language capability, with a four-model macro-average accuracy change of -0.30 percentage points across six benchmarks relative to no intervention. These findings highlight the importance of preserving conditional variation when constructing and evaluating representation interventions, rather than relying solely on axis-wise bias measures.

More broadly, CBC reframes representation-based intervention as a construction problem: the effectiveness of downstream composition is bounded by the information encoded in the underlying direction bank. While our study focuses on controlled English lexical contrasts and a fixed set of instruction-tuned checkpoints, it establishes condition-aware construction as a general principle for representation intervention in large language models. Future work can extend this framework to non-binary attributes, broader social contexts, and more complex forms of conditional behavior.

Reproducibility Statement

The appendix specifies the candidate-scoring protocol, prompts, data splits, model identifiers, intervention settings, and public-method adaptations. The self-contained archive `CBC_four_model_publication_repro_20260919.zip` provides sample-level outputs, exact CPU reaggregation code, expected results, checksums, and a clean inference-code snapshot. The CPU replay recomputes article statistics from retained model outputs. Re-running model inference additionally requires the upstream model weights and the fitted controller states documented by the archive. These are distinct reproducibility requirements. Directions and calibration coefficients use only the training split; the reported target metrics use the validation split. General-language gold labels follow each dataset's native index semantics, and per-example predictions are retained for recomputing accuracy.

Ethics Statement

This work measures associations between model probabilities and coarse, socially constructed race, binary-gender, and binary-age terms. These labels do not represent the full diversity of race, gender, or age, and the benchmark does not measure structural discrimination or lived experience. A larger Pair4 reduction is not evidence of fairness in a deployed system. The intervention is evaluated only on a synthetic lexical task and should not be treated as a deployment-ready safeguard. The benchmark is intended for reproducibility and bias-mitigation research, not for inferring individual attributes or making decisions about people. Any downstream use requires community-informed evaluation, utility and safety testing, group-specific regression monitoring, and meaningful human oversight.

AI Use Statement

Generative AI tools were used as assistive tools during the development of this work and preparation of the manuscript. They supported code modification and debugging, consistency checking of experimental outputs, and improvements to manuscript structure, clarity, and language. The authors conceived the research questions, designed the methodology and experiments, interpreted the results, and made all final scientific and methodological decisions. All AI-assisted code, analyses, and text were reviewed and verified by the authors, including implementation testing and independent validation of the reported results. The authors take full responsibility for the accuracy, integrity, and conclusions of the work.

References

Greta R. Bauer, Siobhan M. Churchill, Mayuri Mahendran, Chantel Walwyn, Daniel Lizotte, and Alma Angelica Villa-Rueda. Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM – Population Health*, 14:100798, 2021. doi: 10.1016/j.ssmph.2021.100798. URL https://doi.org/10.1016/j.ssmph.2021.100798.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces. In *The Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VGw1viYliK.

Tessa E. S. Charlesworth, Kshitish Ghate, Aylin Caliskan, and Mahzarin R. Banaji. Extracting intersectional stereotypes from embeddings: Developing and validating the flexible intersectional stereotype extraction procedure. *PNAS Nexus*, 3(3):pgae089, 2024. doi: 10.1093/pnasnexus/pgae089. URL https://doi.org/10.1093/pnasnexus/pgae089.

Patricia Hill Collins. Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 41: 1–20, 2015. doi: 10.1146/annurev-soc-073014-112142. URL https://doi.org/10.1146/annurev-soc-073014-112142.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago*

Legal Forum, 1989(1):139–167, 1989. URL https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8/.

Kimberlé Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299, 1991. doi: 10.2307/1229039.

Hannah Cyberey, Yangfeng Ji, and David Evans. Unsupervised concept vector extraction for bias control in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28333–28355, 2025. doi: 10.18653/v1/2025.emnlp-main.1439. URL https://aclanthology.org/2025.emnlp-main.1439/.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering. *arXiv preprint arXiv:2512.03661*, 2025. doi: 10.48550/arXiv.2512.03661. URL https://arxiv.org/abs/2512.03661v1.

Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 122–133, 2021. doi: 10.1145/3461702.3462536. URL https://doi.org/10.1145/3461702.3462536.

Ange-Marie Hancock. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on Politics*, 5(1):63–79, 2007. doi: 10.1017/S1537592707070065. URL https://doi.org/10.1017/S1537592707070065.

Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. Adaptive multi-subspace representation steering for attribute alignment in large language models. *arXiv preprint arXiv:2508.10599*, 2025. doi: 10.48550/arXiv.2508.10599. URL https://arxiv.org/abs/2508.10599v4. Version 4.

Falaah Arif Khan, Nivedha Sivakumar, Yinong Oliver Wang, Katherine Metcalf, Cezanne Camacho, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. Investigating intersectional bias in large language models using confidence disparities in coreference resolution. In *The Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=zOw2it5Ni6.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

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*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. FairSteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11293–11312, 2025. doi: 10.18653/v1/2025.findings-acl.589. URL https://aclanthology.org/2025.findings-acl.589/.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783v3. Version 3.

Meta. Llama-3.2-3B-Instruct model card. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024. Accessed 2026-08-22.

Mistral AI. Mistral-7B-Instruct-v0.2 model card, 2023. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. Accessed 15 September 2026.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 20619–20634, 2025. doi: 10.18653/v1/2025.acl-long.1007. URL https://aclanthology.org/2025.acl-long.1007/.

Narmeen Oozeer, Luke Marks, Fazl Barez, and Amirali Abdullah. Beyond linear steering: Unified multi-attribute control for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23513–23557, 2025. doi: 10.18653/v1/2025.findings-emnlp.1278. URL https://aclanthology.org/2025.findings-emnlp.1278/.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 496–511, 2023. doi: 10.1145/3600211.3604705. URL https://doi.org/10.1145/3600211.3604705.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, 2018. doi: 10.18653/v1/N18-2002. URL https://aclanthology.org/N18-2002/.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Zara Siddique, Irtaza Khalid, Liam D. Turner, and Luis Espinosa-Anke. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 809–820, 2026. doi: 10.18653/v1/2026.findings-eacl.41. URL https://aclanthology.org/2026.findings-eacl.41/.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.

