DOI: 10.1111/cgf.70135
Eurographics Conference on Visualization (EuroVis) 2025
W. Aigner, N. Andrienko, and B. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*

Interactive Discovery and Exploration of Visual Bias in Generative Text-to-Image Models

Johannes Eschner[1], Roberto Labadie-Tamayo[2], Matthias Zeppelzauer[2] and Manuela Waldner[1]

[1]TU Wien, Austria
[2]St. Pölten University of Applied Sciences, Austria

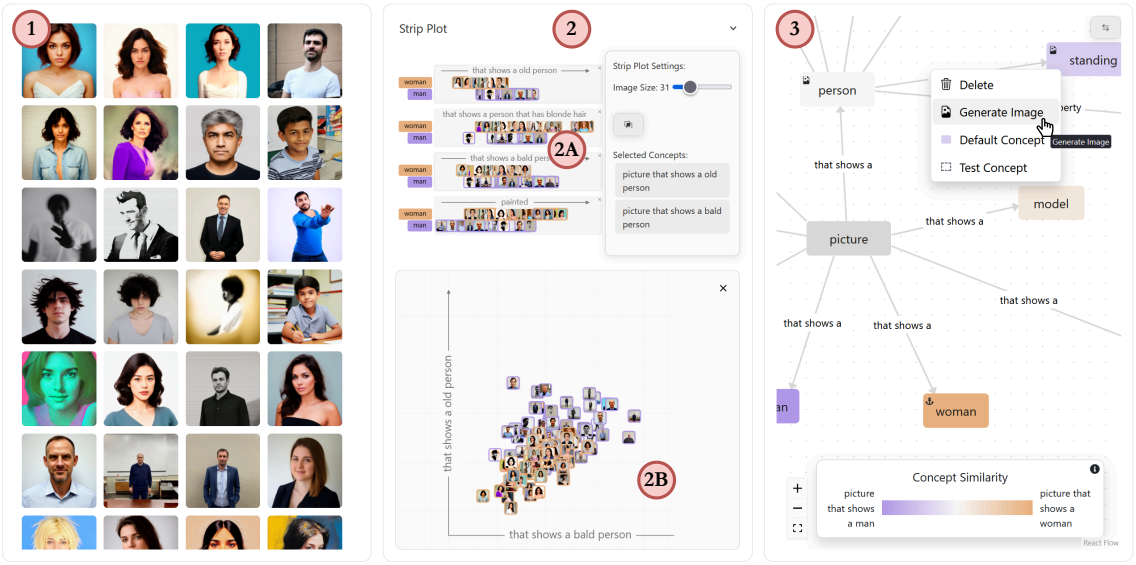


Figure 1: *ViBEx interface:* ***data view*** *(1) with a collection of generated images corresponding to concepts in (3).* ***Bias visualizations****: (2)* ***strip plots*** *(2A) showing image-to-text similarities for selected test concepts and* ***intersectional bias scatterplot*** *(2B) showing the bivariate distribution of two intersected test concepts.* ***Prompting tree*** *(3) for externalization of approximate visual bias and as a prompting interface.*

Abstract

Bias in generative Text-to-Image (T2I) models is a known issue, yet systematically analyzing such models' outputs to uncover it remains challenging. We introduce the Visual Bias Explorer (ViBEx) to interactively explore the output space of T2I models to support the discovery of visual bias. ViBEx introduces a novel flexible prompting tree interface in combination with zero-shot bias probing using CLIP for quick and approximate bias exploration. It additionally supports in-depth confirmatory bias analysis through visual inspection of forward, intersectional, and inverse bias queries. ViBEx is model-agnostic and publicly available. In four case study interviews, experts in AI and ethics were able to discover visual biases that have so far not been described in literature.

CCS Concepts

• ***Human-centered computing*** → ***Visualization;*** • ***Computing methodologies*** → ***Artificial intelligence;***

1. Introduction

Generative Text-to-Image (T2I) models have been shown to amplify demographic stereotypes and perpetuate social biases in their outputs [NN23, LAMJ23, WLD*23, VAHM23, BKD*23]. Prior work has, for instance, observed biases in the depiction of gender and ethnicity across various occupations [WSO*24]. Tradi-

tional approaches to bias exploration of T2I systems mostly focus on statistically analyzing known or anticipated biases [WLD*23, DPM*24, MLL23]. Similarly, bias mitigation methods focus on reducing pre-identified biases in T2I outputs [Nic, DPM*24]. Next to mitigating already discovered biases, the discovery and thorough documentation of previously unknown bias dimensions is essential for creating fairer and more balanced models. Automated methods may uncover new bias dimensions [DPM*24, CSB*24] but may also reveal unrelated entanglements, miss certain biases, or identify biases differently than a human observer. Therefore, we argue that human oversight is essential to determine what constitutes a bias, especially when it comes to different socio-cultural contexts.

Visual Analytics (VA) systems for interactive bias exploration mainly exist in the context of word embeddings [GHM21] and text representations [KLZ24, HGE22]. For T2I models, however, there is a gap when it comes to the interactive bias analysis and discovery of new bias, as previous work focuses either on fully automatic and non-interactive pipelines [DPM*24, LZL*24] or on interactively showcasing known biases [Luc23]. This gap may be rooted in the difficulty of providing an interactive system given the higher computational demands of image generation systems over text models.

In this work, we focus on how bias manifests visually in AI-generated images. Our aim is to support both the *exploration* of potentially biased concepts to discover unexpected bias and the more in-depth *confirmation* of expected or observed visual bias. To this end, we present a workflow for interactive bias discovery together with a reference implementation proposing tailored visualization components to showcase bias as well as performance and (expert) user studies of our approach. Based on this workflow, we propose ViBEx, a novel visual bias exploration approach to support the discovery of bias through interactive exploration and confirmation. ViBEx enables efficient, flexible analysis through several unique interface components. We discuss and show how these components ensure that ViBEx meets the key requirements for exploratory visual analysis systems, such as flexibility and real-time interactivity. In summary, our main contributions are:

1. The conceptualization of a user-driven visual bias exploration workflow (ViBEx), incorporating a novel interaction and visualization design to facilitate flexible, trustworthy, and rapid exploration and confirmation of biases in T2I systems.
2. A first model-agnostic reference implementation of the ViBEx workflow as a publicly available online resource to inspect visual bias in a real-time web application: `https://vibex.jde.cg.tuwien.ac.at`
3. A set of previously unidentified visual biases in Stable Diffusion 3 (SD3) [EKB*24]. These were discovered in case studies by experts in the fields of AI, media ethics, and digital humanism using ViBEx and validated via established bias quantification.

2. Background and Related Work

We follow the formal bias definition by D'Incà et al. [DPM*24] who consider a model as *unbiased* if, in a class-agnostic context t, the set of classes C exhibit a uniform distribution. Conversely, a generator is *biased* if it is more likely to produce content of one class $c_i \in C$ (e.g., *"man"*) given a neutral prompt t (e.g., *"A picture of a doctor"*). In our work, we assume that users explore potential bias with respect to *concepts*. We will refer to the classes C to be inspected as *anchor concepts* and to the context t as *test concept*. *Intersectional* bias occurs when biases associated with multiple test concepts interact to create unique representational harms [CEH*19, TC19]. Biases in T2I outputs are grounded in social harms (e.g. under- or misrepresenting groups) [WSO*24], while entanglements are *incidental correlations* without inherent harm [CSB*24]. For example, the concept *"man"* may be predominantly depicted with a beard. We argue that, ultimately, it should be the user who decides what constitutes a bias versus an expected entanglement.

2.1. Bias Measures and Mitigation

Bias evaluation in T2I systems generally involves creating curated datasets and applying bias evaluation metrics. Curated datasets are designed to target specific biases by carefully selecting prompts or using datasets like FairFace [KJ21], Flickr 30k [YLHH14] or MS COCO [LMB*14], which provide image captions suited for demographic or contextual prompts. Naik and Nushi [NN23] highlight social biases using demographic prompts, while others compare output distributions to training datasets, revealing amplification or reversals of biases [FBS*23]. Typical bias measures include classification-based methods, such as analyzing feature frequencies using vision-language models like CLIP (Contrastive Language-Image Pretraining) [RKH*21] or Visual Question Answering (VQA) [ZJTY23], and embedding-based approaches that assess association scores in embedding space [MLL23, WLD*23].

Prior work on bias measures for T2I systems mainly focuses on *gender*, *race*, and *skin tone*. Luccioni et al. [LAMJ23] find that popular T2I systems underrepresent marginalized identities. They highlight these biases by comparing profession-based image prompts to U.S. labor statistics. Similarly, Naik and Nushi [NN23] show that DALL·E 2 [RPG*21] and Stable Diffusion v1 [RBL*22] amplify social biases in gender, ethnicity, age, and geography, often depicting underrepresented regions in adverse conditions. Recent work expands these known biases by incorporating appearance attributes (e.g., grooming, accessories) [LZL*24] and factors like activity, object size, and emotion [DPM*24].

To minimize the presence of known biases in T2I outputs, different mitigation strategies are employed. The majority of bias mitigation happens at inference time by modifying prompts [WSO*24]. While this allows for correcting existing models, the approach is not fully controllable and may lead to model over-correction [WC24]. Furthermore, prompt modification can only work for known biases. Similarly, mitigation strategies that modify the model weights require prior knowledge of biases. The bias mitigation in DALL·E 2 [Nic] balances the data in the training dataset to compensate for previously known biases.

2.2. Bias Exploration Systems

Interactive VA systems for bias exploration are mostly studied in the context of textual representations, e.g., word embeddings. FAIRVIS [CEH*19], for example, uses multiple coordinated views to showcase model performance for different subgroups (e.g. races)

in the outputs of a classifier. WordBias [GHM21] uses word embeddings to visually explore intersectional bias by intersecting subgroups (e.g. *"male"* and *"Islam"*).

In contrast to text, in the field of image generation, bias exploration systems are either limited to pre-defined prompts [Luc23] or have limited interactivity [DPM*24,LZL*24,CSB*24]. Bias exploration systems, such as OpenBias [DPM*24] or TIBET [CSB*24], operate autonomously with an open set of biases created via LLM-generated prompts. The LLM-based biases (obtained by prompting the model to provide a list of potential biases), along with images generated from a prompt database, are fed into a VQA system. This system analyzes the frequency with which a proposed bias occurs in the images, resulting in the final bias assessment. While this approach allows for arbitrary biases to be inspected, it relies on an LLM (a non-user-controlled black box that itself may be biased) to have an understanding of the socio-cultural contexts from which bias may arise.

