偏见在基于编码器的视觉语言模型中传播：从内在测量到零样本检索结果的系统分析

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

为了构建公平的人工智能系统，我们需要了解基础编码器为基础的视觉语言模型（VLMs）固有的社会群体偏见如何在下游任务中表现为偏见。在这项研究中，我们展示了 VLM 表示中的固有偏见系统性地“传递”或传播到零样本检索任务中，揭示了深层根植的偏见如何影响模型的输出。我们引入了一个受控框架，通过将（a）表示空间中的固有偏见测量与（b）零样本文本到图像（TTI）和图像到文本（ITT）检索中的外在偏见测量相关联来衡量这种传播。结果显示固有偏见和外在偏见之间的显著相关性，平均为 $\rho = 0.83 \pm 0.10$。这个模式在 114 项分析中是一致的，包括两个检索方向、六个社会群体和三个不同的 VLMs。值得注意的是，我们发现更大或性能更好的模型表现出更大的偏见传播，这一发现令人担忧，因为 AI 模型趋向于越来越复杂。我们的框架引入了基线评估任务，以衡量群体和价度信号的传播。调查显示，代表性不足的群体体验到的传播较不稳健，进一步扭曲了与他们相关的模型结果。

1 介绍

现代基于编码器的视觉-语言模型（VLMs），如 CLIP (Radford et al., 2021) 和 BLIP-2 (Li et al., 2023)，在将图像和文本映射到共享表示空间方面表现优异。这些模型是从物体识别到图像检索 (Rombach et al., 2022; Briggs and Laura, 2022; Taesiri et al., 2022; Bui et al., 2023; Barraco et al., 2022; Pirom, 2022) 等众多零样本应用的基础。然而，这些表示可能包含内在偏见已经越来越清楚，例如，系统地将某些社会群体与负面概念联系在一起。这些偏见在下游应用中显现的潜力可能会破坏跨社会群体的公平和道德对待 (Ghosh and Caliskan, 2023; Wolfe and Caliskan, 2022a)。人们对基础编码器型 VLMs，诸如 CLIP (Radford et al., 2021) 和 BLIP2 (Li et al., 2023) 内在偏见，如何传播到它们在未经再训练或微调的情况下轻松执行的任务上，理解上存在显著差距。具体来说：当前的研究通过比较（a）这些模型学习的表示空间中的内在偏见与（b）下游零样本任务的结果，特别是文本到图像（TTI）和图像到文本（ITT）检索来量化编码器型 VLMs 偏见的传播。

在我们的研究中，“传播”一词指的是偏见从模型表示向下游任务结果的方向性流动。通过采用严格控制的实验设计，我们将外部混杂因素降到最低，从而能够清晰地观察偏见如何在基于编码器的 VLM 中传递。虽然我们认识到这并不能完全建立因果关系，但我们的研究结果提供了系统性偏见传递的有力证据，这证明使用“传播”一词是合理的。

为了进行我们的分析，我们策划了由芝加哥脸部数据库（CFD）(Ma et al., 2015) 和开放情感标准化图像集（OASIS）(Kurdi et al., 2017) 中的图像组成的平衡数据集，分别捕捉社交群体信号和效价信号。效价是我们分析中的主要维度，因为它捕捉的是理解 AI 中偏见表现至关重要的偏见态度 (Toney and Caliskan, 2021; Osgood et al., 1957)。与此同时，我们使用语义中立的句子模板 (Tan and Celis, 2019) 开发实验性文本数据集，结合了表示 NRC-VAD 词典中的效价信号 (Mohammad, 2018) 或来自先前研究中的社交群体信号 (Charlesworth et al., 2022, 2024) 的目标词。我们通过量化编码器基础的视觉语言模型中的内在和外在的偏见来测量偏见传播。内在偏见使用单类别嵌入关联测试（SC-EAT）(Caliskan et al., 2017; Caliskan, 2023) 评估，该测试评估模型的表示空间中社交群体和属性关联信号之间的关联。外在偏见通过下游零样本任务（例如 ITT 和 TTI 检索）进行评估，以评估这些内在偏见在实际应用中的表现。图 1 展示了在 ITT 设置中测量基于效价的偏见传播的方案。

我们的研究独特地考察了基于编码器的视觉语言模型（VLMs）在受控的零样本设置下的社会群体和情感偏见的传播，解决了多模态环境中跨多个社会群体的偏见复杂性。乍一看，似乎模型内部表示中的内在偏见会自然出现在其下游任务中，因为它们共享相同的嵌入空间。然而，正如我们将展示的，这种关系并不简单。不同的社会群体（例如，“黑人女性”与“白

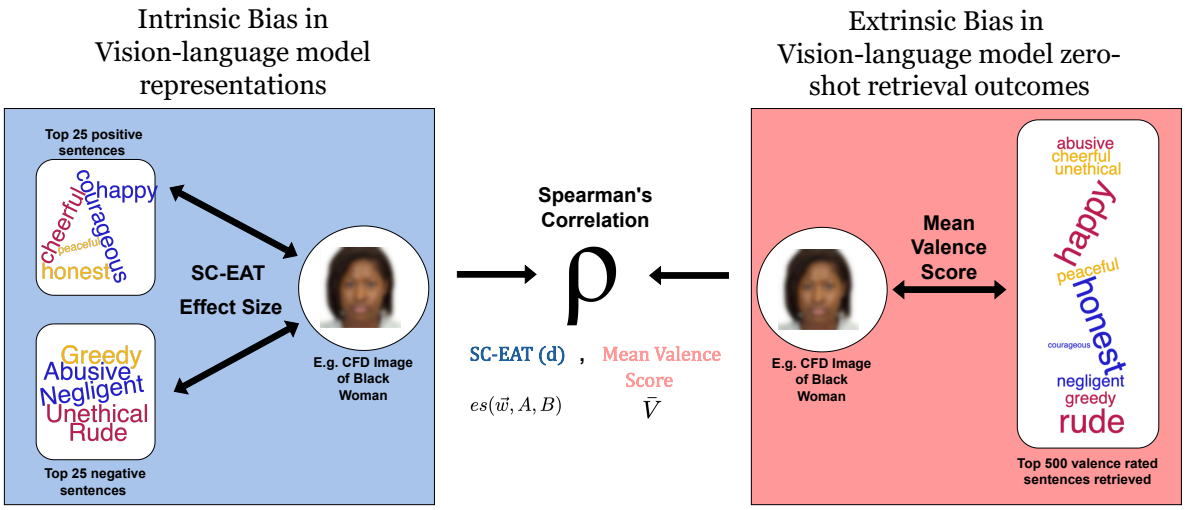

Figure 1: 用于分析图像到文本检索中基于情感的偏差传播的设置。传播是通过内在偏差（由情感 SC-EAT 测量）与外在偏差（图像所属社会群体的检索文本的平均情感）之间的相关性来衡量的。内在 SC-EAT 效应大小是基于每个 CFD 图像与句子模板中的前 25 个积极词汇和前 25 个消极词汇的差异关联来计算的。外在指标源于每个图像检索的句子模板中的前 500 个词的平均情感。真实情感分数来自人类评级来源 (Mohammad, 2018) 。Spearman 的 ρ 用于计算内在和外在指标之间的相关性。

人男性”）可能产生显著不同的相关性。这些系统性差异反映了复杂的社会线索，如“数据标记”和群体代表性不足 (Newton, 2023; Wolfe and Caliskan, 2022c) 如何影响偏见的流动。因此，当前的工作也为特定群体和情感信号如何在基于编码器的 VLMs 中影响偏见传播的程度提供了经验上的清晰性。

我们的研究做出了以下贡献。[1]

