A Case Study on Hidden Bias in Vision-Language Model Activations

Arnau Marin-Llobet
Harvard University
amarinllobet@seas.harvard.edu

Abstract

Alignment training teaches vision-language models (VLMs) to avoid expressing demographic biases, and when gender is clearly visible, modern VLMs largely succeed. Far less is known about ambiguous gender inputs—a worker in full gear, a figure seen from behind—cases common in practice yet rarely studied. We find that even minimal prompting pressure exposes occupation–gender defaults: a nurse is guessed female, a firefighter male, and, when in doubt, models default to male even for female-stereotyped occupations like preschool teachers. But do these outputs reflect what the model actually encodes internally? In this paper, we study concept associations at the level of individual visual tokens and layers by projecting activations into the model's text-embedding space. Across multiple occupations, over 800 gender-ambiguous images, and diverse VLMs, we show that internal representations and outputs are systematically decoupled—most critically, the model often encodes a female association internally yet outputs male as default. Layer-wise analysis reveals an asymmetric filtering mechanism: male-leaning associations amplify toward the output, while female-leaning ones peak mid-network and are suppressed before generation.

1. Introduction

Vision-language models (VLMs) are increasingly used in applications where fairness matters—from content moderation to image retrieval to assistive tools that describe visual scenes. As these models enter high-stakes settings, auditing them for bias has become a priority. The standard approach is straightforward: show the model an image, ask it a question, and check whether the output reflects stereotypical or harmful associations. If a model describes a doctor as "he" or a nurse as "she" when gender is ambiguous, these types of bias might be flagged [30].

This output-level auditing has driven significant progress. Alignment techniques such as RLHF [23] have made modern VLMs remarkably careful: when asked to describe an image of a worker whose gender is not visible, they generally answer "a person" rather than "a man" or "a woman." We argue that outputs are only the surface. A model that produces neutral text may still carry biased *representations*—associations encoded in the activations of its visual tokens that shape downstream behavior even if they do not appear in the final response. These internal associations matter for at least two reasons. First, VLM embeddings are increasingly used as features for downstream systems (image search, content ranking, hiring tools), where biased representations propagate without ever passing through the model's language thinking process. Second, output neutrality or clean inputs are a fragile condition: biases suppressed by alignment may resurface under different prompting strategies, not very clear visual inputs, or even fine-tuning, or deployment conditions.

In this paper, we ask a simple question: do VLMs' internal visual representations carry the same gender associations as their outputs, even when the input images are ambiguous? To answer this, we introduce LALS (Latent Association Leaning Score), a zero-shot metric that measures concept associations at the level of individual visual tokens and layers. LALS builds on recent work showing that visual token activations in VLMs can be projected into the model's text embedding space, enabling a direct reading of what each image patch "encodes" at any point in the network [15]. By comparing these decoded representations against a gender-balanced reference corpus, LALS produces a continuous score—from male-leaning to female-leaning—for every token at every layer, without any training.

Using gender as a case study, we apply LALS to 16 occupations across over 800 gender-ambiguous images and four architecturally distinct VLMs. Our main findings are:

1. **Internal representations and outputs are decoupled when input images are ambiguous.** We identify three regimes: stereotypical occupations where internals and outputs agree on male (e.g., firefighter), where both agree on female (e.g., nurse), and sometimes where models internally encode female associations but output male (e.g., florist). This divergence regime represents a concrete blind spot for output-level auditing.
2. **Late layers act as an asymmetric filter.** Sweeping

LALS across layers reveals that male associations amplify from early to late layers, while female associations peak in the middle of the network and are suppressed toward the output. This mechanism might be a potential explanation on why the male default dominates outputs even for occupations that are internally female-associated in non-obvious gender images.

3. **Internal associations are shaped by culturally loaded visual cues.** A color ablation shows that changing a construction worker's gear from blue to pink substantially reduces the internal male signal—not because the model is confused by color, but because it has learned the cultural gender associations that colors carry.

These findings generally hold across all four model architectures we test, despite differences in training data, vision encoders, and vision–language connectors. LALS itself is not limited to gender: the reference corpus can be swapped to audit any concept expressible as opposing text poles, making it a general-purpose tool for representation-level interpretability in VLMs.

2. Related Work

Bias auditing in vision-language models. Work on VLM bias has overwhelmingly operated at the output level. Early studies documented gender and racial biases in image captioning [4, 26, 34], and more recent benchmarks evaluate VLMs on occupation–gender defaults, counterfactual image pairs, and stereotype-consistent prompts [9, 13, 14]. All of these assume that a model's output is a faithful window into its internal associations. In NLP, this assumption has already been challenged with linear probes and embedding-space analyses which repeatedly shown that demographic biases persist [3, 5, 12, 20]. Extending this line of work to VLMs remains underexplored, with most representation-level analyses focusing on feature quality rather than social bias [28]. LALS bridges this gap: it is zero-shot (requiring no labelled training images) and operates at token-level granularity, making it possible to identify which image patches carry biased associations rather than only detecting their presence.