Beyond automated systems, interactive approaches have been proposed for model explainability. Human-in-the-loop interfaces for concept discovery [WLG23, ZXSR22, HMKB22] focus on explaining the model behavior of large vision-language models, though their analyses are typically performed on static, non-generative datasets. VLSlice [SKL23] offers an interactive interface for open-ended bias discovery in vision-language models by leveraging scatterplot-based visualizations to explore bias tendencies. However, it is limited in its flexibility as it only allows for limited human control over the bias dimensions. Next to the main prompt only a single initial baseline text is provided by the user. Furthermore, while it also utilizes a vision-language model for computing image-text similarities, there is no direct mechanism for users to investigate the trustworthiness of the model outputs.

2.3. User Interfaces for T2I Systems

A core contribution of ViBEx is its novel prompting interface. To put it in context, we survey different approaches to interacting with T2I systems. The traditional text-to-image pipeline involves handwritten prompts provided by the user to the T2I system. This has led to research on how to effectively convey user intent to the image generator (i.e. prompt engineering) [LC22, MGSM24, Opp23]. A common approach to prompt engineering is to employ LLMs for prompt generation or prompt expansion, paired with a visualization of the generated images in some embedding space [FWW*23, BWS*23] or with direct manipulation of model attention for image refinement [WHS*24]. Alternatively, the prompting journey of the user is visualized as a graph that contains the previous output images in conjunction with weights showcasing the influence certain words have on a prompt [GSL*24]. While the above-mentioned VA systems rely on (augmented) text input, some other approaches propose different input modalities. Almeda et al. [AZPK*24], for example, employ a spreadsheet-based interface with commands for stylization or automatic prompt modification.

Using graphs as an input modality has been explored for LLM prompting. Sensecape [SMPX23], an interface for sensemaking and exploration via LLMs, uses a graph-based layout to organize prompt variations for complex information tasks. Similarly, ChainForge [ASV*24] provides users with a graphical, node-based interface for testing LLM robustness. In the context of T2I systems, however, graphs are so far only used as an intermediate data structure for scene organization via scene graphs [SMS*23, FYC*23] and not as a direct input modality. To the best of our knowledge, the combination of an interactive graph-based prompting interface with bias exploration is a novel approach.

3. ViBEx

We introduce ViBEx by first discussing design challenges (Section 3.1) and the anticipated workflow (Section 3.2). Based on this, we describe our bias probing approach (Section 3.3), and the application design (Section 3.4).

3.1. Design Challenges

During exploratory visual analysis, users iteratively question the data, visually inspect the data, and refine their questions and hypotheses accordingly. Here, we list the most essential general characteristics of exploratory visual analysis systems, as discussed by Battle and Heer [BH19], and elaborate on how they pose design challenges for exploratory analysis of visual bias.

C1: Flexibility. Exploratory visual analysis systems need to support systematic and flexible queries so that all possible measures, dimensions, and features of the dataset can be studied by the user [PS08]. For visual bias exploration, this means that users should be able to flexibly test concepts that go beyond a preselected set of prompts. A special challenge, thereby, is that any user queries need to be correctly interpreted by the system.

C2: Real-time interactivity. None of the previous works on bias in T2I systems report to operate in real-time, thereby not supporting the rapid pace of exploration needed in interactive visual exploration systems [BH19]. The main challenge here is to develop a strategy for efficient evaluation of a large number (i.e. hundreds) of images in real-time. It has been shown that immediate approximate visual feedback is preferable over a slow and blocking visualization interface [ZGC*17]. Thus, it is essential to show at least an approximation of a response to the user's bias query without noticeable latency.

C3: Trustworthiness. Understanding the correctness of the visualized data is crucial to exploratory VA systems [AZL*19]. This notion of trust is especially important when designing applications for generative AI [WHM*24]. Bias metrics are not always fully reliable since the bias evaluation system itself may be biased [WSO*24, CSB*24] or because the sample based upon which the bias is computed is not representative [FPNK22]. In addition, a systematic visual difference observed for a test concept t may not necessarily constitute a bias [DPM*24, BBIW20]. In the end, it is up to the user to decide which concepts represent a bias [BBIW20]. Following a quick approximate response to a user's bias query ($\rightarrow$ **C2**), the user, therefore, needs to be able to verify the system response through deeper inspection.

C4: Generalizability. Any exploratory VA system should generalize beyond a single dataset. Similarly, ViBEx should be capable of analyzing images from different T2I models.

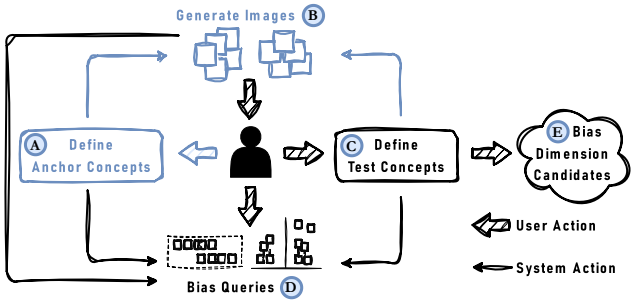


Figure 2: *The ViBEx workflow: Users define anchor and test concepts and subsequently perform bias queries, resulting in bias candidates. Elements in blue represent a non-real-time operation.*

3.2. The ViBEx Workflow

The first step of the ViBEx workflow (Figure 2) is to (A) define two **anchor concepts** c_1 and c_2, with $c_j \in C$ (e.g., *"woman"* and *"man"*) between which to test for systematic visual differences. We generate n sample images per anchor concept (A) → (B). In practice, we chose $n = 50$ because prior work used the same number of images [NN23] and our bias measures can be computed reasonably fast using this number (see Section 5.2 for performance measures). The image generation is a blocking operation that is currently not possible in real-time due to its high computational demands. However, all actions up to this point are performed only once per exploration session.

Users start their exploratory analysis by inspecting the generated images in the **data view** (Section 3.4.1) and iteratively adding, refining, or potentially removing **test concepts** $t \in T$ that describe visual properties of the image content in the **prompting tree** (Section 3.4.2) (C). For example, a test concept could state that the person depicted in the image is smiling. ViBEx automatically visualizes an approximate bias quantification based on **zero-shot bias probing** (Section 3.3) (C) → (D). Users perform in-depth confirmatory analysis by visually comparing the zero-shot similarity distributions between images representing the two anchor concepts and selected test concepts in different visualizations (Section 3.4.3) (D).

Since our workflow builds upon image samples representing anchor concepts, there is a trade-off to consider: trustworthiness (**C3**) vs. real-time interactivity (**C2**). The more images, the more reliable the bias quantification, but the longer they need to compute. In some cases, the test concept of interest may be underrepresented or not even depicted at all in any of the generated anchor concept images. For example, in Figure 5, the user tests for *"black person"*, however, no persons with black skin color are present in the anchor concept images. Thus, for testing T2I-models, ViBEx also supports inverse bias queries: instead of testing whether a test concept t is biased towards one of the anchor concepts in C, we let users generate m images representing a test concept (e.g., *"An image of a smiling person"*) (C) → (B) and, finally, visualize, for each generated test image, the relative similarity to the anchor concepts (B) → (D). Practically, the number of generated images m representing the test concept will be considerably smaller than n to reduce waiting times. For ViBEx, we chose $m = 5$. Based on the insights gained from the bias visualizations the user finally forms a selection of bias dimension candidates (E) as a final output.

3.3. Zero-Shot Bias Probing

Our approach to bias quantification is built upon measuring the compatibility between a (generated) image $\mathcal{I}$ and a natural language text $\mathcal{T}$ representing a concept. To achieve this, ViBEx leverages the contrastively pre-trained model CLIP [RKH*21], which aligns visual and textual information in a shared embedding space.

Let $\mathbf{e}_\mathcal{I}$ and $\mathbf{e}_\mathcal{T}$ be two d-dimensional embeddings for an image $\mathcal{I}$ and a text $\mathcal{T}$, respectively. We define the similarity between these two embeddings as a normalized dot product:

$$s(\mathcal{I},\mathcal{T}) = \frac{\mathbf{e}_\mathcal{I} \cdot \mathbf{e}_\mathcal{T} + 1}{2}. \tag{1}$$

A similarity value of 0 indicates no correlation, while 1 denotes perfect alignment.

ViBEx supports two types of bias queries. For **forward bias queries** (FBQs), we compute the similarity $s(\mathcal{I}_{jk}, t)$ between a textual test concept t (e.g., *"An image of a smiling person"*) and the k'th image associated with anchor concept c_j (e.g., generated from the prompt *"An image of a woman"*). For **inverse bias queries** (IBQs), we compute $s(\mathcal{I}_{tk}, c_j)$ for the k'th generated image representing the test concept t (e.g., generated from the prompt *"An image of a smiling person"*) against a natural language representation of anchor concept c_j (e.g., *"An image of a man"*). These similarity values serve as a foundation for our bias visualizations for in-depth confirmatory visual analysis based on images (Section 3.4.3).

To compute an approximate bias of a test concept t towards an anchor concept c_j based on forward bias queries, we use Bayes' theorem, which expresses the probability of the anchor concept c_j to be true under the condition that the test concept t is true:

$$P(c_j \mid t) = \frac{P(t \mid c_j) \cdot P(c_j)}{P(t)}. \tag{2}$$

Since we generate the same number of n images per anchor concept, the prior is uniformly set to $P(c_j) = 1/|C|$. Therefore, with two anchor concepts $P(c_j) = 0.5$. The likelihood $P(t \mid c_j)$ is computed as the average similarity between test concept t and all images generated for anchor concept c_j:

$$P(t \mid c_j) = \sum_{k=1}^{n} \frac{s(\mathcal{I}_{jk}, t)}{n}. \tag{3}$$

The evidence $P(t)$, finally, describes the overall average similarity of the test concept t across all $2 \cdot n$ generated images. Intuitively, if the difference $|P(c_1 \mid t) - P(c_2 \mid t)|$, is small, then the test concept t does not strongly favor one anchor over the other, suggesting lower bias. We use Equation 2 to instantly show test concept biases directly in the prompting tree via color coding (Section 3.4.2).

Zero-shot bias probing tackles multiple design challenges: 1) flexibility to formulate any test concept as a query to the set of anchor images (**C1**) and 2) foundation to provide instant bias quantification based on a simple dot product computation (**C2**). Since CLIP can encode any input image, it is independent of the T2I model. Thus, zero-shot bias probing is also 3) a model-agnostic approach (**C4**).