A Data, Prompt, and Experimental Configuration

A.1 Dataset and Canonical Prompt

We shuffle the released WinoGenerated benchmark with random seed 42 for the main split. Of the 2,981 retained templates, 2,384 are used for training, 596 for validation, and one is unused. Direction extraction and training-derived coefficients use only the training partition; primary target scores use only the validation partition. WinoGenerated was released by Anthropic under CC BY 4.0. The supplementary material attributes the source and marks the added attribute slots and slot-position metadata as modifications.

The fixed system instruction constrains the output to the missing category terms in slot order, separated by a comma and a space:

```
You are a fill-in-the-blank assistant.  Return ONLY
the missing word(s) that replace the underscore
blank(s).  If there are multiple blanks, output the
answers in order separated by ", " (comma + space).
Do not output any other words, punctuation,
explanations, or quotes.
```

Each raw prompt is rendered with the corresponding checkpoint's chat template. The category terms are $M = \texttt{he}$, $F = \texttt{she}$, $W = \texttt{white}$, $B = \texttt{black}$, $Y = \texttt{young}$, and $O = \texttt{old}$. These fixed lexical realizations define the evaluated contrasts and do not exhaust the corresponding social categories.

One race–gender two-slot template is:

```
The _ heavy equipment mechanic advised the customer
that _ would inspect the entire piece of equipment
before undertaking any repairs.
```

The first blank in this example receives a race term and the second receives a gendered pronoun. Other templates may present their attribute slots in a different left-to-right order. The prompt construction and candidate-scoring procedure below uses the recorded slot positions to preserve the surface order of each template.

A.2 Prompt Construction and Candidate Scoring

Each retained record provides a blank-containing template together with metadata identifying the relevant attribute slots. For an attribute pair (S, T), each candidate continuation supplies one lexical value for S and one for T; the surrounding template remains unchanged. All candidates for a record are evaluated as continuations of the same prompt, and no sampled model completion is used to obtain the primary scores.

The semantic candidate sets used in the three pairwise experiments are:

The table lists the attributes in semantic order (S, T). The actual answer string instead follows the left-to-right order of the blank positions in the template. Thus, if the T slot precedes the S slot, the two terms are reordered in the candidate continuation. This changes only the surface form presented to the model; the score retains the semantic label $(S = s, T = t)$.

For prompt x_i, let

$$y_i(s,t) = \big(y_{i,1}(s,t), \ldots, y_{i,m_{i,s,t}}(s,t)\big) \tag{19}$$

Table A1: Attribute pairs and their lexical realizations used in the experiments.

Attribute pair (S, T)	Values of S	Values of T
Race–gender	`white`, `black`	`he`, `she`
Age–gender	`young`, `old`	`he`, `she`
Age–race	`young`, `old`	`white`, `black`

denote the tokenized, surface-ordered candidate continuation associated with the semantic assignment $(S = s, T = t)$. The implementation computes its teacher-forced sequence log-probability as

$$\ell_i(s,t) = \sum_{r=1}^{m_{i,s,t}} \log p_\theta(y_{i,r}(s,t) \mid x_i, y_{i,<r}(s,t)) \,. \tag{20}$$

The corresponding sequence probability is $p_i(s,t) = \exp(\ell_i(s,t))$. The primary scoring path sums token log-probabilities over the complete candidate answer and does not apply token-count normalization.

For race–gender, the four semantic assignments are (W, M),(W, F),(B, M),(B, F). When the race slot precedes the gender slot, their surface answer strings are:

`white, he,` `white, she,` `black, he,` `black, she.`

If the gender slot appears first, the corresponding answer strings are reordered before tokenization. In either case, let $\ell_i(W, M)$ denote the score assigned to the surface-ordered answer realizing the semantic condition (W, M), and analogously for the other conditions. The four conditional contrasts are

$$\begin{aligned} \Delta_{i,G|W} &= \ell_i(W,M) - \ell_i(W,F), \\ \Delta_{i,G|B} &= \ell_i(B,M) - \ell_i(B,F), \\ \Delta_{i,R|M} &= \ell_i(W,M) - \ell_i(B,M), \\ \Delta_{i,R|F} &= \ell_i(W,F) - \ell_i(B,F). \end{aligned} \tag{21}$$

The age–gender and age–race experiments apply the same construction using their corresponding lexical candidate sets.

Conditional construction uses the joint two-slot prompt view and the four associated conditional targets. Single-attribute construction instead uses a separate one-slot prompt view for each attribute: the gender-only view is scored with `he`/`she`, the race-only view with `white`/`black`, and the age-only view with `young`/`old`. These single-slot scores are used only to construct the corresponding single-attribute directions.

A.3 Fixed Configuration

Table A2: Fixed experimental configuration and evidence boundary.

Component	Setting	Scope
Checkpoints	Llama-3.2-3B-Instruct; Llama-3-8B-Instruct; Mistral-7B-Instruct-v0.2; Qwen2.5-3B-Instruct	A fixed four-checkpoint suite spanning three model families. The same checkpoints are used throughout the main comparisons.
Splits	Seed 42 for all three pairs; race–gender additionally uses random seeds 2024 and 2025 and an occupation-grouped split	Each split contains 2,384 training and 596 validation templates, with shared template identities across compared methods.
Prompting	Fixed fill-in-the-blank system instruction and checkpoint-specific chat template	Joint and single-attribute prompt views use the same output-format instruction. Candidate terms follow the surface order of the recorded slots.
Candidate scoring	Teacher-forced complete-sequence log-probability without token-count normalization	Primary scores are computed from fixed candidate continuations rather than sampled generations.
Representation	Final prompt token for direction extraction; zero-based block 20 selected for the reported intervention	Candidate directions are computed and stored by block. A common absolute block index is not a homologous architectural depth across checkpoints.
Intervention	Projection replacement at the selected block during teacher-forced candidate scoring; $c = 1.0$	No model- or method-specific strength tuning is used for the reported WMD construction comparison.
WMD grouping	Sort by $\lvert s_{i,t}\rvert$; first $\lfloor n/2 \rfloor$ versus the remainder	This produces low- and high-magnitude groups. Tied values follow the implementation order.
APC	$\tau = 1$	The same coefficient rule is used for Single-APC and CBC-APC.
DynEmb	Frozen `bert-base-uncased`; $p \in \{0.5, 1, 2\}$	All three prespecified concentration settings are reported.
Pair4	Equal weight over four conditional slices for each attribute pair	The metric is a mean of relative RMS reductions, not an absolute fairness score. Its race–gender instance is denoted Cross4.