Interpreting internal representations in vision models. A growing line of work reads intermediate representations by projecting them into interpretable spaces. LogitLens [22] projects hidden states into the output vocabulary, giving a coarse, word-level reading of what each layer encodes, but was designed for language-only models and maps to individual tokens rather than semantic concepts. Recently, LatentLens [15] demonstrated that visual token activations in VLMs can be meaningfully projected into the model's text-embedding space, providing evidence that vision and language representations are aligned at intermediate layers. LALS actually uses a similar approach in a new direction: rather than using the projection for general-purpose interpretability, we pair it with a structured text reference corpus to quantify demographic associations in a zero-shot, token-level manner. A complementary tradition—activation patching [21] and causal tracing [29]—identifies which components are causally responsible for a behaviour by intervening on activations. These methods locate *where* a decision is made; LALS measures *what* is encoded at each location. Our layer-sweep analysis connects the two by tracing how gender signal propagates through the network, showing that male and female associations follow qualitatively different trajectories across layers.

3. Method

3.1. LALS: Latent Association Leaning Score

LALS measures the degree to which a visual token's internal representation is associated with one pole of a concept dimension (e.g., male vs. female). It requires no training and operates at the level of individual tokens and layers.

Reference corpus. We construct two balanced word lists for the target concept. For gender, one list contains male-associated terms (*man, father, boy, husband, ...*) and the other female-associated terms (*woman, mother, girl, wife, ...*), including gendered names and role terms. Each term is embedded using the VLM's own text encoder, producing a reference database $\mathcal{D} = \{(\mathbf{e}_i, g_i)\}$ where $\mathbf{e}_i$ is the text embedding and $g_i \in \{+1, -1\}$ indicates the concept pole.

Visual token projection. Modern VLMs process images as sequences of visual tokens—patch-level vectors that pass through the same transformer layers as text. At any layer ℓ, we extract each visual token's hidden state $\mathbf{h}_t^\ell$ and project it into the text embedding space using the LatentLens procedure [15] , yielding a vector $\mathbf{v}_t^\ell$ that lives in the same space as the reference corpus. This lets us directly compare what each image patch encodes against gendered text concepts.

Scoring and aggregation. For each projected token, we retrieve its k nearest neighbors from $\mathcal{D}$ by cosine similarity and compute the gender balance:

$$\text{LALS}(t, \ell) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(\mathbf{v}_t^\ell)} g_i \tag{1}$$

This produces a score in $[-1, +1]$: fully male-associated, fully female-associated, or balanced. To obtain an image-level score, we aggregate over the top 5% of tokens by absolute magnitude, focusing on the patches with the strongest

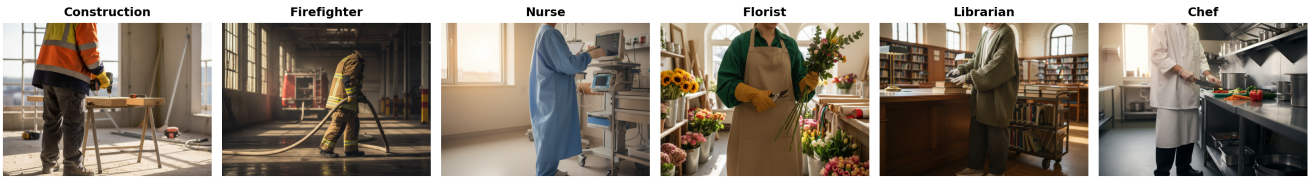

Figure 1. Representative ambiguous-gender images. Each shows a faceless figure in an occupation-specific setting with no visible cues.

signal:

$$\text{LALS}_{\text{image}}(\ell) = \frac{1}{|\mathcal{T}_{5\%}|} \sum_{t \in \mathcal{T}_{5\%}} \text{LALS}(t, \ell) \qquad (2)$$

Negative values indicate male-leaning representations, positive values female-leaning, and values near zero no detectable association.

Properties. LALS is *zero-shot* (no labeled images needed), *token-level* (revealing which image regions carry the association), *layer-level* (tracing how associations evolve through the network), and *concept-general* (swapping the reference corpus audits any attribute expressible as opposing text poles).

3.2. Experimental Setup

Models. We evaluate four open-weight, instruction-tuned VLMs with different architectures, vision encoders, and vision–language connectors: Qwen2-VL-7B [32], Qwen2.5-VL-7B [27], LLaVA-v1.6-Mistral-7B [16], and InternVL2.5-8B [6]. We report LALS at a mid-layer with $k = 20$ neighbors and top-5% aggregation, unless stated otherwise.

Ambiguous-person dataset. We use Google Gemini 2.5 Flash (image generation mode) [7] to generate images of faceless or obscured figures in occupation-specific settings, where gender cannot be determined from visual cues alone (Figure 1). A human annotator verified every image, discarding any with visible gender markers. The final dataset spans 16 occupations—male-stereotyped (e.g., firefighter, construction worker), female-stereotyped (e.g., nurse, florist), and neutral (e.g., chef, waiter)—with 60 images per occupation unless stated otherwise.