(1) 我们提供了一个受控的实验框架，用于系统地量化偏见传播。这个框架加强了之前的研究，这些研究显示出基于编码器的 VLM 的表示空间中存在偏见。同时，这个框架首次显示了与内在偏见直接相关的 TTI 和 ITT 检索任务中下游零样本的偏见。此外，该框架和这里展示的方法在支持大规模研究偏见传播方面是独特的——在 114 个场景中显示出一致的结论，涵盖了 6 个社会群体、3 个模型和 8 个实验，我们显示出相关性持续高且显著（Spearman’s ρ 为 0.83 ± 0.10）。

(2) 我们引入了基线内在评估实验，这些实验评估了零样本检索任务中价数和群体信号的直接传播。这些基线指出了准确检索的程度，可以作为基准来比较重点实验中价数到群体和群体到价数传播中的偏见。结果显示价数传播的基线相关性很强，Spearman’s ρ 为 0.86 ± 0.04，对于群体信息传播，相关性为 ρ 为 0.85 ± 0.12。

(3) 该受控框架将支持未来对 AI 偏见动态的经验和理论理解。特别是，结果强调了在不同社会群体间偏见传播的差异。那些在数据中被代表性不足的边缘群体经历了较不稳健的传播和更偏斜的模型相关结果。我们还显示，随着模型规模和性能的扩大，它们倾向于显示更高的偏见传播——这一发现具有重大影响，尤其是在快速 AI 扩展的时代。

[1]代码和数据可在 https://github.com/kshitishghate/bias_prop 获取

2 相关工作

视觉-语言模型中的偏见。关于视觉-语言模型偏见的最新研究超越了单独的语言或视觉模型，探讨了这两种模态何时相交。例如，Srinivasan and Bisk (2022) 将基于文本的性别偏见研究（如 (Su et al., 2019) ）扩展到了多模态设置中。他们使用 VL-BERT 证明了视觉-语言模型倾向于强化和夸大性别偏见和刻板印象。Wolfe and Caliskan (2022a) 评估了在 CLIP 模型中使用嵌入关联测试 (Caliskan et al., 2017) 和 CFD (Ma et al., 2015) 而得出的将美国人与白人进行关联的偏见，这类似于基础认知科学研究 (Devos and Banaji, 2005) 中的内容。他们观察到，相比于亚洲、拉丁裔和黑人面孔，白人面孔的图片与集体的内群体词汇有更高的关联性。Zhou et al. (2022) 使用 VLStereoSet 评估了多个视觉-语言模型，发现了普遍存在的性别和种族刻板印象和偏见，这些在视觉-语言模型中比在预训练语言模型中更为复杂。Wolfe et al. (2023) 专注于训练自网络爬取数据的视觉-语言模型如何经常延续对女性和女孩的性物化问题。

相关研究的一部分通过检查预训练因素的影响来调查 VLMs 中的偏见起源。Berg et al. (2022) 比较了 9 个 CLIP 模型中的性别偏见，发现通常与更好的下游性能相关联的较大预训练数据集也展示了最少的偏见。Ghate et al. (2025) 在超过 131 个 CLIP 模型，超过 20 种架构和预训练数据集中扩展了这项研究，发现选择预训练数据集和用于过滤它的数据过滤策略是 CLIP 基础的 VLMs 固有偏见的最重要预测因素，超过诸如架构选择、预训练数据集大小和模型参数数量等变量。Hong et al. (2024) 审查了基于 CLIP 的预训练数据过滤策略，发现与历史上良好代表的西方中心人口相比，数据相关的边缘化群体受到更高的排除率。

预训练模型中的偏见传播。有多项研究探讨了预训练语言模型中固有的偏见如何影响下游模型的结果。Cao et al. (2022) 强调了在监督设置下上下文语言模型的固有公平性指标和外在公平性指标之间的弱相关性。Steed et al. (2022) 调查了偏见转移假设，该假设认为大型语言模型 (LLMs) 中的偏见会在微调后转移到有害的特定任务行为中。他们的研究结果表明，在微调之前减少固有偏见对减轻歧视性结果几乎没有作用。然而，需要注意的是，此类过去工作中所研究的下游任务并不是零样本的。相反，这些任务涉及在特定数据集上进行微调，这可能导致模型学习到这些数据集中固有的虚假关联。Feng et al. (2023) 探索了语言模型中的政治偏见传播。他们的工作揭示了预训练语言模型具有政治倾向，这些倾向受训练语料库中的政治影响，从而影响下游任务结果。Cabello et al. (2023) 表明在基于编码器的视觉语言模型 (VLMs) 微调任务中，固有偏见与外在偏见没有一致的相关性。他们提出了性别中立预训练作为一种缓解策略，并证明了其减少群体差距的能力。然而，他们的工作也集中于微调结果，缺乏直接研究偏见传播的系统框架。

相反，我们专注于零样本任务，在这些任务中，模型在没有额外微调的情况下直接被检验内在到外在层次的偏差传播，模型直接依赖于其预训练表示。我们引入了一个受控框架来调查多个 VLMs 及社会群体之间的偏见。这使我们能够揭示以前被忽视的偏差传播中的系统趋势。

我们使用经过精心策划的数据集，这些数据集提供实验控制和真实值。

芝加哥面孔数据库 (CFD) (Ma et al., 2015)：CFD 数据集由 597 个自我识别的男女参与者图像组成，分别属于四个不同的种族人口统计类别，即黑人、白人、拉丁裔和亚裔。选择该数据集是因为其具有可靠的、自我识别的人脸图像，并具有明确的群体特征，这与 FairFace (Karkkainen and Joo, 2021) 不同，后者缺乏自我识别标签且包含视觉噪声较多的图像，可能导致注释偏倚。我们首先扩展数据集，以创建组内图像变形的方法来促进实验的规模和泛化，具体方法如附录 A.1.1 中所述的 (Wolfe et al., 2022) 。然后我们在没有替换的情况下随机抽样图像。

开放情感标准化图像集 (OASIS) (Kurdi et al., 2017)：提供 900 张人类评定的形象价值的图像，使得能够研究 VLMs 如何处理来自自然图像输入的非群体相关的价值信号。这些图像描绘了广泛的主题，如动物、人类、物品和场景。

文本模板和词汇集：我们的研究需要受控的文本数据集。语义中立的句子模板能够帮助理解目标词汇在其中的价性和社会群体特性是如何影响 VLM 中的偏见。因此，我们使用 6 种从 (May et al., 2019) 中提取的模板中嵌入的词汇集来对含有价性或社会群体信号的词语/短语如何影响外在 VLM 输出的偏见进行受控分析。对于使用价性评分句子的实验，我们使用 NRC-VAD 词汇集 (Mohammad, 2018)，该词汇集包含大约 20,000 个英文单词。一个从价性词汇 “happy” 中创建的例句是 “这是词语 ‘happy’。” 对于带有群体信号的实验，我们使用从 Charlesworth et al. (2022) 提取的 864 个群体标签短语的精选列表。从群体标签 “Black Woman” 创建的一个例句是 “这是词语 ‘Black Woman’。” 在我们的分析中，我们关注 6 个交叉种族和性别群体：“Asian Men,” “Asian Women,” “Black Men,” “Black Women,” “White Men,” 和 “White Women”。

CFD、OASIS、NRC-VAD 词典和组标签的汇总统计见表 **??** 。有关其描述和在我们实验中的使用详情，请参阅附录 A 。

3 方法

在本研究中，社会群体偏见被定义为群体信号（例如，黑人女性的表现）与效价信号（例如，积极与消极的表现）之间的关联。这种方法以 AI 伦理学领域中的既定方法为动力，例如应用于嵌入空间的嵌入关联测试（EAT），为研究偏见提供了一个科学合理的框架。效价也是我们分析的核心，因为它反映了与群体表现相关的正面或负面含义，并且是评估偏见的基本维度。