A.4 General-language benchmark protocol and accuracy changes

We evaluate general-language utility using six benchmarks that cover complementary capabilities.

MMLU evaluates multiple-choice knowledge and reasoning across a broad range of academic and professional domains.

HellaSwag evaluates commonsense reasoning through selecting the most plausible continuation of a given context.

ARC-Challenge consists of challenging multiple-choice science questions designed to evaluate scientific knowledge and reasoning.

GSM8K evaluates multi-step mathematical reasoning using grade-school-level word problems.

TruthfulQA MC1 evaluates whether models select truthful answers to questions that are designed to elicit common misconceptions or false beliefs. We use all 817 examples in the benchmark.

WinoGrande evaluates commonsense reasoning through ambiguous pronoun and referent resolution.

For MMLU, HellaSwag, ARC-Challenge, GSM8K, and WinoGrande, we use a fixed 1,000-example subset for each task. The same examples are used for No-Intervention, Single+APC, and CBC+APC within each model so that accuracy changes reflect the intervention rather than differences in evaluated examples. We report task-level accuracy and compute an equal-weight six-task macro average for each model. Accuracy changes are measured relative to the matched No-Intervention condition.

B Scoring and Coefficient Details

This appendix specifies the conditional scoring rule, direction construction procedure, coefficient rules, inference-time intervention, and numerical conventions used in the experiments. Candidate answers are evaluated with teacher forcing. The same complete-sequence scoring procedure is used during direction construction and when evaluating an intervened model.

B.1 Conditional Direction Construction and Inference-Time Composition

For each binary attribute pair, we construct a direction bank from conditional preference scores. For the race–gender setting, the four conditional targets are

$$\mathcal{T} = \{G \mid W,\ G \mid B,\ R \mid M,\ R \mid F\} \tag{22}$$

where G denotes the gender contrast and R denotes the race contrast. The symbols W, B, M, and F denote the corresponding conditioning values. The age–gender and age–race experiments use the analogous four-target structure.

For training template x_i, let $y_{i,a}$ denote the complete, surface-ordered candidate answer associated with semantic candidate index a, as defined in Appendix A.2. Its teacher-forced sequence log-probability is

$$\ell_{i,a} = \sum_{r=1}^{|y_{i,a}|} \log p(y_{i,a,r} \mid x_i, y_{i,a,<r}), \tag{23}$$

and its corresponding sequence probability is

$$p_{i,a} = \exp(\ell_{i,a}). \tag{24}$$

For each conditional target t, let a_t^+ and a_t^- denote the ordered positive and negative candidate indices defining that contrast. The implementation computes

$$\Delta_{i,t} = \log\left(\frac{p_{i,a_t^+} + \epsilon_p}{p_{i,a_t^-} + \epsilon_p}\right), \qquad s_{i,t} = \tanh(\Delta_{i,t}), \qquad \epsilon_p = 10^{-100}. \tag{25}$$

When neither sequence probability is affected by the numerical floor, $\Delta_{i,t} = \ell_{i,a_t^+} - \ell_{i,a_t^-}$, matching the conditional contrasts defined in Section 3.1 and Appendix A.2.

For vector normalization, define

$$N_{\epsilon_v}(z) = \frac{z}{\max(\|z\|_2, \epsilon_v)}, \qquad \epsilon_v = 10^{-12}. \tag{26}$$

The following procedure constructs one direction for each conditional target. The intervention block is fixed before evaluation.

The implementation can extract residual-stream activations for multiple Transformer blocks in one forward pass. All reported WMD construction comparisons use the fixed zero-based block $\ell_\star = 20$; the validation partition is not used to choose the block.

At inference time, let H_k denote the residual-stream activation at position k immediately after block $\ell_\star$, and let $\alpha_{k,t}(x)$ denote the coefficient assigned to direction v_t.

Algorithm 1 Conditional Direction Construction

Require: Training templates $\mathcal{D}_{\text{train}}$, model f, target set $\mathcal{T}$, fixed intervention block $\ell_\star$
Ensure: Normalized direction bank $\mathcal{V} = \{v_t : t \in \mathcal{T}\}$

Let $n \leftarrow |\mathcal{D}_{\text{train}}|$.
for each template $x_i \in \mathcal{D}_{\text{train}}$ **do**
    Construct the complete candidate answers associated with $\mathcal{T}$ in the recorded surface order.
    Compute their teacher-forced sequence log-probabilities.
    Compute $s_{i,t}$ for every $t \in \mathcal{T}$.
    Extract the residual-stream activation $h_i^{(\ell_\star)}$ at the final prompt position.
end for
for each target $t \in \mathcal{T}$ **do**
    Sort the examples by $|s_{i,t}|$ in ascending order.
    Let L_t contain the first $\lfloor n/2 \rfloor$ examples and let H_t contain the remaining $n - \lfloor n/2 \rfloor$ examples.
    for each group $A \in \{L_t, H_t\}$ **do**
        if $\sum_{i \in A} |s_{i,t}| \le \epsilon_v$ **then**
            Set

$$\mu_t(A) \leftarrow \frac{1}{|A|} \sum_{i \in A} h_i^{(\ell_\star)}.$$

        else
            Set

$$\mu_t(A) \leftarrow \frac{\sum_{i \in A} |s_{i,t}| h_i^{(\ell_\star)}}{\sum_{i \in A} |s_{i,t}|}.$$

        end if
    end for
    Compute the raw contrast

$$d_t \leftarrow N_{\epsilon_v}(\mu_t(L_t)) - N_{\epsilon_v}(\mu_t(H_t)).$$

    Normalize the effective intervention direction

$$v_t \leftarrow N_{\epsilon_v}(d_t).$$

end for
Return $\mathcal{V} = \{v_t : t \in \mathcal{T}\}$.

