Biases Propagate in Encoder-based Vision-Language Models: A Systematic Analysis From Intrinsic Measures to Zero-shot Retrieval Outcomes

Kshitish Ghate
Carnegie Mellon University
kghate@andrew.cmu.edu

Tessa Charlesworth
Northwestern University
tessa.charlesworth@kellogg.northwestern.edu

Mona Diab
Carnegie Mellon University
mdiab@andrew.cmu.edu

Aylin Caliskan
University of Washington
aylin@uw.edu

Abstract

To build fair AI systems we need to understand how social-group biases intrinsic to foundational encoder-based vision-language models (VLMs) manifest in biases in downstream tasks. In this study, we demonstrate that intrinsic biases in VLM representations systematically "carry over" or propagate into zero-shot retrieval tasks, revealing how deeply rooted biases shape a model's outputs. We introduce a controlled framework to measure this propagation by correlating (a) intrinsic measures of bias in the representational space with (b) extrinsic measures of bias in zero-shot text-to-image (TTI) and image-to-text (ITT) retrieval. Results show substantial correlations between intrinsic and extrinsic bias, with an average $\rho = 0.83 \pm 0.10$. This pattern is consistent across 114 analyses, both retrieval directions, six social groups, and three distinct VLMs. Notably, we find that larger/better-performing models exhibit greater bias propagation, a finding that raises concerns given the trend towards increasingly complex AI models. Our framework introduces baseline evaluation tasks to measure the propagation of group and valence signals. Investigations reveal that underrepresented groups experience less robust propagation, further skewing their model-related outcomes.

1 Introduction

Modern encoder-based vision-language models (VLMs) such as CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023) excel at mapping images and text into a shared representational space. These models are the backbone of many zero-shot applications, from object recognition to image retrieval (Rombach et al., 2022; Briggs and Laura, 2022; Taesiri et al., 2022; Bui et al., 2023; Barraco et al., 2022; Pirom, 2022). However, it is increasingly clear that these representations may contain intrinsic biases (Caliskan et al., 2017; Charlesworth and Banaji, 2022; Caliskan et al., 2022) for instance, systematically associating certain social groups with negative concepts. The potential for these biases to manifest in downstream applications can undermine fair and ethical treatment across social groups (Ghosh and Caliskan, 2023; Wolfe and Caliskan, 2022a). There exists a significant gap in understanding how biases intrinsic to the representational spaces of foundational encoder-based VLMs like CLIP (Radford et al., 2021) and BLIP2 (Li et al., 2023) propagate to tasks they perform effortlessly without retraining or fine-tuning. Concretely: the current research quantifies the propagation of biases in encoder-based VLMs by comparing (a) intrinsic bias in the representational space learned by these models to (b) the outcomes of downstream zero-shot tasks, specifically text-to-image (TTI) and image-to-text (ITT) retrieval.

In our study, the term "propagation" refers to the directional flow of biases from model representations to outcomes in downstream tasks. By employing a rigorously controlled experimental design, we minimize external confounding factors, enabling a clear observation of how biases are transferred within encoder-based VLMs. While we recognise this does not fully establish causality, our findings provide robust evidence of systemic bias transfer, justifying the use of the term "propagation."

To conduct our analysis, we curate balanced datasets comprising images from the Chicago Face Database (CFD) (Ma et al., 2015) and the Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017), capturing social group signals and valence signals, respectively. Valence is a primary dimension in our analysis, as it captures biased attitudes critical to understanding bias manifestation in AI (Toney and Caliskan, 2021; Osgood et al., 1957). In parallel, we develop experimental text datasets using semantically-neutral sentence templates (Tan and Celis, 2019), incorporating target words representing either valence signals from the NRC-VAD lexicon (Mohammad, 2018) or social group signals from prior research (Charlesworth

Findings of the Association for Computational Linguistics: ACL 2025, pages 18562–18580

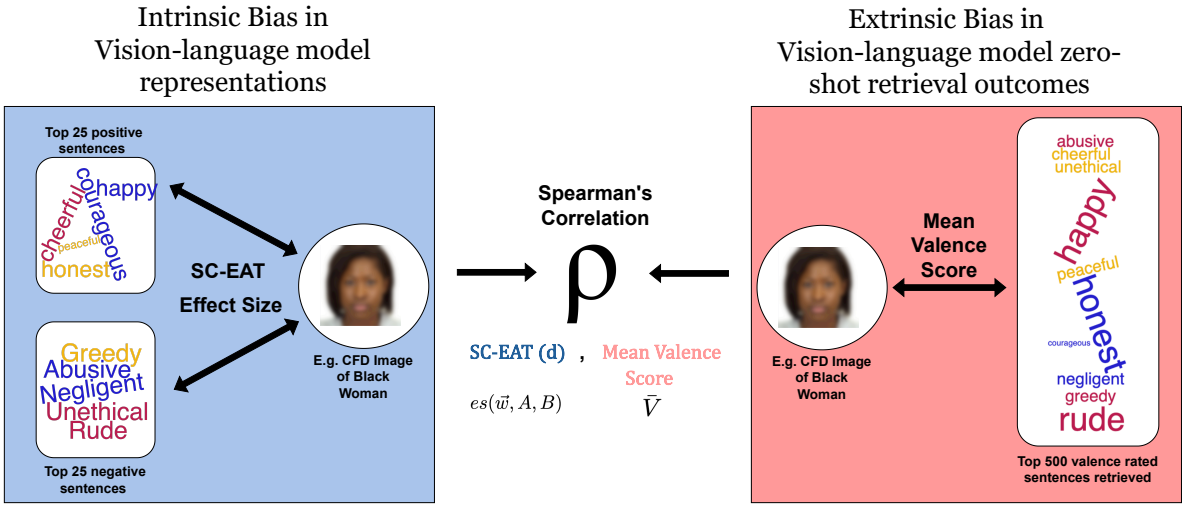

Figure 1: The setup to analyze valence-based bias propagation in Image-to-Text retrieval. Propagation is measured by the correlation between the intrinsic bias (measured by valence SC-EAT), and extrinsic bias (measured by mean valence of retrieved text) for images belonging to a social group. The intrinsic SC-EAT effect size is computed based on the differential association of each CFD image with the top 25 positive and top 25 negative words in sentence templates. The extrinsic metric is derived from the mean valence of the top 500 words in sentence templates retrieved by each image. Ground truth valence scores come from human rated sources (Mohammad, 2018). Spearman's ρ is used to compute the correlation between intrinsic and extrinsic metrics.

et al., 2022, 2024). We measure bias propagation by quantifying both intrinsic and extrinsic biases in encoder-based VLMs. Intrinsic bias is assessed using the single category-embedding association test (SC-EAT) (Caliskan et al., 2017; Caliskan, 2023), which evaluates associations between social groups and attribute association signals within the model's representational space. Extrinsic bias is evaluated through downstream zero-shot tasks, such as ITT and TTI retrieval, to assess how these intrinsic biases manifest in practical applications. Figure 1 demonstrates the setup to measure valence-based bias propagation in the ITT setting.

Our study uniquely examines the propagation of both social group and valence biases in encoder-based VLMs within controlled zero-shot settings, addressing the complexity of bias in multimodal contexts across multiple social groups. It may seem intuitive that intrinsic biases in a model's representations would naturally emerge in its downstream tasks, simply because they share the same embedding space. However, as we will demonstrate, this relationship is not trivial. Different social groups (e.g., "Black Women" vs. "White Men") can yield significantly divergent correlations. These systematic differences reflect how complex social cues such as "data marking" and group underrepresentation (Newton, 2023; Wolfe and Caliskan, 2022c) can affect the flow of biases. Accordingly, the present work also contributes empirical clarity to how specific group and valence signals shape the degree of bias propagation in encoder-based VLMs.

Our study makes the following contributions.[1]
(1) We provide a controlled experimental framework for systematically quantifying bias propagation. The framework reinforces previous work showing biases in the representational spaces of encoder-based VLMs. It also, for the first time, shows biases in the downstream zero-shot tasks of TTI and ITT retrieval as they directly relate to intrinsic biases. Moreover, the framework and methods shown here are unique in enabling the study of bias propagation at scale - showing consistent conclusions across 114 scenarios, incorporating 6 social groups, 3 models, and 8 experiments, we show correlations are consistently high and significant (Spearman's ρ of 0.83 ± 0.10).
(2) We introduce baseline intrinsic evaluation experiments that evaluate the direct propagation of valence and group signals in zero-shot retrieval tasks. These baselines indicate the extent of accurate retrieval and can serve as benchmarks against which to compare the results of the key experiments examining bias in valence-to-group and group-to-valence propagation. Results show a strong base-

[1]The code and data is available at https://github.com/kshitishghate/bias_prop

line correlation for valence propagation, with a Spearman's ρ of 0.86 ± 0.04, and for group information propagation, with a ρ of 0.85 ± 0.12.
(3) The controlled framework will support future empirical and theoretical understanding of AI bias dynamics. Specifically, the results highlight variation in bias propagation across social groups. Marginalised groups that correspond to being underrepresented in data experience less robust propagation and more skewed model-related outcomes. We also show that models have a tendency to show higher bias propagation as they scale in size and performance — a finding with critical implications during an era of rapid AI expansion.

