IS WHAT YOU ASK FOR WHAT YOU GET? INVESTIGATING CONCEPT ASSOCIATIONS IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models are increasingly used in impactful real-life applications. As such, there is a growing need to audit these models to ensure that they generate desirable, task-appropriate images. However, systematically inspecting the associations between prompts and generated content in a human-understandable way remains challenging. To address this, we propose *Concept2Concept*, a framework where we characterize conditional distributions of vision language models using interpretable concepts and metrics that can be defined in terms of these concepts. This characterization allows us to use our framework to audit models and prompt-datasets. To demonstrate, we investigate several case studies of conditional distributions of prompts, such as user defined distributions or empirical, real world distributions. Lastly, we implement Concept2Concept as an open-source interactive visualization tool facilitating use by non-technical end-users.

Warning: This paper contains discussions of harmful content, including CSAM and NSFW material, which may be disturbing to some readers.

1 INTRODUCTION

Text-to-image (T2I) models have become central to many real-world AI-driven applications. However, the complexity of these models makes it difficult to understand how they associate concepts in images with textual prompts. Existing works have shown that T2I models can resolve prompts in unexpected ways (Bianchi et al. (2023)). Furthermore, the training datasets for T2I models are often large, uncurated, and may contain undesirable prompt to image associations that models can learn to internalize (Birhane et al. (2024)). Thus, without robust auditing frameworks that help us detect these undesirable associations, we risk deploying T2I models that generate unexpected and inappropriate content for a given task.

However, auditing T2I models is challenging because it is difficult to systematically, efficiently, and intuitively explore the vast space of prompts and possible outputs. Because raw pixel values alone are difficult to semantically reason about, previous works learn mappings from raw inputs to high-level concepts. This can be achieved post-hoc (Kim et al. (2018); Zhou et al. (2018); Ghorbani et al. (2019)) or as an intervention during training (Koh et al. (2020);Chen et al. (2020). Although these methods were designed for classification networks, it is this general intuition which motivates our work.

In this paper, we propose a framework for producing interpretable characterizations of the conditional distribution of generated images given a prompt, $p(\texttt{image}|\texttt{prompt})$. We do so by extracting high-level concepts from each image and summarizing $p(\texttt{image}|\texttt{prompt})$ in terms of such concepts. Here, we define concepts as a class of objects/nouns, ideas, open vocabulary detected classes or labels.

Our contributions are as follows:

(1) We propose **an interpretable framework** for concept-association based auditing of conditional distributions. Specifically, in our framework: we sample images from a T2I model under audit, given a prior distribution over prompts–either a user defined distribution or an empirical real-world distribution. Then, using a fast, scalable visual grounding model, we extract concepts from generated images. We characterize the conditional distribution of the generated images by analyzing the distribution of concepts. This framework allows users to systemically investigate associations of conditional distributions at varying levels of granularity, from broad concept trends, co-occurrences, to detailed visual features. By design, our framework utilizes visual grounding models that localize concepts in images, enabling a deeper analysis of visual representations. Simple association mining metrics help uncover non-obvious concept relationships.

(2) We demonstrate a wide range of concrete use-cases for our framework, by applying it to audit models and prompt datasets. In addition to demonstrating the effectiveness of the framework, our analysis unearthed **new findings** that are independently interesting. In particular, *we discovered child-sexual abuse material (CSAM)* in a human-preferences prompt dataset and misaligned classes in a synthetically generated ImageNet dataset. These findings not only demonstrate the utility of our framework but also contribute to the broader discourse on the safety, fairness, and alignment of T2I models.

(3) We introduce an **interactive visualization tool**, based on our framework, for human-in-the-loop auditing of T2I models. Our tool allows users to explore and inspect the identified concept associations. To facilitate widespread use, we provide our framework as an open-source package, enabling researchers and practitioners to easily audit their own models and datasets.

