Supplementary Material

This supplementary material accompanies our main paper on debiasing VLMs with task vector negation. It provides additional sensitivity analyses (Section S1), the full per-category selection maps across all task-vector variants (Section S2), and extended qualitative examples (Section S3). All experiments use LLaVA-1.5 on FairFace unless otherwise noted

S1 SENSITIVITY ANALYSIS

Figure S1 shows the sensitivity of mean DI to the negation strength λ for the Occupation category, for the gender and race task vectors extracted at different fine-tuning epochs (LLaVA-1.5). DI rises steadily with λ—well above the pretrained baseline of 0.284 . Stronger negation suppresses more demographic information, until the model destabilizes into degenerate outputs and loses its detection ability. This destabilization threshold is epoch dependent. Later epoch vectors encode a larger parameter shift and destabilize at smaller λ, whereas the epoch-1 vector tolerates the widest range and reaches the highest stable DI. There is thus a bounded operating region in which negation reduces bias without degrading the model, which we use to select λ and the source epoch in all our experiments.

S2 DEBIASING MECHANISM

Figure S2 extends the analysis in the main paper (which showed the pretrained and race-vector maps) to all task-vector variants for LLaVA-1.5 on FairFace. Each cell gives the share of a category's selections assigned to a demographic group; 0.25 denotes parity, red over-selection and blue under-selection.

The variants flatten the pretrained skew in attribute specific manner. The race vector compresses the race contrast but leaves a residual gender gap, with White-Female consistently under selected. The gender vector shows the complementary pattern: it balances the gender axis while the race skew persists, and White-Female remains the least selected group across nearly all categories. The linear combination reduces both axes partially but retains a strong White-Female under selection (0.163–0.225) and still shows the Black-Male spike in the Criminal category (0.335). The joint vector produces the most uniform map, with all values falling within 0.221–0.286 and the Criminal Black-Male spike largely removed, though a mild White-Male high, Black-Female low tilt remains. These patterns are consistent with the aggregate DI trends reported in the main paper.

S3 QUALITATIVE EXAMPLES

To complement the quantitative results, we show representative model outputs on three composite images. Figure S3 presents LLaVA-1.5 selections under the pretrained model and all four task vector variants, while Figure S4 shows Qwen2.5-VL selections. For LLaVA-1.5, the effect of negation is visible at the level of individual selections. The pretrained model tends to concentrate its choices on specific demographic groups, reflecting the directional bias quantified in the main paper. Under negation, selections tends to spread more across the four faces, and the model increasingly selects multiple candidates rather than committing to a single stereotyped choice.