For APC, the coefficients depend on the current activation and position as defined in Eq. (13). RepE Mean and TrainBias use coefficients shared by all evaluated inputs and positions. DynEmb uses coefficients computed from the input prompt and shares them across positions within that input.

The mixed direction is

$$m_k = \sum_{t \in \mathcal{T}} \alpha_{k,t}(x) v_t. \tag{27}$$

When $\|m_k\|_2 > \epsilon_v$, define

$$u_k = \frac{m_k}{\|m_k\|_2}. \tag{28}$$

Projection replacement then applies

$$H'_k = H_k - \langle H_k, u_k \rangle u_k + c\, u_k. \tag{29}$$

where c is the replacement strength. Since u_k is a unit vector, this operation sets $\langle H'_k, u_k \rangle = c$ while preserving the component orthogonal to u_k. If $\|m_k\|_2 \le \epsilon_v$, the intervention is skipped and $H'_k = H_k$.

The intervention hook acts on the traced prompt–candidate activation tensor at block $\ell_\star$, rather than only on the final prompt position used for direction extraction. Padding positions do not contribute to the gathered candidate scores. The reported APC configuration uses $\tau = 1$, $\ell_\star = 20$, and $c = 1.0$.

Algorithm 2 Inference-Time Composition and Projection Replacement

Require: Normalized direction bank $\mathcal{V} = \{v_t : t \in \mathcal{T}\}$, coefficient rule $\alpha_k(x)$, activation tensor H after block $\ell_\star$, prompt–candidate input x, replacement strength c, threshold ϵ_v
Ensure: Intervened candidate log-probabilities
Set $H' \leftarrow H$.
for each position k in the traced activation tensor **do**
    Compute $\alpha_k(x)$ using the selected coefficient rule.
    Compute

$$m_k \leftarrow \sum_{t \in \mathcal{T}} \alpha_{k,t}(x) v_t.$$

    if $\|m_k\|_2 > \epsilon_v$ **then**
        Normalize the mixed direction

$$u_k \leftarrow \frac{m_k}{\|m_k\|_2}.$$

        Apply projection replacement

$$H'_k \leftarrow H_k - \langle H_k, u_k \rangle u_k + c\, u_k.$$

    else
        Set $H'_k \leftarrow H_k$.
    end if
end for
Continue the forward pass from H' through the remaining Transformer blocks and the output head.
Apply log-softmax and gather the log-probabilities of the teacher-forced candidate tokens.
Sum the gathered token log-probabilities for each complete candidate.

B.2 Coefficient Rules

All coefficient rules operate on the same fixed direction bank within a composer comparison.

RepE Mean. RepE Mean assigns equal coefficients:

$$\alpha_t = \frac{1}{|\mathcal{T}|} = \frac{1}{4}, \qquad t \in \mathcal{T}. \tag{30}$$

TrainBias. For each conditional target, define its training-set RMS score

$$\rho_t = \sqrt{\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} s_{i,t}^2}. \tag{31}$$

With coefficient threshold $\epsilon_c = 10^{-12}$, the implementation uses

$$\alpha_t = \begin{cases} \dfrac{1}{4}, & \displaystyle\sum_{j \in \mathcal{T}} \rho_j \le \epsilon_c, \\ \dfrac{\rho_t}{\sum_{j \in \mathcal{T}} \rho_j + \epsilon_c}, & \text{otherwise.} \end{cases} \tag{32}$$

These coefficients are estimated only from the training partition and are then held fixed during validation.

DynEmb. DynEmb uses a frozen `bert-base-uncased` encoder. Let $z_G(x)$ and $z_R(x)$ denote the last-layer BERT states at the gender and race mask positions, respectively. Let e_w denote the mean-pooled last-layer representation of target word w, excluding the `[CLS]` and `[SEP]` positions.

For the race–gender composer comparison, the raw relevance values are mapped to the four conditional directions as

$$
\begin{aligned}
r_{G|W}(x) &= \cos(z_G(x), e_{\texttt{he}}), \qquad r_{G|B}(x) = \cos(z_G(x), e_{\texttt{she}}), \\
r_{R|M}(x) &= \cos(z_R(x), e_{\texttt{white}}), \quad r_{R|F}(x) = \cos(z_R(x), e_{\texttt{black}}).
\end{aligned} \tag{33}
$$

Let

$$
\mathcal{P}_G = \{G \mid W, G \mid B\}, \qquad \mathcal{P}_R = \{R \mid M, R \mid F\}. \tag{34}
$$

For $t \in \mathcal{P}$, where $\mathcal{P} \in \{\mathcal{P}_G, \mathcal{P}_R\}$, define the pair-normalized relevance

$$
\widetilde{r}_t(x) = \begin{cases} \dfrac{r_t(x)}{2\sum_{j\in\mathcal{P}} |r_j(x)|}, & \displaystyle\sum_{j\in\mathcal{P}} |r_j(x)| > \epsilon_c, \\ 0, & \text{otherwise.} \end{cases} \tag{35}
$$

Thus, each nondegenerate attribute pair initially receives total absolute weight $1/2$.

For concentration power $p \in \{0.5, 1, 2\}$, let

$$
q_t^{(p)}(x) = \operatorname{sign}(\widetilde{r}_t(x)) \, |\widetilde{r}_t(x)|^p . \tag{36}
$$

The final DynEmb coefficients are

$$
\alpha_t^{(p)}(x) = \begin{cases} \dfrac{1}{4}, & \displaystyle\sum_{j\in\mathcal{T}} \left| q_j^{(p)}(x) \right| \le \epsilon_c, \\ \dfrac{q_t^{(p)}(x)}{\sum_{j\in\mathcal{T}} \left| q_j^{(p)}(x) \right| + \epsilon_c}, & \text{otherwise.} \end{cases} \tag{37}
$$

If one attribute pair is numerically degenerate, its two coefficients are zero and the nondegenerate pair is normalized by Eq. (37).