3.1 内在和外在偏差测量

内在偏差通过单类别嵌入关联测试（SC-EAT）(Caliskan et al., 2017) 效应大小 (Cohen, 2013) 进行量化，这允许对模型表示中的偏差进行原则性评估。SC-EAT 定量评估单一目标刺激

（例如，代表社交群体的图像或词语）与一组属性刺激（例如，具有情感或群体内容的词或图像）之间的关联。该关联通过目标和属性刺激嵌入之间的平均余弦相似性衡量，并通过所有刺激相似性的标准差进行归一化。这被量化为 SC-EAT Cohen 的 d 效应大小，表达为：

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

$\vec{w}$ 表示目标刺激的嵌入，在我们的案例中，这可以表示一个感兴趣的句子或图像，而 A 和 B 是由效价或群体内容决定的一组属性。

通过从量化与价性和内容相关的任务结果偏差的研究中得出的两个措施来评估 ITT 和 TTI 中的外在偏差 (Charlesworth et al., 2024; Kong et al., 2024)：

1. 检索到的带情感文本/图像的均值：该指标计算模型对给定提示检索的文本/图像的平均情感值：

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

其中 $\bar{V}$ 是平均效价，V_i 是第 i 个检索项的效价，N 是检索项的总数。这个指标量化了模型输出在文本描述或图像检索中偏向于积极或消极效价的程度。该指标有助于理解模型的内部偏见如何转化为它所检索的效价内容中的偏见。

2. 检索属于某社会群体的文本/图像比例：该测量评估模型在响应给定提示时，检索到与特定社会群体相关的文本或图像的可能性或比例，计算公式为：

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

，其中 P_{grp} 是预计在各组之间保持不变的比例。当提示被预期与所有组平等关联时，不等的 P_{grp} 显示了模型偏差。

3.2 衡量偏差传播的框架

为了评估视觉语言模型（VLMs）中的固有偏差如何传播到下游检索任务，我们提出了一个统一框架（算法 1），其中定义了以下变量：

数据（D）：输入可以是我们的任何精选数据集（CFD/OASIS 图像或使用 NRC-VAD/组标签的文本模板）。

检索方向（R）：指定零样本检索任务，分为 ITT 或 TTI。

内容类型（C）：声明当前的偏见分析是针对情感（正面/负面）还是群体（例如，种族和性别）。

相关性（ρ）：在内在和外在指标上计算的最终 Spearman 相关性。

我们现在逐步概述在 1 中构成我们框架的功能，随后提供一个示例。

1. CalculateIntrinsicAssociations（D，C）：此函数基于内容类型 C 和数据 D 计算内在关联度量。如果 C 是“valence”，则使用 SC-EAT 对效价信号进行度量计算（例如，从 OASIS 图像数据集或 NRC-VAD 文本词汇中）。如果 C 是“group”，SC-EAT 则应用于组信号（例如，从图像的 CFD 数据集或预定义文本组标签中）。

2. RetrieveTextFromImage（D）：使用 VLM 检索与输入 D 相关的前 k 个句子。

3. RetrieveImagesFromText（D）：使用 VLM 检索与输入 D 相关的前 k 个图像。

4. CalculateExtrinsicOutput（$retrieved\_items$，C）：此函数基于检索到的项目和内容类型 C 来计算外在偏差度量。如果 C 是“情感”，则度量是检索到项目的平均情感。如果 C 是“群体”，则度量是检索到的群体项目所占的比例。

5. SpearmanCorrelation（$intrinsic\_metric$，$extrinsic\_metric$）：该函数计算内在和外在度量之间的斯皮尔曼相关性 ρ，反映偏见传播的程度。我们选择使用斯皮尔曼 ρ 是因为其非参数性质以及捕捉非线性关系的能力 (Spearman, 1961)。

小例子：假设我们想看看针对“黑人女性”的情感偏向是否体现在 ITT 检索中（如图 1 所示）。这里，输入 D：来自 CFD 的 6000 张图像（每个社会群体 1000 张）；来自 NRC-VAD 的句子模板中的 20,000 个词。C：“情感”；R：ITT。具体来说：

1. 内在关联：我们将 SC-EAT 应用于来自 CFD 的 1000 张黑人女性的图像。每个图像的嵌入与句子模板中排名前 25 的积极词与消极词（来自 NRC-VAD）进行比较。结果是每个图像的内在效价效应大小。

2. ITT 检索：然后，我们将这 1000 张图像中的每一张输入到我们的 VLM 中，并检索出前 500 个文本项（例如，“这是单词 happy”，“这是单词 sad”，等等）。

3. 外在指标：对于每张图像，我们测量检索到的文本的平均情感价值（这些词大多是积极的还是消极的?）。

4. 相关性：我们测量效价效应大小与 1000 张图像的平均检索效价之间的相关性。如果具有高负 SC-EAT 分数的图像也主要检索到负面的文本描述，这表明存在强烈的偏差传播（可能是一个高 ρ）。

Algorithm 1 统一偏差传播框架

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

3.3 通过价和组信号测量偏差传播

我们实验设计的核心在于分析偏差传播，包括 8 个实验，全部遵循算法 1 的框架，并均匀地划分为不同的：（1）方向（即，4 个实验测试 ITT 关联和偏差传播，而 4 个测试 TTI 关联和偏差传播）；以及（2）效价与群体内容（即，4 个实验测试效价的传播，而 4 个测试群体信号的传播）。算法 1 的应用将在以下传播实验中解释。对于所有实验，相关性是基于每个模型针对目标图像或文本群体的内在和外在度量得分来计算，分别针对 ITT 和 TTI。具体细节见附录图 Figure 8 。

基线效价-效价信号传播 (1*-a 和 1*-b)：这是一个基线实验，用于测量内在效价信号如何传播到效价结果。框架参数为 – D ： OASIS 图像和 NRC-VAD 词典句子；C ：“效价”。我们将 R ： ITT 方向标记为 (1*-a) 和 R ： TTI 方向标记为 (1*-b)。内在指标：对顶级效价图像/文本的 SC-EAT 关联。外在指标：前 500 个检索图像/文本的平均效价评分。对于所有实验，我们选择检索 k 的 500 个条目，因为它提供了一个多样且在计算上可行的重要样本量。

基线组-组信号传播 (2*-a 和 2*-b)：这是一个基线实验，用于测量内在组信号如何通过使用 CFD 图像和预定义组标签传播到组结果。框架参数为——D ： CFD 图像和组标签句子；C ：“组”。我们将 R ：ITT 方向标记为 (2*-a) 和 R ： TTI 方向标记为 (2*-b)。内在指标：SC-EAT 组关联。外在指标：在检索到的前 500 张图像/文本中正确识别的组表征的比例。

价态到群体信号传播 (1-a 和 1-b)：评估价态信号如何影响与群体相关的检索结果。框架参数为——我们将 R ： ITT 方向标记为 (1-a)，输入为 D ： CFD 图像和 NRC-VAD 词典句子；C ：“价态”。将 R ：TTI 方向标记为 (1-b)，输入为 D ：群体标签句子和 OASIS 图像；C ：“价态”。内在指标：群体关联的图像/文本的 SC-EAT 价态效应量。外在指标：在前 500 个检索到的图像/文本中群体识别内容的平均价态或比例。图 1 视觉化地概述了这些实验的 ITT 版本。

组对价信号传播 (2-a 和 2-b)：探讨组信号对价相关检索结果的影响。框架参数是——我们将 R : ITT 方向标记为 (2-a)，其输入为 D : OASIS 图像和组标签句子；C ： “组”。R : TTI 方向标记为 (2-b)，其输入为 D : NRC-VAD 词典句子和 CFD 图像；C ：“组”。内在指标：对价图像/文本的 SC-EAT 组关联。外在指标：前 500 个检索到的图像/文本中的平均价或与组相关内容的比例。