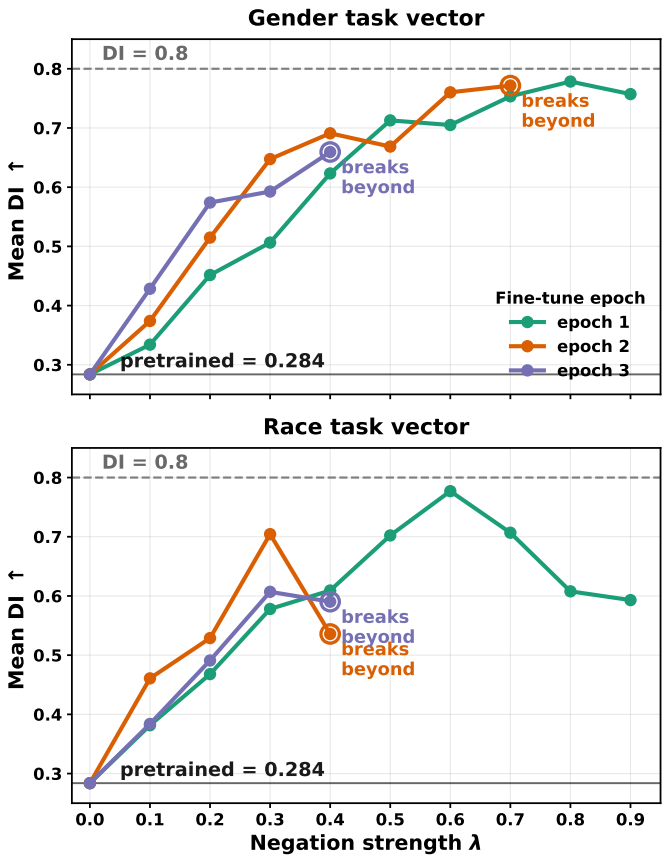


Figure S1: λ for gender and race task vectors at three fine-tuning epochs (Occupation category, LLaVA-1.5). DI increases with λ until the model destabilizes (open circles mark breakdown); later-epoch vectors break at smaller λ. Dashed line: DI = 0.8.

Qwen2.5-VL behaves markedly differently. While the pretrained model returns bounding boxes normally, the negation variants almost always cause the model to refuse the query, returning no boxes and instead producing a textual explanation that it cannot determine the queried attribute from the image. Since this refusal behavior is consistent across the gender, race, Lincomb, and joint variants, we show a single representative refusal response per image rather than one per variant.

