SCALABLE AUDIT AND MITIGATION OF INTERSECTIONAL FAIRNESS BUGS IN VLMS WITH COMBINATORIAL INTERACTION TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Intersectionality analysis is critical for fairness, since individuals hold multiple and overlapping identities whose intersections shape how they are perceived and treated, bringing distinct disadvantages as well as privileges. This matters even more for algorithmic systems such as Vision-Language Models (VLMs), which increasingly serve as auditors of other models, or as backbones of downstream applications. However, VLM fairness is often evaluated one axis at a time, leaving failures that emerge only in higher-order intersectional subgroups undetected: we frame this problem as intersectional fairness bugs. Such bugs remain hard to find: the number of attribute combinations grows combinatorially, making it prohibitively expensive to collect and label data for every subgroup. We propose a scalable and tunable audit for VLMs built on Combinatorial Interaction Testing (CIT) from software testing. Given a set of parameters and their potential values, CIT aims to test all interactions between at most t parameters, instead of testing all combinations of parameter values, thus controlling the combinatorial explosion. Using CIT in conjunction with diffusion-based synthesis, we show that it is possible to both audit VLMs for intersectional fairness bugs and mitigate them by focusing on specific hard subgroups or augmenting training data with synthetic images representative of subgroups. Our empirical evaluation of CIT-based audit and mitigation shows that 1) 60 subgroup accuracies on synthetic images correlate with those on real images for four VLMs (Pearson $r = 0.59$–0.84; Spearman $\rho = 0.56$–0.77), and 2) when fine-tuning CLIP, audit-guided data allocation improves the mean accuracy of the worst 10% of subgroups ($\mathrm{CVaR}_{0.1}$) by up to 4.0 pp under small budgets, and mixing CIT images into real images improves it by up to 2.2 pp under tail data scarcity.

1 INTRODUCTION

Current fairness research acknowledges that bias emerges from complex, layered, and interlocking systems of oppression instead of a single axis of inequality. As established in the framework of intersectionality (Crenshaw, 2013, originally published 1989), social identities like race and gender are not independent variables but are intrinsically linked within a broader social hierarchy; thus, analyzing these axes in isolation fails to capture the intertwined power relations that define them. For instance, Hamidieh et al. (2024) have demonstrated that Vision-Language Models (VLMs) often exhibit biases where the term 'Homemaker' is not only disproportionately associated with women over men, but also overwhelmingly linked to Indian women compared to other racial groups. Further, Shukla et al. (2025) suggest that efforts to de-bias a model for a single axis (e.g., gender) can inadvertently amplify biases along another (e.g., race), highlighting the critical need for a holistic, intersectional perspective.

Despite the consensus on its importance, evaluating intersectional fairness presents a critical challenge rooted in the tension between categorical specificity and statistical reliability, as subdividing attributes tends to result in a combinatorial explosion, in turn leading to a prohibitively large number of potential intersections. This increased granularity inevitably makes direct data collection for each group hard and expensive. Further, as noted by several previous studies (Himmelreich et al., 2025; Wang et al., 2022; Islam et al., 2023; Molina & Loiseau, 2022), the more specific a subgroup

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.

becomes, the smaller its sample size becomes. Such extreme sparsity reduces these intersectional groups to isolated data points, which in turn lose their representativeness. Consequently, these underrepresented identities remain as systemic blind spots. Due to these challenges, existing fairness benchmarks often limit their scope to a single axis, or combinations of two or three axes at best. We recast this evaluation as a search for *fairness bugs*: systematic recognition failures that depend on how attributes co-occur. Our contributions[1] are as follows:

- **A scalable and tunable generative audit.** We combine diffusion-based synthesis with t-way Combinatorial Interaction Testing (CIT) (Kuhn et al., 2004). Covering arrays collapse over 362K attribute combinations into a tractable prompt set of 69, 469, and 2,847 prompts for $t = 2, 3, 4$, which is less than 1% of the exhaustive Cartesian product.
- **Fairness bugs across a granularity–cost tradeoff.** Raising t localizes fairness bugs more finely, at a higher generation and query cost: the image budget grows $6.8\times$ from t=2 to t=3 and $6.1\times$ from t=3 to t=4. Fairness bugs replicate on held-out prompts (t=2, 3) and split-half validation (t=4), partially at t=2 and fully at t=3 and t=4. Rankings over the 60 gender×age×race&eth cells of this synthetic audit track the same cells in real images across four VLMs (Pearson $r = 0.59$–0.84; Spearman $\rho = 0.56$–0.77).
- **Mitigation that reuses the audit.** A fine intersectional taxonomy splits an already small pool of labeled data into many tail cells with only a handful each. We reuse the audit in two complementary ways: at small budgets, as a targeting signal that reallocates real supervision toward the hard cells, raising CLIP's $\mathrm{CVaR}_{0.1}$ by up to 4.0 pp; once the budget exceeds the unique real photographs in the rarest cells, as mixed fine-tuning data that pads those tail cells with the synthetic images themselves, raising it by up to 2.2 pp.

2 Background

Intersectionality, a term coined by Crenshaw (2013, originally published 1989), captures a concept long articulated by earlier scholars of color to describe how identities such as race, gender, and class overlap to create unique modes of discrimination and privilege. There have been many reported cases of technologies interacting with pre-existing societal inequities, resulting in outcomes that disproportionately burden historically marginalized communities (Buolamwini & Gebru, 2018; Obermeyer et al., 2019; Rodriguez, 2023; Rodà et al., 2025). In response, researchers have begun to adopt intersectionality as a theoretically-grounded methodology to mitigate outcome bias and inspire machine learning research (Kearns et al., 2018; Islam et al., 2023).

Combinatorial Interaction Testing, or Combinatorial Testing, was empirically motivated by Kuhn et al. (2004). Empirical analysis of fault data across diverse domains demonstrates that most software failures are triggered by combinations of only one or two specific system input parameters. Based on the observation, if all faults in a system are triggered by a combination of t or fewer inputs, then testing all t-way combinations becomes effectively equivalent to exhaustive testing of combinations between all inputs, since a single test case can contain a significantly large number of t-way combinations. A single test case consisting of 20 input parameters simultaneously covers $\binom{20}{4} = 4845$ different 4-way combinations. By constructing a small *Covering Array* that encompasses all possible t-way interactions, we can achieve comprehensive fault detection with a manageable and optimized number of test cases.

3 Previous Work

3.1 Vision-Language Model Fairness

Extensive research has focused on the fairness of VLMs, revealing significant biases regarding identities such as gender, skin tone, age, and ethnicity (Wu et al., 2024). Addressing these internal biases is a fundamental necessity for the integrity of the broader AI ecosystem, as VLMs are increasingly being adopted in two critical, yet high-risk, capacities.

[1] Code will be made publicly available upon publication.

Table 1: Identity axes and attributes used for intersectional prompt generation. Italicized entries (marked †) were used in generation but dropped from the audit after visual inspection (Sec. 4.3).

Axis (Abbr.)	Axis Type	Attributes
Race & Ethnicity (ETH)	Protected	European, African, East-Asian, South-Asian, Latino
Faith†		*Christian, Muslim, Hindu, Buddhist, Jewish, Unreligious*†
Socioeconomic Status†		*Upper Class, Middle Class, Lower Class*†
Age (AGE)		Infant, Child, Teenager, Adult, Middle-aged, Elderly
Gender (GEN)		Male, Female
Disability Status (DIS)		Without a disability, Who uses a wheelchair; *With a visual impairment, With autism, With cerebral palsy, With a hearing impairment, With a prosthetic limb*†
Appearance (APP)	Contextual	Fat, Thin; *Attractive, Unsightly, Strong, Weak*†
Location (LOC)		Urban area, Suburban area, Rural area, Remote area, Neighborhood, City, Workplace, School

First, as evaluative frameworks, VLMs are frequently utilized to audit other models when obtaining Ground-Truth (GT) labels is cost-prohibitive. For instance, Zhang & Chunara (2024) use CLIP to analyze associations between images and sensitive attributes, while Shukla et al. (2025) use VLM-based automated labeling to evaluate Text-to-Image models. Allowing bias in VLMs risks masking or distorting the unfairness of the target models, leading to false assurances of safety.

Second, as foundational components, VLMs are integrated into diverse downstream applications. In this role, their internal biases are not static but actively transferred to derivative systems, embedding systemic prejudices into the ecosystem. Consequently, ensuring the internal fairness of VLMs is essential to prevent cascading bias.

3.2 Intersectional Evaluation and Mitigation

Recent fairness research emphasizes intersectionality, noting that bias often manifests at the intersection of multiple axes where marginal fairness fails to capture granular disparities (Filippi et al., 2023). This perspective has been applied across diverse domains, from multimodal clinical prediction to recommendation systems (Ramachandranpillai et al., 2024; Wang et al., 2024; Suk & Han, 2025; Wastvedt et al., 2024). To cope with the data sparsity of intersectional subgroups, prior strategies have focused on leveraging shared statistical patterns (Wang et al., 2022), robust fairness estimation via batch-based learning (Islam et al., 2023), or establishing statistical bounds between marginal and intersectional fairness (Molina & Loiseau, 2022).

To address the same sparsity without those statistical assumptions, two other lines of work generate or search for the failing cases themselves. One constructs counterfactual or parallel images to probe, and sometimes mitigate, bias (Fraser & Kiritchenko, 2024; Howard et al., 2024). The other discovers unlabeled failure slices after the fact (Eyuboglu et al., 2022).

Closest to our audit, Metzen et al. (2023) combine combinatorial testing with text-to-image synthesis to find systematic errors of image classifiers on rare subgroups. Those subgroups are mostly scene factors—viewpoint, color, weather, background—under which an ImageNet class is misread; a demographic case study is incidental, and they stop at discovery. We instead treat overlapping identities as the unit of analysis, audit intersectional recognition in VLMs at a chosen covering strength, and reuse that covering both to target real supervision and to pad tail cells.

4 Synthetic Image Generation with CIT

We first generate intersectional synthetic images, which are used for the audit (Sec. 5) and mitigation (Sec. 6). We specify the identity axes that define intersectional subgroups. CIT then reduces the full Cartesian product to a covering set of prompts; a diffusion model renders those prompts as images. Finally, after visual inspection, we determine which axes can be targets of the audit.

4.1 Identity Axes Specification

Table 1 lists the full eight-axis candidate set used to generate prompts and images; italicized entries are later excluded from scoring, not from the image set, to maintain CIT coverage. Protected versus contextual roles and the rationale for each axis are in Appendix A.1.

4.2 Combinatorial Testing for Prompt Reduction and Image Generation

The full Cartesian product of all axes results in an intractably large number of images and probes to VLMs. To efficiently explore the vast space of attribute combinations, we employ CIT. By leveraging CIT's ability to detect failures triggered by sparse attribute interactions, we systematically sample representative intersectional cases while keeping the prompt set tractable.