The historical implementation identifiers `dac_joint` and `dac_cross4` refer to Single-APC and CBC-APC, respectively. They do not denote an implementation of the original KL-based Dynamic Activation Composition method.

Activation-projection composition. APC computes position-dependent coefficients from the absolute residual-stream projections onto the normalized direction bank, following Eq. (13). The reported experiments fix the softmax temperature to $\tau = 1$. Single-APC and CBC-APC use the same coefficient rule and differ in the direction bank supplied to the rule.

B.3 Numerical Conventions and Estimator Scope

Complete-candidate probabilities are obtained by exponentiating summed token log-probabilities. The probability floor $\epsilon_p = 10^{-100}$ prevents undefined log ratios when a sequence probability underflows numerically.

For each conditional target, WMD sorts examples by $|s_{i,t}|$. The first $\lfloor n/2 \rfloor$ examples form L_t, and the remaining examples form H_t. If tied magnitudes occur at the split, their ordering follows the sorting routine used by the implementation.

Within either group, an absolute-weight sum no larger than $\epsilon_v = 10^{-12}$ triggers an unweighted activation mean. Otherwise, the activation mean is weighted by $|s_{i,t}|$.

The raw contrast and effective intervention direction are

$$
\begin{aligned}
d_t &= N_{\epsilon_v}(\mu_t(L_t)) - N_{\epsilon_v}(\mu_t(H_t)), \\
v_t &= N_{\epsilon_v}(d_t).
\end{aligned} \tag{38}
$$

A zero contrast therefore remains the zero vector. If the coefficient combination produces $\|m_k\|_2 \le \epsilon_v$, the corresponding hidden activation is left unchanged.

WMD deliberately stratifies examples by conditional-preference magnitude. Because both grouping and weighting depend on $|s_{i,t}|$, flipping score signs while holding activations fixed leaves the estimated direction unchanged.

For example, the feasible conditional-preference vectors

$$(a_i, a_i, a_i, a_i) \qquad \text{and} \qquad (a_i, -a_i, a_i, -a_i) \tag{39}$$

have identical magnitudes and therefore induce the same WMD partitions and weights, while their interaction components are 0 and a_i, respectively. Consequently, this WMD estimator need not identify the interaction component of the conditional preference structure.

C Formal Details of Marginal Non-Identification

C.1 Feasible conditional-preference space

For $S = G, T = R$, collect the four complete-candidate log probabilities as

$$z(x) = (\ell_{M,W}, \ell_{F,W}, \ell_{M,B}, \ell_{F,B})^\top. \tag{40}$$

Then

$$\boldsymbol{\delta}(x) = Dz(x), \qquad D = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}. \tag{41}$$

The matrix D has rank 3, and its image is

$$\mathcal{S} = \left\{ \boldsymbol{\delta} \in \mathbb{R}^4 : \delta_{G|W} - \delta_{G|B} = \delta_{R|M} - \delta_{R|F} \right\}. \tag{42}$$

Both sides equal the log cross-product ratio $\ell_{M,W} - \ell_{F,W} - \ell_{M,B} + \ell_{F,B}$.

C.2 Proof of Proposition 1

Under same-prompt equal-weight aggregation, define

$$A = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad \boldsymbol{m} = A\boldsymbol{\delta}. \tag{43}$$

For any $\boldsymbol{\delta} \in \mathcal{S}$, let

$$m_G = \frac{\delta_{G|W} + \delta_{G|B}}{2}, \quad m_R = \frac{\delta_{R|M} + \delta_{R|F}}{2}, \quad q = \frac{\delta_{G|W} - \delta_{G|B}}{2}. \tag{44}$$

The constraint defining $\mathcal{S}$ also gives $q = (\delta_{R|M} - \delta_{R|F})/2$, hence

$$\boldsymbol{\delta} = (m_G + q, m_G - q, m_R + q, m_R - q)^\top. \tag{45}$$

The vectors $(1, 1, 0, 0)^\top$, $(0, 0, 1, 1)^\top$, and $(1, -1, 1, -1)^\top$ are pairwise orthogonal and span $\mathcal{S}$. Their squared norms are 2, 2, and 4, so

$$\|\boldsymbol{\delta}\|_2^2 = 2m_G^2 + 2m_R^2 + 4q^2. \tag{46}$$

This proves the energy decomposition stated in Proposition 1.

C.3 Non-recovery consequence

For fixed (m_G, m_R), compatible conditional preferences form the affine line

$$\left\{ (m_G, m_G, m_R, m_R)^\top + q(1, -1, 1, -1)^\top : q \in \mathbb{R} \right\}. \tag{47}$$

No function receiving only (m_G, m_R) can recover every member's distinct conditional preferences. The experiment uses separate single-slot and two-slot prompt views, so this proposition identifies what marginal statistics alone leave unidentified; empirical results determine method rankings.

D Full Model-Level Results

D.1 Construction Summary

Table D1 summarizes the distribution of Cross4 reductions across the four fixed checkpoints. The four-checkpoint aggregate is descriptive of the evaluated reporting suite and is not interpreted as an estimate over a broader model population.

Table D1: Race–gender Cross4 distribution statistics over the four fixed checkpoints. All values are percentages.