2 Related Work

Biases in T2I models. There is a body of works that have qualitatively investigated biases in T2I models, focusing on social biases related to gender, race, and other identity attributes. For example, Bianchi et al. (2023) qualitatively demonstrated a range of social biases in T2I models, including biases related to basic traits, social roles, and everyday objects. Similarly, Ungless et al. (2023) manually analyzed images generated by T2I models and found that certain non-cisgender identities were misrepresented, often depicted in stereotyped or sexualized ways. Through several focus groups, Mack et al. (2024) found that T2I models repeatedly presented "reductive archetypes for different disabilities". Qualitative evaluations play a critical role in exposing instances where generative models can be biased. However, given the large space of possible prompts and images, instance-based bias probing alone cannot paint a systematic picture of how T2I models may (mis)behave in application.

A number of works have focused on automating bias detection at scale. For example, in Cho et al. (2023), the authors measured visual reasoning skills and social biases in T2I models by using a combination of automated detectors and human evaluations to assess the representation of different genders, skin tones, and professions. Likewise, Luccioni et al. (2024) employed Visual Question Answering (VQA) models and clustering-based evaluations to measure correlations between social attributes and identity characteristics. TIBET (Chinchure et al. (2023)) and OpenBias (D'Incà et al. (2024)) dynamically generate axes of bias, either based on a single prompt or collection of input prompts. However, these works either do not operate on the general concept-level (e.g. only specifically probe for concepts related to social attributes) and/or do not leverage the rich information in the concept co-occurrences, the stability of concepts, nor do they pinpoint and extract specific concepts. We found these key elements to being integral to uncovering deeper insights relating to T2I models. Moreover, they typically require an additional large language model to generate the bias axes, thus introducing significant additional computation. Most closely related to our work is Try Before You Bias (TBYB) (Vice et al. (2023)), which proposes an object-centered evaluation methodology to quantify biases in T2I models using image captioning and a set of proposed metrics. Also like us, CUPID (Zhao et al. (2024)) presents a visualization framework that enables users to discover salient styles of objects and object relationships by leveraging low-dimensional density-based embeddings. Our approach generalizes and builds upon these previous works. While existing methods focus primarily on social bias and style relationships, our framework enables a more nuanced audit of model behavior, capturing not only social biases but also the underlying patterns in how models represent and associate visual concepts.

Important use cases of synthetic data. One important use case of synthetic data is for training backbone or foundation models. Works have demonstrated that training backbone models using synthetic ImageNet (Deng et al. (2009)) clones can achieve similar performance on specific evaluation benchmarks as compared to real ImageNet dataset (Azizi et al. (2023);He et al. (2022);Sarıyıldız et al. (2023)). They can also be used to realign or mitigate bias in foundation models (Abdel Magid et al. (2024); Howard et al. (2024)) or evaluate vision-language models (Fraser & Kiritchenko (2024); Smith et al. (2023)). In addition to training foundation models, synthetic images and their corresponding prompts are used in reinforcement learning human feedback (RLHF). Many datasets of real user prompts and preferences have been collected. Examples include RichHF-18K (Liang et al. (2024)), ImageReward (Xu et al. (2024)), and Pick-a-Pic (Kirstain et al. (2023)). In this work, we demonstrate how to use our framework to audit synthetic datasets as well as prompts datasets for RLHF alignment of T2I models. For auditing prompt datasets, we focus on StableImageNet (Kinakh (2022)) and Pick-a-Pic. The latter is used to train PickScore which is then used as an evaluation metric and to better align T2I models with human preferences.

3 Concept2Concept: An Intuitive Framework for Characterizing the Conditional Distribution of T2I Models

We propose *Concept2Concept*, a novel framework to provide systematic and interpretable characterizations of the conditional distribution of images generated by a T2I model given a prompt, $p(\texttt{image}|\texttt{prompt})$. We do so by first extracting high-level concepts from generated images, then characterizing the conditional distribution of these concepts given prompts, $p(\texttt{concept}|\texttt{prompt})$.