2 Related Work

Bias in Vision-Language Models. Recent research on bias in VLMs moves beyond standalone language or vision models to investigate when these two modalities intersect. For instance, Srinivasan and Bisk (2022) extended text-based gender bias studies (e.g., (Su et al., 2019)) to the multi-modal setting. Using VL-BERT, they demonstrate that VLMS tend to reinforce and exaggerate gender biases and stereotypes. Wolfe and Caliskan (2022a) evaluated biases associating American with White, as seen in foundational cognitive science studies (Devos and Banaji, 2005) in CLIP using embedding association tests (Caliskan et al., 2017) with the CFD (Ma et al., 2015). They observed that pictures of White faces (versus Asian, Latina/o, and Black faces) had a higher association with collective ingroup words. Zhou et al. (2022) assessed multiple VLMs with VLStereoSet, and found pervasive gender and race stereotypes and biases, which are more complex in VLMs than in pre-trained language models. Wolfe et al. (2023) focused on how VLMs, trained on web-scraped data, often perpetuate the sexual objectification of women and girls.

A related strand of work by investigates the origins of bias in VLMs by examining the influence of pretraining factors. Berg et al. (2022) compare gender bias across 9 CLIP models and find that larger pretraining datasets often associated with better down-stream performance, also show the least bias. Ghate et al. (2025) extend this study across 131 CLIP models, more than 20 architectures and pretraining datasets and find that the choice of the pretraining dataset and data filtering strategy used to filter it is the most significant predictor of intrinsic bias in CLIP-based VLMs, over and above variables such as architecture choice, size of the pretraining dataset and number of model parameters. Hong et al. (2024) audit CLIP-based pretraining data filtering strategies and find that data related marginalized groups are subject to higher rates of exclusion compared to historically well-represented Western-centric demographics.

Bias Propagation in Pretrained models. Several studies have investigated how biases intrinsically present in pretrained language models can influence downstream model outcomes. Cao et al. (2022) highlighted a weak correlation between intrinsic and extrinsic fairness metrics for contextualized language models in supervised settings. Steed et al. (2022) investigated the bias transfer hypothesis, which posits that biases in large language models (LLMs) transfer into harmful task-specific behavior after fine-tuning. Their findings indicate that reducing intrinsic bias before fine-tuning does little to mitigate discriminatory outcomes. However, it is crucial to note that the downstream tasks examined in such past work are not zero-shot. Instead, these tasks involve fine-tuning on specific datasets which can lead the model to learn spurious associations inherent in those datasets. Feng et al. (2023) explored the propagation of political biases in language models. Their work revealed that pretrained language models possess political leanings, influenced by politics in training corpora, which affects outcomes in downstream tasks. Cabello et al. (2023) showed intrinsic biases do not consistently correlate with extrinsic biases in fine-tuned tasks of encoder-based VLMs. They proposed gender-neutral pretraining as a mitigation strategy and demonstrated its ability to reduce group disparities. However, their work also focused on finetuning outcomes and lacked a systematic framework to directly investigate bias propagation.

In contrast, we focus on zero-shot tasks, where models are directly examined for bias propagation from intrinsic to extrinsic levels without additional fine-tuning, and models directly rely on their pretrained representations. We introduce a controlled framework to investigate biases across multiple VLMs and social groups. This enables us to uncover previously overlooked systematic trends in bias propagation.

3 Data

We employ carefully curated datasets that provide experimental control and ground truth values.

Table 1: Summary Dataset Statistics and Usage in Intrinsic and Extrinsic Measures. Further details in Appendix A.

Dataset	Size	Intrinsic Measure	Extrinsic Measure
CFD	597 images	Group-based images SC-EAT	Group representation in retrieved images
OASIS	900 images	Valenced images in SC-EAT	Valence ratings in retrieved images
NRC-VAD Lexicon	20,000 words	Valenced Text SC-EAT	Valence ratings of retrieved text
Group Labels	864 phrases	Group-based text in SC-EAT	Group representation in retrieved text

Chicago Face Database (CFD) (Ma et al., 2015): The CFD dataset consists of 597 self-identified images of men and women participants belonging to 4 different racial demographics, namely, Black, White, Latino/a and Asian. This dataset was selected for its reliable, self-identified human face images with clear group signals, unlike FairFace (Karkkainen and Joo, 2021), which lacks self-identified labels and includes visually noisy images, potentially biasing annotations. We first expand the dataset to facilitate the scale and generalization of our experiments by creating within-group image morphs following the approach in (Wolfe et al., 2022) as described in Appendix A.1.1. We then randomly sampled images without replacement.

Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017) : Provides 900 human valence-rated images, enabling the study of how VLMs handle non-group-related valence signals from naturalistic image inputs. The images depict a broad spectrum of themes such as animals, humans, objects, and scenes.

Textual Templates and Lexica: Our research requires controlled text datasets. Semantically neutral sentence templates enable understanding how the valence and social group properties of target lexica embedded in them influence biases in VLMs. As such, we employ lexica embedded in 6 such templates taken from (May et al., 2019) for controlled analysis of how words/phrases, particularly those with valence or social group signals, influence biases in extrinsic VLM outputs. For experiments using valence-rated sentences, we employ the NRC-VAD lexicon (Mohammad, 2018), comprising approximately 20,000 English words. An example sentence created from the valence lexicon "happy" is, "This is the word happy." For experiments with group signals, we employ a curated list of 864 group-label phrases derived from Charlesworth et al. (2022). An example sentence created from the group-label "Black Woman" is "This is the word Black Woman." We focus on 6 intersectional race and gender groups: "Asian Men," "Asian Women," "Black Men," "Black Women," "White Men," and "White Women" in our analyses. Summary statistics of CFD, OASIS, NRC-VAD lexicon, and group-labels are present in Table 1. For their description and usage details in our experiments, please refer to the Appendix A.

4 Approach

Social group biases in this study are defined as the association between group signals (e.g., the representation of Black women) and valence signals (e.g., the representation of positivity vs. negativity). This approach is motivated by established methods in the field of AI ethics, such as the Embedding Association Test (EAT) applied to embedding spaces (Caliskan et al., 2017), providing a scientifically robust framework for studying biases. Valence is also integral to our analysis as it reflects the positive or negative connotations associated with group representations, and is the basic dimension along which biases are evaluated (Eagly and Chaiken, 1998).

4.1 Intrinsic and Extrinsic Bias Measurement

Intrinsic bias is quantified using the Single Category-Embedding Association Test (SC-EAT) (Caliskan et al., 2017) effect sizes (Cohen, 2013) which allows for a principled assessment of biases within model representations. SC-EAT quantitatively evaluates the association between a single target stimulus (e.g., an image or word representing a social group) and sets of attribute stimuli (e.g., words or images with valence or group content). The association is measured by the mean cosine similarity between the embeddings of the target and attribute stimuli, normalized by the standard deviation of these similarities across all stimuli. This is quantified as the SC-EAT Cohen's d effect size given by:

$$es(\vec{w}, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

$\vec{w}$ represents the embedding of the target stimulus, which in our case can represent either a sentence or image of interest, and A and B are sets

of attributes determined either by valence or group content.

Extrinsic bias in ITT and TTI is evaluated through two measures derived from work quantifying bias in task outcomes related to valence and content (Charlesworth et al., 2024; Kong et al., 2024) :

1. **Mean of Valenced Text/Images Retrieved:** This measure calculates the average valence of retrieved text/image by the model for a given prompt:

$$\bar{V} = \frac{1}{N} \sum_{i=1}^{N} V_i$$

Where $\bar{V}$ is the average valence, V_i is the valence of the i-th retrieved item, and N is the total number of retrieved items. The measure quantifies the degree to which the model's outputs, in terms of textual descriptions or image retrieval, skew towards positive or negative valence. This measure helps understand how the model's internal biases translate into biases in the valence content it retrieves.