To navigate the combinations, we apply t-way combinatorial testing. This approach reduces the exhaustive search space of over 362K possible combinations (the product of the attribute counts in Table 1) into computationally feasible subsets while preserving intersectional coverage up to order t. We generate t-way covering arrays for $t \in \{2, 3, 4\}$, resulting in 69, 469, and 2,847 prompts, respectively (50 images per prompt; 3,450, 23,450, and 142,350 images). By constructing these covering arrays, we achieve a pseudo-exhaustive evaluation of complex attribute interactions at less than 1% of the original cost. This saving grows with the taxonomy: the Cartesian product grows exponentially in the number of axes, whereas a t-way covering array grows only logarithmically. All images are generated via SDXL (Podell et al., 2024); for detailed technical configurations, please refer to Appendix B. Examples of the generated images are shown in Fig. 4.

4.3 Refining Audit Scope

To check whether the diffusion model faithfully renders the prompted attributes, we preliminarily reviewed 690 images (10 per prompt from the 2-way covering array), splitting the sample among authors, and noted which intended attributes were visually realized. This yields two refinements:

Axis exclusion: *Faith* and *Socioeconomic Status* are not scored. They may have visual correlates, but they are not realized consistently enough to serve as ground truth.

Attribute exclusion: On *Disability Status* and *Appearance* we score only "Who uses a wheelchair / Without a disability" and "Fat / Thin". The remaining attributes are not recognizably generated. After this restriction, alignment is 94.00% for Disability Status (from 31.74%) and 89.50% for Appearance (from 56.96%).

To preserve t-way coverage, we only exclude unreliable axes and attributes from the audit, not the images themselves: an image whose prompt includes, e.g., a faith attribute still counts toward the age–gender audit. On the retained axes, alignment is 92.12% overall. The per-axis breakdown is in Appendix A.2. We treat prompt attributes on these axes as the audit ground truth.

5 Bias Auditing

Our evaluation framework adapts the concept of *Explicit Bias*. Luo et al. (2024) extended it to T2I, where explicit bias is the failure to generate an image matching the protected attributes explicitly specified in the prompt. Conversely, we define it for VLMs as the failure to recognize a person as being an attribute that is explicitly present in the image. We examine how this failure varies across intersectional identities (e.g., are an elderly woman and a girl equally recognized as being female?).

The audit proceeds from axes to subgroups. We first measure *axis-level* interactions on the CIT-guided synthetic images of Sec. 4: in lived terms, does simultaneously being one kind of person affect whether they are recognized as another; in formal terms, does variation on one axis change recognition on another? These interactions pick out the axes where a recognition disparity concentrates. Mitigation does not then balance those axes against one another; it aims to close the gap among subgroups. We therefore target gender×age×race&eth, the three-axis interaction with the highest MPD at t=3, as a mitigation scenario (Sec. 6). Finally, we check that the synthetic subgroup ranking tracks the real subgroup ranking, so that it can serve as a targeting signal for mitigation.

5.1 Models under Audit

We audit four VLMs that differ in architecture, training paradigm, and scale: CLIP ViT-B/32 (∼151M) (Radford et al., 2021), ViLT-B/32 (VQA fine-tuned; ∼113M) (Kim et al., 2021), PaliGemma 2 3B mix (224) (Steiner et al., 2024), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). These models span contrastive dual-encoders, encoder-based VQA, and generative VLMs.

5.2 BIAS METRICS

The atomic unit of the audit is a sample-level decision $\hat{y}_{i,k} \in \{0, 1\}$: whether the person in image I_i from the synthetic evaluation set $\mathcal{D}$ is recognized as being the GT attribute $a_{i,k}$ on axis A_k. To check whether the person in an intersectional image is recognized as being every attribute explicitly present, we probe each axis separately. Each model also yields a confidence $\tilde{p}_{i,k} \in [0, 1]$: the normalized Yes probability (VQA) or the ground-truth caption's softmax probability (CLIP); details are in Appendix D.

Attribute Recognition Accuracy (ACC) aggregates $\hat{y}$ at the axis level, $\mathrm{ACC}(A_k) = \mathbb{E}_{i\sim\mathcal{D}}[\hat{y}_{i,k}]$. ACC is a univariate baseline for one axis: it asks whether that axis is recognizable at all, not whether that recognition is stable under intersection. The mean of the gender, age, and race accuracies inside one cell is a different quantity, α_c^{syn} or α_c^{real}, introduced in Sec. 5.3.2.

5.2.1 MATCHED-CONTEXT METRICS

The two metrics below share a matched-context design. For a target axis A_{target} and a set $\mathcal{S}$ of $t-1$ other axes, written $\mathcal{S} \to A_{target}$, every axis of $\mathcal{S}$ takes the role of A_{vary} in turn, with the remaining axes $S_{fixed} = \mathcal{S} \setminus \{A_{vary}\}$ held fixed. The $S \to A_{target}$ structure is similar to the pairwise bias-interaction analysis of BiasConnect (Shukla et al., 2025).

Mean Pairwise Delta (Yitzhaki et al., 2003) measures how much model certainty about A_{target} moves when one axis of $\mathcal{S}$ changes. Let $\mu_{m,S} = \mathbb{E}[\tilde{p}(I, a_{target}) \mid A_{vary} = g_m, S_{fixed}]$ be the mean affirmative probability of the ground-truth target attribute under attribute g_m of the varying axis. Axis-level MPD is the mean absolute difference of confidences over all pairs of attributes on A_{vary}:

$$\mathrm{MPD}(A_{target} \mid \mathcal{S}) = \mathbb{E}_{A_{vary}\sim\mathcal{S}}\, \mathbb{E}_{S_{fixed}} \left[\mathbb{E}_{g_m, g_n \sim A_{vary},\, m\neq n} \left| \mu_{m,S} - \mu_{n,S} \right| \right]. \tag{1}$$

A large MPD means the confidence in A_{target} depends on $\mathcal{S}$. If $\mathcal{S} = \{\mathrm{Age}\}$ and Gender is the target, that might be a gap between the probability of Female for adult women and for elderly women.

MinMax Fairness measures the floor of target-axis recognition: the worst-case accuracy that min-max group fairness seeks to maximize (Martinez et al., 2020; Sagawa et al., 2019). Let $\alpha_{m,S} = \mathbb{E}[\hat{y} \mid A_{vary} = g_m, S_{fixed}]$ be the accuracy of the ground-truth target attribute under attribute g_m of the varying axis:

$$\mathrm{MinMax}(A_{target} \mid \mathcal{S}) = \mathbb{E}_{A_{vary}\sim\mathcal{S}}\, \mathbb{E}_{S_{fixed}} \left[\min_{g_m \in A_{vary}} \alpha_{m,S} \right]. \tag{2}$$

MinMax uses the hard decision $\hat{y}$, not the probability $\tilde{p}$. If $\mathcal{S} = \{\mathrm{Age}\}$ and Gender is the target, it is the worst gender accuracy among age attributes: high if even infant and elderly faces are still classified as female or male; low if gender recognition collapses for some age.

Both are reported as the excess over a permutation chance level, and for each model and covering strength t the weak points are the axis combinations in the top 10% by excess after empirical-Bayes shrinkage. The chance adjustment, the ranking, and their validation are in Appendix F.

5.3 FAIRNESS BUGS ON A GRANULARITY–COMPUTATIONAL COST TRADEOFF

5.3.1 AXIS-LEVEL INTERACTIONS ON SYNTHETIC IMAGES

When looked at one axis at a time, it is seemingly obvious which axes are easy and which are hard for CLIP: Gender (0.97), Disability Status (0.91), and Appearance (0.88) are recognized relatively accurately, whereas Location (0.54) and Age (0.47) are not. Race & Ethnicity (0.70) falls between them. The full per-axis accuracies for all four models are in Table 7. In CIT framework, individual axes would correspond to covering arrays with $t = 1$.

To investigate how the interactions between multiple axes result in intersectional fairness bugs, we evaluate combinations with higher t. Fig. 1 reports CLIP's weakest $S \to A_{target}$ items at each covering strength, measured in Mean Pairwise Delta (Eq. 1) and MinMax Fairness (Eq. 2), with chance correction. At $t = 2$, age can shift gender recognition with a statistically significant difference: being female or male is not equally recognized across different age groups (de-

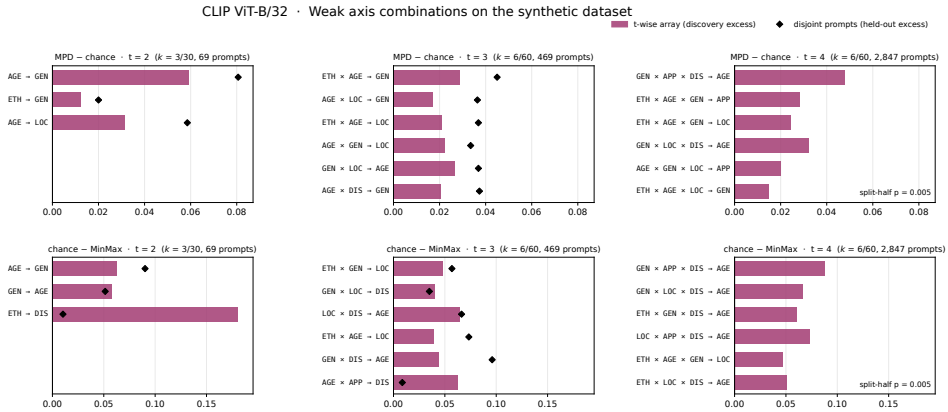


Figure 1: CLIP fairness bugs on synthetic images. Each panel is the top 10% by excess of $S \to A_{target}$ items at that t, ranked by the shrunken chance-corrected excess. Bars: chance-adjusted excess on the t-wise array. Diamonds: the same excess on held-out prompts (t=2, 3); the t=4 panel has no diamonds (split-half only). Selection, k, and checks: Appendix F.

noted AGE → GEN). With $t = 3$, we can observe finer intersectionalities: ethnicity and age together can shift gender (ETH × AGE → GEN). At t=4, a new bug appears, hidden among axes the coarser cuts did not put together: split along gender, appearance, and disability, age recognition moves (GEN × APP × DIS → AGE). The combination with the biggest disparity at $t = 4$, GEN × APP × DIS, is not necessarily the combination with the biggest disparity at $t = 3$ plus one more axis. We report the other models in Appendix E.

5.3.2 A SUBGROUP-LEVEL MITIGATION SCENARIO: GENDER×AGE×RACE&ETH

We choose the subgroup to focus our mitigation based on Sec. 5.3.1. While the combination that leads both MPD and MinMax for CLIP at t=4 is GEN × APP × DIS → AGE, it was not possible to obtain photographs labeled on gender, appearance, disability, and age, at a usable scale, highlighting the need to address intersectional fairness. Consequently, we choose the three-axis interaction with the highest MPD at t=3, ETH × AGE → GEN, and instantiate mitigation on gender×age×race&eth.