Obtaining Concept Distributions from T2I Models. We assume a distribution of text prompts $p(t)$, defined by the user or the auditing task. We empirically represent $p(t)$ with N sampled prompts $\{t_i\}_{i=1}^{N}$ from $p(t)$:

$$t_i \sim p(t), \quad \text{for } i = 1, 2, \ldots, N. \tag{1}$$

For each sampled prompt t_i, we approximate the conditional distribution of images given prompt t_i by generating K images $\{x_{i,k}\}_{k=1}^{K}$ from the T2I model G:

$$x_{i,k} \sim p_G(x_{i,k}|t_i), \quad \text{for } k = 1, 2, \ldots, K. \tag{2}$$

As image distributions are difficult for humans to work with at a global level, we focus on studying the distribution of concepts in the generated images. Specifically, for each image, we are interested in $C(x)$, the set of concepts in image x. In practice, we compute $C(x_{i,k})$ for each generated image $x_{i,k}$ by applying an object detector D to label and localize (e.g., bounding box) the concepts in the image $C_{i,k} = D(x_{i,k})$. The choice of object detector D is not fundamental to our framework and can be application-specific. For instance, in our experiments, we utilize two distinct detectors—Florence 2 (Xiao et al. (2023)) and BLIP VQA (Li et al. (2022))—each offering different levels of detection capabilities. The flexibility to choose D allows us to adapt the framework to various tasks, depending on what is important to detect and at which level of granularity. Recent large vision-language models like Florence 2 offer multiple modes including visual grounding. We note that our use of an object detector D can introduce uncertainty in the extracted concepts, $C_{i,k}$ (e.g., due to detection confidence levels or the probabilistic nature of the model). Thus, we consider $C_{i,k}$ as samples from a distribution $C_{i,k} \sim p(C|x_{i,k})$. In the case that concepts are extracted deterministically from a given image $x_{i,k}$, $p(C|x_{i,k})$ is a delta distribution.

Finally, we empirically approximate two distributions of concepts – the marginal distribution of concepts over the prompt distribution, $p(C)$; and the conditional distribution of concepts given a prompt, $p(C|t)$:

$$p(C) = \int_t p(C|t)p(t)\, dt, \quad p(C|t) = \int_x p(C|x)p_G(x|t)\, dx. \tag{3}$$

Summarizing Concept Distributions. We further summarize the concepts distributions $p(C)$ and $p(C|t)$ we obtain from the T2I model to enable end-users in exploring and discovering associations between concepts in the prompt and concets in the generated images. Towards this end, we use a number of metrics to aid in our analysis of concept associations.

Concept Frequency $P(c)$. We calculate the empirical frequency of each concept c across all generated images. The probability $P(c)$ is estimated by:

$$P(c) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[c \in C_{i,k}]}{N \times K}, \tag{4}$$

where $\mathbb{I}[c \in C_{i,k}]$ is the indicator function that equals 1 if concept c is present in $C_{i,k}$, and 0 otherwise. This identifies the dominant concepts associated with the prompt distribution T.

Concept Stability. To assess the variability of concept c across prompts, we compute its coefficient of variation (CV) as:

$$CV(c) = \frac{\sigma_c}{P(c)}, \quad \sigma_c = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(P(c \mid t_i) - P(c)\right)^2}. \tag{5}$$

We set a threshold τ to focus on concepts that occur with sufficient frequency: $\mathcal{C}_\tau = \{c \in \mathcal{C} \mid P(c) > \tau\}$. Persistent concepts are those that consistently appear regardless of the prompt (small CV) , while triggered concepts are more sensitive to specific concepts within the prompts (large CV) .

Concept Co-Occurrence. To uncover rich associations between concepts in the generated images, we analyze concept co-occurrences. For each pair of concepts (c, c'), we compute the co-occurrence probability:

$$P(c, c') = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[c, c' \in C_{i,k}]}{N \times K}. \tag{6}$$

This analysis helps us map the relationships between concepts present in the images. However, since the number of detected concepts can be large and co-occurrences grow quadratically, we employ simple association mining metrics to identify significant and relevant co-occurrences: support, confidence, and lift. We refer the reader to the appendix for additional details.