2. **Proportion of Text/Images Belonging to a Social Group Retrieved:** This measure evaluates the likelihood or proportion of the model retrieving text or images associated with specific social groups in response to given prompts, calculated as:

$$P_{grp} = \frac{\text{Number of group-related items retrieved}}{\text{Total number of retrieved items}}$$

where P_{grp} is the proportion expected to remain constant across groups for a prompt. An unequal P_{grp} when the prompt is expected to be associated with all groups equally shows model bias.

4.2 Framework to Measure Bias Propagation

To evaluate how intrinsic biases in VLMs propagate to downstream retrieval tasks, we propose a unified framework (Algorithm 1) where the following variables are defined:

Data (D): The input can be any of our curated datasets (CFD/OASIS images or text templates using NRC-VAD/group labels).

Retrieval Direction (R): Specifies the zero-shot retrieval task, either ITT or TTI.

Content Type (C): Declares whether the current bias analysis targets valence (positivity/negativity) or group (e.g., race and gender).

Correlation (ρ): The final Spearman's correlation computed over intrinsic and extrinsic metrics.

We now provide a step-by-step overview the functions that constitute our framework in 1 followed by an example.

1. CalculateIntrinsicAssociations(D, C): This function computes the intrinsic association metric based on the content type C and data D. If C is "valence," the metric is calculated using the SC-EAT on valence signals (e.g., from the OASIS dataset for images or the NRC-VAD lexicon for text). If C is "group," the SC-EAT is applied to group signals (e.g., from the CFD for images or predefined group labels for text).

2. RetrieveTextFromImage(D): Retrieves top k sentences associated with input D using a VLM.

3. RetrieveImagesFromText(D): Retrieves top k images associated with input D using a VLM.

4. CalculateExtrinsicOutput($retrieved\_items$, C): This function calculates the extrinsic bias metric based on the retrieved items and the content type C. If C is "valence," the metric is the mean valence of the retrieved items. If C is "group," the metric is the proportion of retrieved group items.

5. SpearmanCorrelation($intrinsic\_metric$, $extrinsic\_metric$): This function calculates the Spearman's correlation ρ between the intrinsic and extrinsic metrics, reflecting the degree of bias propagation. We choose to employ Spearman's ρ due to its non-parametric nature and ability to capture nonlinear relationships (Spearman, 1961).

Mini-Example: Suppose we want to see whether valence bias toward "Black Women" is reflected in ITT retrieval (depicted in Figure 1). Here, input D: 6000 images of (1000 per social group) from CFD; 20,000 words in sentence templates from NRC-VAD. C: "Valence"; R: ITT. Concretely:

1. Intrinsic Association: We apply SC-EAT to 1000 images of Black Women from CFD. Each image's embedding is compared with the top-25 positive vs. negative words in sentence templates (from the NRC-VAD). The result is an intrinsic valence effect size per image.

2. ITT Retrieval: We then feed each of those 1000 images into our VLM and retrieve the top-500 text items (e.g., "This is the word happy," "This is the word sad," etc.).

3. Extrinsic Metric: For each image, we measure the average valence of the retrieved text (are the words mostly positive or negative?).

4. Correlation: We measure the correlation between the valence effect sizes and the mean retrieved valence over the 1000 images. If images with high negative SC-EAT scores also retrieve mostly negative text descriptors, it suggests strong bias propagation (likely a high ρ).

Algorithm 1 Unified Bias Propagation Framework

Input: Data D, Retrieval Direction R, Content Type C
Output: Correlation ρ between intrinsic and extrinsic metrics
$intrinsic\_metric \leftarrow$ **CalculateIntrinsicAssociations**(D, C)
if $R ==$ image_to_text **then**
  $retrieved\_items \leftarrow$ **RetrieveTextFromImage**(D)
else if $R ==$ text_to_image **then**
  $retrieved\_items \leftarrow$ **RetrieveImagesFromText**(D)
end if
$extrinsic\_metric \leftarrow$ **CalculateExtrinsicOutput**($retrieved\_items$, C)
$\rho \leftarrow$ **SpearmanCorrelation**($intrinsic\_metric$, $extrinsic\_metric$)
return ρ

4.3 Measuring Bias Propagation Via Valence and Group Signals

The core of our experimental design to analyze bias propagation consists of 8 experiments that all follow the framework in Algorithm 1, and are evenly divided to differ in: (1) direction (i.e., 4 experiments test ITT association and bias propagation, while 4 test TTI association and bias propagation); and (2) valence versus group content (i.e., 4 experiments test the propagation of valence, while 4 test the propagation of group signals). The application of Algorithm 1 is explained in the following propagation experiments. For all experiments, the correlation is computed over intrinsic and extrinsic metric scores for target group of images or text per model for ITT and TTI respectively. Exact details can be found in Appendix Figure 8.

Baseline Valence-Valence Signal Propagation (1*-a and 1*-b): This is a baseline experiment to measure how intrinsic valence signals propagate to valence outcomes. The framework parameters are – D: OASIS images and NRC-VAD lexicon sentences; C: "valence." We label the R: ITT direction as (1*-a) and the R: TTI direction as (1*-b). Intrinsic metrics: SC-EAT associations for top valenced images/text. Extrinsic metrics: Mean valence ratings of top-500 retrieved images/text. For all experiments, we choose a retrieval k of 500 items as it provides a significant sample size that is diverse while being computationally feasible.

Baseline Group-Group Signal Propagation (2*-a and 2*-b): This is a baseline experiment to measure how intrinsic group signals propagate to group outcomes using CFD images and predefined group labels. The framework parameters are – D: CFD images and group label sentences; C: "group." We label the R: ITT direction as (2*-a) and the R: TTI direction as (2*-b). Intrinsic metrics: SC-EAT group associations. Extrinsic metrics: Proportion of correctly identified group representations in the top-500 retrieved images/text.

Valence-to-Group Signal Propagation (1-a and 1-b): Assesses how valence signals influence group-related retrieval outcomes. The framework parameters are – We label the R: ITT direction as (1-a) with the inputs D: CFD images and NRC-VAD lexicon sentences; C: "valence." The R: TTI direction is labeled as (1-b) with the inputs D: group label sentences and OASIS images; C: "valence." Intrinsic metrics: SC-EAT valence effect sizes for group-associated images/text. Extrinsic metrics: Mean valence or proportion of group-identified content in the top-500 retrieved images/text. Figure 1 outlines the ITT version of these experiments visually.

Group-to-Valence Signal Propagation (2-a and 2-b): Investigates the influence of group signals on valence-related retrieval outcomes. The framework parameters are – We label the R: ITT direction as (2-a) with the inputs D: OASIS images and group label sentences; C: "group." The R: TTI direction is labelled as (2-b) with the inputs D: NRC-VAD lexicon sentences and CFD images; C: "group." Intrinsic metrics: SC-EAT group associations for valenced images/text. Extrinsic metrics: Mean valence or proportion of group-related content in the top-500 retrieved images/text.

5 Experiments and Results

5.1 Evaluation of Intrinsic Biases in VLMs

To lay the groundwork for our study, we initially assess intrinsic biases present in the representational spaces of encoder-based VLMs using valence and group-based SC-EATs. Details of the experimental setup for measuring these biases in

experiments 1-a, 1-b (see Figure 1 for a visualization of the setup for ITT), 2-a, and 2-b will be elaborated later in this section, ensuring a clear linkage between initial findings and their implications in broader model behaviors. We choose to evaluate encoder-based VLMs, specifically, OpenAI versions of CLIP-B-32, CLIP-L-14, and Salesforce's BLIP-2 (the encoder-decoder versions) detail the choice of models used in our experiments due to their foundational role and extensive usage in the community (over 50 million downloads on Huggingface [2]), and availability of principled methods Steed and Caliskan (2021) to measure biases within their representations. Further model details are elaborated in the Appendix B.

Our analysis highlights significant intrinsic biases within VLMs. Specifically, the aggregate valence-based SC-EAT effect sizes for different social groups obtained in experiments 1-a and 1-b reveal that "Black Women," followed by "Black Men" are associated with the most negative valence (with -1.43 and -1.31 z-scored effect sizes (d) respectively). "White Men" and "White Women" are similarly placed with 0.44 and 0.45 z-scored effect sizes respectively, while "Asian Women" are most associated with positive valence with 1.21 z-scored effect size, indicating a strong valence-based representation bias. 2-a and 2-b demonstrate an overrepresentation of "White Men" and "White Women" group associations to valenced text and images (with 0.48 and 0.46 z-scored effect sizes respectively) as opposed to "Black Women" (with -0.58 z-scored effect size). The detailed results, including figures highlighting these findings are presented in the Appendix D.1.