4 实验与结果

4.1 基于视觉语言模型的内在偏差评估

为了为我们的研究奠定基础，我们首先使用情感和基于群体的 SC-EATs 评估基于编码器的 VLMs 的表示空间中存在的内在偏见。本节后面将详细说明实验 1-a、1-b（参见图 1 以获取 ITT 设置的可视化）、2-a 和 2-b 中衡量这些偏见的实验设置，确保初步发现与其在更广泛的模型行为中的影响之间有明确的联系。我们选择评估基于编码器的 VLMs，特别是 OpenAI 版本的 CLIP-B-32、CLIP-L-14，以及 Salesforce 的 BLIP-2（编码器-解码器版本），详细说明了在我们实验中使用这些模型的选择，因为它们在社区中具有基础性作用和广泛使用（在 Huggingface 上下载超过 5000 万次 [2] ），并且能够利用系统化的方法 Steed and Caliskan (2021) 来衡量其表示中的偏见。更多的模型细节在附录 **??** 中详细说明。

我们的分析突显了 VLMs 中显著的内在偏见。具体来说，在实验 1-a 和 1-b 中获得的不同社会群体的总情感倾向 SC-EAT 效应大小显示，“黑人女性”，其次是“黑人男性”与最负面的情感倾向相关（分别为-1.43 和-1.31 的 z 分

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

数效应大小 (d))。“白人男性”和“白人女性”则类似地关联到 0.44 和 0.45 的 z 分数效应大小，而“亚洲女性”与最正面的情感倾向相关，z 分数效应大小为 1.21，表明一种强烈的基于情感倾向的表现偏见。2-a 和 2-b 展示了“白人男性”和“白人女性”群体在有情感倾向的文本和图像中被过度代表（分别为 0.48 和 0.46 的 z 分数效应大小），而“黑人女性”为-0.58 的 z 分数效应大小。详细结果，包括突显这些发现的图表，在附录 **??** 中展示。

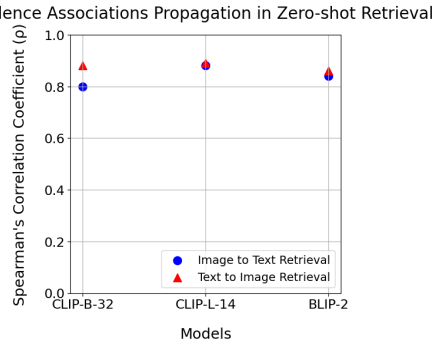

Figure 2: 实验 1*-a 和 1*-b 展示了 Spearman 的 ρ，测量了内在效价关联与外在效价结果在零样本检索任务中的相关性。

我们现在列出第 3.3 节中传播实验的主要发现。附录 C 包含更多的发现和可复现性细节。

基线价-价信号传播 (1*-a 和 1*-b)：在 1*-a 和 1*-b（图 2）中，模型在 1*-a 中表现出 0.84 ± 0.04 ρ 的强大的整体 Spearman 相关性，在 1*-b 中为 0.87 ± 0.02 ρ。此外，我们观察到较大的模型如 CLIP-L-14 一贯显示出较高的相关性（可达 $\rho = 0.88$），相比之下，较小的模型如 CLIP-B-32 的相关性为（$\rho = 0.80$）。

基线组-组信号传播 (2*-a 和 2*-b)：2*-a 和 2*-b 表明 VLMs 能够稳健地传播社会群体信号，这由所有模型中观察到的强平方级 Spearman's 相关（0.94 ± 0.04 ρ 在 2*-a 中，0.76 ± 0.11 ρ 在 2*-b 中）所指示 (图 3)。

价态到群体信号传播 (1-a 和 1-b)：图 4 中的 1-a 和 1-b 结果表明，在模型与社会群体中，斯皮尔曼相关性显著正相关，1-a 的相关性为 0.81 ± 0.10 ρ，而 1-b 的相关性为 0.78 ± 0.10 ρ。

组到效价的信号传递（2-a 和 2-b)：图 5 中的 2-a 和 2-b 结果展示了模型和社会群体之间在 2-a 中显著的正斯皮尔曼相关性，总体相关系数为 0.91 ± 0.02 ρ，在 2-b 中为 0.78 ± 0.07 ρ。

5 讨论

我们的研究首先量化了基于编码器的 VLM 中的内在偏见。接下来，在 8 个实验和 114 个[3]分析中，当前的工作一致地发现，在所有方法中，VLM 表示空间中的内在偏见与零样本 ITT 和 TTI 检索任务中的外在偏见输出之间存在正相关且普遍较强的相关性（Spearman's ρ 为 0.83 ± 0.10），其中与社会群体身份相关的系统性差异尤为显著。

自身偏见在 VLMs 中的评估。我们关于初步量化自身偏见的实验证明了交叉性别假设 (Ghavami and Peplau, 2013)，其中男女之间的偏见在“白人男性”和“白人女性”之间最为相似。此外，跨实验 2-a 和 2-b 的群体 SC-EAT 效应量表明“白人男性”和“白人女性”在 VLMs 中对有价内容的群体关联度最高，这意味着他们在模型结果中被过度代表。相反，“黑人女性”在 2-a 和 2-b 中展现出极少的自身群体关联，同时在 1-a 和 1-b 中与最负面的有价内容相关联。这与关于群体身份偏见的研究 (Devos and Banaji, 2005; Wolfe and Caliskan, 2022a) 是一致的。

基线传播情感和群体信号。实验 1*-a 和 1*-b 表明，VLM 牢固地编码并传播基本的情感信号，这表明情感在视觉和语言领域是一个重要信号 (Wolfe and Caliskan, 2022d; Toney and Caliskan, 2021)。然而，VLM 在传播不同社会群体表示时显示出明显的偏向，例如，“黑人女性”群体展示了特别一致的高相关性（例如，2*-b 中对于 BLIP-2 的 $\rho = 0.86$），而“白人女性”群体展示了较低的相关性（例如，2*-b 中对于 BLIP-2 的 $\rho = 0.53$）。这一结果可能归因于数据集中某些身份的“标记”，尤其是非默认（即非白人）群体的标记。标记和“他化”可能导致更明显的过拟合描绘，增加了传播并加重定型化风险 (Newton, 2023; Wolfe and Caliskan, 2022c)。

支持这一结果，我们发现根据谷歌 Ngram (https://books.google.com/ngrams/) 统计，过去一个世纪里“大女人”在英语中出现的频率比“白女人”更高。“白女人”较低的相关性提出了一个复杂的问题。在 AI 中，对一个人的默认类别是白人 (Wolfe and Caliskan, 2022a)。因此，女人的默认是白女人 (Wolfe and Caliskan, 2022c)。结果是，提到“白女人”时通常只包括“女人”一词。因此，我们使用输入“白女人”进行的分析可能较之“黑人女人”捕获到了较弱的表征，因为它们还没有明显到如“黑

[3]我们澄清，由于我们的价-价传播基线中没有基于群体的区分，我们进行了 114 次分析（2 个情景 * 3 个模型 + 6 个情景 * 3 个模型 * 6 个社交群体）。