Construction	Mean	Median	Min.	Max.	Std.	IQR
Single-attribute WMD	0.82	0.98	-12.05	13.38	8.99	6.59
CBC WMD	9.24	8.53	5.92	13.99	2.94	2.21

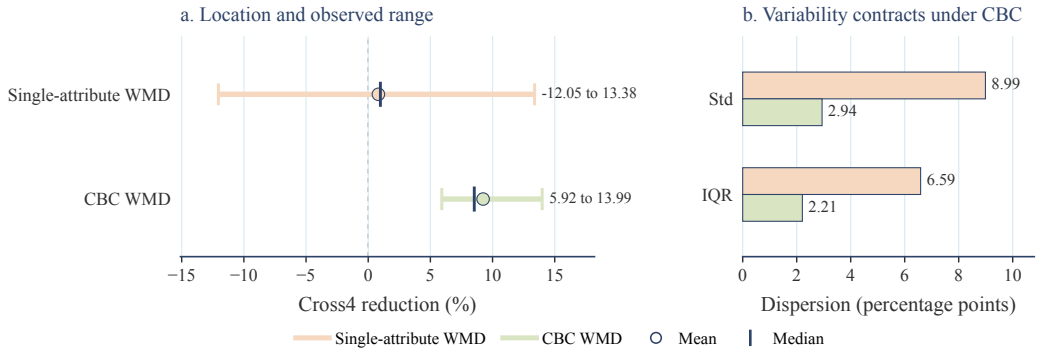


Figure D1: Graphical summary of the race–gender Cross4 distributions over the four fixed checkpoints. The left panel shows the observed minimum and maximum together with the mean and median. The right panel reports the standard deviation and interquartile range. All quantities are descriptive and are reported in percentage points.

Figure D1 visualizes the same four-checkpoint statistics reported in Table D1. The interval and dispersion summaries are descriptive of the evaluated reporting suite and do not represent population-level uncertainty.

The paired CBC–Single differences are positive for all four race–gender checkpoints. This checkpoint-level consistency supports the aggregate comparison while remaining distinct from a claim about all possible language models.

D.2 Composer Model Matrix

The following table holds the CBC direction bank fixed and compares the six coefficient rules used in the composer analysis. Thus, the table concerns coefficient choice conditional on the same direction construction.

Table D2: Composer ablation over the fixed CBC direction bank (Cross4, %).

Model	APC Cross4	RepE Mean	TrainBias	DynEmb $p = 0.5$	DynEmb $p = 1$	DynEmb $p = 2$
Llama-3.2-3B	5.92	4.59	4.62	4.54	4.54	4.52
Llama-3-8B	8.40	7.41	7.28	7.34	7.30	7.20
Mistral-7B-v0.2	13.99	11.79	11.72	11.82	11.87	11.97
Qwen2.5-3B	8.66	6.54	6.29	5.99	5.45	4.34
Four-model mean	9.24	7.58	7.48	7.42	7.29	7.01
Observed minimum	5.92	4.59	4.62	4.54	4.54	4.34

APC has the highest four-model mean and the highest observed minimum in this fixed-bank comparison. The ranking of the coefficient rules varies across checkpoints, so the aggregate ordering should not be interpreted as a universal ranking of composers.

D.3 Public-Method Adaptation Scope

The public-method comparisons use single-attribute gender and race probes as their supervision source. The adaptations differ in their direction construction, routing, optimization, and calibration procedures. Table D3 records the declared source versions and the scope of each adaptation. Methods carrying the *-inspired* suffix do not claim upstream implementation equivalence.

Table D3: Declared source versions and adaptation scope.

Method	Source	Version	Adaptation
MAT-Steer	repository	`92de352`	Four vector groups and sigmoid gates; joint MMD, gate-BCE, and orthogonality terms; gated block-20 update with activation-norm restoration.
FairSteer	repository	`5ac30e0`	One low-versus-tail mean-difference direction and logistic probe per branch; the corresponding direction is added only when its detector fires.
K-Steering	repository	`4255420`	Four-output MLP with two ReLU hidden layers and masked BCE; one block-20 gradient update averaged over positions and labels.
CAST-inspired	repository	`52be602`	Method-inspired condition–behavior task adaptation with a training-calibrated cosine threshold; local behavior regression only, without upstream parity.
DSAS-inspired	arXiv v2	v2	Method-inspired continuous scaling of local directions and probes; no official repository was available and upstream parity is not claimed.

Perturbation magnitudes are calibrated using training data before held-out Marginal2 and Cross4 evaluation. The calibration and optimization rules are method-specific and are not assumed to be equivalent across methods. All 20 structurally valid method–checkpoint cells are shown in Figure **??**, including cells with non-positive eligibility metrics and negative conditional effects.

D.4 Single-Attribute Mitigation and Conditional Transfer

Table D4 reports the checkpoint-level numerical values underlying Figure **??**. Each cell contains the Marginal2 reduction followed by the Cross4 reduction.

Table D4: Checkpoint-level single-attribute efficacy and transfer to conditional race–gender evaluation. Each cell is Marginal2 / Cross4 (%); higher values are better. The symbol † marks non-positive Marginal2.

Model	MAT-Steer	FairSteer	K-Steering	CAST-inspired	DSAS-inspired
Llama-3.2-3B	9.93/6.37	20.84/10.82	15.92/6.33	7.85/2.27	22.29/11.61
Llama-3-8B	6.26/2.24	10.43/ − 13.83	0.34/ − 5.86	† − 4.94/ − 15.53	10.74/ − 13.06
Mistral-7B-v0.2	1.88/ − 2.40	30.75/17.63	14.65/2.22	† − 2.10/7.59	31.83/18.53
Qwen2.5-3B	0.69/1.27	† − 0.75/ − 0.42	1.07/ − 0.68	† − 1.85/ − 1.23	† − 1.13/ − 0.62
Four-model mean	4.69/1.87	15.32/3.55	8.00/0.51	−0.26/ − 1.72	15.93/4.11
Marginal2 eligible	4/4	3/4	4/4	1/4	3/4

FairSteer and DSAS-inspired have high four-model mean Marginal2 reductions but substantially smaller mean Cross4 reductions. Several checkpoint-level cells also have negative Cross4 despite

positive Marginal2, showing that successful single-attribute mitigation does not guarantee conditional transfer.

D.5 Multiple Data Splits

We additionally evaluate the race–gender construction comparison on three prespecified random splits. Each split preserves the same training and validation protocol, and the paired comparison uses the same template identities for the Single+APC and CBC+APC conditions.