As a target real dataset, we merge UTKFace (Zhang et al., 2017) with the train and validation splits of FairFace (Karkkainen & Joo, 2021) into a benchmark we call UTK-FF, and align UTKFace, FairFace, and the synthetic attribute spaces to the taxonomy of Table 1 (Appendix C.1). UTK-FF is relatively balanced along gender and race (max/min ratios of 1.06:1 and 1.92:1 on the one-way counts), but severely imbalanced along age (15.75:1). Details about the dataset are in Appendix C.

Within this scenario, each of the $|\mathcal{C}| = 60$ gender×age×race&eth intersections of attributes is one unit of analysis and of mitigation, and we now measure bias *across subgroups*. For each cell c we report α_c^{syn} on the CIT images restricted to those axes and α_c^{real} on UTK-FF, each the mean of the within-cell accuracies on gender, age, and race. α_c^{syn} is the targeting signal used in Sec. 6. The subgroup-level MPD and $\text{CVaR}_{0.1}$ of Sec. 6.2 are computed from α_c^{real} on the UTK-FF evaluation split of that section, not the full pool.

5.3.3 CAN THE SYNTHETIC AUDIT TARGET REAL SUBGROUPS?

α_c^{syn} comes from CIT-guided images. A 4-way covering array guarantees that every 4-way combination appears at least once, but it also pairs each gender×age×race&eth cell with attributes of the other five axes, which may confound it. Before using the score to allocate UTK-FF, we check that additional nuance does not change which subgroups are hard when scores are aggregated to the subgroup level. The comparison is to UTK-FF. Both sides of Fig. 2 use the pretrained (zero-shot) models; α_c^{real} is computed on the full merged pool, before the train/evaluation cut of Sec. 6.

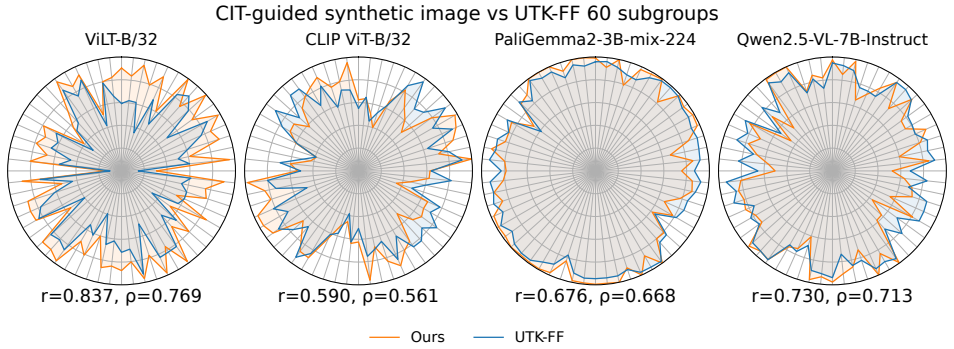


Figure 2: Per-cell α_c^{syn} versus α_c^{real} on the 60 gender×age×race&eth subgroups. Each spoke is one cell. Both are the mean of the gender, age, and race accuracies in that cell. Pearson r and Spearman ρ are annotated per model.

Fig. 2 shows that the synthetic ranking tracks UTK-FF: Pearson r (Spearman ρ) is 0.84 (0.77) for ViLT, 0.73 (0.71) for Qwen2.5-VL, 0.68 (0.67) for PaliGemma 2, and 0.59 (0.56) for CLIP. All four correlations are substantial. We treat the synthetic ranking as good enough to decide which subgroups to intervene on, so the UTK-FF images can be reserved for fine-tuning.

6 From Diagnosis to Mitigation

Within the gender×age×race&eth scenario of Sec. 5.3.2, we connect the diagnosis in Sec. 5.3.3 to mitigation. We propose two strategies for bias mitigation. We aim to show that the CIT audit can be reused for mitigation: the same audit that exposes intersectional failures also tells us where to spend supervision and augment synthetic data. We show that the former helps when the per-cell budget is small, and the latter helps once that budget outgrows the unique UTK-FF images in the tail cells.

6.1 Mitigation Strategies

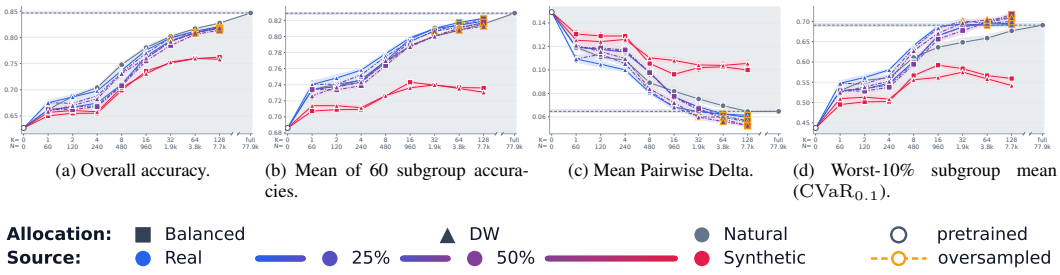
We explore two strategies for budget allocation and image sources, Difficulty-Weighted Sampling and Mixed, against simple baselines described below. We name runs by their image source and budget allocation, e.g., Real-Balanced, Mixed-Balanced 50%, or Mixed-DW 25%.

Budget Allocation: Our strategy is **Difficulty-weighted sampling (DW)**, which uses CLIP's subgroup-level audit on Synthetic images as the targeting signal. For each of the 60 subgroups we take α_c^{syn} from Sec. 5.3.2, the mean cell accuracy of gender, age, and race axes defines the cell difficulty as $d_c = 1 - \alpha_c^{\text{syn}}$: we allocate the training budget in proportion to d_c, giving more images to harder cells. For baselines, we adopt **Natural** and **Balanced**. Natural follows the empirical subgroup distribution of UTK-FF (Sec. 5.3.2): frequent cells receive more images, with no rebalancing across the 60 gender $\times$ age $\times$ race subgroups. Balanced assigns the same number of images to every subgroup, so a budget of K shots per cell uses $N = 60 \times K$ images in total.

Image Sources: Our strategy is **Mixed**, which is to keep $N = 60 \times K$ fixed and replace a fraction of the UTK-FF images with synthetic images. In our experiments, we set the mix rate at 25% and 50%. For baselines, we adopt **Real** and **Synthetic**. Real only uses UTK-FF photographs, with ground-truth gender, age, and race attributes, and no synthetic samples. Synthetic, on the other hand, trains on those CIT-guided synthetic images only.

6.2 Experimental Setup

Dataset. We fine-tune CLIP on the UTK-FF pool of Sec. 5.3.2, but we do not keep FairFace's original validation split. Some cells in that official split contain as few as 6 images. We therefore restratify UTK-FF by the 60 gender $\times$ age $\times$ race&eth subgroups and also by source dataset. For each subgroup c of size n_c we assign $\max(50, 0.2n_c)$ images to evaluation and the rest to train-



(a) Overall accuracy. (b) Mean of 60 subgroup accuracies. (c) Mean Pairwise Delta. (d) Worst-10% subgroup mean ($\mathrm{CVaR}_{0.1}$).

Figure 3: CLIP fine-tuning metrics as the training set grows. The x-axis is $N{=}60 \times K$ for a per-cell budget K. (a)–(b) and (d) are higher-better; (c) is lower-better. DW allocates by the audit; Mixed and Synthetic reuse CIT images. Bands are 20-seed CIs. Pretrained is CLIP before fine-tuning. Yellow marks oversampling: some cells, mainly Infant, have fewer unique UTK-FF photos than K and are repeated. Grey is worse than Natural; leaving it outperforms that baseline.

ing. This yields 60 subgroups and 19,587 evaluation images, with 50–1,459 images per subgroup (median 203). The attribute map is in Appendix C.1, and the exact per-cell cut is in Appendix C.2.

Training. Of the four models audited in Sec. 5.1, we fine-tune CLIP ViT-B/32. CLIP is the usual workhorse for VLM fairness interventions. Following Kavuri et al. (2025), we update the vision encoder and keep the text tower frozen; full hyperparameters are in Appendix G. Labeled UTK-FF images for rare intersectional cells are scarce, so we also vary the per-cell budget from a few-shot regime upward: $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, for a training set of $N = 60 \times K$ images. This lets us see how Budget Allocation and Image Sources behave as more real data becomes available.

Evaluation Metrics. On these 60 cells we report MPD from Sec. 5.2 over subgroups rather than attributes, and replace the MinMax floor with $\mathrm{CVaR}_{0.1}$. Because an intervention succeeds only if it changes what the model decides, we evaluate mitigation on hard decisions, so both metrics are computed from α_c^{real} of Sec. 5.3.2 on the UTK-FF evaluation split. Unlike the axis-level MPD of the audit, which reads the confidence $\tilde{p}(I, a_{target})$, subgroup MPD therefore reads $\hat{y}$: it is the mean absolute difference of accuracies over all pairs of cells,

$$\mathrm{MPD}(\mathcal{C}) = \mathbb{E}_{c,c' \sim \mathcal{C},\, c \neq c'} \left| \alpha_c^{\mathrm{real}} - \alpha_{c'}^{\mathrm{real}} \right|. \tag{3}$$

A literal $\min_{c \in \mathcal{C}} \alpha_c^{\mathrm{real}}$ can be noisy due to small sample sizes, given the sample imbalance along the age axis. We replace it with a Conditional Value-at-Risk floor at level 0.1 (Rockafellar et al., 2000; Williamson & Menon, 2019). Let $\alpha_{(1)} \leq \cdots \leq \alpha_{(|\mathcal{C}|)}$ be the cell scores α_c^{real} in increasing order. Then $\mathrm{CVaR}_{0.1}(\mathcal{C}) = \frac{1}{m} \sum_{j=1}^{m} \alpha_{(j)}$ with $m = \lceil 0.1\,|\mathcal{C}| \rceil$. With $|\mathcal{C}| = 60$, this is the mean accuracy of the six worst cells. Below, we use $\mathrm{MPD}(\mathcal{C})$ and $\mathrm{CVaR}_{0.1}$ as intervention objectives, and the per-cell α_c^{syn} as the targeting signal.

6.3 When targeting helps vs. when mixing helps

Fig. 3a shows that Natural sampling often wins on overall accuracy (e.g., 82.7% at $K{=}128$ versus 81.8% for Real-Balanced). That plot is a frequency-weighted mean over images, not a fairness metric: Natural leaves the empirical attribute imbalance intact, so frequent cells dominate the average. Fig. 3b is the macro accuracy over the 60 cells, where Real-Balanced reaches 82.3% at $K{=}128$ against Natural's 81.6%. Figs. 3c and 3d are the fairness metrics of Sec. 6.2: Mean Pairwise Delta and $\mathrm{CVaR}_{0.1}$. Training on Synthetic only is weak on all three, because the generated images carry diffusion artifacts, resulting in a domain shift relative to UTK-FF (at $K{=}128$, Synthetic-Balanced is 76.2% overall and 55.9% $\mathrm{CVaR}_{0.1}$, against 81.8% and 69.5% for Real-Balanced).