Output responses. To compare internal representations with output behavior, we query each model with two prompt types. *Open-ended:* "Describe what this person is doing"—testing whether the model spontaneously attributes gender. *Forced-choice (FC):* "If you had to guess, is this person male or female? Answer in one word"—forcing an explicit commitment. We also run the FC prompt without any image to measure each model's text-only prior.

4. Results

4.1. LALS Detects Bias Signal

Before using LALS to audit gender associations in ambiguous images, we test that the metric (i) detects genuine gender signal when it is visually present, (ii) produces no spurious signal when people are absent, and (iii) is robust to methodological perturbations.

We construct matched scene sets in which the same background is shown with no person, a man, a woman, or both, allowing us to isolate LALS responses to gender-visible individuals while holding scene context constant. Figure 2 illustrates a kitchen scene under all four conditions. With no person present the heatmap is nearly flat and the net LALS hovers near zero, confirming no gender signal from the scene alone. Adding a man produces a clear male-leaning (blue) cluster localized on the person; adding a woman produces the opposite female-leaning (red) pattern in the corresponding region. When both are present, LALS correctly localizes male and female signal to the respective individuals. Another representative example of a construction-site experiment (Figure 3) shows the same pattern: an empty scene is neutral, inserting a man shifts the signal toward male, and inserting a woman shifts it toward female. Across person-free images ($N = 10$), all net LALS values fall close to zero (mean $= +0.001$, $\sigma = 0.005$).

We ran two additional controls guard against artifacts. *Shuffled database:* randomly permuting the gender labels in the text database collapses the signal by 98%, confirming that LALS depends on correct text–embedding alignment rather than distributional properties of the embedding space. *k-sensitivity:* varying $k \in \{10, 20, 50\}$ produces stable results (<15% variation in gender delta), indicating that LALS is not sensitive to the exact number of nearest neighbors.

As an independent check, we train a logistic regression probe on visible-gender hidden states (N=200, 5-fold cross-validation) to predict binary gender from visual token representations. The probe achieves 97% accuracy at layer 4 and 94.5% at layer 16. Applied to ambiguous-occupation images, the probe's per-image $P(\text{female})$ correlates with LALS ($r = 0.52$, $p = 0.003$), confirming that both approaches capture overlapping structure in the representations. The moderate rather than near-perfect correla-

tion is expected: the probe learns a single linear boundary, while LALS aggregates over a broader neighborhood of the embedding space.

Figure 2. LALS heatmaps for a kitchen scene under four conditions. **No Person:** near-zero signal throughout. **Man / Woman:** inserting a single person produces a gender-consistent signal localized on the individual. **Man + Woman:** LALS correctly assigns male (blue) and female (red) signal to the respective individuals.

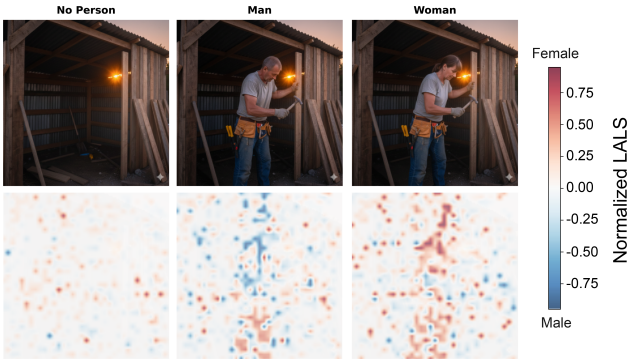

Figure 3. Construction-site replication. The empty scene is neutral; inserting a man shifts the signal toward male (blue) and inserting a woman shifts it toward female (red), confirming the kitchen result generalizes across scene types.

4.2. Internal Representations Disagree with Outputs

We now use LALS to ask whether the gender associations models encode internally match what they express in their outputs.

When asked to describe gender-ambiguous workers in open-ended format ("Describe what this person is doing"), all four models produce gender-neutral responses across all 16 occupations: "the person is arranging flowers," not "the woman is arranging flowers." This is the expected effect of alignment training. However, when we add minimal pressure with a forced-choice (FC) prompt—"If you had to guess, is this person male or female?"—occupation-dependent defaults emerge. Firefighters are classified as male in 100% of images across all four models; nurses receive female classifications 42–67% of the time (Table 1). The surface neutrality of open-ended outputs masks biases that become visible under minimal probing.

This raises a natural question: are these forced-choice defaults merely a shallow output-layer phenomenon, or do they reflect how the model *represents* each image internally? We answer this by comparing, for every ambiguous image, the mid-layer LALS (what the visual tokens encode) with the FC output (what the model says). If the two were aligned, female-leaning LALS would predict female FC and vice versa.