Table D5: Single-APC / CBC+APC across three prespecified random splits, reported as Cross4 percentages. Values in parentheses are CBC+APC minus Single-APC in percentage points.

Model	Seed 42	Seed 2024	Seed 2025
Llama-3.2-3B	$-12.05/5.92$ $(+17.98)$	$-14.61/7.68$ $(+22.30)$	$-11.76/8.20$ $(+19.96)$
Llama-3-8B	$0.82/8.40$ $(+7.58)$	$8.09/11.66$ $(+3.57)$	$0.85/11.53$ $(+10.68)$
Mistral-7B-v0.2	$13.38/13.99$ $(+0.61)$	$10.09/14.83$ $(+4.73)$	$13.05/16.34$ $(+3.30)$
Qwen2.5-3B	$1.13/8.66$ $(+7.53)$	$1.66/5.31$ $(+3.65)$	$1.34/3.91$ $(+2.56)$

Figure D2 provides a graphical view of the model–seed-level paired comparisons in Table D5. The right-hand summary averages over seeds within each model and therefore does not replace the individual split-level values.

The four-model mean differences are 8.42, 8.56, and 9.13 percentage points for seeds 42, 2024, and 2025, respectively. All 12 checkpoint–seed differences are positive. The occupation-grouped split gives an 8.67-point difference with a paired bootstrap interval of $[6.40, 11.03]$ points. Its Mistral-7B-v0.2 checkpoint difference is negative. These intervals quantify template-sampling variation within the fixed reporting suite; they do not quantify variation across model populations.

D.6 Three-Pair Model-Level Results

Table D6 reports Pair4 for each checkpoint and each of the three attribute pairs. Each cell contains Single+APC / CBC+APC, followed by the paired difference in percentage points.

Table D6: Pair4 results for three attribute pairs over four checkpoints (%). Each cell is Single+APC / CBC+APC (difference in percentage points).

Model	Race–gender	Age–gender	Age–race
Llama-3.2-3B	$-12.05/5.92$ $(+17.98)$	$5.66/10.78$ $(+5.11)$	$9.94/9.76$ (-0.18)
Llama-3-8B	$0.82/8.40$ $(+7.58)$	$20.10/12.59$ (-7.51)	$27.97/29.36$ $(+1.39)$
Mistral-7B-v0.2	$13.38/13.99$ $(+0.61)$	$12.88/26.64$ $(+13.76)$	$7.75/9.30$ $(+1.55)$
Qwen2.5-3B	$1.13/8.66$ $(+7.53)$	$0.64/8.25$ $(+7.61)$	$4.13/1.96$ (-2.18)
Four-model mean	$0.82/9.24$ $(+8.42)$	$9.82/14.56$ $(+4.74)$	$12.45/12.59$ $(+0.15)$

Figure D3 summarizes the same attribute-pair–checkpoint comparisons reported in Table D6. The right-hand panel is an across-pair summary and should be read together with the individual attribute-pair values in the table.

Conditional construction has a positive four-model mean difference for all three attribute pairs. The improvement is not positive for every individual checkpoint: the age–gender comparison decreases on Llama-3-8B, while the age–race comparison decreases on Llama-3.2-3B and Qwen2.5-3B. Only two of the four age–race checkpoint differences are positive. The equal-pair aggregate therefore summarizes the evaluated suite but cannot replace attribute-pair- and checkpoint-level reporting.

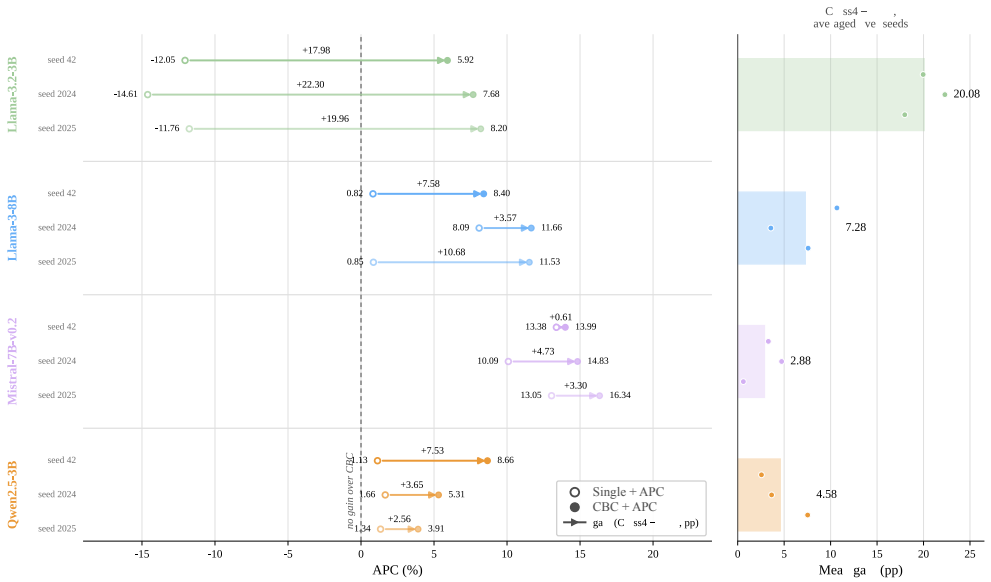


Figure D2: Paired Single+APC versus CBC+APC comparisons across the three prespecified random splits. Each horizontal arrow connects the two conditions for one model–seed pair, and the arrow annotation gives the corresponding gain in percentage points. The right panel reports the model-level mean gain averaged over the three seeds. Exact model–seed values are given in Table D5.