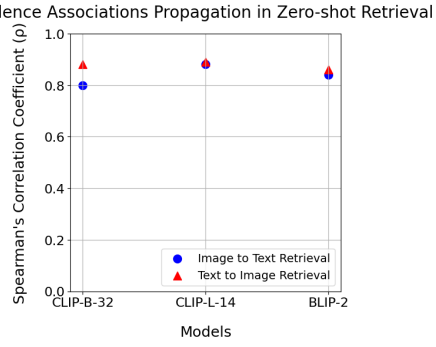

Figure 2: Exp. 1*-a and 1*-b illustrate Spearman's ρ measuring the correlation of intrinsic valence associations with extrinsic valence outcomes in zero-shot retrieval tasks.

[2]https://huggingface.co/openai

5.2 Bias Propagation Via Valence and Group Signals

We now list the key findings from our propagation experiments of Section 4.3. Appendix D has further findings and reproducibility details.

Baseline Valence-Valence Signal Propagation (1*-a and 1*-b): In 1*-a and 1*-b (Figure 2), the models show a strong aggregate Spearman's correlation of 0.84 $\pm$ 0.04 ρ in 1*-a, 0.87 $\pm$ 0.02 ρ in 1*-b. Furthermore, we observed larger models like CLIP-L-14 consistently showing higher correlations (up to $\rho = 0.88$) compared to smaller models like CLIP-B-32 ($\rho = 0.80$).

Baseline Group-Group Signal Propagation (2*-a and 2*-b): 2*-a and 2*-b demonstrate that VLMs can robustly propagate social group signals, as indicated by the strong aggregate Spearman's correlations (0.94 $\pm$ 0.04 ρ in 2*-a, 0.76 $\pm$ 0.11 ρ in 2*-b) observed across all models (Figure 3).

Valence-to-Group Signal Propagation (1-a and 1-b): 1-a and 1-b results in Figure 4 demonstrate significant positive aggregate Spearman's correlations across models and social groups with 0.81 $\pm$ 0.10 ρ in 1-a, and 0.78 $\pm$ 0.10 ρ in 1-b.

Group-to-Valence Signal Propagation (2-a and 2-b): 2-a and 2-b results in Figure 5 demonstrate significant positive aggregate Spearman's correlations across models and social groups with 0.91 $\pm$ 0.02 ρ in 2-a, and 0.78 $\pm$ 0.07 ρ in 2-b.

6 Discussion

Our study first quantifies the intrinsic biases in encoder-based VLMs. Next, across 8 experiments and 114[3] analyses, the current work finds consistently across all approaches, positive and generally strong correlations (Spearman's ρ of 0.83 $\pm$ 0.10) between intrinsic biases in VLM representational spaces and extrinsic bias outputs in zero-shot ITT and TTI retrieval tasks, with notable systematic differences relating to social group identity.

Evaluation of Intrinsic Biases in VLMs. Our experiment on initially quantifying intrinsic bias demonstrates the intersectional gender hypothesis (Ghavami and Peplau, 2013) where biases between men and women are most similar in the case of biases between "White Men" and "White Women." Additionally, aggregate group SC-EAT effect sizes

[3]We clarify that we have 114 analyses due to no group-based distinctions in our valence-valence propagation baseline (2-scenarios * 3-models + 6-scenarios * 3-models * 6-social groups).

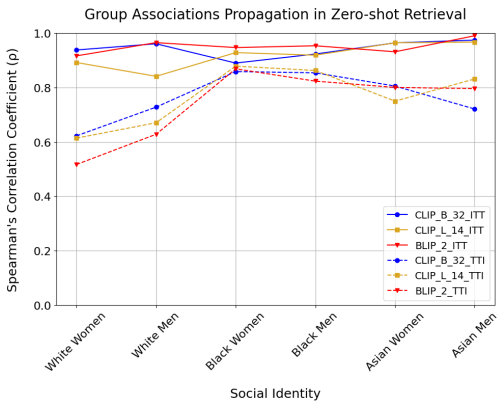

Figure 3: Exp. 2*-a and 2*-b illustrate Spearman's ρ measuring the correlation of intrinsic group signal propagation to extrinsic group outcomes in zero-shot retrieval tasks, stratified by social groups.

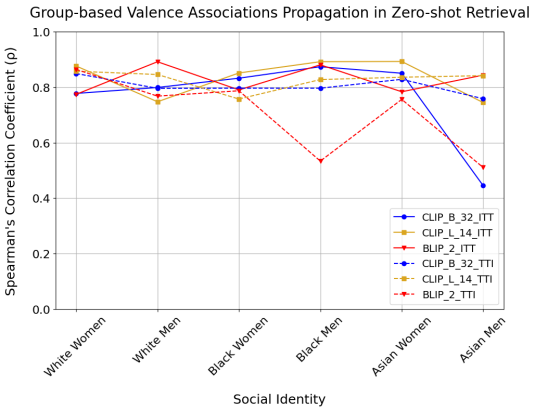

Figure 4: Exp. 1-a and 1-b illustrate Spearman's ρ measuring the correlation of intrinsic valence-based bias propagation to extrinsic valence outcomes in zero-shot retrieval tasks, stratified by social groups.

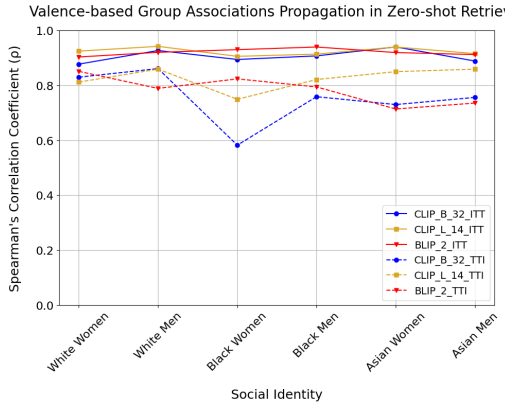

Figure 5: Exp. 2-a and 2-b illustrate Spearman's ρ measuring the correlation of intrinsic valence-based group bias propagation to extrinsic group outcomes in zero-shot retrieval tasks, stratified by social groups.

across experiments 2-a and 2-b indicate "White Men" and "White Women" have the highest levels of group associations to valenced content in VLMs, meaning they are overrepresented in model outcomes. In contrast, "Black Women" demonstrate minimal intrinsic group associations in 2-a and 2-b while being associated with the most negatively valenced content in 1-a and 1-b. This is consistent with research on group identity bias (Devos and Banaji, 2005; Wolfe and Caliskan, 2022a)

Baseline Propagation of Valence and Group Signals. Experiments 1*-a and 1*-b show that VLMs robustly encode and propagate fundamental valence signals indicating that valence is a significant signal in vision and language domains (Wolfe and Caliskan, 2022d; Toney and Caliskan, 2021). However, VLMs show distinct biases in how they propagate different social group representations, for instance, the "Black Women" group showed a particularly consistently high correlation (e.g. $\rho = 0.86$ in 2*-b for BLIP-2), while the "White Women" group showed a lower correlation (e.g. $\rho = 0.53$ in 2*-b for BLIP-2). This result could be attributed to the "marking" of certain identities in datasets, especially to non-default (i.e., non-White) groups. Marking and "othering" can lead to more distinct overfitted portrayals that increase propagation and risks entrenching stereotypes (Newton, 2023; Wolfe and Caliskan, 2022c).

To support this result, we found that the bigram "Black Women" occurs more frequently in English language compared to "White Women" according to Google Ngram (`https://books.google.com/ngrams/`) statistics from the last century. The lower correlation for "White Women" presents a complex issue. The default category for a person in AI is White (Wolfe and Caliskan, 2022a). Accordingly, the default for Woman is a White Woman (Wolfe and Caliskan, 2022c). Consequently, references to "White Women" typically just include the word women. As a result, our analysis using the input "White Women" likely captures a weaker representation compared to "Black Women." as they are not distinctly "different" enough to be as marked as groups like "Black Women," yet they are not represented as the "default" group like "White Men," causing the models to learn underfitted representations of them. Following existing audits of VLM-based data filtering methods (Hong et al., 2024), we leave the analysis of disentangling further societal defaults in AI bias analysis to future work.

We also note that propagation in all cases is not 100% due to variable template content and confounding signals in the text and images, and this is seen more so in TTI as opposed to ITT. The generally high correlations we observe suggest that when models operate purely on their learned representations, particularly without the intervening layer of task-specific fine-tuning that can introduce new associations, the biases embedded in the representational space directly manifest in outputs.