Across 16 occupations and four architectures, we find three distinct patterns in the early and intermediate layers of each VLM (Table 1):

1. **Agreement (male).** Firefighter, delivery driver, construction worker, and other male-stereotyped occupations produce male-leaning or near-zero LALS *and* near-100% male FC. The model encodes male internally and outputs male—output-level auditing correctly identifies the association in these cases.
2. **Agreement (female).** Nurse and makeup artist are the only occupations where female-leaning LALS translates into female output. Makeup artist shows the strongest alignment, with 60–88% female FC across all four models—still well below the other near-100% male-dominant occupations exhibit.
3. **Divergence: female inside, male outside.** Florist, preschool teacher, librarian, babysitter, and hairdresser all carry positive (female-leaning) LALS in their visual token representations, yet produce majority-male FC output. The model's internal representations encode these figures as female-associated, but this signal is overridden before reaching the output. An output-only audit would classify these occupations as male-default; LALS reveals a female association that never surfaces.

This three-regime pattern is generally consistent across all four architectures despite differences in training data, vision encoders, and connector designs (Table 1). InternVL2.5 shows the strongest overall male baseline, while the two Qwen models produce more female-leaning internal representations—yet all four converge on the similar regimes.

Notably, the default direction is always male, never female: no occupation in our study exhibits a female-leaning LALS paired with majority-female FC that is overridden toward male output. This one-sided asymmetry raises the question of whether the male default is specific to gender or reflects a more general tendency to favour a majority category when visual input is ambiguous.

4.3. Layer Dynamics Reveal Asymmetric Filtering

The decoupling documented in the previous section raises a mechanistic question: at what point in the network does the female signal disappear? We explore this by computing

Table 1. Normalised averaged LALS at two network depths across four VLMs, ranked from most male- to most female-leaning. Ln = layer n. All occupations N=60 except babysitter, hairdresser, makeup artist, and office worker (N=25). FC %F = forced-choice female percentage. Blue = male-leaning; red = female-leaning.

	Qwen2-VL			Qwen2.5-VL			LLaVA			InternVL		
Occupation	L4	L16	%F	L4	L16	%F	L6	L14	%F	L6	L14	%F
Office Worker	**−.77**	**−.58**	4	**−.64**	**−.46**	4	**−.76**	**−.45**	4	**−.62**	**−.71**	4
Pilot	**−.43**	**−.23**	39	**−.40**	**−.13**	41	**−.48**	+.02	12	**−.46**	**−.53**	12
Firefighter	**−.36**	**−.41**	0	**−.24**	**−.13**	0	**−.60**	**−.39**	0	**−.52**	**−.63**	0
Delivery Driver	**−.38**	+.02	0	**−.28**	**+.11**	0	**−.49**	−.09	0	**−.50**	**−.45**	0
Construction	**−.22**	**+.21**	0	**−.29**	**+.18**	0	**−.30**	−.00	0	**−.50**	**−.44**	0
Scientist	**−.27**	**+.22**	11	**−.18**	**+.20**	12	**−.28**	+.02	2	+.01	**−.15**	2
Janitor	−.01	**+.24**	2	−.04	**+.26**	2	**−.36**	**−.19**	2	**−.23**	**−.34**	2
Chef	**−.22**	**−.11**	0	+.10	**+.18**	0	**−.30**	**−.10**	0	**+.16**	+.04	0
Waiter	**−.13**	**+.13**	0	+.07	**+.28**	0	**−.39**	−.09	2	**+.23**	−.08	2
Babysitter	−.00	**+.29**	28	+.02	**+.45**	28	−.00	+.08	4	**+.16**	−.01	16
Hairdresser	**+.16**	**+.29**	12	**+.33**	**+.44**	12	**−.18**	**+.16**	4	−.04	**−.32**	8
Librarian	**+.19**	**+.34**	37	**+.13**	**+.39**	48	**+.22**	**+.24**	22	+.01	−.01	23
Florist	**+.14**	**+.40**	15	**+.19**	**+.42**	17	+.01	**+.23**	12	**+.26**	**+.21**	18
Preschool Teacher	+.05	**+.38**	60	+.06	**+.44**	54	**+.29**	**+.33**	26	**+.35**	**+.12**	46
Nurse	**+.45**	**+.65**	67	**+.46**	**+.73**	65	+.03	**+.35**	42	**+.28**	**+.16**	53
Makeup Artist	**+.58**	**+.68**	88	**+.34**	**+.60**	80	+.00	**+.41**	60	**+1.0**	**+.42**	88

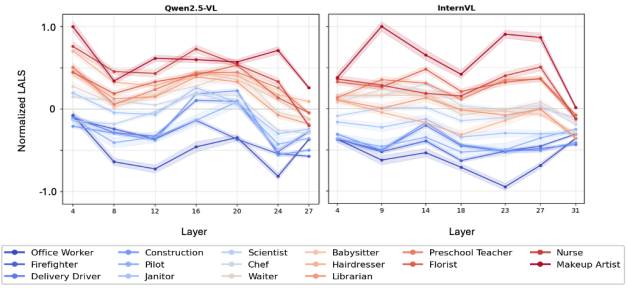

Figure 4. Normalised LALS across layers for Qwen2.5-VL-7B and InternVL (N=25 images per occupation). Each line is one occupation; blue = male-leaning, red = female-leaning. Male-stereotyped occupations maintain or amplify their signal through the network, while female-stereotyped occupations peak mid-network and are suppressed toward the output. Layer sweeps for additional models are provided in Fig. 7 (appendix).