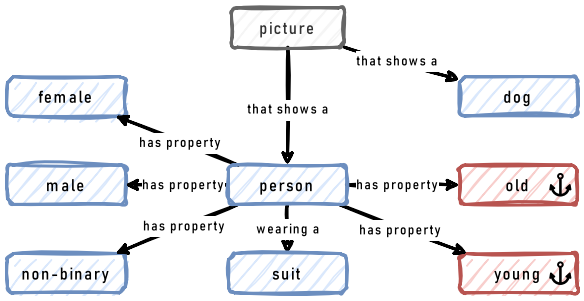


Figure 3: *Schematic overview of the prompting tree. The root node* "picture" *is part of all prompts. Anchor concepts (in red) represent the classes* $c_1, c_2 \in C$*, for which we probe different test concepts* $t_i \in T$ *(blue) for potential bias. The relation type between two concepts is indicated by the edge label. From this tree we may parse prompts such as* "picture that shows a young person" *or* "picture that shows a female person wearing a suit".

3.4. ViBEx Application Design

To showcase the real-world applicability of ViBEx, we built a reference implementation for a bias exploration interface in line with the ViBEx workflow. The web-based interface consists of multiple coordinated views (see Figure 1): the data view, the bias visualizations, and the prompting tree. Adding a new test concept automatically refreshes all visualizations and newly generated images instantly populate relevant views. Brushing and linking across all components enable quick filtering and highlighting. Additionally, users can switch between sessions with different data sources, supporting diverse scenarios across multiple T2I models (**C4**).

3.4.1. Data View

The data view (Figure 1 (1)) is a simple scrollable grid view, allowing the user to browse all generated images. Being able to view the dataset, the user may perform some free, unguided exploration here as a starting point for defining test concepts in the prompting tree. Interacting with the data view, the user will see the prompt from which an image was generated on hover. The images displayed in the data view are filtered or highlighted with a red outline based on interactions in other views. For instance, when the user hovers over a node in the prompting tree, all images generated from the node's prompt will be highlighted.

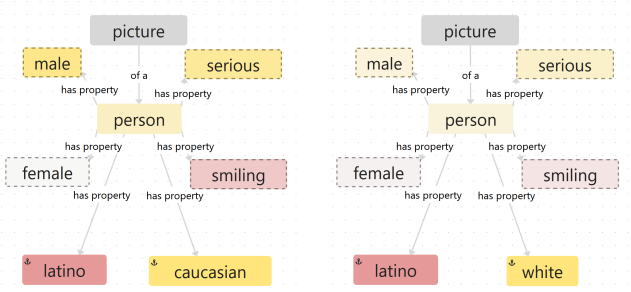
3.4.2. Prompting Tree

The prompting tree (Figure 1 (3)) is the central user interface element of ViBEx. It serves two main purposes: First, as input method, it supports flexible and systematic bias exploration by gradually adding and adjusting nodes that represent test concepts (**C1**). It thereby helps users keep track of the bias candidates they have already investigated. Second, the prompting tree also serves as output method to provide instant zero-shot bias probing feedback (**C2**).

Formally, the prompting tree is a directed acyclic graph $G = (V, E)$ where $V = \{v_0\} \cup \{v \mid v \in C \text{ or } v \in T\}$ are k concepts representing the anchor concepts C and a variable number of test concepts T, with v_0 being the root node. $E = \{(v_i, r, v_j) \mid v_i, v_j \in V, r \in R\}$ are labeled edges between two nodes with labels contained in the relation set R. The relation set R contains the connecting words, which allow for parsing the tree structure into natural language. For the example tree in Figure 3 the relation set is defined as $R = \{$*"has property"*, *"that shows a"*, *"wearing a"*$\}$. The *"has property"* relation is the default edge label, which is used to attach adjectives to concepts. Anchor icons within the nodes indicate anchor concepts, and image icons indicate the test concepts from which images were generated by the user for inverse bias queries. Additionally, as the user may define test concepts for more in-depth bias probing (see Section 3.4.3), the respective nodes will be marked with dashed outlines (see Figure 4).

Every concept expressed as a node in the tree can be parsed into a natural language representation $\mathcal{T}$ that serves as input to our similarity computation (Equation 1), as well as potential prompt to an image generator for inverse bias queries. A text representation of a concept $\mathcal{T}$ is constructed by concatenating node and edge labels along a branch of the tree, starting at the root node. For example, in Figure 3, the branch ending with the concept *"female"* would be parsed in the following way: ("picture" $\oplus$ "that shows a" $\oplus$ "female" $\oplus$ "person"), where $\oplus$ denotes a concatenation. If multiple branches are selected, they will be parsed into a combined text. If one node has multiple *"has property"* relations, the corresponding labels are chained together (e.g., *"young male person"*), while all other relations at the same level are connected with "and" (e.g., *"picture that shows a person and a dog"*). With this definition, nodes generally represent nouns and adjectives, while the edges are verbs and prepositions. Thus, the prompting tree adheres to design guidelines for T2I prompting, which suggest that effective prompting should "*focus on subject and style keywords over connecting words*" [LC22]. While a list of concepts could encode the same information as the prompting tree, the tree-based representation avoids redundancies by allowing for concept reuse: a comprehensive list would contain $|V|$ prompts with the root node being redundantly represented in each sentence. In the case of inverse bias queries, it is possible to combine multiple branches of the tree into a single prompt, leading to a combinatorial explosion of possible prompts. Furthermore, due to the ability to expand the tree at any node, it is possible to "flip" a concept (e.g. adding *"no"* to the concept *"hair"* using the relation *"has property"* to indicate baldness). A potential alternative representation would be a Word Tree [WV08], although here, the more constrained structure would hinder manual re-ordering and custom spatial organization through, for example, clustering of similar concepts by the user.

To visually encode the result of the zero-shot bias probing, we color-code the test concept nodes according to the computed probabilities from Equation 2. We use a diverging color scale, where the endpoints represent the colors associated with the two anchor concepts. The anchor concept colors can be selected by the user from a pre-defined list of colors. Gray represents an unbiased test concept, where $P(c_1 \mid t) \sim P(c_2 \mid t) \sim 0.5$. If an anchor concept's probability nears 1, the test concept's color will closely match the anchor concept's color. With that, the user can very quickly grasp which test concepts are similar to an anchor concept according to zero-shot bias probing. For example, Figure 4a shows *"serious"* as more similar to *"caucasian"*, while *"smiling"* is closer to *"latino"*.



(a) *Generated Images (SD3)* **(b)** *Images from FairFace*

Figure 4: *Two prompting trees with the same concepts but differing data sources. The data loaded from the FairFace dataset is gender-balanced. Thus, gender is expected to be neutral. The minimal imbalance hints at a bias in CLIP. The SD3 data shows a pronounced bias toward* "caucasian" *for the test concept* "male".

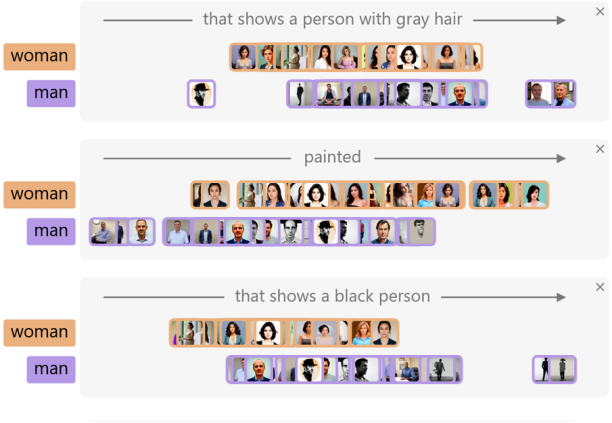


Figure 5: *Strip plots for three test concepts (*"gray hair", "painted picture", "black person"*). The univariate distribution of image text similarities is plotted for each anchor concept $c_i \in C$, here* "woman" *and* "man". *Note how the highest scoring image for* "black person" *is a silhouette picture.*

3.4.3. Bias Visualizations

The advantage of zero-shot bias probing based on CLIP is that it provides instant bias estimates. The downside is that these estimates are not always reliable because CLIP itself is biased [HZG*24, SC21]. This is illustrated in Figure 4: using the prompting tree, we discovered a bias of the test concept *"male"* towards the anchor concept *"caucasian"*. We tested this discovery against a hand-picked gender- and race-balanced sample from FairFace [KJ21], where images of male persons are evenly distributed across images labeled as *"latino"* and *"white"*. Our test concept bias metric (Equation 2) also indicates a weak bias towards images labeled as *"white"* when testing against *"picture of a male person"*, although this bias is not present in the anchor images. This could be caused by CLIP interpreting images of Caucasian persons to be "more male" than images of Hispanic people. Another

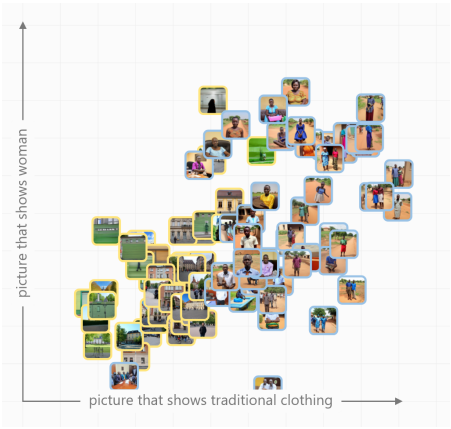


Figure 6: *Intersectional bias plot for* "woman" *and* "traditional clothing" *with the anchor concepts* "Germany" *and* "Nigeria". *Images in the top right are Nigerian women in traditional clothing, while for Germany no traditional clothing is present.*

problem is that text representations of concepts may be interpreted differently by the vision-language model than intended by the user. For example, in Figure 5, the test concept *"picture that shows a black person"* has very strong similarities with images showing a dark silhouette, while people with dark skin are not even present in the anchor concept image set. Similarly, we could not use the FairFace label *"white"* to generate the anchor images for the example in Figure 4: SD3 would generate images of people with their faces painted white, so we used *"caucasian"* instead.

To allow for real-time interactivity (**C2**) while maintaining trustworthiness (**C3**), it is therefore essential to keep the human in the loop. We support a human-centered bias exploration loop through multiple visualizations (Figure 1 (2)) offering in-depth confirmation of observed or suspected biases. For in-depth bias probing, users select one or multiple nodes in the prompting tree to generate a prompt of a test concept to be investigated in more detail.

For in-depth visual inspection of forward bias queries, we provide juxtaposed **strip plots**, where the x axis corresponds to the similarity distribution of all anchor concept images towards the selected test concept (computed according to Equation 1), separated by anchor concept on the y axis. The strip plots in Figure 5 confirm that only anchor images for *"man"* depict persons with gray hair and that primarily images associated with *"woman"* are painted.