Multidimensional Nature of Bias in Propagation through Intersection of Valence and Group Signals. VLMs effectively process simple valence signals, their handling of complex social group signals is inconsistent, which would imply skewed representations and degrading performance. For instance, in 1-b, "White Women" showed strong correlations in all models, with the highest being in BLIP-2 at 0.86 ρ. In contrast, "Asian Men" and "Black Men" exhibited lower correlations in some models, particularly in BLIP-2 (0.51 ρ and 0.53 ρ, respectively). These low correlations were weaker than "White Men" and also "Asian Women" or "Black Women," respectively, suggesting the intersectional combination of men with a marginalized racial identity may reduce Valence-to-Group propagation in TTI. This may be attributed to the underrepresentation of "Asian Men" and "Black Men" in the selected models. Google Ngram statistics suggest that mentions of these terms are far less common than groups such as "White Men."

Figures 4 and 5 highlight the subtle yet important differences in how biases observed in the initial intrinsic bias assessment propagate to zero-shot retrieval tasks. Notably, "Black Women," who exhibit the least magnitude in group associations to valenced content in 2-a and 2-b show significantly varied outcomes in Figure 5's Group-to-Valence association propagation. This variability can be linked to the inconsistent representations of the complex intersectional interactions of group identity and valence in VLMs. For instance, lower correlations may signal a model's inability to consistently represent content associated with certain social groups and impact downstream performance.

Scaling of Models and Bias Propagation. We found that model size and complexity play a role in bias propagation. As seen in 1*-a and 1-a, larger and better-performing models like CLIP-L-14, which regularly outperforms CLIP-B-32 on numerous zero-shot benchmarks (Radford et al., 2021), demonstrated stronger correlations in bias propagation. Similarly, in the case of "White Women," the CLIP-B-32 model showed a Spearman's correlation of 0.78 ρ in 1-a, while CLIP-L-14 model demonstrated a higher correlation of 0.88 ρ suggesting that as models scale, the degree of valence and group bias propagation increases. While this enhanced propagation capability can be advantageous in terms of the "accuracy" of the models for simple Valence-to-Valence or Group-to-Group tasks, it also implies a heightened risk of replicating and amplifying biases in more complex scenarios such as in 1-a. This finding of strong bias propagation in larger models in previous work that demonstrate language models that perform better on standard benchmarks have a greater risk of toxic generations (Longpre et al., 2024), and encoder-based CLIP models exhibit greater intrinsic bias as they scale (Ghate et al., 2025). Targeted bias mitigation strategies are thus essential for larger models through training and reinforcement learning-based approaches where fair intrinsic group representation objectives are prioritized. Future work can also look to develop flexible learning algorithms that will inhibit and steer models to equitable outcomes.

7 Conclusion

This research investigated intrinsic biases propagation in ITT and TTI retrieval of encoder-based VLMs. Across 114 analyses focusing on six social groups, we demonstrated that bias propagate with high and systematically varying correlations (Spearman's ρ of 0.83 $\pm$ 0.10 on average). Larger and more complex models showed greater propagation of biases, a finding that is particularly salient given the deployment of increasingly larger AI models. Our results imply detrimental downstream performance implications for marginalised groups that are misrepresented in these models.

8 Limitations

The image datasets used, notably the Chicago Face Database (CFD) and OASIS, provide a substantial foundation for analyzing biases as they provide self identification information and human-rated scores. However, they are far from representing the full complexity of human identities and valenced stimuli. In particular, the reliance on these datasets created in the United States could inadvertently reinforce a Western-centric perspective, and fail to account for cultural differences in how facial images and valenced stimuli are experienced.

In this study, we focus on 3 well-known encoder-based VLMs; CLIP-B-32 and CLIP-L-14 to capture a realistic scaling effect (150M vs. 430M parameters), plus BLIP-2 for architectural diversity. Future work that evaluates more encoder-based VLMs and even newly emerging models is straightforward with our controlled framework, although outside the current scope of the study's computational limits. Furthermore, our focus on English-only VLMs is an added limitation, as the interpretation of social groups and valence can vary across different languages and cultures (Charlesworth et al., 2023). Such an English-centric perspective could limit understanding of the global impact of AI biases, highlighting a need for multi-lingual and culturally diverse research in future studies. Furthermore, our

The use of 'propagation' in our study denotes the directional flow of biases from model representations to outcomes in downstream tasks. This terminology is supported by the rigorously controlled experimental design that minimizes external confounding factors, allowing a clear observation of how biases are transferred within VLMs. Our comprehensive analysis across 114 scenarios reveals consistent correlations between intrinsic and extrinsic biases, substantiating the directional influence implied by propagation. Pinpointing the exact origin of these biases (e.g., training data vs. architecture) is distinct from measuring if and how biases appear in outputs. Our focus is on the latter which is to systematically demonstrate that biases measured intrinsically do not stay hidden in the intrinsic representations, but are in fact manifested in real zero-shot tasks. Future research may further delineate the precise causal mechanisms along the lines of Ghate et al. (2025), for instance, the nature of text to image correspondence, pretraining data composition, model objective, and architectural constraints, which are outside the scope of our current study. Using our study's findings to develop methods to debias and steer model representations to be fairer and validated in further downstream tasks such as text-to-image generations presents another important avenue for exploration.

Our methodology to measure intrinsic associations focuses on SC-EAT because it is a principled and validated embedding-association technique which isolates how one target category associates with two sets of attributes. The method is grounded in cognitive science and social psychology literature (Greenwald et al., 1998) while other existing methods are ad hoc (Blodgett et al., 2020) and require further grounding. Additionally, zero-shot retrieval is a primary use-case for these models in practice, which we choose to evaluate using our chosen extrinsic metric. In future work, we hope to see the exploration of ML tasks and downstream settings that may correspond to other bias and fairness metrics (for instance, metrics that exist for fairness in binary classification). However, our choice of extrinsic bias metric here is most aligned with measuring representational harms from zero-shot retrieval tasks.

Finally, our approach to intersectionality, while an attempt to address complex social identities, still only captures a selection of 6 intersectional identities. There are hundreds of potential intersecting identities, including identities related to health, occupation, class, and more (Nicolas and Fiske, 2023). Future studies could explore more intersections beyond gender and race. It is important to note, however, that our methods to quantify bias propagation are generalizable and can be easily adapted to analyze any group associations.

9 Ethical Considerations

As the use of VLMs becomes more widespread in various industries, including object tracking, robot training, advertising, and marketing, (Briggs and Laura, 2022; Barraco et al., 2022; Bui et al., 2023; Taesiri et al., 2022) it is important to consider the ethical implications of these models. Our study highlights the intrinsic biases of current models in equitably representing different social identities. Such disparities in accuracy could have adverse effects on equity for how certain groups can use and engage with AI systems.

Moreover, there is a major ethical concern that these models may perpetuate negativity against certain groups, such as Black and Asian individuals, and perhaps especially intersectional groups (e.g., Black and Asian women). Given that the intrinsic biases embedded in the representational spaces of VLMs also lead to differential extrinsic outputs, using VLMs for tasks involving these groups content result in further marginalization and discrimination against these groups. Indeed, when AI users get outputs that reinforce (rather than challenge) their biases, they may come to view those biases as more acceptable and normative (Vlasceanu and Amodio, 2022).

In conclusion, our study highlights that intrinsic

representations related to social group identity and valence are biased in pretrained VLMs, and that they propagate such biases to downstream zero-shot applications. It is crucial to develop and use models that accurately represent all social identities in a fair and unbiased manner to avoid perpetuating stereotypes and biases against certain groups and to prevent further discrimination and marginalization. An important ethical consideration in this context is the balance between accuracy and fairness. It may be necessary to deliberately interrupt the propagation of biases from intrinsic representational spaces to achieve fairness, even if this intervention compromises some aspects of the model's performance. This trade-off reflects a crucial ethical stance where the value of fairness and inclusivity is prioritized over optimizing for accuracy alone. Such a perspective is essential in developing AI technologies that are not only advanced in their capabilities but are also aligned with broader societal values and ethical principles.

10 Acknowledgments

This work was supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB23D194. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST.

References

Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4662–4670.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

James Briggs and Laura Laura. 2022. Zero shot object detection with openai's clip.

Huy-Giap Bui, Minh-Huy Trinh, Canh-Toan Le, Quoc-Lam Vu, and Khac-Trieu Vo. 2023. Zero-shot video retrieval using clip with temporally ordered multi-query scoring. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, pages 938–944.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore. Association for Computational Linguistics.