At small K, Real-DW improves faster than Real-Balanced on macro accuracy, MPD, and $\mathrm{CVaR}_{0.1}$, raising $\mathrm{CVaR}_{0.1}$ by up to 4.0 pp (64.1% versus 60.1% at $K{=}8$). As K grows, Mixed improves even though Mixed 50% uses UTK-FF images for only half of N. The problem is not oversampling in general, but oversampling of *tail cells*: rare intersections that do not have enough unique UTK-FF photographs to fill a large per-cell budget. CIT images pad those cells with new samples rather than

repeating the same photos. At $K{=}128$, Mixed-Balanced 50% oversamples only the bottleneck cell Male / Infant / Latino (42 unique UTK-FF photographs) by $1.52\times$, whereas Real-Balanced repeats that cell $3.05\times$. Image Sources help once that budget outgrows the unique UTK-FF photographs in the tail cells: Mixed-Balanced 50% raises $\mathrm{CVaR}_{0.1}$ by 2.2 pp over Real-Balanced at $K{=}128$ (71.7% versus 69.5%).

7 Discussion and Limitations

Using the audit to build mitigation data. The axis-level audit names combinations where a disparity concentrates. It does not say which subgroup fails. Sec. 6 therefore reuses the audit on the one intersection with labeled photographs at a usable scale, allocating real and synthetic supervision across the 60 gender×age×race&eth cells by the synthetic per-cell scores.

A one-shot targeting signal saturates. Difficulty-weighted allocation beats Balanced at small K, but its advantage disappears as K grows (Fig. 3). The extra budget DW assigns to the audit-flagged cells has decreasing gain, while the other cells fall behind Balanced due to decreased budgets. This resembles the effect Shukla et al. (2025) report across axes: an intervention tuned to a fixed measurement trades gains on the targeted groups for losses elsewhere. A one-shot targeting signal is therefore best spent at small budgets; at larger budgets the targeting signal may need to be re-estimated after each intervention, which is practical here because the audit requires no labeled real data.

What synthetic data is for. Synthetic-only training is weak on all three metrics, so CIT images do not substitute for UTK-FF. However, they help avoid oversampling in rare cells that cannot fill large budgets, whether it is a uniform per-cell budget or a difficulty-weighted one. Their value is sample diversity in the tail rather than realism, giving a concrete switching rule: mix once a cell's assigned budget exceeds the unique UTK-FF images available in the rarest cells.

Limitations. This study has several limitations that bound our claims. The scored axes are not exhaustive: *Faith* and *Socioeconomic Status* remain in the prompts but are not audited, so they may still confound the scored axes. Diffusion models are not bias-free either; we observed structural warping in markers such as wheelchairs (Appendix B), and computational constraints prevented a full regeneration that would isolate those artifacts. The axis-level tests show that a disparity is not sampling noise and that it replicates inside this generation pipeline. They do not separate a bias of the VLM from a combination the generator renders less faithfully, or from a combination that is hard to recognize, such as the gender of an infant. Covering arrays also require a finite attribute set, so every kept axis is discretized—age attributes, wheelchair vs. not, fat/thin, location, race, gender. VQA probes leave some of that overlap in: some adults can look elderly, so both "Is this person Adult?" and "Is this person Elderly?" can return yes. CLIP cannot use that slack and it picks exactly one attribute. We treat that as a limitation of the CLIP probe (Appendix D). Coverage is also incomplete: gender is binary in this audit, erasing non-binary, intersex, and genderqueer people, and the race taxonomy omits Middle Eastern and Southeast Asian attributes.

As noted by Ovalle et al. (2023), much of the current literature—including this work—treats intersectionality primarily as optimizing fairness metrics over demographic subgroups. We report recognition disparities. A VLM can be part of a larger system, in which those disparities may be amplified on downstream tasks, but we do not analyze that propagation and leave it to future work.

8 Conclusion

We present a scalable generative auditing framework for intersectional bias in VLMs. Combining **Combinatorial Interaction Testing** with diffusion-based synthesis turns an intractable space of attribute combinations into a reasonable number of covering prompts. The covering strength t sets a tradeoff between how much context a comparison holds fixed and the generation and query budget. Fairness bugs replicate on held-out prompts ($t{=}2$, 3) and split-half validation ($t{=}4$). The synthetic audit also flags the same hard cells as UTK-FF: per-cell recognition on our images tracks that on UTK-FF for all four audited VLMs. The same audit can be reused for mitigation—as a targeting signal that reallocates real supervision toward hard cells at small budgets, and as mixed fine-tuning data for tail cells as the budget grows.

AI USE STATEMENT

In this work, we used generative AI tools for generating the synthetic audit dataset (Sec. 4), dataset reformatting, implementing methods, proposing the statistical validation procedure, quantitative data analysis and statistical testing, providing feedback on the experimental design, and interpreting results. We have not used generative AI tools for developing the conceptual framework, or proposing or refining hypotheses. Apart from the statistical validation procedure above, the CIT-based audit and mitigation framework, the hypotheses, the methodology, and the experimental design were developed by the authors. We also have not used generative AI tools for qualitative or thematic data analysis. Formulating and proving mathematical claims and translation are not applicable to this work. Additionally, we used generative AI tools for creating and modifying scientific figures, writing and editing software code, drafting parts of the paper, and editing the paper to improve readability. We have reviewed all AI-assisted work. A sample of 690 synthetic images was screened by four authors after generation (Sec. 4.3). An LLM was also used to enlarge the pool of attribute candidates (Appendix A.1) and to iterate on the generation prompt wording (Appendix B.2). AI-assisted method implementations were reviewed by one to two authors. The statistical procedure, and the outputs of AI-assisted data analysis and statistical testing, were reviewed by one author. Every AI-assisted interpretation of results was verified by two authors against the underlying experimental outputs. All AI-assisted drafts of the paper were read and revised in full by two authors, and the final manuscript was reviewed by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies intersectional recognition failures in VLMs by synthesizing attribute-labeled images and, on a subset of axes, asking whether a model recognizes the depicted person as being each attribute. Our aim is to audit that recognition, not to build a demographic classifier: the VQA probe asks only whether the ground-truth label applies and never asks the model to choose among labels. CLIP is the exception. Because it does not reliably process negation, it cannot compare a label with its negation, so its probe is necessarily a closed-set classifier over the labels of each axis (Appendix D). Both probes are audit instruments, not a recommendation to infer race, gender, age, or disability in deployment. Likewise, the fine-tuning in Sec. 6 aims to narrow gaps between subgroups, not to improve demographic inference for its own sake. We do not advocate production use of such classifiers, including in settings restricted by rules on biometric categorization (e.g., the EU AI Act).

Generating images from prompts that name race, gender, age, disability, faith, socioeconomic status, and appearance can reproduce stereotypes and visual artifacts, including warped depictions of wheelchair users. We retain every generated image in the covering array, including prompts whose faith or class attributes we later exclude from scoring, because dropping rows would break t-way coverage of the remaining axes. We treat those images as test fixtures. Attributes that should not be read from appearance—sexual orientation is the running example—were not used as generation targets. Which remaining attributes ought to be rendered, versus when an audit should abstain rather than infer, is a contextual ethical choice, not a solved technical one.

Visual inspection of 690 images was performed by the authors, who share a similar cultural background; that review is not a substitute for annotators from the communities being depicted. The scored taxonomy is itself an ethical cut: gender is operationalized as binary, erasing non-binary, intersex and genderqueer people, and the race axis omits Middle Eastern and Southeast Asian attributes. Photographic supervision uses UTK-FF, built from the public UTKFace and FairFace corpora under their existing licenses; we did not collect new images of people. Code will be released for research on fairness auditing and mitigation, not as a toolkit for inferring identity from faces in the wild.

The disparities we report are recognition errors on an audit, not measured distributive harm to the groups named in the prompts. Mitigation experiments reallocate training budget toward cells the audit flags as hard; they are not a claim that fine-tuning on synthetic identities remediates real-world harm.

Reproducibility statement

The covering-array construction, prompt template, and SDXL generation hyperparameters are given in Appendix B, and the exact probe wording for each model is in Appendix D. The attribute mapping and train/evaluation split of UTK-FF are described in Appendix C, and the fine-tuning hyperparameters and sampling procedures are in Appendix G.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect?, 2024. URL https://arxiv.org/abs/2406.02965.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

David M. Cohen, Siddhartha R. Dalal, Michael L. Fredman, and Gardner C. Patton. The aetg system: An approach to testing based on combinatorial design. *IEEE Transactions on Software Engineering*, 23(7):437–444, 1997.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pp. 23–51. Routledge, 2013.

Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Giulio Filippi, Sara Zannone, and Adriano Koshiyama. Intersectional fairness: A fractal approach. *arXiv preprint arXiv:2302.12683*, 2023.

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)*, pp. 690–713, 2024.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 547–561, 2024.

Johannes Himmelreich, Arbie Hsu, Ellen Veomett, and Kristian Lum. The intersectionality problem for algorithmic fairness. In *Workshop on Algorithmic Fairness Through the Lens of Metrics and Evaluation*, pp. 68–95. PMLR, 2025.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

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

Rashidul Islam, Kamrun Naher Keya, Shimei Pan, Anand D Sarwate, and James R Foulds. Differential fairness: an intersectional framework for fair ai. *Entropy*, 25(4):660, 2023.

Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

Vivek Hruday Kavuri, Vysishtya Karanam Karanam, Venkamsetty Venkata Jahnavi, Kriti Madumadukala, Balaji Lakshmipathi Darur, and Ponnurangam Kumaraguru. Freeze and reveal: Exposing modality bias in vision-language models. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pp. 17–26, 2025.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

D Richard Kuhn, Dolores R Wallace, and Albert M Gallo. Software fault interactions and implications for software testing. *IEEE transactions on software engineering*, 30(6):418–421, 2004.

Hanjun Luo, Ziye Deng, Ruizhe Chen, and Zuozhu Liu. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *arXiv preprint arXiv:2405.17814*, 2024.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6755–6764. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/martinez20a.html.

Jan Hendrik Metzen, Robin Hutmacher, N Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5041–5050. IEEE, 2023.

Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. *Advances in Neural Information Processing Systems*, 35:16796–16807, 2022.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

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.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*, 2016.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Resmi Ramachandranpillai, Kishore Sampath, Ayaazuddin Mohammad, and Malihe Alikhani. Fairness at every intersection: Uncovering and mitigating intersectional biases in multimodal clinical predictions. *arXiv preprint arXiv:2412.00606*, 2024.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Antonio Rodà et al. The compas case: An educational journey for explaining fairness in ai-based applications. In *AIMMES*, 2025.