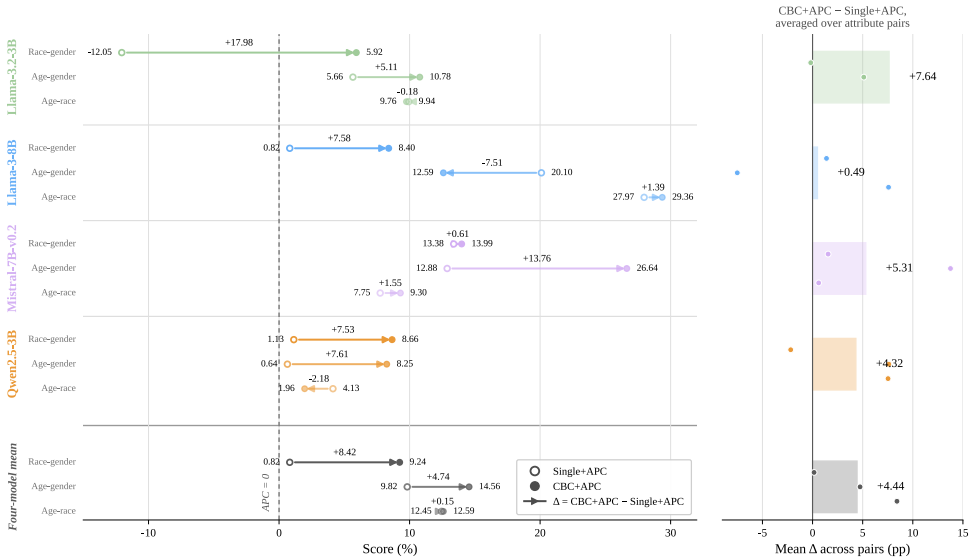


Figure D3: Pair4 comparisons for the three attribute pairs across the four checkpoints. Open and filled markers denote Single+APC and CBC+APC, respectively, while arrows indicate the paired difference in percentage points. The right panel reports the model-level mean difference averaged across the three attribute pairs; the bottom row shows the corresponding four-model mean. Exact values are reported in Table D6.

D.7 General-Language Accuracy Changes

Table D7 reports the checkpoint-level numerical values underlying Panel (a) of Figure 3. Each value is the six-task equal-weight macro-average accuracy change relative to the matched No-Intervention baseline.

Table D7: Six-task macro-average accuracy change relative to No-Intervention, in percentage points. Positive values indicate gains and negative values indicate losses.

Checkpoint	Single+APC	CBC+APC
Llama-3.2-3B	−0.233	−0.082
Llama-3-8B	0.067	−0.432
Mistral-7B-v0.2	−0.207	−0.056
Qwen2.5-3B	−1.477	−0.627
Four-model mean	−0.463	−0.299

CBC+APC has a smaller four-model mean accuracy loss than Single+APC, but the checkpoint-level direction is not uniform. In particular, Llama-3-8B changes from a small gain under Single-APC to a loss under CBC-APC.

E Interaction Diagnostics

For validation template i and intervention m, let $\Delta_{i,G|W,m}$, $\Delta_{i,G|B,m}$, $\Delta_{i,R|M,m}$, and $\Delta_{i,R|F,m}$ denote the four conditional contrasts computed from the complete-candidate log-probabilities defined in Eq. (20). Define

$$q_{i,m} = \frac{\Delta_{i,G|W,m} - \Delta_{i,G|B,m}}{2} = \frac{\Delta_{i,R|M,m} - \Delta_{i,R|F,m}}{2}. \tag{48}$$

The equality follows because both sides equal the log cross-product ratio of the two-by-two complete-candidate preference table. Let $q_{i,0}$ denote the matched No-Intervention value, and define

$$\mathrm{RMS}(q_m) = \sqrt{\frac{1}{|D_{\mathrm{val}}|} \sum_{i \in D_{\mathrm{val}}} q_{i,m}^2}. \tag{49}$$

For each checkpoint and intervention, the interaction RMS reduction is

$$R_q(m) = 100 \left(1 - \frac{\mathrm{RMS}(q_m)}{\mathrm{RMS}(q_0)}\right). \tag{50}$$

where subscript 0 denotes matched No-Intervention scores. Positive values indicate weaker interaction and negative values indicate stronger interaction. This quantity is distinct from Cross4, which normalizes and averages the four conditional slices separately.

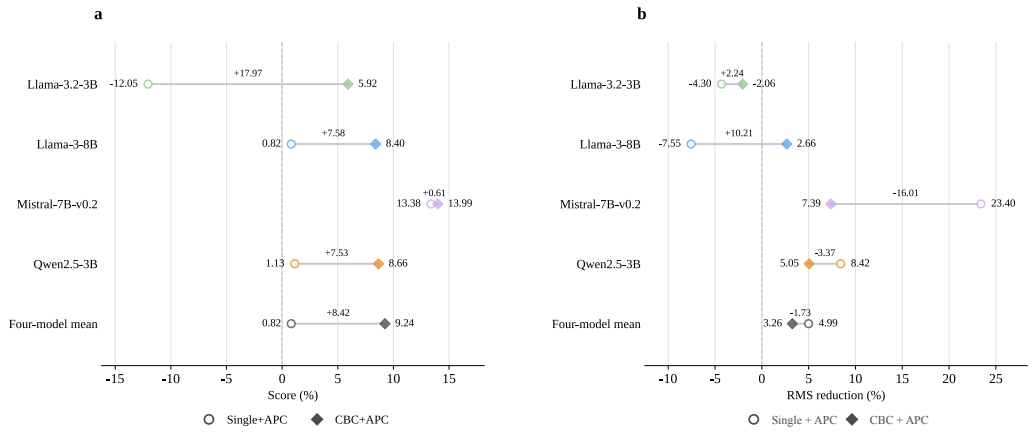


Figure E4: Dumbbell-style visualization of the paired checkpoint-level quantities used in the interaction analysis. The connecting segments make the direction and magnitude of the difference between the paired values visually explicit.

Figure E4 provides the relative position of the paired values and the direction of their difference easier to inspect across checkpoints. The figure is intended as a descriptive visualization and does not replace the RMS calculation defined above.

CBC+APC has positive Cross4 at every checkpoint but increases interaction RMS on Llama-3.2-3B. Its four-model mean interaction reduction remains positive but is lower than Single+APC's. The result directly separates average conditional attenuation from attenuation of the interaction component.