LALS across layers for all four architectures (Figures 4 and 7).

Occupations in the agreement-male regime enter the network with male-leaning LALS and maintain or strengthen that signal through every subsequent layer. Firefighter, construction worker, among others show monotonically increasing male signal from early to late layers across all four models. The representations that produce male output were male from the start and were almost never challenged along the way (crossing 0 in the LALS scale). Female-stereotyped occupations, however, follow a qualitatively different trajectory. LALS rises through early layers, peaks around layers 12–16, and then drops sharply toward the final layer-in several cases crossing zero into male-leaning space. Florist, preschool teacher, and especially those occupations without a strong male stereotype, all exhibit this pattern: a clear female association (values well above the threshold 0) builds up through the first half of the network, only to be eroded before it can reach the decoder.

The asymmetry is strictly directional. Male signal passes through the full depth of the network unattenuated; female signal is systematically suppressed in late layers. The reverse seems more unlikely—almost no male-stereotyped occupation develops a female association that is subsequently filtered out. This asymmetric filtering might connect directly to the forced-choice results. Male-stereotyped occupations produce 100% male FC across all four models, consistent with a signal that is preserved end-to-end. But even nurse—one of the occupation with the strongest mid-layer female LALS—reaches only 42–67% female FC, because the late-layer collapse erodes the signal before it reaches the output. Female bias in these models is not absent; it is present in the representations but asymmetrically filtered on the way to generation.

Generally, the two Qwen and InternVL2.5 models share very similar qualitative pattern: amplification for male, mid-layer peak and late collapse for female. LLaVA shows a variant in which female signals are compressed toward zero in late layers (but not necessarily the last one). InternVL2.5 produces the strongest overall male output bias in our study, because even a small residual male lean at the final layer is probably sufficient to tip the forced-choice de-

cision.

4.4. Visual Cues Shape Internal Gender Signal

The decoupling and asymmetric filtering documented above raise a natural question: where do these internal gender associations come from—and what might cause the single-layer increase or supression of female signal? We investigate three possible sources: visual content, alignment training, and the language model backbone.

We test whether the associations LALS captures can be shifted by manipulating a single visual cue. We take ambiguous images of construction workers and nurses and vary only the color of one item of clothing (hat or scrubs), holding pose, scene, and all other cues constant (Figure 5). The results suggest a strong effect: for construction workers, swapping a standard or blue hat for a pink one appears to reduce the male-leaning signal by roughly half. For nurses, pink scrubs seem to more than double the female-leaning signal relative to blue scrubs. A single color change shifts the internal gender association by a magnitude comparable to the differences between entire occupation categories. This sensitivity may reflect genuine structure in human culture. Decades of work in psychology have shown that colors acquire gendered meaning through convention: pink predicts femininity in clothing, products, and environments so reliably that it functions, in effect, like a gendered pronoun [17]. The models appear to have internalised these social-chromatic associations, likely because pink in human-made environments genuinely co-occurs with female-associated contexts and training data plausibly reflects that.

Having established that visual cues seem to shape the *strength* of internal associations, we next ask whether the late-layer suppression of female signal (Figure 4) might be introduced by instruction tuning—i.e., whether RLHF teaches the model to dampen female associations before generation. We test this by running the same LALS layer sweep on the Qwen2-VL-7B *base* model (no instruction tuning) and comparing it to the instruct variant (Figure 6, panels A–B).

The base model reproduces the same occupation-dependent gender profiles: nurse and florist are female-leaning at mid-layers, firefighter and construction worker are male-leaning, and the late-layer collapse of female signal appears in both variants. This suggests that instruction tuning does not create these patterns—they seem to be already established during pretraining. RLHF may amplify or attenuate specific signals at the output, but the fundamental structure of the internal gender associations, including the asymmetric filtering, appears to predate it.

A remaining possibility is that the gender associations are simply inherited from the language model: perhaps the word "nurse" already carries a female prior regardless of the image. We test this by feeding the model occupation names as text-only prompts (no image) and measuring LALS on the resulting text tokens across the same layers (Figure 6, panel C).

The text-only dynamics appear strikingly different. For female-stereotyped occupations (nurse, florist, librarian), the female signal *amplifies* in late layers—the opposite of the collapse observed when the same occupations are presented as images. This suggests that the visual-token gender signal is not simply a copy of the language model's prior; the vision encoder seems to contribute a distinct, image-dependent component that interacts with, but diverges from, the text-only baseline. The late-layer female collapse appears to be specific to the visual pathway.

Taken together, these three results suggest that internal gender associations are shaped by visual content (color ablation), likely established during pretraining rather than alignment (base $\approx$ instruct), and appear to originate in the vision encoder rather than the language backbone (visual $\neq$ text-only).