Users may also visualize bivariate distributions of forward bias queries for two test concepts using the **intersectional bias scatterplot**. All images generated from the anchor concepts C are placed at the (x,y) coordinate corresponding to their similarities to the two selected test concepts. Each image is outlined with the color of its associated anchor concept. The resulting distribution allows the user to observe possible correlations between the two selected test concepts with respect to the anchor concepts. For example, Figure 6 illustrates an intersection discovered by an expert in our case study between *"woman"* and *"traditional clothing"*, which only applies to the anchor concept *"Nigeria"*.

Finally, results of inverse bias queries can be inspected in the

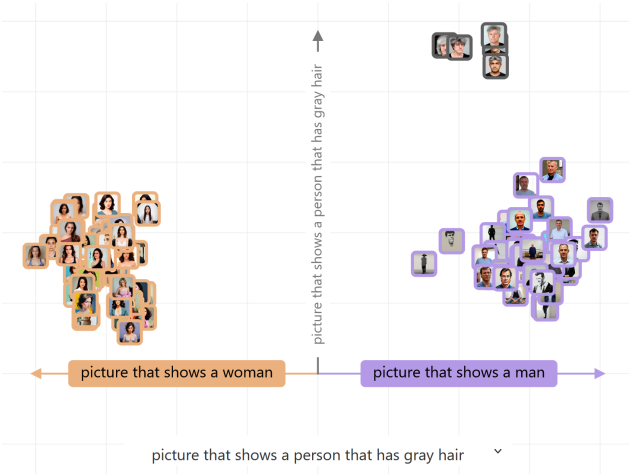


Figure 7: *Inverse bias query plot for* "picture that shows a person with gray hair" *for the anchor concepts* "woman" *and* "man". *Images generated for the test concept have a gray outline. We see that* "gray hair" *is predominantly associated with* "man".

inverse bias scatterplot. In this scatterplot, we show all anchor concept images, as well as images generated for the selected test concept (e.g., *"picture of a person with gray hair"*). Anchor concept images are framed by the color associated with their respective anchor concept. Test concept images have a gray frame. The x position of each image $\mathcal{I}_i$ represents the relative similarity to the textual representation of both anchor concepts:

$$x(\mathcal{I}_i) = s(\mathcal{I}_i, c_2) - s(\mathcal{I}_i, c_1). \tag{4}$$

We draw a vertical line at $x = 0$ to visually separate the two anchor concepts. The y axis represents the forward bias query towards the selected test concept (i.e., $s(\mathcal{I}_i, t)$). It can be expected that the anchor concept images are clearly separated horizontally, while the test concept images are separated from the anchor concept images vertically. This pattern is clearly visible in Figure 7, where all anchor concept images for *"woman"* (orange) are on the left, and all images for *"man"* (purple) are on the right. Also test concept images showing *"a person with gray hair"* (gray) are clearly more similar to their corresponding text representation and therefore shown on top. Most interestingly, all test concept images are clearly more similar to the anchor concept *"man"* than to *"woman"*, thereby confirming the observed forward bias in Figure 5 that gray hair is more strongly associated with *"man"*.

4. Implementation

The focus of our prototype implementation of ViBEx is to address the design challenge of real-time interactivity (**C2**), while supporting the user interaction discussed in Section 3.4. ViBEx is implemented as a client-server application. The back-end is a Python server, which handles session management and computes image-text similarities according to Equation 1. Whenever the front-end reports changes to the prompting tree, the server computes and stores the similarity values between the text representation of the affected prompting tree node and all anchor images. While the server is running, we keep the CLIP model in memory and also cache the image embedding vectors, such that for a new similarity calculation, we only need to generate the embedding of the test concept and compute the similarity measure. When the new similarity values have been computed, the array of similarity values is synced with the front-end, which then updates the forward bias values for the affected nodes in the prompting tree using Equation 2. Upon image generation (which is performed on a dedicated server), an inverse bias query is triggered, updating the inverse bias plot. Since image generation can take up to a few minutes, the front-end shows m placeholder images with a progress indicator in the data view until it gets notified that image generation is complete.

The ViBEx front-end is a web application using the React Framework [FC24], with D3.js [BC24] for the bias visualizations while the prompting tree is based on React Flow [GC24]. The back-end Python server uses CLIP `laion/CLIP-ViT-bigG-14-laion2B-39B-b160k` [LA24]. With the system being model-agnostic, we can utilize different image generation APIs either from local models (e.g., SD3) or from repositories such as *Hugging Face*.

5. Evaluation

We performed multiple evaluations to test ViBEx against our design challenges formulated in Section 3.1. To test for flexibility (**C1**), we performed a pilot study to test how confidently and correctly users can express observed and expected biases in the prompting tree (Section 5.1). To evaluate real-time interactivity (**C2**), we measured the computation times of zero-shot bias queries using CLIP (Section 5.2). Trustworthiness (**C3**) was evaluated through a combination of CLIP-based similarity results compared to a ground truth (Section 5.2) and an expert case study, where the discovered biases were a-posteriori validated using traditional non-interactive bias measures (Section 5.3). Finally, we cross-validated confirmed biases with prior work (including an automated method) to filter out which visual biases experts could discover using ViBEx that have not yet been reported in literature.

For all evaluations, we utilize SD3 as a state-of-the-art diffusion model. SD3 does not employ explicit bias mitigation techniques, such as prompt injection, allowing for direct evaluation of its intrinsic bias. All evaluations are based on three selected anchor concept pairs, which were already investigated in prior work. Our scenarios were selected based on prior work by Naik and Nushi [NN23]. In their paper, they investigated the anchor concept dimensions *gender, age, race*, and *geographical location* with respect to neutral test concepts, which were different occupations, personality traits, and everyday situations. We selected the scenarios *gender* (S_{gender}), *race* (S_{race}), and *geographical location* (S_{loc}) and derived corresponding anchor concepts listed in Table 1. In addition, we created a scenario S_{train} for the training session in our two user studies (Section 5.1 and Section 5.3). The corresponding anchor concept images can be found in Appendix E.

For these scenarios, using Stable Diffusion v1 and DALL·E as image generators, Naik and Nushi [NN23] reported a clear bias towards images being situated in Western countries and depictions of poor economic conditions being associated with Nigeria (S_{loc}). For the neutral prompt *"person"* they furthermore reported a general bias towards white people and a model-specific gender imbalance.

Scen.	$\mathcal{T}(c_1)$	$\mathcal{T}(c_2)$
S_{train}	drawing of a old-timer car	drawing of a futuristic car
S_{gender}	picture that shows a woman	picture that shows a man
S_{race}	picture of a Latino person	picture of a Caucasian person
S_{loc}	picture taken in Germany	picture taken in Nigeria

Table 1: *Anchor concept prompts used as evaluation scenarios.*

5.1. Prompting Tree Pilot Study

We conducted a pilot study to evaluate the flexibility and reliability of our prompting tree as *input* method (**C1**) in an early phase of the design process. The goal of the study was to assess whether users can flexibly express test concepts using the prompting tree and whether the test concepts can be parsed into a useful text representation following the procedure described in Section 3.4.2.

Nine volunteer students and employees from a local university participated in the study (five females and four males, aged 24-36, all with a computer science background). No participant was familiar with the planned ViBEx application. Participants were equally distributed across the three scenarios listed in Table 1, resulting in three participants per scenario. For each anchor concept, we generated $n = 50$ images. The users' task was to describe all observed or suspected biases as labeled nodes in the prompting tree. Prior to the main task, we conducted a training task using scenario S_{train} to explain how to add and remove nodes and relations, as well as how to define new relation types. For this study, we only showed the data view and used the prompting tree as the sole input method without visual encoding of zero-shot bias probing.

After the study, we automatically parsed each node in the created prompting trees into a text representation. Two independent coders then categorized each parsed text into three categories: *correct* (the text is meaningful), *concept problem* (e.g., a typo or a missing word in one or more nodes), and *relation problem* (e.g., an illogical connecting word or a grammatical error in one or more relations).

Results: A total of 121 nodes were created during the study, with seven to 22 per user (13 on average). The coders found 44 problems in total, but the majority of problems were contributed by two outlier users. One user had a 100% error rate as they built a hierarchical graph that could not directly be translated into natural language. This resulted in parsed texts like *"picture with classic architecture focus content"*. The other user did not adjust the relation qualifiers, keeping "has property" for all relations. This resulted in parsed texts like *"picture of a formal wear person"*. Excluding these outliers, 78.7% of texts were classified as *"correct"*, 7.8% as *"concept problem"* and 13.5% as *"relation problem"*. Relation problems were primarily caused by default *"has property"* relations that were not adjusted appropriately.

We compared a selected set of texts with relation problems to a corrected version with respect to CLIP-based similarity scores and resulting images when used as prompts for SD3 (see Appendix A). It can be observed that both models have a certain level of resilience to grammatical errors, which confirms prior work that subject and style keywords are more important than connecting words [LC22].

The test concepts expressed in the prompting trees primarily contained rather obvious entanglements, such as a tendency towards *"dark skin"* for the test concept *"Nigeria"* in scenario S_{loc}, towards *"man"* for *"beard"*, and towards *"woman"* for *"long hair"* (S_{gender}). More interestingly, all three participants inspecting scenario S_{race} reported a gender bias, which was also previously described by Naik and Nushi [NN23]. Three users also reported systematic differences for test concepts related to clothing style in scenarios S_{loc} and S_{race}, respectively. This has also been reported from previous bias studies [NN23, LZL*24, DPM*24].

When asked about special difficulties in the post-experiment interview, participants noted that the cognitive load required to parse and verify the text representation associated with a prompting tree node was challenging. Overall, however, all participants considered the prompting tree to be a very engaging interface.

We conclude that it is challenging to use the prompting tree with complete accuracy without training. However, commonly observed relation problems seem to have little influence on the model performance. To reduce the cognitive load for the user, we now display the fully parsed text when hovering over a node. This way, users can easily check and potentially correct their prompting tree.

5.2. Zero-Shot Bias Probing Performance

We performed a limited evaluation of CLIP testing its ability to detect biases as well as its time performance. The details of this evaluation can be found in Appendix B. In summary, we found that CLIP reliably detects an expected entanglement, while it shows no significant bias for a balanced image characteristic. This indicates that our zero-shot bias probing approach can provide sufficiently reliable results for a first approximate bias estimation (**C3**). We conducted the experiments using an NVIDIA H100 GPU with 80GB of memory.

System logging revealed that a single CLIP-based similarity measure between one concept and 100 anchor images (with cached embedding vectors) requires 1.47 ± 0.18 seconds to compute. This does not entirely satisfy our design challenge of real-time interactivity (**C2**) since the impression of instant feedback usually requires response times below 0.1 seconds [Nie94]. However, since we compute similarities on the server (see Section 4), these computations are not blocking, and visual feedback is automatically updated one to two seconds after the user has modified the prompting tree.