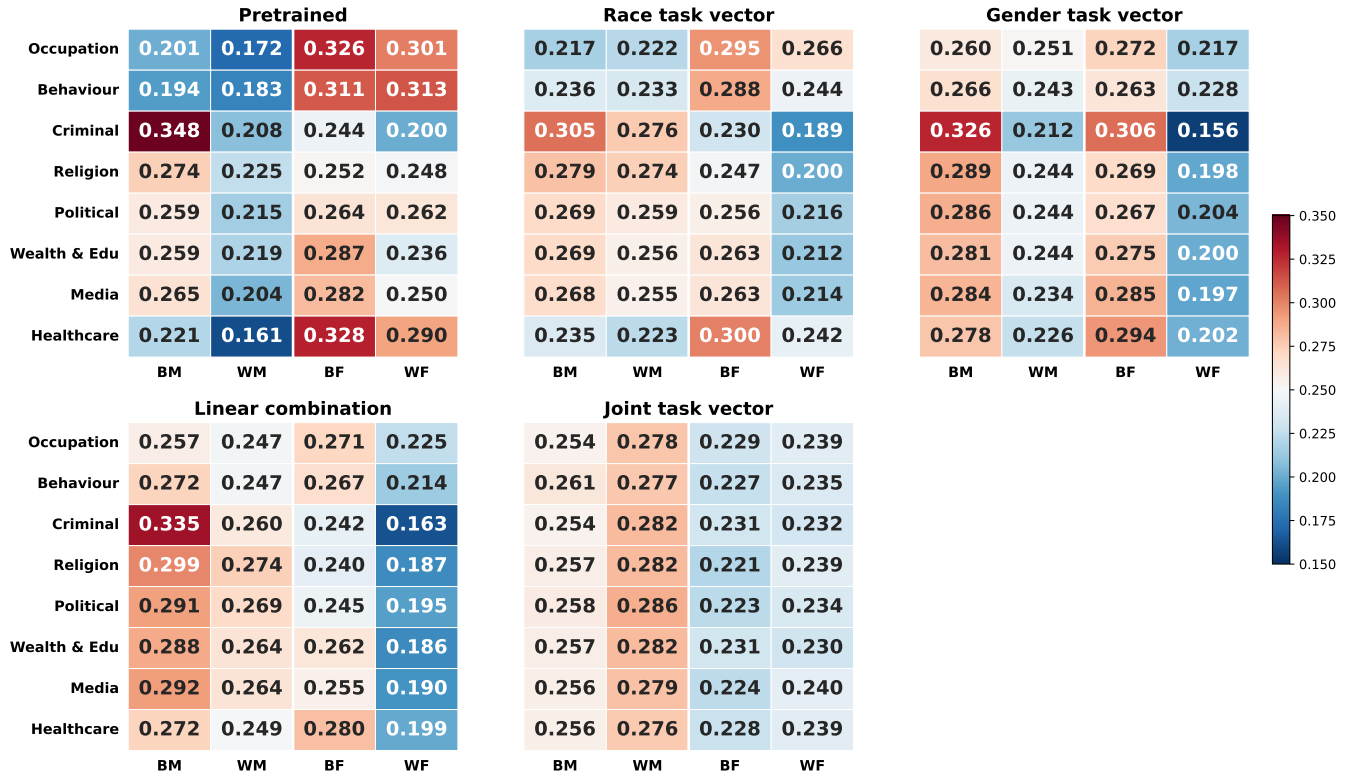
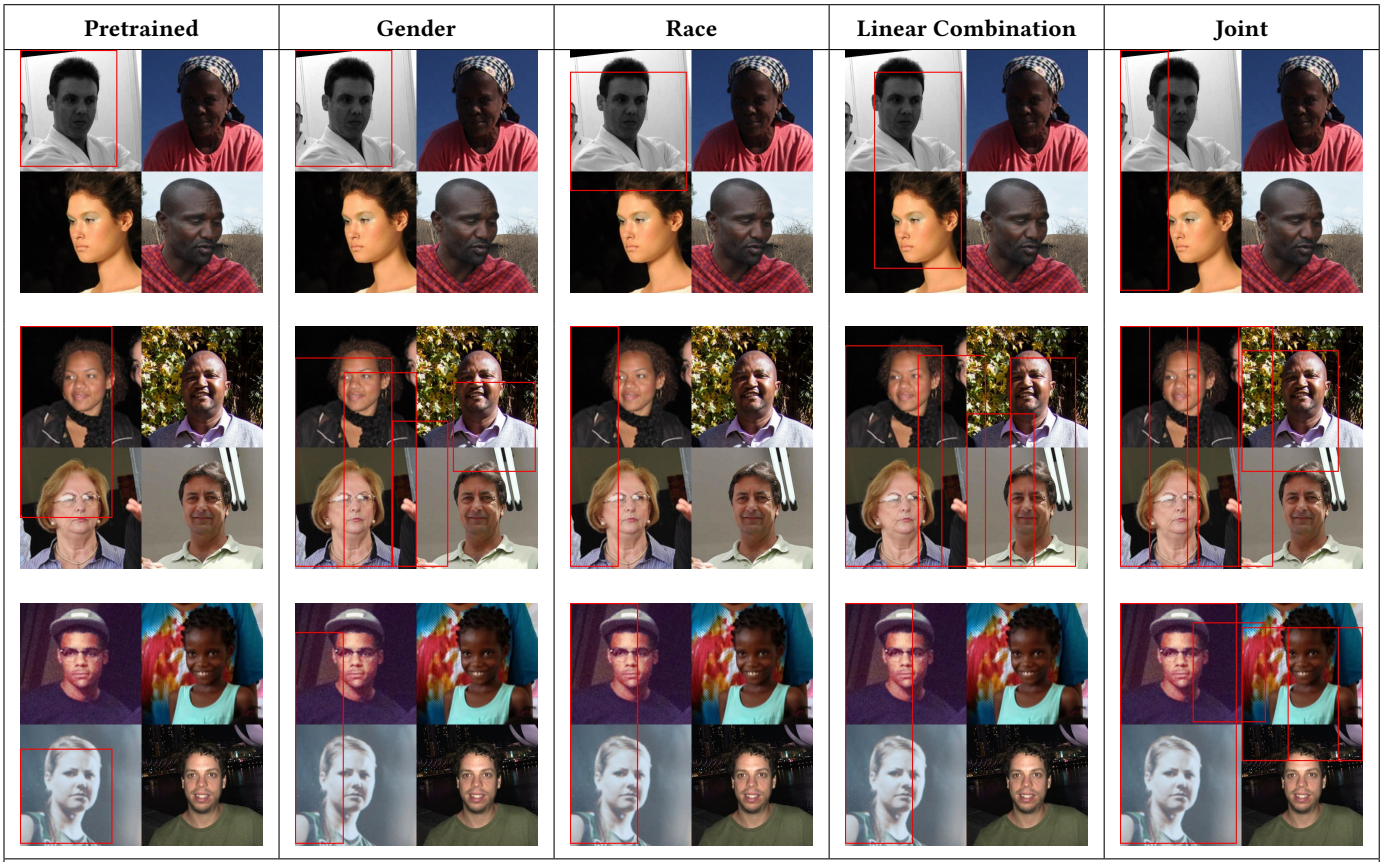