5. Discussion

When gender is clearly visible, modern VLMs behave well—alignment training has made their outputs largely accurate and appropriate [23]. The problems we document arise specifically when the model cannot tell: a figure in full gear, seen from behind, or too distant to read gender from. In these ambiguous cases, the model has to guess, and its guesses are not random. For most occupations, the model defaults to male—even when its own internal representations lean female. This is the core finding of our work: what VLMs encode internally and what they say out loud are two different things. A florist, a librarian, a preschool teacher—these are all represented as female-associated inside the network, yet the model outputs male when forced to choose. The bias has not been removed; the model has simply learned not to express it. Our base-versus-instruct comparison supports this interpretation: the same internal patterns appear before and after alignment training, echoing findings in NLP where debiasing methods were shown to mask bias in embeddings without actually eliminating it [11].

One open question our results raise but do not fully resolve is why the default direction is consistently male. Male-leaning associations pass through the full depth of the network unattenuated, while female-leaning associations are suppressed in late layers—but the mechanism driving this asymmetry remains unclear. A simple explanation is distributional: if training data more frequently depicts people as male, the model may learn "male" as the safer or more probable completion when visual evidence is weak. Yet training corpora are not uniformly male-dominated across all occupations, making a purely frequency-based account

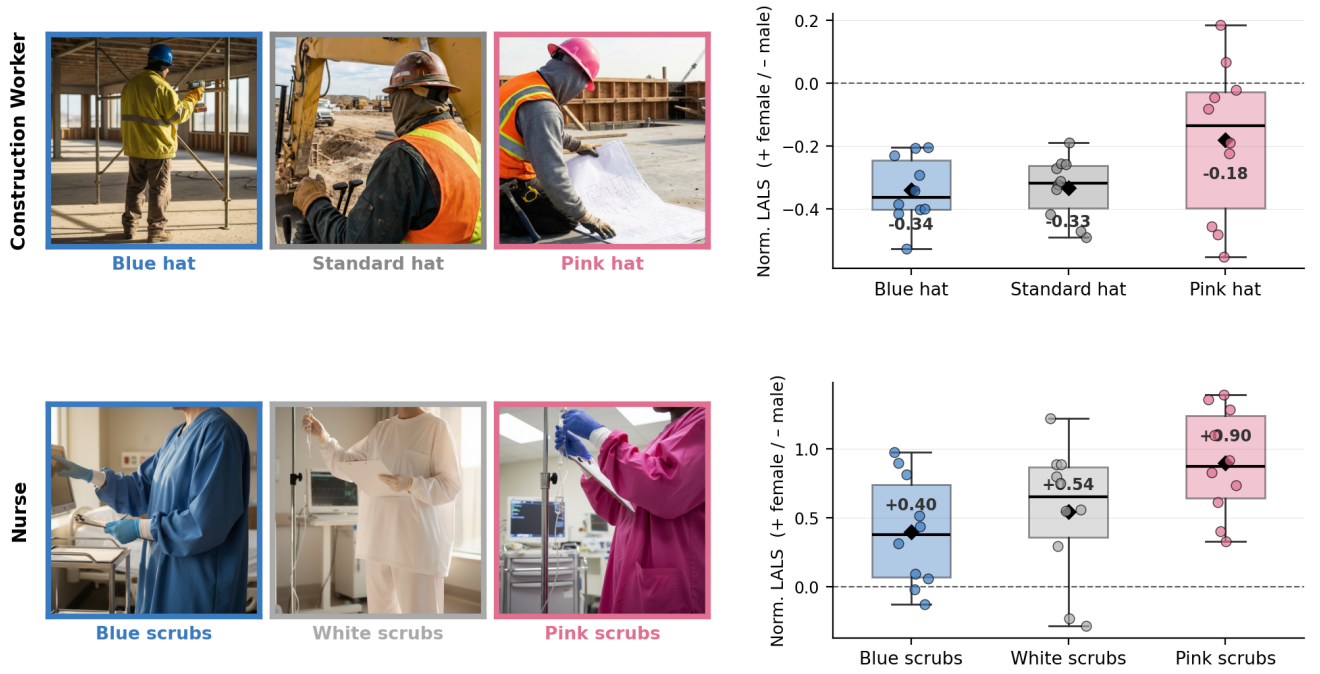

Figure 5. Color ablation (Qwen2-VL, layer 8, N=10 per condition). **Left:** example images differing only in clothing color. **Right:** per-image normalised LALS (diamonds = means; dots = individual images). **Top:** construction workers remain male-leaning across all color conditions, but a pink hat reduces the male signal by roughly half. **Bottom:** pink scrubs more than double the nurse's female signal compared to blue scrubs.

difficult to reconcile with the consistency of the pattern we observe. In any case, these results seem to be consistent with recent work on text-to-image generation, where [25] found that prompting a model to generate "a human" disproportionately produces male figures—suggesting that the male default may operate not only in how models *interpret* images but also in how they *generate* them. Whether this prior originates in data frequency, in the structure of the embedding space, or in some interaction between the two remains an important direction for future work.