5.3. Expert Case Study

Finally, we evaluated whether expert users were able to discover visual bias dimensions with the support of ViBEx that were so far unknown to them. We invited four domain experts in the fields of AI, media ethics, and digital humanism to perform a bias exploration case study using the ViBEx interface. Users could select one or multiple scenarios to investigate from Table 1. Three experts (E1, E2, and E4) explored S_{gender}, E3 and E4 also examined S_{loc}. We first performed a pre-experiment interview about their experience with generative AI after which the participants performed a training task as described in Section 5.1. For all scenarios, we prepared $n = 50$ generated images per anchor concept and a prompting tree containing the two anchor concepts. We then asked the experts to

Scenario	Test Concept	Query	Tendency	Expert
S_{gender}	**person**	IBQ	men	E1
	beautiful	IBQ	women	E1, E4
	naked shoulders	FBQ	women	E1, E4
	bare skin	FBQ	women	E2
	long hair	FBQ	women	E1, E4
	dark skin	IBQ	women	E1
	asian*	FBQ	women	E2
	happy	IBQ	men	E1
	standing	FBQ	men	E1, (E4)
	doctor*	FBQ	men	E2
	nurse*	FBQ	women	E2
	black and white	FBQ	men	E4, (E2)
	bright colors	FBQ	women	E2
	old*	FBQ	men	E2
	young*	FBQ	women	E2
	business look*	FBQ	men	E4
	professional*	FBQ	men	E2
	boss*	FBQ	men	E2
	serious*	FBQ	balanced	E2
	child*	FBQ	balanced	E2
S_{loc}	traditional clothing*	FBQ	Nigeria	E3, E4
	sand-colored tones	FBQ	Nigeria	E4
	greenery	FBQ	Germany	E4
	classic architecture*	FBQ	Germany	E4

Table 2: *Biases discovered via ViBEx during the expert studies. FBQ and IBQ show whether the concept was tested through a forward or inverse bias query; the tendency indicates towards which anchor concept the bias is expressed. The * marker signifies concepts that are covered by the bias axes automatically determined by TIBET [CSB*24], while the tendencies for **bold** concepts were confirmed by our subsequent analysis (only performed for S_{gender}).*

extend the prompting tree with test concepts based on what they observed in the anchor concept images and based on their expectations, respectively. The participants were encouraged to verbally express their actions by thinking aloud, which was recorded and transcribed for qualitative evaluation. After the study, we asked participants to list all test concepts that they considered to represent a bias. The individual sessions lasted between 60 and 90 minutes.

5.3.1. Results

The test concepts for the discovered biases are summarized in Table 2, while all prompting trees are displayed in Appendix F. The three experts who explored S_{gender} found 20 test concepts with potential bias. The majority (16) of these concepts were forward bias queries confirmed through a strip plot. E1 and E4 also performed inverse bias queries (for E2 there was a problem with the image generation server, so no inverse bias queries were conducted). Notably, especially through the inverse queries, some unexpected biases were found: SD3 has a general gender bias towards *"male"* when prompted to generate a *"picture that shows a person"*. Furthermore, *"happy"* is skewed towards *"man"* and both, *"beautiful"* as well as *"dark skin"*, toward *"woman"*. All four experts used the intersectional bias plot to check for expected correlations.

E3 found a potential intersectional bias in S_{loc}, where Nigerian women appeared to be more frequently shown in traditional clothing (see Figure 6). The same tendency was also discovered by E4 in a forward bias query. Besides the bias candidates discovered through interaction with the interface, general observations of the images resulted in the conclusion that the *"man"* images showed more diverse body types and framing in S_{gender} (E4), while E1 noticed through their *"person"* query that this ungendered prompt resulted in melancholic depictions of mostly light-skinned men.

All four experts noted that the tool was engaging and made them more aware of the problem biases in T2I pose. They mentioned that, although some of the tendencies were visible even without bias probing, it was the interactivity and instant feedback of the prompting tree that kept them engaged. The prompting tree still posed a challenge, with some grammatically incorrect relations (e.g., Appendix F Figure 23c). However, as discussed in Section 5.2, the system proved to be robust to malformed text input. Apart from using the tool in their own research, experts also suggested that it would be well-suited for journalists and decision-makers to learn about biases in T2I models. Furthermore, they argued for ViBEx to be used as a basis for model auditing.

5.3.2. Validation of Discovered Biases

To verify that the biases discovered by experts constitute actual imbalances, we utilize the FairFace classifier [KJ21] as a trusted evaluation entity because it is trained on a balanced dataset. We pick those 20 biases from the case study that align with the FairFace classes (age, gender, and race), generate 50 new images per bias (1000 in total), and classify them. We generate new images to also compensate for a potential sample bias in the case studies. The images generated for our validation step are available on osf.io.

We show the results of our validation step in Table 2, where we highlight all confirmed biases. The FairFace classifier only operates on images with visible faces, thus we only classify such images. All four inverse bias queries (*"person"*, *"beautiful"*, *"dark skin"*, *"happy"*) are confirmed by the FairFace validation, indicating that the small sample of five images provides reliable insight. Furthermore, the expected gender bias for *"nurse"*/*"doctor"* by E2 is confirmed. For the forward bias queries with test concepts tending toward *"woman"*, most tendencies are not confirmed, as they instead showcase a bias toward depicting men. This is probably caused by a general gender bias of SD3 towards men, which can be observed in Appendix C Figure 15. The test concept *"young"* may be too ambiguous: while the experts used it to label young adults, SD3 produced images of kids. The concept *"naked shoulders"*, on the other hand, may be too specific and, therefore, not well-represented in the training data. We discuss a potential strategy to alleviate the problem of these "false positives" in Section 6. Only for *"long hair"*, *"dark skin"* and *"bright colors"*, the bias toward *"woman"* is also clearly present in the validation. While *"long hair"* was also reported in the pilot study based on image observations only, *"dark skin"* and *"bright colors"* were only discovered by the experts through the ViBEx workflow. To the best of our knowledge, these biases have not been reported by prior work.

To further test whether an automated bias discovery system could find these bias dimensions, we queried TIBET [CSB*24] with the image generation prompts from S_{gender} and S_{loc}. We performed the first step of TIBET's bias discovery pipeline, which uti-

lizes GPT 3.5 to come up with bias dimensions to be tested by the system. We observed that TIBET suggested generic dimensions, such as *"age group"*, *"occupation"*, or *"appearance"* for S_{gender}. However, it missed many of the experts' more specific observations, such as *"happy"* or *"bright colors"*. This might be because TIBET creates bias dimensions prior to the image generation step, just from the textual prompt. A full list of the automatically suggested bias dimensions can be found in Appendix D. We marked the test concepts that are covered by one of the bias axes from TIBET with a * in Table 2. Overall, for S_{gender} and S_{loc}, only 50% of the test concepts found by the experts are also covered by TIBET.

6. Discussion and Conclusions

Our expert study highlights that ViBEx successfully supports users discovering visual biases in T2I systems. Users reported SD3's bias towards *"woman"* for the test concepts *"beautiful"*, *"dark skin"*, and *"bright colors"* – findings not previously documented in literature or via automated methods. We now revisit our design challenges for further discussion and suggestions for future work.

Flexibility through prompting tree: Our pilot study and the prompting trees constructed by the experts showed that the prompting tree supports flexible expression of test concepts. However, users also faced difficulties correctly expressing their observed and suspected biases. In particular, defining and using appropriate relation types was a frequently observed difficulty. Here, we could illustrate that both CLIP and SD3 are, to a certain extent, resilient to grammar errors, and imperfect textual inputs may still lead to usable results (see Appendix A). In the future, text parsing from the prompting tree could be enhanced by a language model to transform the parsed texts so that the respective model can interpret them more reliably. Such reformulations to less ambiguous or more well-understood concepts could potentially help reduce the presence of "false positives" where the user's intent of a test concept is not correctly interpreted by CLIP. In addition, as shown in Section 5.3.2, some generic bias dimensions could easily be detected automatically. A future iteration of our workflow could, therefore, take a mixed-initiative approach where an automated system proposes and tests a large number of bias dimensions while the human user gets to test for more nuanced and specific bias dimensions.

Instant zero-shot bias feedback: We showed that a bias measure based on image-text similarity in a multimodal embedding space can be computed within about a second. With server-side calculation and caching, zero-shot bias probing did not lead to any disruptive system lag. In our case study, users explicitly appreciated the *"instant feedback"* and confirmed that the quick bias estimations made the discovery process very engaging.

Improving zero-shot trustworthiness: Interactive bias exploration may introduce bias itself, for instance through a *sample selection bias* by the image generator or a *confirmation bias* by the user. In addition, the multimodal embedding space used for automatic zero-shot bias probing may be inherently biased. Our initial zero-shot bias probing performance experiments did not reveal any concerning results, but we occasionally observed that text input was not interpreted by CLIP as intended (see, for instance, Figure 5 bottom) and CLIP indicating bias for test concepts that were in fact not depicted in the anchor concept images (e.g., *"doctor"* and *"nurse"*, as investigated by E2). To address transparency and ambiguity of vision-language models in the future, the data view could additionally show saliency maps [GB24] of images on demand, thereby illustrating the image features that constitute the similarity to a selected test concept. If the user observes spurious correlations, they could then decide to exclude unrepresentative images from the zero-shot bias probing.

Trustworthiness through confirmatory visualization: In the expert case study, we observed that all users consulted our provided visualizations for more in-depth confirmatory analysis. Similarly important are inverse bias queries to confirm the observed bias with a clear focus on the T2I model. For example, in the case of test concepts *"doctor"* and *"nurse"*, an a-posteriori inverse bias query clearly confirmed the existence of a gender bias in these two occupations for SD3. Unfortunately, inverse bias queries are slow due to the necessary image generation step. In the future, progressive image generation for faster approximate feedback could be a promising solution. This would allow for the image data to be delivered and processed in chunks, followed by updating the bias visualizations via progressive visual analytics strategies such as extension [FFS24, UAF*24].

Scalability: A challenge not explicitly addressed in our reference implementation is scalability with respect to the number of sample images and anchor concepts. The ViBEx workflow and zero-shot bias probing conceptually allow for an extension to an arbitrary number of anchor concepts and sample images. However, the visualizations will have to be adjusted: Both the strip plot and inverse bias scatterplot are trivially scalable. Here, to reduce clutter, a kernel density plot could be used instead of showing the images directly. In the prompting tree an option would be to employ pairwise color codings and only show the similarity values between the top two anchor concepts per node. For the inverse bias scatterplot, a radar chart or a parallel coordinates representation with an axis per anchor concept would be suitable candidates.