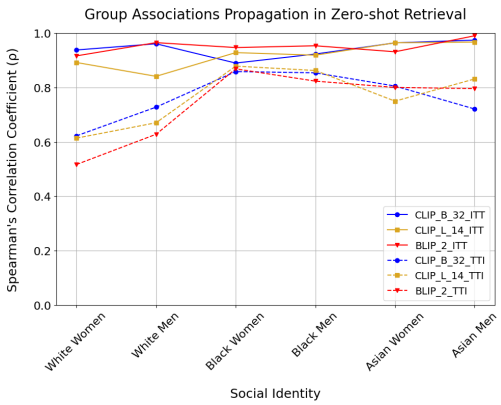

Figure 3: 实验 2*-a 和 2*-b 展示 Spearman's ρ，测量内在群体信号传播与外在群体在零样本检索任务中结果的相关性，按社会群体划分。

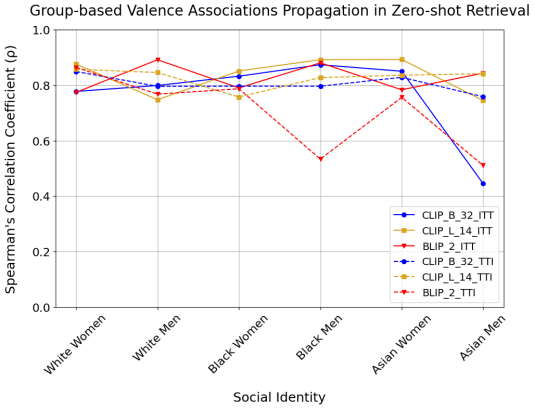

Figure 4: 实验 1-a 和 1-b 展示了 Spearman 的 ρ，它用于衡量内在效价偏差传播与外部效价结果在零样本检索任务中的相关性，按社会群体分层。

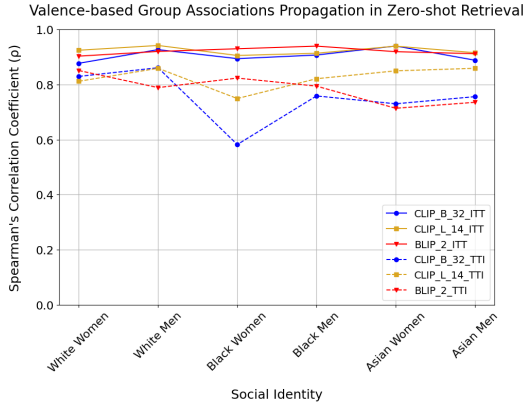

Figure 5: 实验 2-a 和 2-b 展示了 Spearman 的 ρ，用于测量内在基于效价的群体偏见传播与外在群体结果在零样本检索任务中的相关性，并按社会群体进行分层。

人女人”这样的群体以至于能够被标记出来，但它们也不像“白人男性”那样被表征为“默认”群体，导致模型对它们学习到的是欠拟合的表征。根据现有关于基于 VLM 的数据过滤方法的审计 (Hong et al., 2024)，我们将 AI 偏见分析中进一步区分社会默认值的工作留待将来研究。

我们还注意到，由于模板内容的变化和文本及图像中混杂的信号，在所有情况下的传播不是 100%，这在 TTI 中比在 ITT 中更为明显。我们观察到的普遍较高的相关性表明，当模型纯粹在它们学习的表示上运行时，特别是在没有引入可能引发新关联的任务特定微调层的情况下，嵌入在表示空间中的偏见直接体现在输出中。

在通过价和群体信号交叉传播时的偏见的多维特性。视觉语言模型 (VLMs) 可以有效地处理简单的价信号，但它们对复杂的社会群体信号的处理不一致，这可能意味着偏斜的表示和性能下降。例如，在 1-b 中，“白人女性”在所有模型中都显示出强烈的相关性，其中在 BLIP-2 中最高，达到 0.86 ρ。相比之下，“亚洲男性”和“黑人男性”在某些模型中显示的相关性较低，尤其是在 BLIP-2 中（分别为 0.51 ρ 和 0.53 ρ）。这些较低的相关性比“白人男性”以及“亚洲女性”或“黑人女性”相对较弱，表明男性与边缘化种族身份的交叉组合可能减少在 TTI 中的价-群体传播。这可能是由于在所选模型中“亚洲男性”和“黑人男性”的代表性不足。谷歌 Ngram 统计显示提到这些术语的情况远不如诸如“白人男性”的群体常见。

图表 4 和 5 突出了初步内在偏见评估中观察到的偏见如何传播到零次获取任务的微妙但重要的差异。值得注意的是，“黑人女性”，在 2-a 和 2-b 中，其群体与价内容的关联幅度最小，在图表 5 的群体与价的关联传播中显示出显著的不同结果。这种可变性可能与 VLM 中群体身份和价的复杂交叉互动的表示不一致有关。例如，较低的相关性可能表明模型无法一致地代表某些社会群体相关的内容，并影响后续表现。

模型规模与偏差传播。我们发现模型的大小和复杂性在偏差传播中起到了一定作用。如在 1*-a 和 1-a 中所示，较大的高性能模型，例如 CLIP-L-14，常常在多个零样本基准测试中表现优于 CLIP-B-32 (Radford et al., 2021)，表现出在偏差传播中的较强相关性。同样，在“白人女性”的情况下，CLIP-B-32 模型在 1-a 中显示了 0.78 的 Spearman 相关性 ρ，而 CLIP-L-14 模型的相关性更高，为 0.88 ρ，这表明随着模型的扩展，效价和群体偏差传播的程度增加。虽然这种增强的传播能力在简单的效价到效价或群体到群体任务中可提高模型的“准确性”，

但在更复杂的场景中，例如在 1-a 中，这也意味着复制和放大偏差的风险更高。这一发现与先前研究中较大的模型在标准基准测试中表现更好的语言模型具有更高的有害生成风险的结果一致 (Longpre et al., 2024)，而基于编码器的 CLIP 模型在扩展时则表现出更大的内在偏差 (Ghate et al., 2025)。因此，通过训练和基于强化学习的方法进行目标偏差缓解策略对于较大模型是必不可少的，在这种情况下，公平的内在群体表示目标优先考虑。未来的工作也可以着眼于开发灵活的学习算法，以抑制和引导模型实现公平的结果。

6 结论

本研究调查了基于编码器的视觉语言模型（VLMs）的 ITT 和 TTI 检索中的内在偏见传播。在针对六个社会群体的 114 项分析中，我们展示了偏见以高且系统性变化的相关性（平均 Spearman 系数为 0.83，标准差为 0.10）传播。更大且更复杂的模型显示出更大的偏见传播，这一发现尤为重要，因为越来越多大型 AI 模型的部署正值此时。我们的结果暗示，对于在这些模型中被错误表示的边缘化群体而言，可能会导致下游性能的不利影响。

7 局限性

所使用的图像数据集，特别是芝加哥面孔数据库（CFD）和 OASIS，为分析偏见提供了坚实的基础，因为它们提供了自我识别信息和人类评分。但是，它们远不能代表人类身份和情感刺激的全部复杂性。特别是，依赖于这些在美国创建的数据集可能无意中强化以西方为中心的视角，而未能考虑到文化差异对面部图像和情感刺激的体验影响。

在这项研究中，我们专注于三个著名的基于编码器的视觉语言模型（VLM）；CLIP-B-32 和 CLIP-L-14 用于捕捉现实的规模效应（150M 对 430M 参数），以及 BLIP-2 提供建筑多样性。即使在本研究计算限制范围之外，通过我们受控的框架，未来对更多基于编码器的 VLMs 甚至新兴模型的评估也是简单直接的。此外，我们对仅限英语的 VLMs 的关注是一个额外的限制，因为社会群体和情感值的解释在不同语言和文化之间可能会有所不同 (Charlesworth et al., 2023)。这样一种以英语为中心的视角可能会限制对 AI 偏见全球影响的理解，强调了在未来研究中需要多语言和文化多样性的研究。此外，我们的