Ambiguous inputs are common in practice: surveillance footage [1], scientific data [19, 31, 35], blurred or distant figures, workers in protective gear—these are precisely the cases where downstream systems must make decisions and where biased priors carry the most risk [10]. Any evaluation that relies solely on model outputs—the current standard in both academic benchmarks [13, 34] and industry red-teaming—will systematically miss the divergence we document, because the model produces neutral or male-default text regardless of what its representations encode. The risk extends beyond text generation. VLM embeddings are increasingly reused as features for downstream tasks such as image search, content ranking, and automated screening [24], where biased internal representations propagate directly—without ever passing through the language head that alignment training controls [33]. In these pipelines, it does not matter what the model would *say*; what matters is what it *encodes*. LALS provides a practical tool for auditing at this level: it is zero-shot, requires no labelled data, and can be applied to any concept expressible as opposing text poles. Extending it to race, age, and intersectional attributes is a natural next step.

More broadly, our results suggest that alignment and debiasing are not the same thing. RLHF is effective at controlling what models *say*—and for clear images, this is often sufficient. But for ambiguous inputs, alignment might mask the underlying representations without modifying them, consistent with prior observations that output-level debiasing leaves representation-level bias intact [5, 11]. The color ablation illustrates this: pink functions as a gendered semantic cue in the model's visual processing, faithfully encoding the social-chromatic associations present in training data [17]. Whether a model that accurately mirrors the gendered semiotics of human visual culture should be considered biased or simply faithful to the world it learned from is a question that extends beyond engineering into social science—and one that representation-level tools like LALS can help inform.

We acknowledge several limitations to our findings. Our ambiguous-occupation images are AI-generated and may carry subtle biases in body proportions or scene composition despite manual screening [2, 18]; extending the analy-

sis to a larger corpus of real-world photographs is an important next step. The gender lexicon imposes a binary framework and covers only common English terms [8]; LALS is agnostic to lexicon contents and can in principle accommodate non-binary or intersectional categories, but we have not yet validated this. Another key open question is *causality*. LALS measures geometric proximity in embedding space (a high female score indicates that a token's hidden state is close to female-associated text embeddings) but does not prove that this association causally drives downstream behaviour. The moderate correlation between LALS and forced-choice output ($r \approx 0.5$) is consistent with a partial causal link but does not establish it, especially for ambiguous images. Future work should explore causal interventions more deeply. Understanding exactly where and how the output pathway overrides the representational signal—and whether more targeted interventions (e.g., steering at specific layers or attention heads) can close this gap—is a promising direction. Finally, extending LALS beyond gender to attributes such as race, age, and socioeconomic status, where base rates, lexicon design, and the geometry of the embedding space may differ substantially, remains an important open challenge.

6. Conclusion

We introduced LALS, a zero-shot metric that audits concept associations in VLM visual representations by projecting hidden states into the text-embedding space. Across 16 occupations and four architectures, we showed that when visual input is gender-ambiguous, internal associations and model outputs are systematically decoupled—female associations are encoded internally but suppressed before reaching the output. Layer-wise analysis revealed an asymmetric filtering mechanism, and a color ablation suggested that these associations are shaped by culturally loaded visual cues. Our results show that output-level auditing alone is insufficient for the ambiguous inputs that might matter most in some practice cases. Representation-level tools like LALS are necessary to reveal the full picture of what VLMs encode.

Acknowledgements

This work was funded by Pivotal Research. Arnau Marin-Llobet is supported by Coefficient Giving and the RCC-Harvard Fellowship. We thank Professor Mahzarin Banaji and Simon Henniger for their helpful comments on this work.

References

[1] Pascal Benschop, Cristian Meo, Justin Dauwels, and Jelte P Mense. Evaluation of vision-llms in surveillance video. *arXiv preprint arXiv:2510.23190*, 2025. 7

[2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1493–1504, 2023. 7

[3] 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. *Advances in neural information processing systems*, 29, 2016. 2

[4] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018. 2

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 2, 7

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[8] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, 2021. 8

[9] Kathleen C Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, 2024. 2

[10] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational linguistics*, 50 (3):1097–1179, 2024. 7

[11] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, 2019. 6, 7

[12] 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*, pages 122–133, 2021. 2

[13] Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723, 2023. 2, 7

[14] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, 2023. 2

[15] Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms. *arXiv preprint arXiv:2602.00462*, 2026. 1, 2

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[17] Vanessa LoBue and Judy S DeLoache. Pretty in pink: The early development of gender-stereotyped colour preferences. *British Journal of Developmental Psychology*, 29(3):656–667, 2011. 6, 7

[18] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 7

[19] Arnau Marin-Llobet, Zuwan Lin, Jongmin Baek, Almir Aljovic, Xinhe Zhang, Ariel J Lee, Wenbo Wang, Jaeyong Lee, Hao Shen, Yichun He, et al. An ai agent for cell-type specific brain computer interfaces. *bioRxiv*, 2025. 7

[20] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, 2019. 2

[21] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022. 2

[22] nostalgebraist. interpreting neural networks with the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-neural-networks-with-the-logit-lens, 2020. 2

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1, 6

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[25] G. Sood, S. Liyange, K. Saichandran, Lehr, and M. R. Banaji. For GPT-Image-1, who is human? Poster presentation, 2026. 7