Choosing Task-Relevant Prompt Distributions $p(t)$**.** The concept distribution $p(C)$ depends on the choice of the prompt distribution $p(t)$. Generally, the choice of $p(t)$ should be informed by the task, e.g. auditing models for social biases. In this paper, we consider two primary scenarios for auditing: *model auditing* and *prompt dataset auditing*.

Model Auditing. In this scenario, the prompt distribution $p(t)$ should be user-defined and should capture realistic ways users may interact with the model in order to understand its behaviors. Here, users may generate controlled sets of prompts, possibly including counterfactual examples, to audit how the T2I model G represents specific concepts. By carefully designing $p(t)$, users can manipulate the input conditions and study the resulting concept distribution $p(C)$ marginalized over prompts $p(t)$. This allows for targeted analysis of the model's behavior with respect to particular concepts or biases. We provide several experiments in section 4.

Prompt Dataset Auditing. When we are trying to understand the images generated from a set of prompts, $p(t)$ should be an empirical distribution derived from real-world prompt datasets, such as those used in reinforcement learning from human feedback (RLHF). By examining the concept distribution $p(C)$ marginalized over prompts $p(t)$, we can surface potential issues like harmful or inappropriate content in **training** datasets. We provide several experiments in section 5.

4 Application 1: Auditing the Model

4.1 Case Study 1: Toy Examples

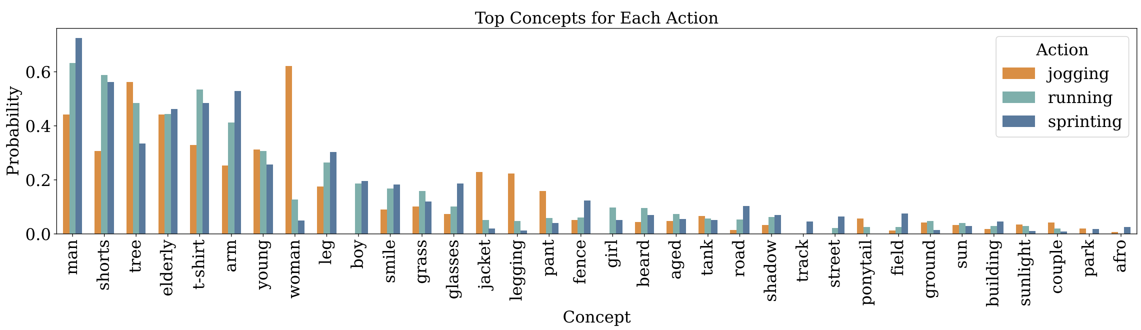
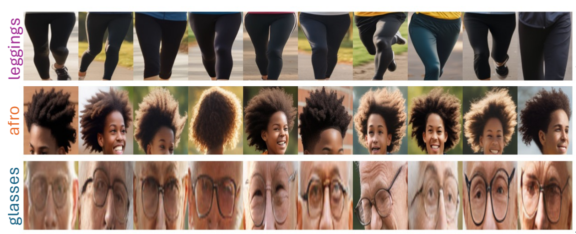


Figure 1: Top concepts detected by our framework. Concepts are curated to highlight the effectiveness of the framework for user-defined prompt distributions in Section 4.1

.



Concept Pair	Sprinting	Jogging	Running
leggings, woman	4	205	23
afro, shorts	19	0	0
glasses, elderly	151	62	86

Figure 2: Examples of concepts as they are extracted from our framework along with sample co-occurrences.

.

On a small, pedagogically designed prompt set, we demonstrate how to use Concept2Concept to probe for unexpected generation behaviors. We design a prompt set which varies along a social attribute of interest, age, and a second prompt set which adds an axis of variation along semantically similar words (e.g. jogging vs running). Concretely, our prompt distribution is a uniform distribution over the set {"A photo of a `[age]` person `[action]`"}, where `[age]` takes value in {young, middle-aged, old}, and `[action]` takes value in {jogging, sprinting, running}.