Aylin Caliskan. 2023. Artificial intelligence, bias, and ethics. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

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

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Tessa ES Charlesworth and Mahzarin R Banaji. 2022. Word embeddings reveal social group attitudes and stereotypes in large language corpora. *Handbook of language analysis in psychology*, pages 494–508.

Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

Tessa ES Charlesworth, Kshitish Ghate, Aylin Caliskan, and Mahzarin R Banaji. 2024. Extracting intersectional stereotypes from embeddings: Developing and validating the flexible intersectional stereotype extraction procedure. *PNAS nexus*, 3(3):pgae089.

Tessa ES Charlesworth, Mayan Navon, Yoav Rabinovich, Nicole Lofaro, and Benedek Kurdi. 2023. The project implicit international dataset: Measuring

implicit and explicit social group attitudes and stereotypes across 34 countries (2009–2019). *Behavior Research Methods*, 55(3):1413–1440.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Thierry Devos and Mahzarin R Banaji. 2005. American= white? *Journal of personality and social psychology*, 88(3):447.

A Eagly and Shelly Chaiken. 1998. Attitude structure. *Handbook of social psychology*, 1:269–322.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Justin D García. 2013. "you don't look mexican!" my life in ethnic ambiguity and what it says about the construction of race in america. *Multicultural Perspectives*, 15(4):234–238.

Kshitish Ghate, Isaac Slaughter, Kyra Wilson, Mona T. Diab, and Aylin Caliskan. 2025. Intrinsic bias is predicted by pretraining data and correlates with downstream performance in vision-language encoders. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2899–2915, Albuquerque, New Mexico. Association for Computational Linguistics.

Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1):113–127.

Sourojit Ghosh and Aylin Caliskan. 2023. 'person' == light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6971–6985, Singapore. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Arnold K Ho, Jim Sidanius, Daniel T Levin, and Mahzarin R Banaji. 2011. Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. *Journal of personality and social psychology*, 100(3):492.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17.

Kimmo Karkkainen and Jungseock Joo. 2021. 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*, pages 1548–1558.

Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. 2024. Mitigating test-time bias for fair image retrieval. *Advances in Neural Information Processing Systems*, 36.

Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. 2017. Introducing the open affective standardized image set (oasis). *Behavior research methods*, 49:457–470.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47:1122–1135.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Veronica A Newton. 2023. Hypervisibility and invisibility: Black women's experiences with gendered racial microaggressions on a white campus. *Sociology of Race and Ethnicity*, 9(2):164–178.

Gandalf Nicolas and Susan T Fiske. 2023. Valence biases and emergence in the stereotype content of intersecting social categories. *Journal of Experimental Psychology: General*.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.

Wasin Pirom. 2022. Object detection and position using clip with thai voice command for thai visually impaired. In *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 391–394. IEEE.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Charles Spearman. 1961. The proof and measurement of association between two things.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 701–713.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Mohammad Reza Taesiri, Finlay Macklon, and Cor-Paul Bezemer. 2022. Clip meets gamephysics: Towards bug identification in gameplay videos using zero-shot transfer learning. In *Proceedings of the 19th International Conference on Mining Software Repositories*, pages 270–281.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.

Autumn Toney and Aylin Caliskan. 2021. ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7203–7218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Madalina Vlasceanu and David M Amodio. 2022. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, 119(29):e2204529119.

Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. 2022. Evidence for hypodescent in visual semantic ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1293–1304.

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

Robert Wolfe and Aylin Caliskan. 2022b. Contrastive visual semantic pretraining magnifies the semantics of natural language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3050–3061.

Robert Wolfe and Aylin Caliskan. 2022c. Markedness in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1269–1279.

Robert Wolfe and Aylin Caliskan. 2022d. Vast: The valence-assessing semantics test for contextualizing language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11477–11485.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1174–1185.

Kankan Zhou, Yibin LAI, and Jing Jiang. 2022. Vl-stereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics.

A Data

We utilize two types of image datasets: one representing social group identity in terms of race and gender, and the other capturing valence (positivity/negativity) as rated by human annotators. We employ valence ratings as a metric to analyze bias due to its nature of generating strong and well understood signals in both image and text (Toney and

Caliskan, 2021; Wolfe and Caliskan, 2022b). Additionally, we employ text datasets to complement the image datasets. The sections below detail the specific characteristics of these datasets, the rationale behind their selection, and how they are used following from Section 3 of the Main text.

A.1 Chicago Face Database (CFD).

We chose the Chicago Face Database (CFD) (Ma et al., 2015) for its reliable, controlled source of self-identified human face images, providing clear group signals—unlike FairFace (Karkkainen and Joo, 2021), which contains visually noisy images and lacks self-identified labels, leading to potential bias in annotations. These images are provided with consent of participants to be shared according to agreed upon terms. All personally identifiable information is anonymised (Ma et al., 2015) before the dataset is released.

The CFD dataset consists of 597 self-identified images of men and women participants belonging to 4 different racial demographics, namely, Black, White, Latino/a and Asian. [4]. The photos in the CFD are high-resolution and standardized so that the face always occupies the same portion of the image and they are all placed against a White background. Subjects were photographed with various facial expressions (happy, angry, and terrified) and with a neutral expression while facing directly to the camera. To maintain consistency with Ho et al. (2011) and Wolfe et al. (2022), we use only images with neutral facial expressions. In order to obtain a dataset of significant sample size that, we first expand the dataset by creating within-group image morphs (Wolfe et al., 2022) as described below. The images are then randomly sampled without replacement.

A.1.1 Generation of morphed images

To begin, we first created all possible pairs of images within each demographic group in the CFD dataset. For example, if we consider the Black men demographic, there were approximately 100 images available. We created 50% morphs of each possible pair of images, which resulted in around 5,000 morphs for each demographic group.

To generate realistic morphed images, we adopt the state-of-the-art StyleGAN2-ADA3 architecture that is pretrained on the high-quality FFHQ dataset. To normalize images, we crop around the facial region, ensuring that the facial features are in positions comparable to those in the pretraining dataset.

Using StyleGAN2-ADA3, we train the generator on the source and target images of each pair to produce a morph sequence. Given the standardised, high-resolution photos we feed the GAN, we train it for 125 iterations per image using the default hyperparameters of StyleGAN2-ADA3, after which no appreciable effect is shown (Wolfe et al., 2022). For each pair, the trained generator creates a source embedding and a target embedding. Subsequently, we use these embeddings to generate the first and second images of each morph sequence.

Finally, we obtain the projected image embedding from CLIP for each morph. This produces a dataset of 30,000 morphed images with corresponding embeddings, each of which is distinct and diverse. Note that due to copyright issues, we are not able to share generated samples from CFD in the Appendix, but can do so upon request.

Operationlisation of CFD We randomly select 1,000 (this number is chosen for its significant sample size as well as computational feasibility) images per intersectional group without replacement, resulting in a total of 6,000 images. In experiments where CFD is used in the group SC-EAT, 140 images per group are selected, including 840 separate images from the SC-EAT group of interest for the two attribute sets. For text-to-image (TTI) retrieval, all 6,000 equally balanced CFD images are used. When CFD is the target dataset for valence SC-EAT and image-to-text (ITT) retrieval, all 6,000 images are utilized.

We want to note that the GAN generative step is used strictly for data augmentation within a group, minimizing the risk of introducing spurious cues about race or gender. While generated images can sometimes introduce artefacts, our approach carefully confines morphs within a single demographic cluster so that any potential artefact is minimal and does not conflate racial and gender categories.

A.2 Open Affective Standardized Image Set (OASIS) Database.

The Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017) comprises 900 color images, depicting a broad spectrum of themes such as humans, animals, objects, and scenes. These images were rated by 822 human participants on valence (degree of positivity or negativity) and arousal (intensity of the valenced response). The

[4]Because of the known ambiguity in racial perception surrounding Latino/a faces, the current research focused only on Black, White, and Asian racial groups (García, 2013)

900 images are used in the current experiment to provide non-group-related valence signals from naturalistic image inputs. All images were standardised to 500 X 400 pixels through scaling/cropping processes similar to Wolfe et al. (2022).

Operationlisation of OASIS All 900 images are used across experiments where OASIS is required for group SC-EAT or ITT retrieval. In experiments involving valence SC-EAT, we use the top 25 pleasant and top 25 unpleasant OASIS images, sorted by valence ratings, as the two attribute sets. For TTI retrieval, all 900 images are employed. Utilizing the entire lexicon and valence spectrum allows us to provide comprehensive insights into bias propagation, with implications extending to the factuality of retrieval results.