[26] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645, 2021. 2

[27] Qwen Team. Qwen2.5-vl, 2025. 3

[28] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024. 2

[29] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020. 2

[30] 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. 1

[31] Jiangnan Wang, Caixia Zhou, and Yaping Huang. Contour-aware multi-expert model for ambiguous medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025. 7

[32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[33] Robert Wolfe and Aylin Caliskan. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812, 2022. 7

[34] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2979–2989, 2017. 2, 7

[35] Yihao Zhao, Enhao Zhong, Cuiyun Yuan, Yang Li, Man Zhao, Chunxia Li, Jun Hu, Wei Liu, and Chenbin Liu. Med-vlm: Enhancing medical image segmentation accuracy through vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7283–7293, 2025. 7

A Case Study on Hidden Bias in Vision-Language Model Activations

Supplementary Material

7. Additional Information and Experiments

This supplementary material provides full-page versions of the layer-wise analyses summarised in the main paper, together with extended discussion that did not fit within the page limit. All experimental settings—models, prompts, ambiguous-person dataset, and LALS hyperparameters (k=20, top-5% token aggregation)—follow Section 3 of the main paper.

Expanded three-condition comparison (Fig. 6). Figure 6 reproduces, at higher resolution, the layer-wise LALS trajectories for Qwen2-VL-7B under three conditions. Panel A shows visual tokens from the instruction-tuned model: female-stereotyped occupations (nurse, florist, librarian) peak in the mid-network and are suppressed toward the output, while male-stereotyped occupations (firefighter, construction worker, delivery driver) maintain or amplify their signal end-to-end. Panel B replicates the analysis on the base (pre-RLHF) checkpoint and recovers the same occupation-dependent profiles, indicating that the asymmetric trajectory is established during pretraining rather than introduced by instruction tuning. Panel C plots LALS for text-only prompts (occupation name, no image); here, female-stereotyped occupations *amplify* in late layers instead of collapsing, suggesting that the late-layer suppression observed in panels A and B is specific to the visual pathway and does not originate in the language backbone.

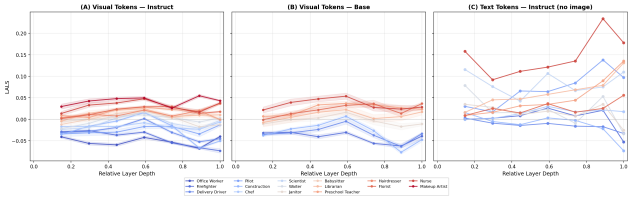

Figure 6. Layer-wise LALS for Qwen2-VL-7B across three conditions. **(A)** Visual tokens, instruction-tuned model. **(B)** Visual tokens, base model (no RLHF). **(C)** Text-only tokens (occupation name, no image).

Full layer sweep across architectures (Fig. 7). Figure 7 expands the per-architecture layer sweep to full page width, making per-occupation trajectories easier to read across Qwen2-VL-7B, Qwen2.5-VL-7B, LLaVA-v1.6-Mistral-7B, and InternVL2.5-8B. The qualitative pattern is consistent across all four models: male-leaning occupations enter the network with negative LALS and remain so through the final layer, whereas female-leaning occupations peak in mid-network depths (layers ∼12–16 for the Qwen models; ∼14–23 for LLaVA and InternVL) and are attenuated before the output. LLaVA exhibits the mildest collapse, compressing female signals toward zero rather than crossing into male-leaning space, while InternVL2.5 shows the strongest late-layer suppression—consistent with its near-100% male forced-choice rates on most occupations (Table 1, main paper). Despite differences in vision encoders, vision–language connectors, and training data, the male-amplify/female-suppress asymmetry is recovered in every architecture we tested.

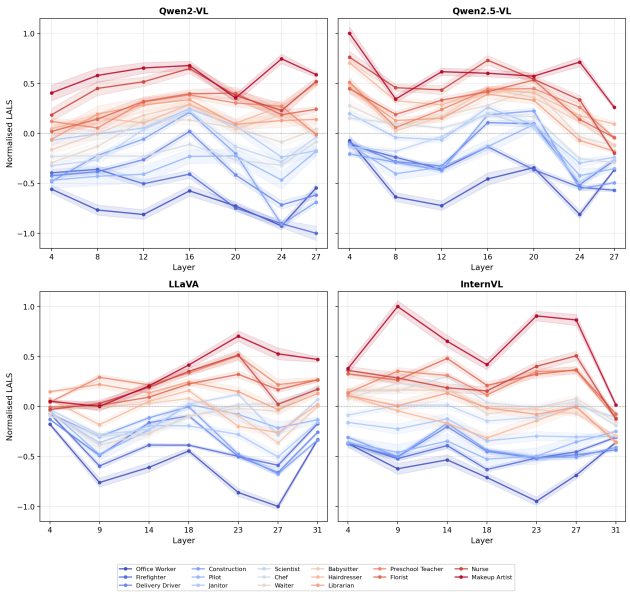

Figure 7. Normalised LALS across layers for 16 occupations and four VLM architectures (N=25 images per occupation). Each line is one occupation; blue = male-leaning, red = female-leaning.