在我们的研究中，" 传播" 的使用指的是偏差从模型表征流向下游任务结果的方向流动。这一术语由严格控制的实验设计作为支持，该实验设计将外部混淆因素最小化，从而可以清晰地观察到偏差在 VLMs 内部是如何传递的。我们在 114 种情境下的综合分析显示了内在和外在偏差之间的持续相关性，证实了由传播所暗示的方向性影响。确认这些偏差的确切来源（例如，训练数据对比构架）与测量偏差是否以及如何出现在输出中是不同的。我们的重点在于后者，即系统性地证明内在测量的偏差并不会隐藏在内在表征中，而是实际上在真实的零样本任务中表现出来。未来的研究可能会进一步沿着 Ghate et al. (2025) 等方面界定确切的因果机制，例如文本到图像的对应关系、预训练数据的组成、模型目标和结构限制等，这超出了我们当前研究的范围。利用我们研究的发现来开发去偏方法及引导模型表征变得更为公正，并在进一步的下游任务如文本到图像生成中加以验证，提供了另一个重要的探索途径。

我们测量内在关联的方法关注 SC-EAT，因为它是一种有原则且经过验证的嵌入关联技术，可以隔离一个目标类别与两组属性的关联。该方法是基于认知科学和社会心理学文献而形成的，而其他现有方法则是临时的，并且需要进一步的理论基础。此外，零样本检索是这些模型在实际应用中的主要使用场景，我们选择用我们选定的外部指标来评估。在未来的研究中，我们希望看到对机器学习任务和下游环境的探索，这可能对应其他偏差和公平性指标（例如，现有的二元分类公平性指标）。然而，我们选择的外部偏差指标与零样本检索任务中代表性损害的衡量最为一致。

最后，我们对交叉性问题的处理虽然试图解决复杂的社会身份问题，但仍然只涵盖了 6 个交叉性身份的选择。还有数百种潜在的交叉身份，包括与健康、职业、阶级等相关的身份 (Nicolas and Fiske, 2023)。未来的研究可以探索超越性别和种族的更多交叉点。然而，值得注意的是，我们用于量化偏差传播的方法是可普遍应用的，并且可以很容易地适应于分析任何群体关联。

8 伦理考量

随着 VLMs 在包括物体跟踪、机器人培训、广告和市场营销等各个行业中的广泛使用，(Briggs and Laura, 2022; Barraco et al., 2022; Bui et al., 2023; Taesiri et al., 2022)，考虑这些模型的伦理影响变得尤为重要。我们的研究强调了当前模型在公平地代表不同社会身份时的内在偏见。这种精确度上的差异可能会对某些群体如何使用和参与 AI 系统的公平性产生不利影响。

此外，一个主要的伦理问题是这些模型可能会持续对某些群体产生负面影响，如黑人和亚

洲人，尤其是交叉群体（例如，黑人和亚裔女性）。鉴于嵌入在 VLM 表征空间中的内在偏见也导致外在输出的差异，使用 VLM 处理涉及这些群体的内容任务可能会进一步边缘化和歧视这些群体。事实上，当 AI 用户获得强化（而非挑战）他们偏见的输出时，他们可能会认为这些偏见变得更可接受和规范化 (Vlasceanu and Amodio, 2022) 。

总之，我们的研究表明，预训练的视觉语言模型（VLM）中与社会群体身份和情感相关的内在表示是有偏见的，并且这些偏见会传播到下游的零样本应用中。开发和使用能够以公正和无偏见的方式准确代表所有社会身份的模型是至关重要的，以避免延续对某些群体的刻板印象和偏见，并防止进一步的歧视和边缘化。在这种情况下，一个重要的伦理考量是准确性与公正性之间的平衡。为了实现公正性，可能需要有意识地阻断来自内在表现空间的偏见传播，即使这会在某种程度上影响模型的性能。这个权衡反映了一种关键的伦理立场，即将公平和包容的价值置于单纯优化准确性之上。在开发不仅在功能上先进且与更广泛的社会价值观和道德原则一致的人工智能技术时，这种观点是必不可少的。

9 致谢

本工作得到了美国国家标准与技术研究院（NIST）60NANB23D194 号资助。本文中所表达的任何观点、发现和结论或建议仅代表作者个人意见，并不必然反映 NIST 的观点。

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.

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, 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, 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.

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.

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 数据

我们使用两种类型的图像数据集：一种代表社会群体身份，具体为种族和性别，另一种捕捉由人工标注者评价的情感（积极/消极）。我们使用情感评分作为分析偏见的一个指标，因为它在图像和文本中产生强烈且易于理解的信号 (Toney and Caliskan, 2021; Wolfe and Caliskan, 2022b) 。此外，我们使用文本数据集来补充图像数据集。下面的部分详细介绍了这些数据集的具体特征、选择依据及其在主文第 3 节中的使用方法。

A.1 芝加哥面孔数据库（CFD）。

我们选择了芝加哥脸部数据库（CFD）(Ma et al., 2015) ，因为它提供了一种可靠且受控的自我识别的人脸图像来源，提供了清晰的群体信号——与 FairFace (Karkkainen and Joo, 2021) 不同，后者包含视觉上嘈杂的图像且缺乏自我识别的标签，可能导致注释上的偏见。这些图像是在参与者同意的情况下，根据约定的条款共享的。所有个人身份信息在数据集发布之前都已被匿名化 (Ma et al., 2015) 。

CFD 数据集由 597 张自我识别的男性和女性参与者的图像组成，这些参与者属于 4 种不同的种族人口群体，即黑人、白人、拉丁裔和

亚裔。[4] 。CFD 中的照片是高分辨率且标准化的，因此人脸总是占据图像的同一部分，并且它们都放置在一个白色背景下。被试者被拍摄时具有各种面部表情（快乐、愤怒和恐惧）以及面对摄像机时的中性表情。为了与 Ho et al. (2011) 和 Wolfe et al. (2022) 保持一致，我们仅使用具有中性面部表情的图像。为了获得一个具有显著样本量的数据集，我们首先通过创建组内图像变体 (Wolfe et al., 2022) 来扩展数据集，如下所述。然后不放回地随机抽取图像。

A.1.1 生成变形图像

首先，我们在 CFD 数据集中的每个人口统计群体内创建了所有可能的图像对。例如，对于黑人男性的人口统计群体，大约有 100 张可用的图像。我们为每对可能的图像创建了 50 个% 变形，这导致每个人口统计群体大约有 5,000 个变形图像。

为了生成真实的变形图像，我们采用了在高质量 FFHQ 数据集上预训练的最先进的 StyleGAN2-ADA3 架构。为了规范化图像，我们在面部区域周围进行裁剪，确保面部特征的位置与预训练数据集中的位置相当。

使用 StyleGAN2-ADA3，我们在每一对的源图像和目标图像上训练生成器以生成变换序列。给定我们提供给 GAN 的标准化、高分辨率照片，我们使用 StyleGAN2-ADA3 的默认超参数对每张图像训练 125 次迭代，此后不再显示显著效果 (Wolfe et al., 2022) 。对于每一对，训练后的生成器创建一个源嵌入和一个目标嵌入。随后，我们使用这些嵌入生成每个变换序列的第一张和第二张图像。

最后，我们从 CLIP 中获取每个变形后的投影图像嵌入。这产生了一个包含约 30,000 个变形图像及其对应嵌入的数据集，每个嵌入都是独特且多样的。请注意，由于版权问题，我们无法在附录中共享 CFD 生成的样本，但可以根据请求提供。

CFD 的操作化：我们随机选择每个交叉组的 1,000 张图片（这个数字是基于其显著的样本规模和计算上的可行性而选择的），不重复，结果得到总共 6,000 张图片。在 CFD 用于 SC-EAT 组的实验中，每组选择 140 张图片，包括 SC-EAT 组中特定兴趣的两组属性集合共 840 张单独的图片。在文本到图像（TTI）检索中，使用所有 6,000 张 CFD 图片以保持平衡。当 CFD 用作情感 SC-EAT 和图像到文本（ITT）检索的目标数据集时，也使用所有 6,000 张图片。