Generalizability: In this work, we employed ViBEx for SD3 and a subset of the FairFace image collection. In the future, we hope to see ViBEx in experiments comparing T2I models trained on different data sets, as well as original and de-biased models.

In conclusion, with ViBEx, we demonstrated how users can flexibly explore visual bias in T2I models at a rapid pace while maintaining trust through confirmatory visualizations. We showed that, with our proposed workflow, users quickly can come up with new bias dimension candidates. We see the greatest potential for future improvements in mixed-initiative bias queries, explainable zero-shot bias probes, and faster inverse bias queries. Beyond exploration and confirmation of visual bias, ViBEx could ultimately be used for finding bias mitigation strategies by using zero-shot bias probing to create efficient prompt injections.

Acknowledgments

We thank Dominik Wolf for his contributions to the software framework and the participants of the studies. This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/P36453 and by the Austrian Research Promotion Agency (FFG), project no. 898085 and FO999904624. Open access funding provided by Technische Universitat Wien/KEMÖ.

References

[ASV*24] ARAWJO I., SWOOPES C., VAITHILINGAM P., WATTENBERG M., GLASSMAN E. L.: ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–18. 3

[AZL*19] ALSPAUGH S., ZOKAEI N., LIU A., JIN C., HEARST M. A.: Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 22–31. Conference Name: IEEE Transactions on Visualization and Computer Graphics. 3

[AZPK*24] ALMEDA S. G., ZAMFIRESCU-PEREIRA J., KIM K. W., MANI RATHNAM P., HARTMANN B.: Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–17. 3

[BBIW20] BLODGETT S. L., BAROCAS S., III H. D., WALLACH H.: Language (Technology) is Power: A Critical Survey of "Bias" in NLP, May 2020. arXiv:2005.14050 [cs]. 3

[BC24] BOSTOCK M., CONTRIBUTORS O. S.: D3.js - data-driven documents, 2024. Accessed: 2024-12-01. URL: https://d3js.org/. 7

[BH19] BATTLE L., HEER J.: Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum 38*, 3 (2019), 145–159. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13678. 3

[BKD*23] BIANCHI F., KALLURI P., DURMUS E., LADHAK F., CHENG M., NOZZA D., HASHIMOTO T., JURAFSKY D., ZOU J., CALISKAN A.: Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago IL USA, June 2023), ACM, pp. 1493–1504. 1

[BWS*23] BRADE S., WANG B., SOUSA M., OORE S., GROSSMAN T.: Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco CA USA, Oct. 2023), ACM, pp. 1–14. 3

[CEH*19] CABRERA A. A., EPPERSON W., HOHMAN F., KAHNG M., MORGENSTERN J., CHAU D. H.: FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Vancouver, BC, Canada, Oct. 2019), IEEE, pp. 46–56. 2

[CSB*24] CHINCHURE A., SHUKLA P., BHATT G., SALIJ K., HOSANAGAR K., SIGAL L., TURK M.: TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models, July 2024. arXiv:2312.01261 [cs]. 2, 3, 9, 14, 15

[DPM*24] D'INCÀ M., PERUZZO E., MANCINI M., XU D., GOEL V., XU X., WANG Z., SHI H., SEBE N.: OpenBias: Open-set Bias Detection in Text-to-Image Generative Models, Apr. 2024. arXiv:2404.07990 [cs]. 2, 3, 8

[EKB*24] ESSER P., KULAL S., BLATTMANN A., ENTEZARI R., MÜLLER J., SAINI H., LEVI Y., LORENZ D., SAUER A., BOESEL F., PODELL D., DOCKHORN T., ENGLISH Z., LACEY K., GOODWIN A., MAREK Y., ROMBACH R.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, Mar. 2024. arXiv:2403.03206 [cs]. 2

[FBS*23] FRIEDRICH F., BRACK M., STRUPPEK L., HINTERSDORF D., SCHRAMOWSKI P., LUCCIONI S., KERSTING K.: Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness, July 2023. arXiv:2302.10893 [cs]. 2

[FC24] FACEBOOK I., CONTRIBUTORS O. S.: React - a javascript library for building user interfaces, 2024. Accessed: 2024-12-01. URL: https://reactjs.org/. 7

[FFS24] FEKETE J.-D., FISHER D., SEDLMAIR M.: Progressive Data Analysis: Roadmap and Research Agenda, 2024. 10

[FPNK22] FABBRIZZI S., PAPADOPOULOS S., NTOUTSI E., KOMPATSIARIS I.: A survey on bias in visual datasets. *Computer Vision and Image Understanding 223* (Oct. 2022), 103552. 3

[FWW*23] FENG Y., WANG X., WONG K. K., WANG S., LU Y., ZHU M., WANG B., CHEN W.: PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11. 3

[FYC*23] FARSHAD A., YEGANEH Y., CHI Y., SHEN C., OMMER B., NAVAB N.: SceneGenie: Scene Graph Guided Diffusion Models for Image Synthesis. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (Paris, France, Oct. 2023), IEEE, pp. 88–98. 3

[GB24] GIULIVI L., BORACCHI G.: Concept Visualization: Explaining the CLIP Multi-modal Embedding Using WordNet. In *2024 International Joint Conference on Neural Networks (IJCNN)* (Yokohama, Japan, June 2024), IEEE, pp. 1–9. 10

[GC24] GMBH W., CONTRIBUTORS O. S.: React flow - a library for building node-based uis, 2024. Accessed: 2024-12-01. URL: https://reactflow.dev/. 7

[GHM21] GHAI B., HOQUE M. N., MUELLER K.: WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, May 2021), ACM, pp. 1–7. 2, 3

[GSL*24] GUO Y., SHAO H., LIU C., XU K., YUAN X.: PromptTHis: Visualizing the Process and Influence of Prompt Editing during Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–12. 3

[HGE22] HOQUE M. N., GHAI B., ELMQVIST N.: Dramatvis personae: Visual text analytics for identifying social biases in creative writing. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (2022), pp. 1260–1276. 2

[HMKB22] HUANG J., MISHRA A., KWON B. C., BRYAN C.: ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–11. 3

[HZG*24] HAMIDIEH K., ZHANG H., GERYCH W., HARTVIGSEN T., GHASSEMI M.: Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2024), vol. 7, pp. 547–561. 6

[KJ21] KARKKAINEN K., JOO J.: FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa, HI, USA, Jan. 2021), IEEE, pp. 1547–1557. 2, 6, 9, 13, 14

[KLZ24] KABIR S., LI L., ZHANG T.: STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–20. 2

[LA24] LAION-AI: Clip-vit-bigg-14-laion2b-39b-b160k. https://huggingface.co/laion/CLIP-ViT-bigG-14-laion2B-39B-b160k, 2024. Accessed: 2024-12-01. 7

[LAMJ23] LUCCIONI A. S., AKIKI C., MITCHELL M., JERNITE Y.: Stable Bias: Analyzing Societal Representations in Diffusion Models, Nov. 2023. arXiv:2303.11408 [cs]. 1, 2

[LC22] LIU V., CHILTON L. B.: Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA, Apr. 2022), ACM, pp. 1–23. 3, 5, 8, 13

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (2014), Springer, pp. 740–755. 2

[Luc23] LUCCIONI S.: StableDiffusionBiasExplorer - a Hugging Face Space by society-ethics, Sept. 2023. URL: https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer. 2, 3

[LZL*24] LIU M., ZHONG Z., LI J., FRANCHI G., ROY S., RICCI E.: Organizing Unstructured Image Collections using Natural Language, Oct. 2024. arXiv:2410.05217 [cs]. 2, 3, 8

[MGSM24] MAHDAVI GOLOUJEH A., SULLIVAN A., MAGERKO B.: Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–13. 3

[MLL23] MANDAL A., LEAVY S., LITTLE S.: Measuring Bias in Multimodal Models: Multimodal Composite Association Score. In *Advances in Bias and Fairness in Information Retrieval*, Boratto L., Faralli S., Marras M., Stilo G., (Eds.), vol. 1840. Springer Nature Switzerland, Cham, 2023, pp. 17–30. Series Title: Communications in Computer and Information Science. 2

[Nic] NICHOL A.: DALL·E 2 pre-training mitigations. URL: https://openai.com/research/dall-e-2-pre-training-mitigations. 2

[Nie94] NIELSEN J.: *Usability engineering*. Morgan Kaufmann, 1994. 8

[NN23] NAIK R., NUSHI B.: Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal QC Canada, Aug. 2023), ACM, pp. 786–808. 1, 2, 4, 7, 8

[Opp23] OPPENLAENDER J.: A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* (Nov. 2023), 1–14. 3

[PS08] PERER A., SHNEIDERMAN B.: Systematic Yet Flexible Discovery: Guiding Domain Experts through Exploratory Data Analysis. *Proceedings of the 13th International Conference on Intelligent User Interfaces* (2008), 109–118. 3

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-Resolution Image Synthesis with Latent Diffusion Models, Apr. 2022. arXiv:2112.10752 [cs]. 2

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G., SUTSKEVER I.: Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. arXiv:2103.00020 [cs]. 2, 4

[RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-Shot Text-to-Image Generation, Feb. 2021. arXiv:2102.12092 [cs]. 2

[SC21] STEED R., CALISKAN A.: Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 701–713. 6

[SKL23] SLYMAN E., KAHNG M., LEE S.: VLSlice: Interactive Vision-and-Language Slice Discovery. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, Oct. 2023), IEEE, pp. 15245–15255. 3

[SMPX23] SUH S., MIN B., PALANI S., XIA H.: Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco CA USA, Oct. 2023), ACM, pp. 1–18. 3

[SMS*23] SHUKLA T., MAHESHWARI P., SINGH R., SHUKLA A., KULKARNI K., TURAGA P.: Scene Graph Driven Text-Prompt Generation for Image Inpainting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Vancouver, BC, Canada, June 2023), IEEE, pp. 759–768. 3

[TC19] TAN Y. C., CELIS L. E.: Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems* (2019), vol. 32, Curran Associates, Inc. 2

[UAF*24] ULMER A., ANGELINI M., FEKETE J.-D., KOHLHAMMER J., MAY T.: A Survey on Progressive Visualization. *IEEE Transactions on Visualization and Computer Graphics 30*, 9 (Sept. 2024), 6447–6467. 10

[VAHM23] VICE J., AKHTAR N., HARTLEY R., MIAN A.: Quantifying Bias in Text-to-Image Generative Models, Dec. 2023. arXiv:2312.13053 [cs]. 1

[WC24] WAN Y., CHANG K.-W.: The Male CEO and the Female Assistant: Gender Biases in Text-To-Image Generation of Dual Subjects, June 2024. arXiv:2402.11089 [cs]. 2