A.3 Textual Templates and Lexica.

Our research requires controlled and balanced text datasets to understand how the psycholinguistic and social group properties of textual content influence biases in VLMs. Specifically, we seek to analyze how words and phrases, with known psycholinguistic ground-truth ratings or specific intersectional identities belonging to one of the 6 groups of interest, can affect the propagation of biases in these models. In this study, we adopt the sentence template approach of May et al. (2019) and Tan and Celis (2019), utilizing "semantically bleached templates." These templates are designed to be semantically neutral, meaning they do not contribute novel semantic information but ensure that the target word phrases are placed within similar syntactic frames. This approach allows the values associated with the target words to convey the psycho-semantic characteristics of the sentence.

A "target word" refers to a specific word inserted into a template, designed to elicit a psycho-semantic response. For example, in the template "This is the word [WORD]," the target word replaces "[WORD]" and serves as the primary variable in our experiment. In our study we use 6 sentence templates derived from May et al. (2019) which are listed in Table 2.

For experiments involving the use of valence-rated sentences, we employ the NRC-VAD psycholinguistic lexicon (Mohammad, 2018), comprising approximately 20,000 English words, with each word rated by 6 human participants on three key dimensions: valence, arousal, and dominance (VAD). This comprehensive lexicon provides us with a robust database of words grounded in reliable, human-rated psycholinguistic data.

For experiments that involve sentences with group signals, we employ a curated list of group labels derived from Charlesworth et al. (2022). Here, the target word is a combination of a race-identifying word followed by a gender-identifying word. An example sentence is, "This is the word Black Woman". The full list of group words is in Table 2.

Operationlisation of NRC-VAD Lexicon Sentences We use 20,000 words across six sentence templates, with all results averaged across templates. All 20,000 sentences per template are used when the NRC-VAD lexicon is required for group SC-EAT or TTI retrieval. In experiments requiring valence SC-EAT, the top 25 pleasant and top 25 unpleasant sentences, sorted by valence ratings, are used as the two attribute sets. For ITT retrieval, all 20,000 sentences per template are employed. The comprehensive use of the lexicon and valence spectrum provides critical insights into bias propagation and its broader implications.

Operationlisation of Group Label Sentences We use 5,184 sentences across six sentence templates, with results averaged across templates. When Group Label sentences are the target dataset for valence SC-EAT and TTI retrieval, all 5,184 sentences are used. For group SC-EAT, where Group Label sentences measure group association, 140 sentences per group are randomly selected, including 840 separate sentences from the SC-EAT group of interest for the two attribute sets. For ITT retrieval, all 5,184 sentences are utilized.

B Models

Following from Section 5 of the Main text, the following details the choice of open-source models used in our experiments.

Vision-Language Models – CLIP (B-32 and L-14). Radford et al. (2021) use the WebImageText corpus (WIT), a web scrape made up of 400 million images and related captions to train CLIP. The query list is created by Radford et al. (2021) by utilising all words that appear at least 100 times in English Wikipedia, as well as word bi-grams with high pointwise mutual information from Wikipedia, the titles of Wikipedia articles, and all WordNet synsets. As part of the creation process, they look for image-text pairs whose texts contain one of a set of 500,000 queries in an effort to cover as many visual notions as feasible. The significance of CLIP,

Table 2: Sentence templates and group words used for bias measurement in VLMs. Sentence templates are adopted from May et al. (2019), and group words are selected based on Charlesworth et al. (2022) These elements form the basis of the text datasets employed in our study to investigate the propagation of social biases.

Type	Content
Templates	"This is the word [WORD]", "That is the word [WORD]", "There is the word [WORD]", "Here is the word [WORD]", "They are the word [WORD]", "Those are the word [WORD]"
Gender Words	woman, daughter, mother, sister, grandmother, niece, female, girl, madam, aunt, maiden, queen, man, son, father, brother, grandfather, nephew, male, boy, sir, uncle, gentleman, king
Race Words	black, blacks, black-american, afro-caribbean, dark-skinned, jamaican, african, africans, ethiopian, ethiopians, african-american, afro-american, white, whites, british, caucasian, caucasians, light-skinned, american, americans, european, europeans, englishman, englishmen, asian, asians, asian-american, asian-americans, chinese-american, japanese-american, chinese, asiatic, japanese, korean, koreans, korean-american

particularly its B-32 and the significantly larger L-14 versions, lies in its zero-shot learning capabilities (achieving zero-shot accuracy of 63.2% and 76.2% respectively) on ImageNet (Radford et al., 2021). These models can process and interpret images and their associated texts without needing task-specific training, demonstrating a level of generalization that is highly applicable in various settings (Briggs and Laura, 2022).

BLIP2. BLIP-2 (Li et al., 2023) is a vision-language pre-trained model that integrates off-the-shelf frozen image encoders and LLMs. It includes a Querying Transformer (Q-Former) trained in two stages: vision-language representation learning from a frozen image encoder and vision-to-language generative learning from a frozen LLM. It's trained on a dataset comprising 129 million images, including a subset of COCO, Visual Genome, CC3M, CC12M, SBU, and LAION400M datasets. We study BLIP-2 due to its efficiency and high performance in zero-shot tasks. For instance, it achieved state-of-the-art zero shot accuracy of 65% on VQAv2 (Li et al., 2023).

C Approach

The following section elaborates on the definition and formulation of the SC-EAT taken from Caliskan et al. (2017) and Steed and Caliskan (2021).

D Details of Experiments

This section supplements the results mentioned in Section 5 in the Main text and contains the details for the reproducibility of all experiments, including specific datasets and parameter settings used. Figure 8 contains the reproduction details of each of the 8 bias and association propagation experiments.

D.1 Initial Assessment of Intrinsic Biases in Vision-Language Models

Figures 6 and 7 highlight significant intrinsic biases within VLMs. Figure 8 presents the aggregate valence-based SC-EAT effect sizes for different social groups obtained in experiments 1-a and 1-b. Figure 9 presents the aggregate group-based SC-EAT effect sizes for high (valence ≥ 0.5) and low (valence < 0.5) valence bands across experiments 2-a and 2-b.

D.2 Propagation of Bias Through Valence and Group Signals

Baseline Propagation of Intrinsic Valence Associations to Extrinsic Valence Outcomes.

This experiment serves as a baseline to measure how isolated valence signals in text and valence signals in images are propagated through VLMs. We follow previous works such as VAST (Wolfe and Caliskan, 2022d) and ValNorm (Toney and Caliskan, 2021) that use intrinsic valence signals to evaluate the intrinsic quality of word embeddings and language models. First, for the image-to-text direction (1*-a), we measure the correlation be-

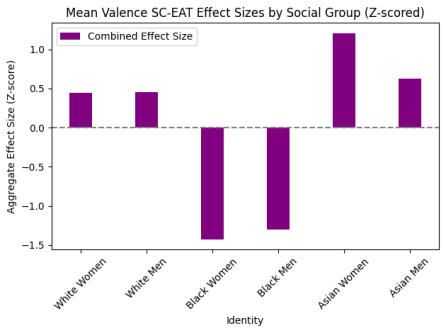

Figure 6: Valence based Intrinsic Bias Visualization through SC-EAT Effect Sizes: This figure presents the normalised mean of SC-EAT effect sizes for different social groups obtained in experiments 1-a and 1-b. A negative effect size indicates a stronger association of the group with negative valence than positive. The data reveals intrinsic biases within the models, where notably, Black Women are depicted with the highest negative association, as indicated by the most negative effect size. In contrast, Asian Men and Women exhibit the least negative associations. This graphically demonstrates the prevalence of intrinsic bias in the representational spaces of VLMs, with varying degrees of negativity associated with each group.