[4]由于围绕拉丁裔面孔的种族感知存在已知的模糊性，目前的研究仅专注于黑人、白人和亚裔种族团体 (García, 2013)

我们想指出，GAN 生成步骤严格用于组内的数据增强，尽量减少引入关于种族或性别的伪线索的风险。虽然生成的图像有时可能引入伪影，我们的方法谨慎地将变形限制在一个单一的人口群体内，以确保任何潜在的伪影最小化，并且不会混淆种族和性别类别。

A.2 开放情感标准化图像集（OASIS）数据库。

Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017) 包含 900 张彩色图像，描绘了广泛主题，如人类、动物、物体和场景。这些图像由 822 名参与者根据效价（积极或消极的程度）和唤醒度（效价反应的强度）进行评分。当前实验中使用这 900 张图像来提供来自自然图像输入的非群体相关的效价信号。所有图像均通过类似于 Wolfe et al. (2022) 的缩放/裁剪过程标准化为 500 X 400 像素。

OASIS 的操作化在需要 OASIS 进行组 SC-EAT 或 ITT 检索的实验中，全部 900 张图像都被使用。在涉及效价 SC-EAT 的实验中，我们使用效价评分排序中前 25 个愉快和前 25 个不愉快的 OASIS 图像作为两个属性集。对于 TTI 检索，使用全部 900 张图像。利用整个词汇和效价范围，使我们能够提供关于偏见传播的全面见解，其影响延伸到检索结果的真实性。

我们的研究需要受控和平衡的文本数据集，以理解文本内容的心理语言学和社会群体特性如何影响视觉语言模型中的偏见。具体来说，我们希望分析具有已知心理语言学真实评分或属于我们感兴趣的六个群体之一的特定交叉性身份的词汇和短语如何影响这些模型中偏见的传播。在这项研究中，我们采用了 May et al. (2019) 和 Tan and Celis (2019) 的句子模板方法，使用“语义漂白模板”。这些模板旨在保持语义中立，这意味着它们不提供新的语义信息，但确保目标词组被放置在类似的句法框架中。该方法允许与目标词相关的值传达句子的心理语义特征。

“目标词”指的是插入到模板中的特定单词，旨在引发心理语义反应。例如，在模板“这是单词 [WORD]”中，目标词替换“[WORD]”，并作为我们实验中的主要变量。在我们的研究中，我们使用了衍生自 May et al. (2019) 的 6 个句子模板，如表 1 所示。

对于涉及使用效价评分句子的实验，我们采用 NRC-VAD 心理语言词典 (Mohammad, 2018) ，该词典包括大约 20,000 个英语单词，每个单词由 6 名参与者在效价、唤醒和支配（VAD）三个关键维度上进行评分。这个全面的词典为我们提供了一个基于可靠的、由人类评分的心理语言数据的稳健词库。

对于涉及群体信号的句子的实验，我们使用来源于 Charlesworth et al. (2022) 的精选群体标签列表。在这里，目标词是一个种族识别词后接一个性别识别词的组合。一个例句是，“这是一个词，黑女人”。群体词的完整列表在表 1 中。

对 NRC-VAD 词典句子的操作化我们在六种句子模板中使用了 20,000 个单词，所有结果都是在模板间平均的。当需要 NRC-VAD 词典用于组 SC-EAT 或 TTI 检索时，每个模板的 20,000 个句子全部使用。在需要效价 SC-EAT 的实验中，按效价评分排序，选择最愉快的前 25 个句子和最不愉快的前 25 个句子作为两个属性集。对于 ITT 检索，每个模板的 20,000 个句子都使用。词典和效价光谱的全面使用提供了有关偏差传播及其更广泛影响的关键见解。

群体标签句子的操作化

我们使用来自六个句子模板的 5,184 个句子，并对模板进行平均。当团队标签句子成为效价 SC-EAT 和 TTI 检索的目标数据集时，使用所有 5,184 个句子。对于团队 SC-EAT，在团队标签句子测量团队关联的情况下，每个团队随机选择 140 个句子，包括来自 SC-EAT 目标团队的 840 个单独句子，用于两个属性集。对于 ITT 检索，利用所有 5,184 个句子。

根据主文第 5 部分，以下详细说明了我们实验中使用的开源模型的选择。

视觉-语言模型——CLIP（B-32 和 L-14）。Radford et al. (2021) 使用 WebImageText 语料库（WIT），该语料库由 4 亿张图像和相关标题组成，用于训练 CLIP。Radford et al. (2021) 创建的查询列表利用了在英文维基百科中至少出现 100 次的所有词，以及来自维基百科的点互信息高的词二元组、维基百科文章标题以及所有 WordNet 词集。作为创建过程的一部分，他们寻找文本中包含 500,000 个查询之一的图文对，以尽可能覆盖尽可能多的视觉概念。CLIP 的重要性，特别是其 B-32 版本和大得多的 L-14 版本，体现在其零样本学习能力上（在 ImageNet (Radford et al., 2021) 上分别达到 63.2 % 和 76.2 % 的零样本准确率）。这些模型可以处理和解释图像及其关联文本，而无需特定任务的训练，展示了一种高度适用于各种场景的泛化能力 (Briggs and Laura, 2022) 。

BLIP2。BLIP-2 (Li et al., 2023) 是一个视觉-语言预训练模型，它集成了现成的冻结图像编码器和 LLM。它包括一个查询变压器（Q-Former），通过两个阶段进行训练：从冻结图像编码器进行视觉-语言表示学习以及从冻结 LLM 进行视觉到语言的生成学习。它在一个包含 1.29 亿图像的数据集上进行训练，其中包括 COCO、Visual Genome、CC3M、CC12M、SBU 和 LAION400M 数据集的子集。我们研究 BLIP-2 是因为它在零样本任务中具有高效和高性能表现。例如，它在 VQAv2 (Li et al., 2023) 上达到了 65 % 的最新零样本准确率。

B 方法

以下部分详细阐述了从 Caliskan et al. (2017) 和 Steed and Caliskan (2021) 获取的 SC-EAT 的定义和公式。

C 实验细节

本节补充了正文第 5 节中提到的结果，并包含了所有实验可重复性的详细信息，包括使用的具体数据集和参数设置。图 8 包含了 8 个偏差和关联传播实验中每一个的复现细节。

图 6 和 7 突出显示了 VLM 中显著的内在偏差。图 8 展示了在实验 1-a 和 1-b 中获得的不同社会群体的基于价值得 SC-EAT 效应量的总和。图 9 则展示了在实验 2-a 和 2-b 中高（价值得 ≥ 0.5）和低（价值得 < 0.5）价值段的群体基础 SC-EAT 效应量的总和。

C.1 偏差通过价态和群体信号的传播

内在价联想向外在价结果的基线传播。

本实验作为一个基准，用于衡量文本中的情感信号和图像中的情感信号在 VLMs 中是如何传播的。我们遵循前人的研究，例如 VAST (Wolfe and Caliskan, 2022d) 和 ValNorm (Toney and Caliskan, 2021)，使用内在情感信号来评估词嵌入和语言模型的内在质量。首先，对于图像到文本的方向 (1*-a)，我们测量图像的内在情感与基于该图像检索到的前 500 句子的平均情感之间的相关性。例如，一个具有高内在积极情感的图像（例如玫瑰、蝴蝶等）理想情况下应该检索到具有类似积极情感的文本（例如，“这是一个美丽的词”）。这里的内在测量是 50 个情感最强的 OASIS 图像（前 25 个积极和前 25 个消极）的 SC-EAT 得分；这里的外在测量是与特定输入 OASIS 图像相关联的前 500 个句子的平均情感评分（来自人类情感评分）。