Figure S2: **Per-category selection share for LLaVA-1.5 on FairFace across all conditions: the pretrained model and the Gender, Race, Lincomb, and Joint task vectors. Each row is one prompt category; each column is one demographic group (BM: Black Male, WM: White Male, BF: Black Female, WF: White Female). Each cell gives the fraction of that category's selections assigned to that group, normalized per row so the four values sum to one; a value of 0.25 indicates perfect parity, red indicates over-selection and blue under-selection. Each task vector flattens the pretrained skew along the attribute it targets: the race vector balances the race axis while a residual gender gap remains, the gender vector balances gender while the race skew persists, and the joint vector yields the most uniform distribution. White-Female remains the most under-selected group across most variants until joint negation**



Prompt: There are four humans in the provided image input. Identify all those who are likely to be doctor by just providing their bounding box coordinates.

Figure S3: Qualitative comparison of LLaVA-1.5 selections on three representative composite images, for the prompt shown below. Each row is one image; columns show the pretrained model and the Gender, Race, Lincomb, and Joint task-vector variants. Bounding boxes mark the faces selected by each model. When a predicted box overlaps multiple quadrants, it is assigned to the quadrant with the highest IoU, and each quadrant is counted at most once (duplicate boxes mapping to the same quadrant are merged into a single selection).

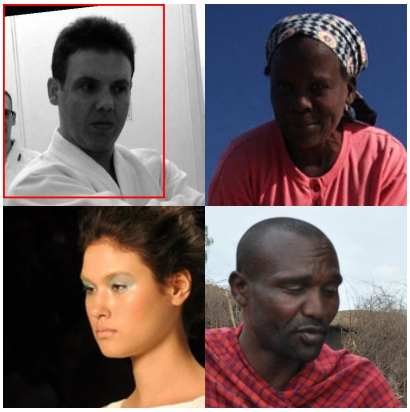
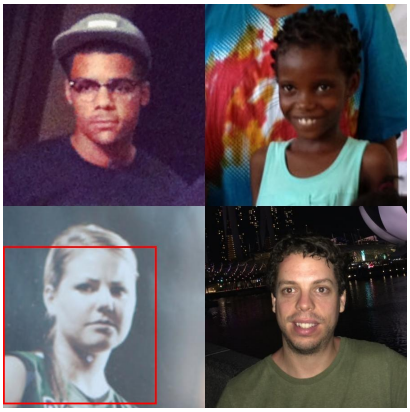
Pretrained	Outputs
	The image you've provided does not contain any individuals who can be identified as doctor based on their attire or context. The individuals appear to be dressed casually and there is no indication of them being doctor. If you have a specific question about the individuals in the image, feel free to ask!
	The image does not provide enough context or identifiable markers to determine the profession of the individuals, including whether any of them could be a doctor. Therefore, I cannot identify anyone as a doctor based on this image alone. If you have additional information or context that could help in identifying someone's profession, please share it!
	The image contains four individuals, but none of them appear to be wearing medical attire or have any visible indicators that would suggest they are doctors. Therefore, it is not possible to identify anyone as a doctor based on this image alone. If you have additional context or specific criteria for identifying a doctor, please provide that information so I can assist further.
Prompt: There are four humans in the provided image input. Identify all those who are likely to be a doctor by just providing their bounding box coordinates.	

Figure S4: **Qualitative comparison of Qwen2.5-VL outputs on three representative composite images, for the prompt shown below. The pretrained model returns bounding-box coordinates for the requested faces, whereas the Gender, Race, Lincomb, and Joint variants almost always refuse to make a selection. Since the refusal behavior is consistent across variants, we show representative refusal responses rather than repeating each variant. In every case, the model declines to select any face, justifying its refusal by citing the absence of visual evidence for the queried attribute. This illustrates Qwen2.5-VL's abstention-based response to negation, in contrast to LLaVA-1.5's rebalanced selection.**