Vanessa Rodriguez. Ethical implications of ai-based algorithms in recruiting processes: A study of civil rights violations under title vii and the americans with disabilities act. 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Pushkar Shukla, Aditya Chinchure, Emily Diana, Alexander Tolbert, Kartik Hosanagar, Vineeth N Balasubramanian, Leonid Sigal, and Matthew Turk. Mitigate one, skew another? tackling intersectional biases in text-to-image models. *arXiv preprint arXiv:2505.17280*, 2025.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Youmi Suk and Kyung (Chris) T Han. Evaluating intersectional fairness in algorithmic decision making using intersectional differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, 50(5):833–862, 2025.

UN Women. Intersectionality resource guide and toolkit, 2022. URL `https://www.unwomen.org/sites/default/files/2022-01/Intersectionality-resource-guide-and-toolkit-en.pdf`. Accessed: 2026-01-30.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 336–349, 2022.

Yifan Wang, Peijie Sun, Weizhi Ma, Min Zhang, Yuan Zhang, Peng Jiang, and Shaoping Ma. Intersectional two-sided fairness in recommendation. In *Proceedings of the ACM Web Conference 2024*, pp. 3609–3620, 2024.

Solvejg Wastvedt, Jared D Huling, and Julian Wolfson. An intersectional framework for counterfactual fairness in risk prediction. *Biostatistics*, 25(3):702–717, 2024.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/williamson19a.html`.

Xuyang Wu, Yuan Wang, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. *arXiv preprint arXiv:2406.17974*, 2024.

Shlomo Yitzhaki et al. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron*, 61(2):285–316, 2003.

Miao Zhang and Rumi Chunara. Configurable fairness: Direct optimization of parity metrics via vision-language models. *arXiv preprint arXiv:2403.10624*, 2024.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5810–5818, 2017.

A Axes Selection

A.1 Rationale for Identity Axes and Attribute Selection

The eight axes and attributes in Table 1 were selected in three stages.

1. **Consolidation of Existing Benchmarks**: We first take *Race & Ethnicity*, *Gender*, and *Age*, the protected axes used in prior fairness benchmarks (Luo et al., 2024; Hamidieh et al., 2024; Shukla et al., 2025), so that cells in this audit stay comparable with that literature. The five Race & Ethnicity attributes in Table 1 follow Luo et al. (2024). *Appearance* is contextual: it changes how a protected identity is depicted and read. Its attributes are adapted from the characteristic antonyms in the same benchmark, which pairs attractive with ugly and strong with weak. We use fat/thin, attractive/unsightly, and strong/weak.
2. **Expansion via Global Frameworks**: We then add *Disability Status*, *Faith*, and *Socioeconomic Status*. UN Women (2022) treats marginalization as occurring along these axes as well as along race & ethnicity and gender, and we count them as protected. *Location* is added in the same stage as a contextual axis: it is the setting in which those identities are depicted. Faith and socioeconomic status remain in the prompts; whether they can be scored is decided by the visual check in Sec. 4.3.
3. **Attribute Selection**: Attributes on each axis come from the sources above. An LLM was used to enlarge the pool of candidates; the attributes in Table 1 are the ones we kept. They are not mutually exclusive or collectively exhaustive. This choice was made before any generated image was inspected. We included attributes that we deemed visually expressible, and the later visual check in Sec. 4.3 decides which of them can be scored.

A.2 Prompt–Image Alignment

The visual check in Sec. 4.3 covers 690 images of the 2-way covering array, 10 per prompt. Four authors split these images among themselves. For every attribute in the prompt, the annotator marked the image as *aligned* (the attribute is visibly realized), *unsure*, or *misaligned* (a different attribute is realized). Table 2 reports the result on the attributes that remain scored after the audit scope is refined. Disability Status and Appearance have smaller N because only images prompted with the retained attributes are counted. The overall row pools all image–attribute pairs.

Table 2: Prompt–image alignment from author annotation, restricted to the attributes scored in the audit. N is the number of annotated images per axis; rows are sorted by alignment.

Axis	Scored attributes	N	Aligned (%)	Unsure (%)	Misaligned (%)
Gender	All (2)	690	99.57	0.43	0.00
Age	All (6)	690	94.93	1.30	3.77
Disability Status	Uses a wheelchair, Without a disability	200	94.00	1.00	5.00
Race & Ethnicity	All (5)	690	90.72	2.32	6.96
Appearance	Fat, Thin	200	89.50	4.00	6.50
Location	All (8)	690	83.48	8.41	8.12
Overall		3160	92.12	3.04	4.84

B Synthetic Image Generation Details

B.1 Combinatorial Interaction Covering Array Generation

Each covering array is built by a randomized variant of AETG's one-row-at-a-time greedy construction (Cohen et al., 1997). For $t = 2$, the procedure starts from an empty set of rows and a pool of every 2-way combination of identity-axis attributes. For $t > 2$, it loads the full $(t-1)$-way array as the initial row set, and the pool contains the t-way combinations still absent from those rows. The procedure then adds one row at a time. A row assigns a single attribute to every identity axis. To choose that assignment, the algorithm draws 1000 candidate rows. Each candidate selects an attribute independently and uniformly at random on every axis, and the candidate that covers the

largest number of combinations still in the pool is retained. A later candidate replaces the current choice only when its coverage is strictly larger, so equal coverage leaves the earlier draw in place. If every candidate covers none of the remaining combinations, a row drawn in the same way is added anyway. Combinations that appear in the retained row are removed from the pool. The procedure stops when the pool is empty or $10{,}000$ new rows have been added. In the arrays used here the pool is empty at termination, so every t-way combination occurs in at least one row. A combination already removed from the pool can still appear in a later row, because a new row is a full assignment and only the pool used for scoring is reduced.

For each interaction strength $t \in \{2, 3, 4\}$, this procedure is run on the axes in Table 1. Each row is then rendered as a natural-language prompt with the template in the next subsection.

B.2 Prompt Template Design

To encode these combinations of attributes into natural language, we adopt a structured prompt template designed for semantic clarity and consistency: *"a photo of a {Appearance} {Race & Ethnicity} {Age&Gender} in a {Location}, who is of {Faith} faith, has a {Socioeconomic status} status, and {Disability status}."* Slots are adjusted where needed so the sentence stays grammatical, including the article *a* or *an*.

Age and *Gender* remain separate axes in the covering array, but the template realizes them as one noun phrase so the sentence stays natural. A female infant is written as "baby girl," and a middle-aged male as "middle-aged man," in place of forms such as "female infant" or "middle-aged male." The slot order and these surface forms were settled by iterating on draft prompts with an LLM until the wording read as ordinary description across the attribute set.

B.3 Image Generation Configuration

Synthetic images are produced via the **Stable Diffusion XL (SDXL 1.0-base)** model (Podell et al., 2024) using the HuggingFace *diffusers* framework.

Hyperparameters. The generation process utilizes 15 inference steps with a classifier-free guidance (CFG) scale of 7.0. All images are generated at a resolution of 1024×1024 pixels. To account for the stochastic nature of the diffusion process and to ensure statistical robustness, we generate 50 independent images per single prompt.

Prompt Engineering. A standardized positive suffix is appended to all CIT-generated prompts to maintain consistent composition across multiple batches:

> *", exposed face, looking at the camera, ultra quality, sharp focus, 8k, photorealistic, intricate details, hdr, high resolution"*

To mitigate common artifacts such as disfigured limbs or excessive cropping that might obscure identity-related features (e.g., mobility aids or ethnic attire), we apply the following negative prompt (Ban et al., 2024):

> *"lowres, bad anatomy, error body, error hands, missing fingers, cropped, worst quality, low quality, jpeg artifacts"*

Hardware. All generation tasks are performed on a single 20GB partition of an NVIDIA A100 GPU (GRID A100D-20C).

B.4 Example generated images

Fig. 4 shows 64 images from prompts built by the procedure above, spanning the eight axes in Table 1.



Figure 4: Example CIT-guided synthetic images.

Across the eight axes, the prompted person is recognizable in most images. A few renders are still awkward: a warped wheelchair, a wheelchair prompt shown as two people, and malformed limbs. In the visual inspection (Sec. 4.3), leftovers of this kind did not seem to make the intended attributes hard for a person to read, so we accepted them at that level.

C UTK-FF Dataset Details

UTK-FF merges UTKFace (Zhang et al., 2017) with both the train and validation splits of FairFace (Karkkainen & Joo, 2021) after both are mapped onto the gender, age, and race attributes in Table 1. Fig. 7 shows 64 example images.

C.1 Mapping UTK-FF onto the synthetic attributes

UTK-FF uses the gender, age, and race attributes of Table 1 . FairFace (Karkkainen & Joo, 2021) and UTKFace (Zhang et al., 2017) use different vocabularies, so both are mapped onto that taxonomy before the two are merged. Images that do not map onto a single attribute are dropped. The resulting attribute space is $2 \times 6 \times 5 = 60$ cells.

Gender is already binary. FairFace `Male` and `Female` are kept as written. UTKFace encodes gender in the filename as 0 (Male) or 1 (Female).

FairFace age is nine bins. Adjacent bins that fall inside one age attribute are collapsed, and UTK-Face's integer age is cut at the same boundaries (Table 3).

Table 3: Age mapping. FairFace bins and UTKFace integer ages are mapped to the six age attributes.

Audit age	**FairFace**	**UTKFace (years)**
Infant	0–2	0–2
Child	3–9	3–9
Teenager	10–19	10–19
Adult	20–29, 30–39	20–39
Middle-aged	40–49, 50–59	40–59
Elderly	60–69, 70+	60+

Race is mapped in Table 4. Five FairFace races map directly onto the audit attributes. Middle Eastern and Southeast Asian have no counterpart in Table 1 and are excluded. UTKFace stores race as an integer: White, Black, Asian, and Indian map to European, African, East-Asian, and South-Asian. UTKFace does not separate East Asian from Southeast Asian, so its Asian code is assigned entirely to East-Asian. Its Others code mixes Hispanic, Latino, and Middle Eastern faces and is discarded, so the Latino cells of UTK-FF come only from FairFace.

Table 4: Race mapping. Rows below the line are excluded from UTK-FF.

Audit race	**FairFace**	**UTKFace code**
European	White	0 (White)
African	Black	1 (Black)
East-Asian	East Asian	2 (Asian)
South-Asian	Indian	3 (Indian)
Latino	Latino_Hispanic	—
excluded	Middle Eastern, Southeast Asian	4 (Others)

C.2 TRAIN/EVAL SPLIT

In Sec. 6, we fine-tune CLIP on the UTK-FF training split. FairFace's official validation split is merged back into that pool before the cut, because some cells in the official split contain as few as 6 images. UTK-FF has 97,455 images in 60 gender $\times$ age $\times$ race cells. For a cell c of size n_c, the evaluation count is

$$n_c^{\mathrm{val}} = \min\big(n_c,\ \max(\mathrm{round}(0.2\,n_c),\ 50)\big), \tag{4}$$

and the remainder is training. That quota is divided between FairFace and UTKFace by the largest-remainder method, in proportion to the two sources inside c. Within each source, paths are sorted and the selected files are taken at even intervals, with no random seed. The same evaluation set is used for every training seed. This yields 19,587 evaluation images (15,102 FairFace, 4,485 UTKFace) and 77,868 training images (59,961 FairFace, 17,907 UTKFace). Evaluation size per cell ranges from 50 to 1,459, and training size from 42 to 5,838. The smallest cell has 92 images, so the floor of 50 leaves 42 for training.