其次，文本到图像 (1*-b) 方向反映了 1*-a，并探讨文本中的内在效价信号如何与基于文本输入检索到的图像的聚合效价相关联。具体而言，内在测量是句子中效价最高的 50 个词 (前 25 个正面词和前 25 个负面词，如 A 节所述) 的 SC-EAT 得分；此处的外在测量是与输入句子相关的前 500 张 OASIS 图像的平均效价评分（来自人类效价评分）。

基线传播内在群体关联到外在群体结果。这组实验旨在建立理解群体内容简单传播的基线。

Table 1: 用于在 VLMs 中进行偏见测量的句子模板和组词。句子模板取自 May et al. (2019)，组词基于 Charlesworth et al. (2022) 选择。这些元素构成了我们研究中用于调查社会偏见传播的文本数据集的基础。

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

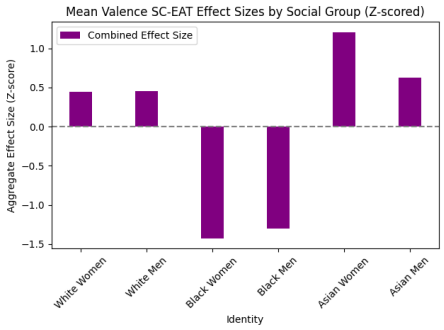

Figure 6: 基于效价的内在偏见通过 SC-EAT 效应量可视化：该图展示了实验 1-a 和 1-b 中针对不同社群获得的 SC-EAT 效应量的归一化平均值。负效应量表示该群体与负效价的关联比正效价更强。数据显示模型内部存在的内在偏见，尤其是，黑人女性被展现为与负效价关联最强的群体，这由最负的效应量所指示。相比之下，亚裔男性和女性展示了最弱的负关联。这以图形方式展示了在 VLMs 的表征空间中内在偏见的普遍存在，各群体所关联的负效价程度不同。

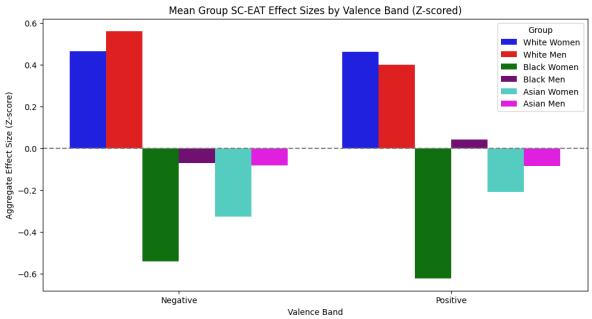

Figure 7: 通过 SC-EAT 效应大小的基于群体的内在偏见可视化：此图展示了在实验 2-a 和 2-b 中获得的高/低效价带的 SC-EAT 效应大小的标准化均值。我们将效价 ≥ 0.5 的图像/文本分类为正向，否则为负向，范围为 0 到 1。正向效应大小表示效价项目与特定群体相比于任意群体具有更强的关联。数据显示模型内在的偏见，特别是白人男性和女性被描绘成极高的正向和负向关联。相比之下，黑人女性表现出与效价图像/文本关联度的最低幅度。这图形化地展示了在 VLMs 的表征空间中存在的基于群体的内在偏见。

在图像到文本的设定中 (2*-a)，我们测试图像中特定社会群体的表现与检索到的文本中的群体表现之间的相关性。这里的内在测量是图像-文本群体 SC-EAT，我们发现给定图像（来自 CFD）与包含单个交叉群体的句子（与表示所有交叉群体的句子相比）之间的关联。此方法本质上类似于一对全部的机器学习验证任务，其中评估模型在表现层面上区分一个群体身份与其它群体身份的能力。在外部，对于相同的 CFD 图像，我们获得检索到的群体识别句子的比例，这些句子是"正确的"（即与文本提示相同的交叉身份）。例如，如果输入图像是"黑人女性"，那么外在评估检索到的文本准确反映黑人女性概念的频率。

在文本到图像的方向（2*-b）中，我们测试代表某个社会群体的文本是否能从 CFD 中检索到具有相同群体表现的图像。内在度量是文本提示与特定群体信号之间的文本-图像 SC-EAT，以及代表单一交叉群体身份的图像（对比于包含所有交叉群体的图像的均匀性分布）。从外部来看，我们记录为文本提示"正确"代表特定社会群体检索的图像的比例。为清楚起见，此处的"正确"意味着检索到的图像准确反映了文本中提到的特定社会群体身份。例如，如果文本提示是关于"亚洲男性"，则外在度量评估检索到的图像中有多大比例确实描

	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: 图详细说明了用于展示内在偏见传播到外在零样本检索任务的 8 个实验的具体情况。实验共享一个用于测量内在偏见的共同框架（例如，偏见效应大小，它捕捉了一个社会群体图像与负面价值句子的关联强度相对于正面价值句子的差异），随后进行外在偏见的评估（例如，针对特定群体图像提示检索到负面价值文本的可能性），然后将这两个指标进行关联。这些实验在偏见传播方向（图像到文本四项，文本到图像四项）和分析的内容类型（四项关于正负情感或正负评级，四项关于群体信号）方面有所不同。

绘了亚洲男性。

与所有其他研究一样，我们使用图像/文本的内在 SC-EAT 组关联与每组正确检索的句子/图像的外在比例之间的 Spearman 相关值来量化内容/组传播。

内群体基于内隐的效价关联向外显效价结果的传播。该实验是我们关于偏见传播的主要分析，因为它关注群体信号（在文本或图像中）与效价属性之间的偏见关联如何通过 VLM 传播。对于从图像到文本的方向（1-a），我们考虑图像中的内隐群体信号如何影响检索到的文本的效价。图像-文本 SC-EAT 测量属于某一群体身份的每个图像的内隐效价效果大小（来自 CFD）。然后，在外显方面，对于同一图像，我们计算检索到的前 500 句子的平均效价，并在模板上取平均值。

在文本到图像的方向（1-b）中，我们研究与社会群体相关的文本如何影响检索到的图像的情感。内在地，通过文本-图像 SC-EAT 来测量，这是通过对每个包含交叉群体身份的句子计算情感关联的效应量，使用情感评级的图像（来自 OASIS）。然后，外在地，对于同样的句子，我们计算检索到的情感评级最高的 500 张 OASIS 图像的平均情感值。

内在基于效价的群体关联对外在群体结果的传播。最后，与我们对群体到效价传播的主要分析相并行，我们考虑效价到群体的传播，或更具体地说，我们研究通过 SC-EAT 测量的图像/文本的内在效价关联如何导致群体信号的差异性检索。在图像到文本的方向（2-a）中，我们研究图像中更强的内在效价评级是否与更高的可能性相关联，以检索与边缘化群体相关的文本。本质上，我们使用图像-文本群体 SC-EAT 来找出给定的效价图像（来自 OASIS）与包含属于单个交叉身份（相对于均匀代表所有交叉群体的文本）的群体术语的文本的关联。外在上，我们获取与相同的 OASIS 图像对于每个 6 个交叉社会群体检索到的群体标识句子的相应比例。

相反，在文本到图像方向（2-b）中，我们测试带有强烈情感评分的文本是否检索出主要以某些社会群体为特色的图像。内在衡量标准是给定情感文本提示与来自特定社会群体的图像（相对于通过随机平衡抽样选择的所有群体图像）之间的 SC-EAT 关联。从外部来看，我们计算为同一个交叉群体文本提示检索到的代表

某一社会群体的图像比例。