[WHM*24] WEISZ J. D., HE J., MULLER M., HOEFER G., MILES R., GEYER W.: Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–22. 3

[WHS*24] WANG Z., HUANG Y., SONG D., MA L., ZHANG T.: PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, May 2024), ACM, pp. 1–21. 3

[WLD*23] WANG J., LIU X., DI Z., LIU Y., WANG X.: T2IAT: Measuring Valence and Stereotypical Biases in Text-to-Image Generation. In *Findings of the Association for Computational Linguistics: ACL 2023* (Toronto, Canada, 2023), Association for Computational Linguistics, pp. 2560–2574. 1, 2

[WLG23] WANG Q., L'YI S., GEHLENBORG N.: DRAVA: Aligning Human Concepts with Machine Learning Latent Dimensions for the Visual Exploration of Small Multiples. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2023), CHI '23, Association for Computing Machinery, pp. 1–15. 3

[WSO*24] WAN Y., SUBRAMONIAN A., OVALLE A., LIN Z., SUVARNA A., CHANCE C., BANSAL H., PATTICHIS R., CHANG K.-W.: Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation, May 2024. arXiv:2404.01030 [cs]. 1, 2, 3

[WV08] WATTENBERG M., VIEGAS F.: The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov. 2008), 1221–1228. 5

[YLHH14] YOUNG P., LAI A., HODOSH M., HOCKENMAIER J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78. 2

[ZGC*17] ZGRAGGEN E., GALAKATOS A., CROTTY A., FEKETE J.-D., KRASKA T.: How Progressive Visualizations Affect Exploratory Analysis. *IEEE Transactions on Visualization and Computer Graphics 23*, 8 (Aug. 2017), 1977–1987. Conference Name: IEEE Transactions on Visualization and Computer Graphics. 3

[ZJTY23] ZHANG Y., JIANG L., TURK G., YANG D.: Auditing Gender Presentation Differences in Text-to-Image Models, Feb. 2023. arXiv:2302.03675 [cs]. 2

[ZXSR22] ZHAO Z., XU P., SCHEIDEGGER C., REN L.: Human-in-the-loop Extraction of Interpretable Concepts in Deep Learning Models. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (Jan. 2022), 780–790. 3

Appendix A: Prompt Sensitivity

Our evaluation showed that the prompting tree may lead to malformed prompts due to inaccurate relations. The literature shows that T2I models emphasize *keywords* (i.e., nodes in the prompting tree) over *connecting words* (edges) [LC22]. We performed tests prompting SD3 with malformed prompts from both the prompting tree study and the expert case study. Results of these tests indicate that inaccurate relations do lead to the same image content but may decrease image quality (Figure 8 - 10).

(a) *Prompt:* "picture that shows a person that shows a nurse".

(b) *Prompt:* "picture that shows a nurse".

Figure 8: *Images generated with SD3 comparing two prompt formulations for* "nurse".

(a) *Prompt:* "picture with modern architecture focus content".

(b) *Prompt:* "picture that shows modern architecture".

Figure 9: *Images generated with SD3 comparing two prompt formulations for* "modern architecture".

(a) *Prompt:* "picture of a formal wear person".

(b) *Prompt:* "picture that shows a person with formal wear".

Figure 10: *Images generated with SD3 comparing two prompt formulations for* "formal wear".

Furthermore, we also computed CLIP similarity scores with malformed prompts to see how bias queries are affected (Figure 11 and 12). Here we observe similar behavior as with SD3, where the general distributions are comparable, but separations are less pronounced. As these tests are only of an exploratory nature, a more thorough evaluation remains to be conducted in future work.

Figure 11: *Distribution of similarities for* "formal wear person" *vs.* "person with formal wear" *in* S_{gender}.

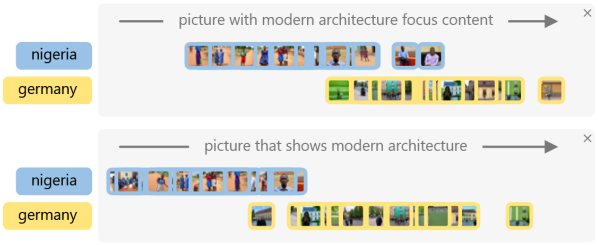


Figure 12: *Distribution of similarities for* "modern architecture focus content" *vs.* "modern architecture" *in* S_{loc}.

Appendix B: CLIP Experiments

We systematically compared the similarity measures produced by CLIP (see Equation 1) for test concepts with known ground truth to validate the reliability of CLIP for zero-shot bias probing (**C3**). We conducted forward bias queries for two scenarios: S_{gender} against the test concept *"beard"* and S_{race} against the test concept *"person"*. The first scenario thereby represents an expected entanglement, where we expect to see a tendency towards *"man"*. In the second scenario, *"person"* should be balanced between *"Latino"* and *"Caucasian"*. 50 images per anchor concept were manually selected from FairFace [KJ21] (see Figures 13 and 14). For the second experiment, we measured the response times for similarity computations of ten different test concepts taken from the prompting tree pilot study and 100 anchor images from S_{loc}.

Results: For the scenario S_{gender}, a Kolmogorov-Smirnov test showed a significant difference between the similarity distributions ($D(50) = .653, p < .001$). For the scenario S_{race}, we do not see a significant difference ($D(50) = .16, p = .548$). This confirms that CLIP can correctly detect an expected and obvious entanglement (i.e., men are more associated with beards than women), while it shows no significant difference for a balanced image distribution (i.e., an equal number of Latino and Caucasian persons).

Figure 13: *Gender- and age-balanced sample from FairFace for race* "latino".

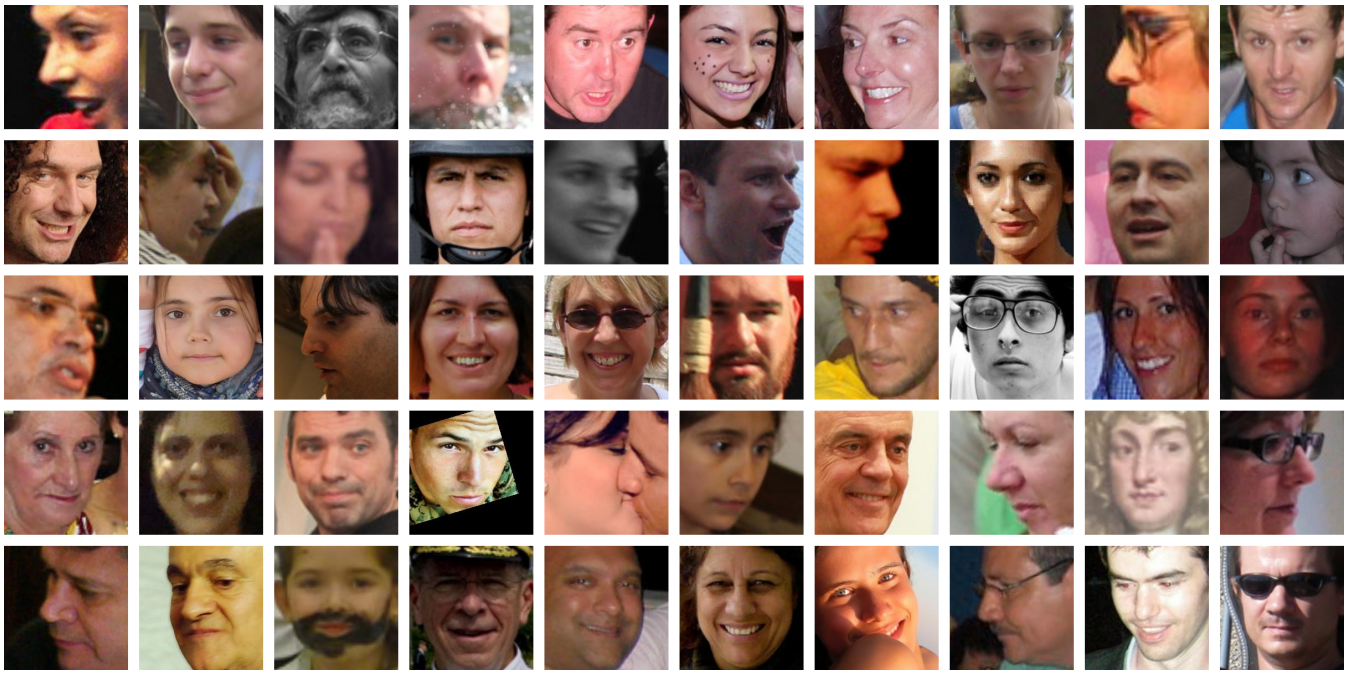


Figure 14: *Gender- and age-balanced sample from FairFace for race* "white".

Appendix C: Discovered Biases

Figure 15 shows the results of the bias validation via FairFace [KJ21]. Here, we plot the absolute frequency of the two anchor concepts for each test concept.

Appendix D: Automated Bias Evaluation Experiment

Tables 3 and 4 show the bias dimensions (formulated as questions) for scenarios S_{gender} and S_{loc} obtained from running the bias axis generation step of TIBET [CSB*24].

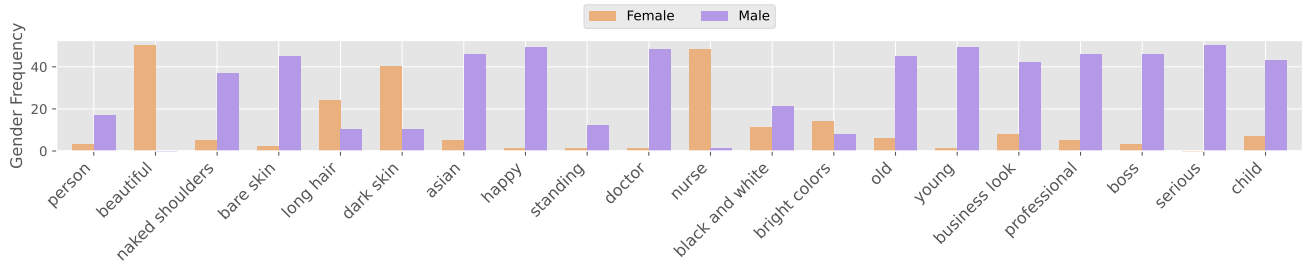


Figure 15: *Results of the bias validation. For each bias from S_{gender} listed in Table 2, we generated 50 new images and plot the frequency of* "male" *and* "female" *as classified by FairFace. Note that FairFace is only able to classify those images with visible faces.*

$\mathbf{S_{gender}}$	
picture that shows a woman	picture that shows a man
What is the gender (male, female, other) of the person in the image?	What is the gender (male, female, other) of the person in the image?
What is cultural in the image?	Describe the appearance in the image.
What age group (young, middle, old) is the person in the image?	Describe the background of the image.
Describe the appearance in the image.	What is the occupation of the person in the image?
What is societal in the image?	What age group (young, middle, old) is the person in the image?
	What is body type in the image?