C.3 COMPARISON OF SYNTHETIC AND UTK-FF

C.3.1 STATISTICS

Fig. 5 shows the pairwise counts. Along race and gender the bars stay near the uniform line, with a max/min ratio of 2.06:1. Age is the unbalanced axis: the mass sits on Adult, and the ratio rises to 19.38:1 for gender$\times$age and 68.38:1 for age$\times$race&eth. Among the 60 cells, the largest is Female / Adult / European (7,297 images) and the smallest is Male / Infant / Latino (92 images).

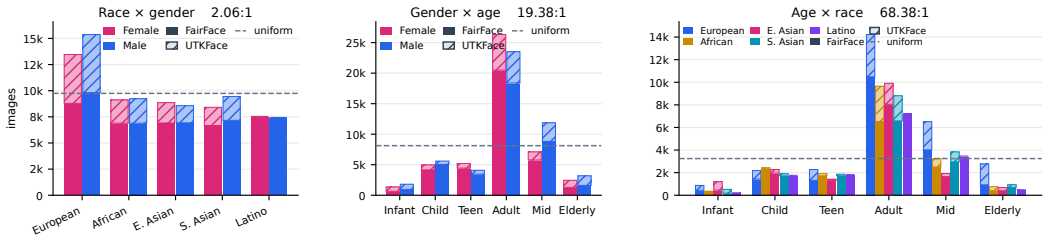


Figure 5: UTK-FF counts on the merged train and validation pool. Color marks gender in the first two panels and race in the third; hatching marks UTKFace against solid FairFace. The dashed line is a uniform split of that panel, and the annotated ratio is the max/min ratio of the subgroups split by that panel's two axes.

The synthetic images in the same 60-cell comparison number $n = 142{,}350$. Fig. 6 shows their pairwise counts on gender, age, and race. The 60 cells sit in a narrow band: the largest are Male / Adult / European and Male / Adult / Latino, both with $2{,}800$ images, and the smallest is Female / Adult / Latino, with $1{,}850$. The cell max/min ratio is $2{,}800/1{,}850 \approx 1.51{:}1$.

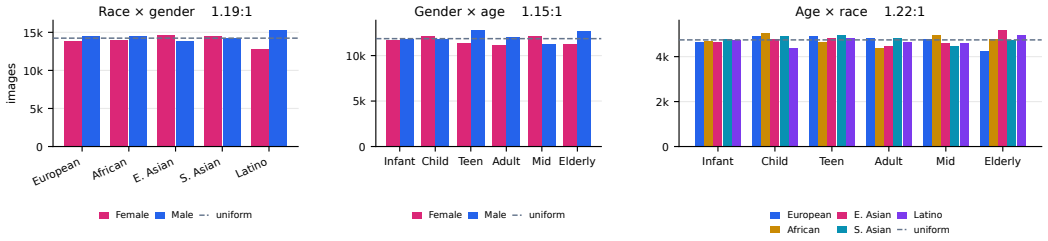


Figure 6: Synthetic image counts on the gender×age×race&eth margin. Color marks gender in the first two panels and race in the third. The dashed line is a uniform split of that panel, and the annotated ratio is the max/min ratio of the subgroups split by that panel's two axes.

C.3.2 QUALITATIVE

Fig. 7 shows photographs from UTK-FF. Fig. 8 pairs one UTK-FF photograph with one CIT-guided image for each of the 60 subgroups. Side by side, UTK-FF photographs are tighter face crops, and some are letterboxed or have the background removed, whereas CIT-guided images leave a larger share of the frame to the background.



Figure 7: 64 example images from UTK-FF. Images to the left of the black line are FairFace examples, and images to the right are UTKFace examples.



Figure 8: One cell per gender×age×race&eth subgroup. Each cell pairs a UTK-FF photograph with a CIT-guided image of the same subgroup.

Note that we are reusing the same synthetic image, so the influence of appearance, location, disability, faith, and socioeconomic status still remains in the picture.

D EVALUATION PROBES

We write the VQA probe schematically as "Is this person $\{a_{i,k}\}$?". That line is a schema. The string sent to the model is a handwritten sentence for that attribute, with the article, preposition, and verb written out. Hyphenated race attributes are spaced (*East-Asian* $\rightarrow$ "East Asian"). The disability attribute "who uses a wheelchair" is asked as "Does the person use a wheelchair?". Faith and socioeconomic status are not probed (Sec. 4.3). Appearance is queried only for Fat and Thin, and disability only for wheelchair use and the absence of a disability.

On the synthetic images the scored axes are gender, age, race, location, appearance, and disability. UTK-FF has attributes only for gender, age, and race, so only those three sentences (and the matching CLIP prompts) are used there. The wording is the same.

VQA models. ViLT, PaliGemma 2, and Qwen2.5-VL share Table 5. Each image is asked only the sentence for its ground-truth attribute on that axis, so the expected answer is always Yes. The Yes/No probabilities are normalized as

$$\tilde{p}_{i,k} = \frac{p_{\text{yes}}}{p_{\text{yes}} + p_{\text{no}}}, \qquad \hat{y}_{i,k} = \mathbb{1}(\tilde{p}_{i,k} \geq 0.5). \tag{5}$$

Qwen2.5-VL appends "Answer yes or no." and compares the next-token logits of `yes` and `no`; the decision is Eq. (5). PaliGemma 2 uses the same comparison after the mix-model prefix "`answer en`". ViLT-B/32 is a VQA classifier; we take the `yes` and `no` probabilities from its answer vocabulary and apply the same Eq. (5).

Table 5: Yes/no questions used for ViLT, PaliGemma 2, and Qwen2.5-VL. One question is asked per scored axis, the one whose attribute matches the image.

Axis	Attribute	Question
Gender	Male	Is the person male?
	Female	Is the person female?
Age	Infant	Is the person an infant?
	Child	Is the person a child?
	Teenager	Is the person a teenager?
	Adult	Is the person an adult?
	Middle-aged	Is the person middle-aged?
	Elderly	Is the person elderly?
Race	European	Is the person European?
	African	Is the person African?
	East-Asian	Is the person East Asian?
	South-Asian	Is the person South Asian?
	Latino	Is the person Latino?
Location	Urban area	Is the person in an urban area?
	Suburban area	Is the person in a suburban area?
	Rural area	Is the person in a rural area?
	Remote area	Is the person in a remote area?
	Neighborhood	Is the person in a neighborhood?
	City	Is the person in a city?
	Workplace	Is the person at a workplace?
	School	Is the person at a school?
Appearance	Fat	Is the person fat?
	Thin	Is the person thin?
Disability	Who uses a wheelchair	Does the person use a wheelchair?
	Without a disability	Is the person without a disability?

CLIP. CLIP ViT-B/32 scores a closed set of captions. Each axis is treated as a closed set: I_i is scored against one caption per attribute $l \in L_k$ (Table 6), and

$$\hat{y}_{i,k} = \mathbb{1}\left(\arg\max_{l \in L_k} s(I_i, l) = a_{i,k} \right), \tag{6}$$

where $s(I_i, l)$ is the image–text similarity and $\tilde{p}_{i,k}$ is the softmax probability of the ground-truth caption. CLIP's audit is therefore closer to closed-set classification than to open VQA. The indefinite article follows the next word, and race captions drop the hyphen. Location uses "in" for every

site, including school and workplace, where the VQA questions use "at". Appearance and disability stay two-way and use the same attribute restriction as the VQA probes. Age is the 6-way set over ordered neighbors (Infant, Child, Teenager, Adult, Middle-aged, Elderly). On the 60-cell comparison (Fig. 2), CLIP's correlation with UTK-FF is the lowest of the four models. A possible reason is that the closed-set probe is brittle on Age. The 60 cells use gender, age, and race only, and of those three, the Age captions are ordered neighbors, so the argmax sometimes lands on an adjacent bin. A VQA probe asks only the ground-truth attribute, so a neighboring age caption does not have to lose for the correct answer to count.

Table 6: CLIP ViT-B/32 captions. One softmax per axis; the sets are 2-way (gender, appearance, disability), 6-way (age), 5-way (race), and 8-way (location).

Axis	Captions
Gender	a photo of a male; a photo of a female
Age	a photo of an infant; a photo of a child; a photo of a teenager; a photo of an adult; a photo of a middle-aged person; a photo of an elderly person
Race	a photo of a European; a photo of a South Asian; a photo of a Latino; a photo of an African; a photo of an East Asian
Location	a photo of a person in a city; a photo of a person in a neighborhood; a photo of a person in a remote area; a photo of a person in a rural area; a photo of a person in a school; a photo of a person in a suburban area; a photo of a person in an urban area; a photo of a person in a workplace
Appearance	a photo of a fat person; a photo of a thin person
Disability	a photo of a person who uses a wheelchair; a photo of a person without a disability

Hardware. VQA evaluation is performed on an NVIDIA RTX 5080 and an RTX 3090.

E Other Model Results

E.1 ACC

Table 7 gives per-axis accuracy on the synthetic images for all four models. Which axes clear chance depends on the model. Gender lies above chance for every model (0.83–0.99; chance 0.5). Appearance and Disability Status, at 0.88 and 0.91 for CLIP, fall near or below chance for ViLT (0.22 and 0.51) and for PaliGemma 2 (0.44 and 0.38). Age accuracy is lowest for CLIP (0.47) and highest for PaliGemma 2 (0.91). Every scored axis of Qwen2.5-VL lies above chance.

Table 7: Attribute recognition accuracy on the synthetic images, aggregated per axis. Chance is $1/|L_k|$ over the scored attributes of that axis, the closed-set rate for CLIP. A yes/no VQA probe has chance 0.5 on every axis.

Axis	ViLT	CLIP	PaliGemma 2	Qwen2.5-VL	Chance (CLIP)
Race & Ethnicity	0.7853	0.7049	0.8000	0.6999	0.2000
Age	0.7227	0.4684	0.9131	0.6964	0.1667
Appearance	0.2195	0.8781	0.4414	0.7386	0.5000
Gender	0.8332	0.9737	0.9932	0.9570	0.5000
Location	0.8809	0.5403	0.8941	0.8319	0.1250
Disability Status	0.5117	0.9076	0.3779	0.8483	0.5000

E.2 MPD/MinMax

Figs. 9, 10, and 11 repeat the axis-level audit of Fig. 1 for ViLT, PaliGemma 2, and Qwen2.5-VL. As in that figure, each panel is the $k_t = \lceil 0.1\, n_t \rceil$ items with the largest shrunken chance-corrected excess on the t-wise array (3 of 30 at t=2; 6 of 60 at t=3 and t=4), the same selection as Tables 8 and 9. Bars are the unshrunken chance-adjusted discovery excess; diamonds are the held-out excess. Held-out checks are in Appendix F.