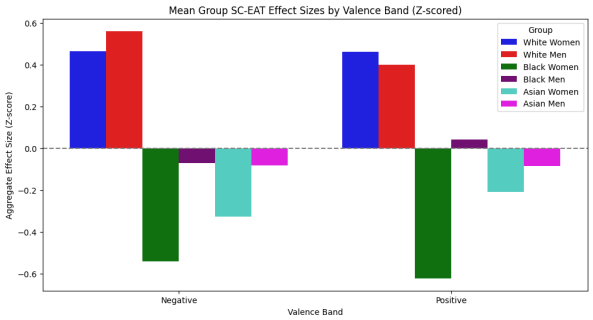

Figure 7: Group-based Intrinsic Bias Visualization through SC-EAT Effect Sizes: This figure presents the normalised mean of SC-EAT effect sizes for high/low valence bands obtained in experiments 2-a and 2-b. We categorise images/text with valence ≥ 0.5 as positive and negative otherwise, on a scale of 0 to 1. A positive effect size indicates a stronger association of the valenced item with the indicated group as opposed to any group at random. The data reveals intrinsic biases within the models, where notably, White Men and Women are depicted with the highest positive and negative associations. In contrast, Black Women exhibit the least magnitude in associations to valenced images/text. This graphically demonstrates the prevalence of group-based intrinsic bias in the representational spaces of VLMs.

tween the intrinsic valence of the image and the mean valence of the top-500 sentences retrieved based on that image. For instance, an image with a high intrinsic positive valence (e.g., roses, butterflies, etc) should ideally retrieve text with similarly positive valence (e.g., "This is the word pretty"). The intrinsic measure here is the SC-EAT score of each of the 50 most valenced OASIS images (top 25 positive and top 25 negative); the extrinsic measure here is the average valence rating (from human valence ratings) of the top 500 sentences associated with the specific input OASIS image.

Second, the text-to-image (1*-b) direction mirrors 1*-a and explores how intrinsic valence signals in text correlate with the aggregate valence of retrieved images based on the text input. Specifically, the intrinsic measure is the SC-EAT score of the 50 most valenced words in sentences (top 25 positive and top 25 negative words, as described in Section A); the extrinsic measure here is the average valence rating (from human valence ratings) of the top 500 OASIS images associated with the input sentence.

Baseline Propagation of Intrinsic Group Associations to Extrinsic Group Outcomes. This set of experiments seeks to establish a baseline for understanding the simple propagation of group content. In the image-to-text setting (2*-a), we test the correlation between the representation of a specific social group in an image and the group representation in the retrieved text. The intrinsic measure here is the image-text group SC-EAT where we find the association between a given image (from the CFD) and sentences including a single intersectional group (versus sentences that represent all intersectional groups). This method essentially parallels a one-vs-all machine learning verification task, where the model's ability to differentiate one group identity from all others at the representational level is evaluated. Extrinsically, for the same CFD image, we obtain the proportion of group-identified sentences retrieved that are "correct" (i.e., the same intersectional identity as the text prompt). For example, if an input image is of a "Black Woman," then extrinsically we assess how often the text retrieved accurately reflects Black women concepts.

In the text-to-image direction (2*-b) we test whether text representing a social group retrieves images with that same group representation from the CFD. The intrinsic measure is the text-image SC-EAT between a given text prompt with a specific group signal and images representing the single intersectional group identity (versus images containing all intersectional groups uniformly). Ex-

	Valence Based Bias Propagation		Content Based Bias Propagation	
	1*. Valence-Valence Interactions	1. Valence-Group Interactions	2*. Group-Group Interactions	2. Group-Valence Interactions
Image to text Retrieval (a)	Intrinsic Metric: SC-EAT(OASIS image O, top25 pleasant Trait words in Sentence Templates, top25 Trait words in Sentence Templates) Extrinsic Metric: mean(valence of top500 Trait words in Sentence Templates) retrieved for OASIS image O Valence Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every OASIS image	Intrinsic Metric: SC-EAT(CFD image C, top25 pleasant Trait words in Sentence Templates, top25 Trait words in Sentence Templates) Extrinsic Metric: mean(valence of top500 Trait words in Sentence Templates) retrieved for CFD image C Valence Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every CFD image	Intrinsic Metric: For each group g belongs to G, SC-EAT(CFD image C, N sentences containing g in sentence templates, N sentences containing uniform representation of all groups G in sentence templates) Extrinsic Metric: proportion(top500 group words in Sentence Templates) retrieved for CFD image C Content Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every CFD image C	Intrinsic Metric: For each group g belongs to G, SC-EAT(OASIS image O, N sentences containing g in sentence templates, N sentences containing uniform representation of all groups G in sentence templates) Extrinsic Metric: proportion(top500 group words in Sentence Templates) retrieved for OASIS image O Content Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every OASIS image O
Text to Image Retrieval (b)	Intrinsic Metric: SC-EAT(Trait word in Sentence Template S, top25 pleasant OASIS images, top25 unpleasant OASIS images) Extrinsic Metric: mean(valence of top500 OASIS images) retrieved for Trait word in Sentence Template S Valence Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every Trait word in Sentence Template S	Intrinsic Metric: SC-EAT(Group word in Sentence Template S, top25 pleasant OASIS images, top25 unpleasant OASIS images) Extrinsic Metric: mean(valence of top500 OASIS images) retrieved for Group word in Sentence Template S Valence Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every Group word in Sentence Template S	Intrinsic Metric: For each group g belongs to G, SC-EAT(Group word in Sentence Template S, N CFD images containing g, N CFD images containing uniform representation of all groups in G) Extrinsic Metric: proportion(top500 CFD images) retrieved for Group word in Sentence Template S Content Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every Group word in Sentence Template S	Intrinsic Metric: For each group g belongs to G, SC-EAT(Trait word in Sentence Template S, N CFD images containing g, N CFD images containing uniform representation of all groups in G) Extrinsic Metric: proportion(top500 CFD images) retrieved for Trait word in Sentence Template S Content Propagation: Spearman's correlation values between intrinsic and extrinsic metrics for every Trait word in Sentence Template S

Figure 8: Figure details the specifics of the 8 experiments used to show the propagation of intrinsic bias to extrinsic zero-shot retrieval tasks. The experiments share a common framework for measuring intrinsic bias (e.g., the effect size of bias that captures the differential association of a social group image with negatively valenced sentences over positively valenced sentences), followed by an assessment of extrinsic bias (e.g., the likelihood of retrieving negatively valenced text for a given group image prompt), and then correlating these two metrics. The experiments vary in their focus on the direction of bias propagation (four each on image-to-text and text-to-image) and the type of content analyzed (four on valence or rating of positivity/negativity, and four on group signals)

trinsically, we then record the proportion of images that "correctly" represent the specific social group retrieved for text prompt. For clarity, "correct" in this context means that the retrieved images accurately mirror the specific social group identity mentioned in the text. For instance, if the text prompt is about "Asian Men," the extrinsic measure evaluates how often the retrieved images indeed depict Asian men.

As in all other studies, we quantify content/group propagation using Spearman's correlation values between the intrinsic SC-EAT group based associations of the image/text with the extrinsic proportion of sentences/images correctly retrieved for each group.

Propagation of Intrinsic Group-based Valence Associations to Extrinsic Valence Outcomes. This experiment is our primary bias propagation analysis as it focuses on how biased associations between group signals (in text or image) and valenced attributes are propagated through VLMs. For the image-to-text direction (1-a), we consider how intrinsic group signals in images influence the valence of retrieved text. The image-text SC-EAT measures the intrinsic valence effect size of each image belonging to a group identity (from the CFD). Then, extrinsically, for the same image, we calculate the mean valence of the top-500 sentences retrieved, averaged over the templates.

In the text-to-image direction (1-b) we examine how text associated with social groups influences the valence of retrieved images. Intrinsically, this is measured through the text-image SC-EAT by computing the effect size of valenced associations for each sentence containing an intersectional group identity with valence-rated images (from OASIS). Then, extrinsically, for the same sentence, we calculate the mean valence of the top-500 valence-rated OASIS images retrieved.

Propagation of Intrinsic Valence-based Group Associations to Extrinsic Group Outcomes. Finally, and in parallel to our primary analysis of

group-to-valence propagation, we consider valence-to-group propagation or, more specifically, how intrinsic valence associations of images/text as measured by SC-EAT lead to differential retrieval of group signals. In the image-to-text direction (2-a), we investigate whether stronger intrinsic valence ratings in images correlate with a higher likelihood of retrieving text related to marginalized groups. Intrinsically, we use the image-text group SC-EAT to find the association of a given valenced image (from OASIS) with text containing group terms belonging to a single intersectional identity (versus text uniformly representing all intersectional groups). Extrinsically, we obtain the corresponding proportion of group-identified sentences that are retrieved for the same OASIS image for each of the 6 intersectional social groups.

In the opposite, text-to-image direction (2-b) we test if text with strong valence ratings retrieves images predominantly featuring certain social groups. The intrinsic measures are the SC-EAT associations between a given valenced text prompt and images from a specific social group (versus images of all groups uniformly that were selected through random balanced sampling without replacement). Extrinsically, we calculate the proportion of images that represent a social group retrieved for the same single intersectional group text prompt.