Table 3: *Automatically generated bias dimensions by TIBET [CSB*24] for scenario S_{gender}.*

$\mathbf{S_{loc}}$	
picture taken in germany	picture taken in nigeria
What is geographic in the image?	What is geographic in the image?
What is cultural in the image?	What is cultural in the image?
What is historical in the image?	What is economic in the image?
What is touristic in the image?	What is urban vs. rural in the image?
	What is political in the image?
	What is historical in the image?

Table 4: *Automatically generated bias dimensions by TIBET [CSB*24] for scenario S_{loc}.*

Appendix E: Bias Exploration Scenarios

Figures 16 - 21 show all images used for our three bias exploration scenarios.

Appendix F: Expert Case Study Prompting Trees

Figures 22 and 23 show all prompting trees authored during the expert case study.

Figure 16: S_{gender}. *Prompt:* "picture that shows a man".

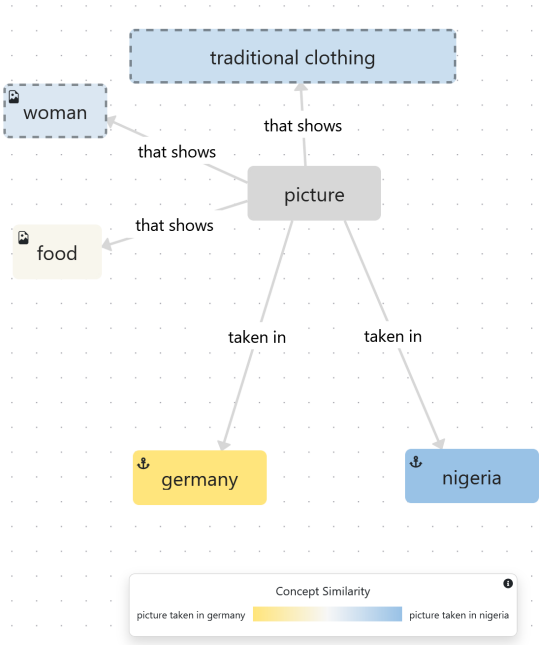
Figure 17: S_{gender}. *Prompt:* "picture that shows a woman".

Figure 18: S_{loc}. *Prompt:* "picture taken in Germany".

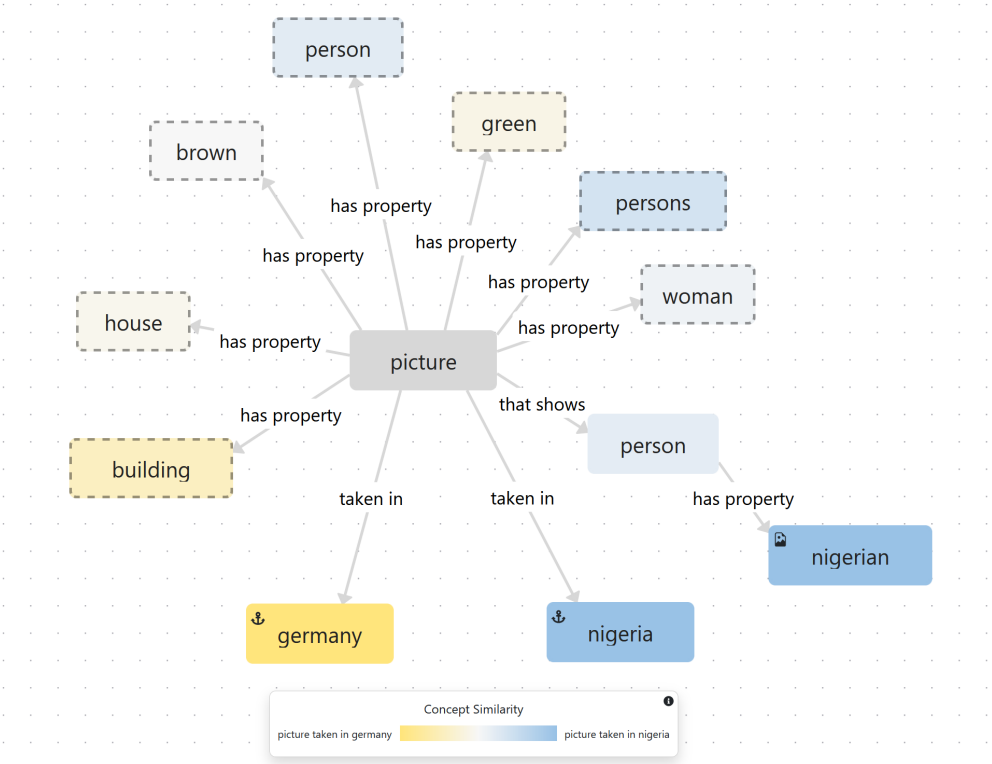
Figure 19: S_{loc}. *Prompt:* "picture taken in Nigeria".

Figure 20: S_{race}. *Prompt:* "picture of a caucasian person".

Figure 21: S_{race}. *Prompt:* "picture of a latino person".

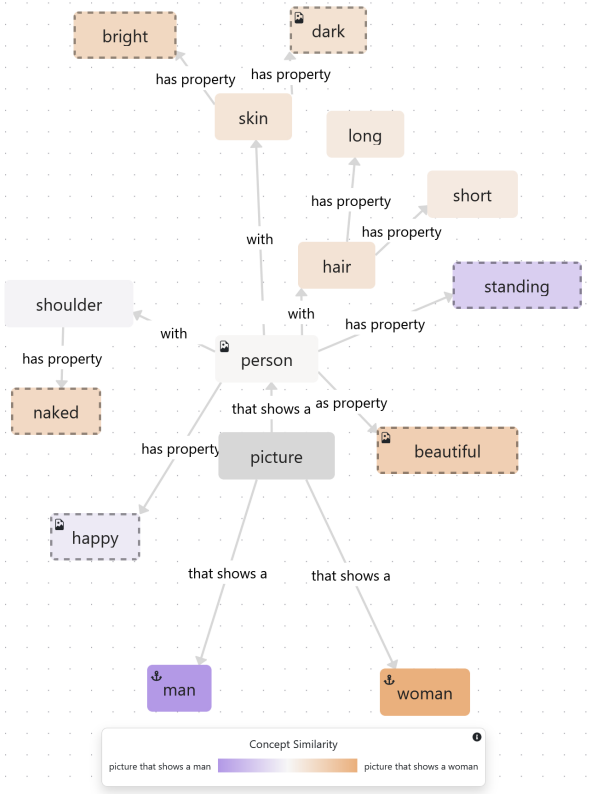


(a) *E3*

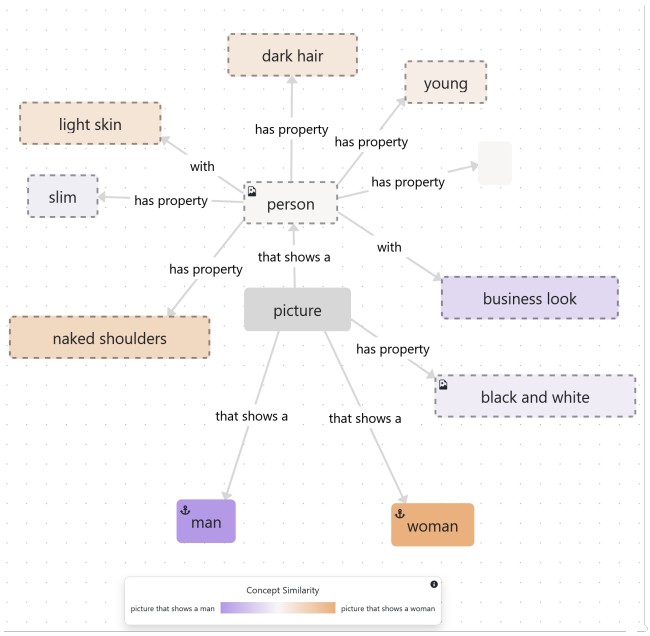


(b) *E4*

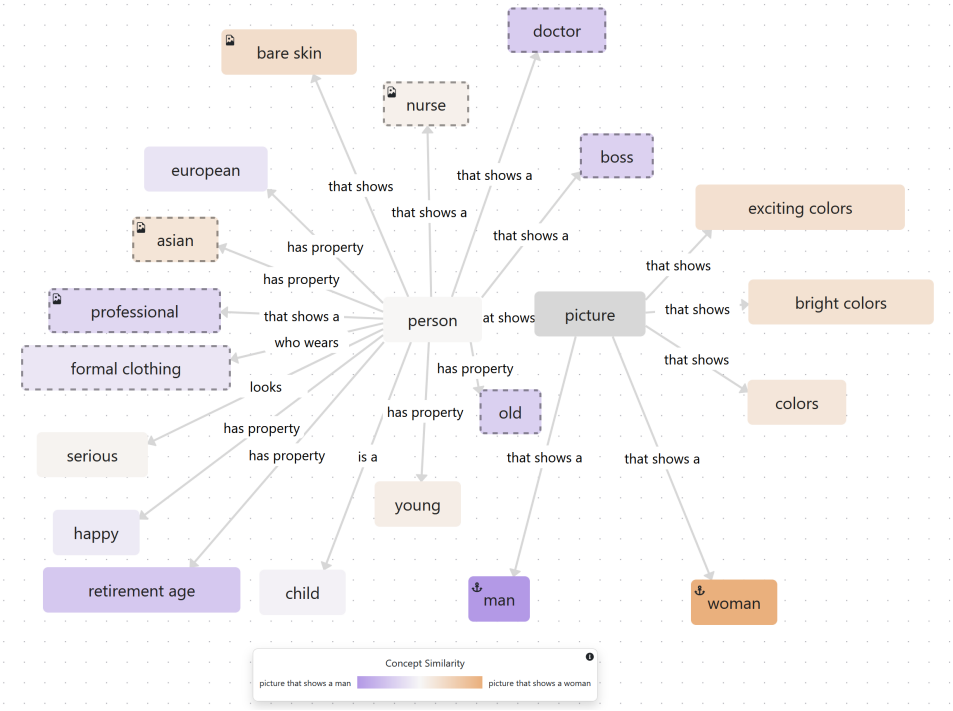
Figure 22: *Prompting trees generated during the case studies for S_{loc}.*



(a) *E1*



(b) *E4*



(c) *E2*

Figure 23: *Prompting trees generated during the case studies for S_{gender}.*