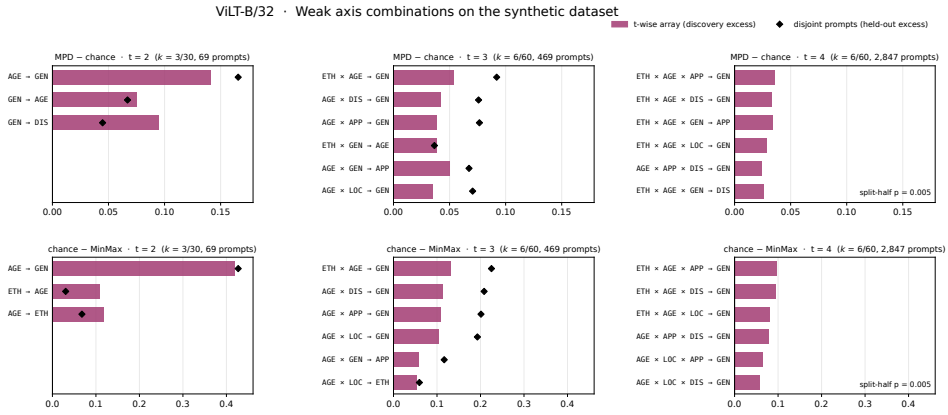


Figure 9: ViLT-B/32 fairness bugs on synthetic images. Same selection, order, and bar convention as Fig. 1 and Tables 8–9: the k_t items with the largest shrunken chance-corrected excess; bars are the unshrunken discovery excess; diamonds are the held-out excess. Split-half $p = 0.005$ at t=4.

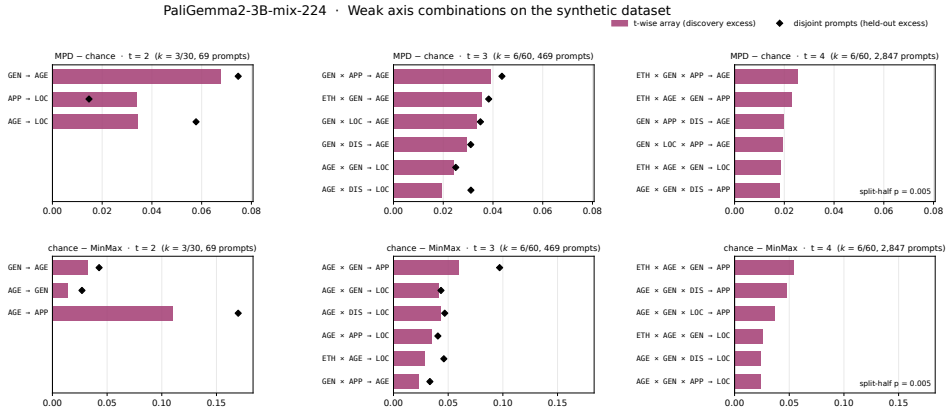


Figure 10: PaliGemma 2 3B fairness bugs on synthetic images. Same selection, order, and bar convention as Fig. 1 and Tables 8–9: the k_t items with the largest shrunken chance-corrected excess; bars are the unshrunken discovery excess; diamonds are the held-out excess. Split-half $p = 0.005$ at t=4.

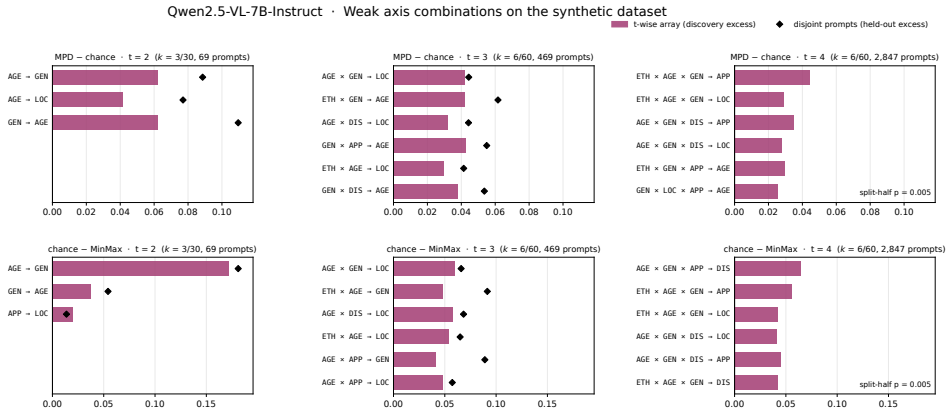


Figure 11: Qwen2.5-VL-7B fairness bugs on synthetic images. Same selection, order, and bar convention as Fig. 1 and Tables 8–9: the k_t items with the largest shrunken chance-corrected excess; bars are the unshrunken discovery excess; diamonds are the held-out excess. Split-half $p = 0.005$ at t=4.

F Validation of the axis-level weak points found by each covering array

Question. For each strength $t \in \{2, 3, 4\}$, the audit ranks the $S \to A_{target}$ items of Sec. 5.2 by their disparity, computed on the prompts of the t-wise covering array only. We ask whether the items it ranks as weakest are genuine: (i) their disparity exceeds what sampling noise produces, (ii) they remain weaker than a randomly chosen item on prompts that were not used to find them, and (iii) the magnitude estimated on the t-wise array is not inflated by selection. The covering arrays are nested, so the prompts outside a given t-wise array were not used to rank it: the t=2 array consists of prompts 0–68 (69 prompts), the t=3 array of prompts 0–468 (469 prompts), and the t=4 array of all 2,847 prompts, each rendered as 50 images. The 50 images of a prompt share every assigned attribute and are strongly correlated, so the prompt is the unit of inference throughout.

Chance adjustment. In this design the sampling noise is large, and it is larger for some items than for others, so the raw metrics cannot be ranked. We use a small t-wise covering array (69, 469, and 2,847 prompts at t=2, 3, and 4), and the 50 images of a prompt are one observation. Each item splits that prompt budget across the attribute combinations of $\mathcal{S} \cup \{A_{target}\}$. Location has 8 attributes, age 6, and ethnicity 5, so a varied axis with many attributes leaves few prompts in each cell. Appearance and disability are scored only on Fat/Thin and wheelchair/without-disability, while the array was built on the full attribute lists, so those targets are estimated from the prompts that carry a scored attribute. Sampling noise therefore inflates MPD and lowers the MinMax floor by an amount that tracks cell size, even when $\mathcal{S}$ has no effect. As with the adjusted Rand index or Cohen's κ, we report each metric relative to its value expected by chance,

$$\mathrm{MPD}^\star = \mathrm{MPD} - \mathbb{E}_\pi[\mathrm{MPD}], \qquad \mathrm{MinMax}^\star = \mathbb{E}_\pi[\mathrm{MinMax}] - \mathrm{MinMax}. \tag{7}$$

For every item j and prompt set, $\mathbb{E}_\pi$ is the mean of the item's value over 1,000 permutations that swap the outcomes (yes-probability and correctness) of whole 50-image prompt blocks between prompts with the same ground-truth attribute of the target. Assigned attributes stay fixed, and swaps stay within the nested bands (prompts 0–68, 69–468, and 469–2,846), so that each t-wise array and its complement are permuted within themselves. Both starred metrics are zero in expectation when $\mathcal{S}$ carries no information about the model's output; larger values are less fair. $\mathrm{MinMax}^\star$ also removes the difficulty of the target itself: it is how far the recognition floor falls below the floor expected for this target at these cell sizes. The excess e_j used below is $\mathrm{MPD}^\star_j$ or $\mathrm{MinMax}^\star_j$, and σ_j^2 is the permutation variance of the unstarred metric under the same null.

Ranking by empirical-Bayes shrinkage. The same cells stay noisy after the chance level is removed. An item whose slice of the covering array contains few prompts has a large permutation

variance, and that variance would place it in the weak tail. The weak points of strength t are the $S \to A_{target}$ items with the largest shrunken excess, computed on the t-wise covering array. There are $n_t = 30$, 60, and 60 such items at t=2, 3, and 4. We rank them by the empirical-Bayes estimate (Efron & Morris, 1975)

$$\tilde{e}_j = \bar{e} + (e_j - \bar{e})\,\frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_j^2},$$

$$\hat{\tau}^2 = \max\Big\{\tfrac{1}{n_t}\textstyle\sum_j (e_j - \bar{e})^2 - \tfrac{1}{n_t}\sum_j \sigma_j^2,\ 10^{-10}\Big\},$$

which shrinks each excess toward the mean $\bar{e}$ of the n_t items of order t in proportion to its sampling variance; $\hat{\tau}^2$ is the method-of-moments estimate of the between-item variance of the true excess. When $\hat{\tau}^2$ reaches its floor, all $\tilde{e}_j$ are close to $\bar{e}$ and the ranking is by $(e_j - \bar{e})/\sigma_j^2$. The weak points of order t are the $k_t = \lceil 0.1\, n_t \rceil$ items with the largest $\tilde{e}_j$ (3 of 30 at t=2, and 6 of 60 at t=3 and at t=4). Selection uses only the prompts of the t-wise array.

Tests. *(A) Within the t-wise array.* The statistic is the mean $\tilde{e}$ of the k_t selected items. In each of the 1,000 permutations the excess, the shrinkage, the ranking, and the selection are recomputed, so the null distribution includes the selection step; $p = (1 + \#\{T^* \geq T\})/1{,}001$ (Phipson & Smyth, 2016).

(B) Held-out prompts, $t \in \{2, 3\}$. Because the arrays are nested, the prompts of the t=4 array outside the t-wise array (2,778 prompts for t=2 and 2,378 for t=3) share no prompt and no image with the discovery prompts. For every item we compute its held-out excess h_j, with the chance level re-estimated on the held-out prompts from the same permutations. We test the mean held-out excess of the selected items (B1) against its permutation distribution on the held-out prompts, i.e., whether the selected disparity is real; (B2) against the means of 10^5 random k_t-subsets of all items, i.e., whether the selection beats chance and the ranking replicates; and (B3) item by item, with a one-sided held-out permutation p-value per selected item, Holm-corrected over the k_t items (Holm, 1979).

(C) Split-half, t=4. The t=4 array has no prompts outside it. We split its 2,847 prompts into two random halves 100 times. On each half, the chance levels and null variances are estimated from 30 permutations; the weak points are selected on one half and their excess is read on the other, in both directions (200 replicates). The reference distribution is the identical pipeline applied to a permuted data set (one permutation per split), so the selection and the finite number of permutations are part of the reference; $p = (1 + \#\{\text{reference} \geq \text{observed mean}\})/(1 + 200)$. We also report the ratio of the held-out mean to the discovery estimate (the mean $\tilde{e}$ of the selected items), which is 1 when the estimate is neither inflated nor attenuated.

Results. Table 8 summarizes the primary analysis (MPD) and the same procedure applied to MinMax. *Ranking.* For MPD, the weak points replicate in all 12 model–order combinations. On the held-out prompts, the three t=2 weak points have 2.2–3.3 times the mean excess of all items (B2: $p = 0.006$–0.045) and the six t=3 weak points 2.3–3.4 times ($p \leq 0.002$). At t=4, the split-half excess exceeds the reference for every model ($p = 0.005$, the smallest attainable value). *Reality of each weak point.* For MPD, the selected items exceed chance on the held-out prompts (B1: $p \leq 0.001$ for every model at t=2 and t=3), and every one of them is confirmed individually (B3: 3/3 at t=2 and 6/6 at t=3 for every model). *Magnitude.* At t=4 the discovery estimate is calibrated (held-out/discovery ratio 1.08–1.22). At t=3 the held-out excess is larger than the discovery estimate (ratio 1.48–2.24), so the estimate is conservative: the excess of an absolute difference is attenuated when the true disparity is small relative to the noise, and the shrinkage adds to this. At t=2 the shrinkage pulls the estimates close to the across-item mean (for CLIP slightly below zero), so the magnitudes of the t=2 weak points should be quoted from the held-out prompts (Table 8). *Within the array alone.* Test (A) is significant at t=3 and t=4 for every model ($p \leq 0.014$), but at t=2 only for ViLT ($p = 0.001$) and PaliGemma2 ($p = 0.012$; CLIP 0.733, Qwen2.5-VL 0.054). The 69 prompts of the t=2 array identify the weak points, but certifying them requires the held-out prompts. *MinMax.* With the same procedure, the MinMax weak points replicate at t=3 and t=4 for every model, but at t=2 only for ViLT and PaliGemma2 (B2: $p = 0.024$ and 0.011; CLIP 0.151, Qwen2.5-VL 0.072).

Which items. Table 9 lists the weak points.

Sensitivity to the design. The primary analysis combines two choices: correcting for chance and weighting the ranking by precision. Table 10 reports tests (B2) and (C) for all 18 combinations of ranking, k, and metric. Uncorrected values replicate for at most one model in 17 of 18 cells, because they rank items largely by their noise floor. With chance correction, the weak points at t=3 and t=4 replicate for at least three models in all 24 cells of the table, and for all four models in every cell once the ranking is shrunken. At t=2, they replicate for all four models only for MPD: the top 10% by excess (excess or shrunken) and the shrunken top 20% by excess. With the shrunken ranking, the top 5 by excess replicates for three of four models; MinMax replicates for at most two.

Limitations. (i) The held-out prompts come from the same t=4 array. They are disjoint in prompts and images but share the generation pipeline. (ii) The primary condition was fixed after comparing the 18 conditions of Table 10. Each of its choices is motivated independently (above), but the weakest case (CLIP at t=2, B2 $p = 0.045$) should be regarded as borderline. (iii) At t=2 the discovery estimates carry little information about magnitude, and held-out magnitudes should be quoted instead. (iv) p-values are bounded below by $1/1{,}001$ (permutations), $1/100{,}001$ (random subsets), and $1/201$ (split-half).

Table 8: Validation of the axis-level weak points ($S \rightarrow A_{target}$ items, ranked by the shrunken chance-corrected excess, top 10% by excess). Discovery uses the t-wise array only. t=2 and t=3: held-out = the remaining prompts of the t=4 array; Sel. (All): mean held-out excess of the selected items (of all items); p: test (B2), selected items vs. 10^5 random subsets of items; Conf.: selected items confirmed individually on the held-out prompts (test B3, Holm). t=4: split-half; Sel. [range]: mean held-out excess of the selected items [2.5th, 97.5th percentile over 200 replicates]; p: test (C), for which 0.005 is the smallest attainable value; Ratio: held-out mean / discovery estimate. Excess: MPD minus its chance level, or the chance level minus MinMax; higher is less fair. All selected sets exceed chance on the held-out prompts (test B1 (t=2, 3), $p \leq 0.001$).

		t=2 (69 → 2,778, $k = 3/30$)			t=3 (469 → 2,378, $k = 6/60$)			t=4 (split-half, $k = 6/60$)		
	Model	Sel. (All)	p	Conf.	Sel. (All)	p	Conf.	Sel. [range]	p	Ratio
MPD	CLIP ViT-B/32	0.053 (0.024)	0.045	3/3	0.038 (0.016)	0.002	6/6	0.009 [0.000, 0.015]	0.005	1.08
	ViLT-B/32	0.092 (0.029)	0.006	3/3	0.070 (0.020)	<0.001	6/6	0.013 [0.010, 0.016]	0.005	1.17
	Qwen2.5-VL-7B	0.092 (0.028)	0.008	3/3	0.050 (0.018)	<0.001	6/6	0.018 [0.012, 0.024]	0.005	1.10
	PaliGemma2-3B	0.049 (0.017)	0.007	3/3	0.034 (0.011)	<0.001	6/6	0.013 [0.010, 0.017]	0.005	1.22
MinMax	CLIP ViT-B/32	0.051 (0.028)	0.151	3/3	0.056 (0.020)	0.002	5/6	0.049 [0.029, 0.069]	0.005	1.31
	ViLT-B/32	0.175 (0.047)	0.024	3/3	0.167 (0.034)	<0.001	6/6	0.033 [0.023, 0.041]	0.005	1.33
	Qwen2.5-VL-7B	0.083 (0.039)	0.072	3/3	0.073 (0.029)	<0.001	6/6	0.020 [0.001, 0.032]	0.005	0.86
	PaliGemma2-3B	0.080 (0.022)	0.011	3/3	0.051 (0.016)	<0.001	6/6	0.020 [0.011, 0.029]	0.005	1.29

Table 9: Weak points ($S \rightarrow A_{target}$, MPD) of each model and order, in decreasing order of the shrunken excess on the t-wise array.

Model	t=2	t=3	t=4
CLIP ViT-B/32	AGE → GEN	ETH×AGE → GEN	GEN×APP×DIS → AGE
	ETH → GEN	AGE×LOC → GEN	ETH×AGE×GEN → APP
	AGE → LOC	ETH×AGE → LOC	ETH×AGE×GEN → LOC
		AGE×GEN → LOC	GEN×LOC×DIS → AGE
		GEN×LOC → AGE	AGE×GEN×LOC → APP
		AGE×DIS → GEN	ETH×AGE×LOC → GEN
ViLT-B/32	AGE → GEN	ETH×AGE → GEN	ETH×AGE×APP → GEN
	GEN → AGE	AGE×DIS → GEN	ETH×AGE×DIS → GEN
	GEN → DIS	AGE×APP → GEN	ETH×AGE×GEN → APP
		ETH×GEN → AGE	ETH×AGE×LOC → GEN
		AGE×GEN → APP	AGE×APP×DIS → GEN
		AGE×LOC → GEN	ETH×AGE×GEN → DIS
Qwen2.5-VL-7B	AGE → GEN	AGE×GEN → LOC	ETH×AGE×GEN → APP
	AGE → LOC	ETH×GEN → AGE	ETH×AGE×GEN → LOC
	GEN → AGE	AGE×DIS → LOC	AGE×GEN×DIS → APP
		GEN×APP → AGE	AGE×GEN×DIS → LOC
		ETH×AGE → LOC	ETH×GEN×APP → AGE
		GEN×DIS → AGE	GEN×LOC×APP → AGE
PaliGemma2-3B	GEN → AGE	GEN×APP → AGE	ETH×GEN×APP → AGE
	APP → LOC	ETH×GEN → AGE	ETH×AGE×GEN → APP
	AGE → LOC	GEN×LOC → AGE	GEN×APP×DIS → AGE
		GEN×DIS → AGE	GEN×LOC×APP → AGE
		AGE×GEN → LOC	ETH×AGE×GEN → LOC
		AGE×DIS → LOC	AGE×GEN×DIS → APP

Table 10: Sensitivity of the replication to the design: number of models (of four) whose weak points replicate ($p < 0.05$ in test B2 at t=2 and t=3, and in test C at t=4; held-out values are always chance-corrected). Ranking: uncorrected value, excess over chance, or shrunken excess (EB). k: top 5, 10%, or 20% by excess of the 30/60/60 $S \rightarrow A_{target}$ items at t=2/3/4. Bold: primary analysis.

		t=2			t=3			t=4		
Metric	Ranking	5	10%	20%	5	10%	20%	5	10%	20%
MPD	value	0	1	1	0	0	1	0	0	0
	excess	3	4	3	3	3	4	3	3	3
	EB	3	**4**	4	4	**4**	4	4	**4**	4
MinMax	value	1	0	0	0	0	2	1	1	1
	excess	2	2	2	4	4	4	4	4	4
	EB	2	2	1	4	4	4	4	4	4

G FINE-TUNING DETAILS

Each run starts from `openai/clip-vit-base-patch32` and fine-tunes for 3 epochs in fp16. Only the vision encoder and the visual projection are updated; the text tower stays frozen. We evaluate the weights after epoch 3, with no early stopping or checkpoint selection. Seeds 0 through 19 each resample the training subset.

Objective. Each image is paired with the caption "a photo of {a/an} {age} {race} {gender}," with the article taken from the age word. Training uses the symmetric CLIP contrastive loss. Rows that share a caption count as positives, so identical captions in a batch are not treated as negatives.

Optimization. The learning rate is 1.6×10^{-5}, obtained by linearly scaling the 10^{-6} rate of Kavuri et al. (2025) from their batch size of 16 to ours of 256. It then decays linearly to zero, with no warmup. The batch size is 256 when the subset has at least 256 images. For $K \in \{1, 2, 4\}$ the whole subset is one batch, and the learning rate is not rescaled. Remaining optimizer settings follow the HuggingFace Trainer defaults.

Sampling. A budget $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ is $N = 60 \times K$ rows from the UTK-FF train split of Sec. 6.2 and Appendix C.2. Natural sampling draws unique images with counts proportional to the training-cell prior: every cell gets at least one, the remainder is allocated by the largest-remainder method, and there is no replacement. Balanced sampling gives every cell exactly K rows, repeating photographs when a cell has fewer than K so that multiplicities differ by at most one. Difficulty-weighted sampling uses the per-cell weights from Sec. 6.1 and is not capped by how many UTK-FF photographs a cell has; synthetic-only DW applies the same weights on the CIT pool. Synthetic balanced sampling likewise assigns K images per cell, spread across as many distinct CIT prompts as that cell's pool contains. Mixed sampling keeps $N = 60 \times K$. For a cell target T it takes $\mathrm{round}(pT)$ synthetic images and the rest real, with $p \in \{0.25, 0.50\}$; real and synthetic rows follow the balanced draws above. Jobs whose realized synthetic count is zero are skipped, which drops Mixed-Balanced 25% at $K \in \{1, 2\}$ and Mixed-Balanced 50% at $K{=}1$. Real-only and synthetic-only epochs are shuffled into batches of 256. Mixed epochs shuffle the two sources separately and fill each batch so the synthetic fraction stays at p.