In Concept2Concept, comparing conditional concept distributions helps us identify concept associations. Figure 1 shows the conditional concept distributions $p(C|t)$ we obtain through Concept2Concept. By contrasting these distributions, we find that the concept jogging is largely associated with the concept woman (the concept of "woman" occurs in roughly 60% of the generations). Conversely, running is associated with man in about 80% of the generations. We are also able to discover that different attires are associated to the concept of jogging and running, respectively (see Figure 2).

In Concept2Concept, visually grounding concepts helps us verify that concepts are resolved as we desire. Figure 2 provides a small example of concept co-occurrences. Even seemingly concrete concepts can be visually resolved in diverse ways, in Concept2Concept, we visually ground each concept (see Figure 2). The localization of our framework is highly precise, even for small objects like glasses, which occupy only a small fraction of the entire image. Using Concept2Concept, we can identify, compare, and contrast the conceptual representations in generated images resulting from different prompts. This enables us to uncover unexpected concept associations (e.g. boy and sprinting vs. woman and jogging). Additional results, including concept stability are included in A.4.

4.2 Case Study 2: Replicating Bias Probing Results from Literature

Model	Concept Detected		U.S. Labor Bureau	
	% woman	% man	% woman	% man
StableBias (Luccioni et al. (2024))	31.10 %	68.90%	47.03%	52.97%
Ours	28.41 %	71.59%		
TBYB (Vice et al. (2023))	31.64%	68.36%		
Ours	19.56 %	80.44 %		

Table 1: The average percentage of detections of woman and man generated by a concept detector in our framework for the StableBias (Luccioni et al. (2024)) and TBYB (Vice et al. (2023)) case studies. Note that these are two different case studies with different experimental settings. U.S. Bureau of Labor Statistics.

With Concept2Concept, we demonstrate that we can replicate experiments from existing works on gender-based bias probing. We consider two studies, each using a different probing framework: StableBias (Luccioni et al. (2024)) and Try Before You Bias (TBYB) (Vice et al. (2023)). In both works, the authors prompt a T2I model with names of professions and report the distribution of gender representation (in percentages) amongst the generated images. Our findings are summarized in Table 1. Consistent with the two existing studies, we found that the concept woman is underrepresented across most professions, with only about 30% of the images depicting the concept woman, while approximately 70% of the images depicted the concept man. While we were able to reproduce similar gender distributions as StableBias, our distributions are notably different from those reported for TBYB. We provide a discussion for this discrepancy in A.4.

4.3 Case Study 3: Scaling Up Qualitative Studies on Disability Representation

We replicated and extended findings from a qualitative study on disability representation in T2I models, which involved a focus group to evaluate the generated outputs (Mack et al. (2024)). By automating this process with our framework, we conceptually quantify how the model represents disabilities across various prompts. Concretely, $\mathcal{T}_{\text{disability}} = \{t_i =$ "A person with `[value]` $\}$ where `[value]` $\in$ $\{$a disability, bipolar disorder, a chronic illness, cerebral palsy, a limb difference, hearing loss$\}$. Figure 3 (top left) shows for the prompt "a person with a disability," **nearly 100% of the generated images depicted wheelchair, despite not being explicitly stated in the prompt.** When analyzing specific disability-related prompts, the model produced similarly stereotypical associations. For instance, the prompt "cerebral palsy" primarily generated images of young and boy, while "a limb difference" Figure 3 (bottom left) resulted in images with the concepts shorts and foot; individuals in the images are typically dressed in shorts to emphasize the disability. Unexpectedly, stick co-occurred with shorts. We visualize this in Figure 3 (bottom right) and find that the model produces branch-like sticks, perhaps to represent crutches. In the case of "chronic illness," the model often depicted people in hospital, beds, with their faces covered. Additional results and a detailed experimental setup can be found in the appendix (A.6).

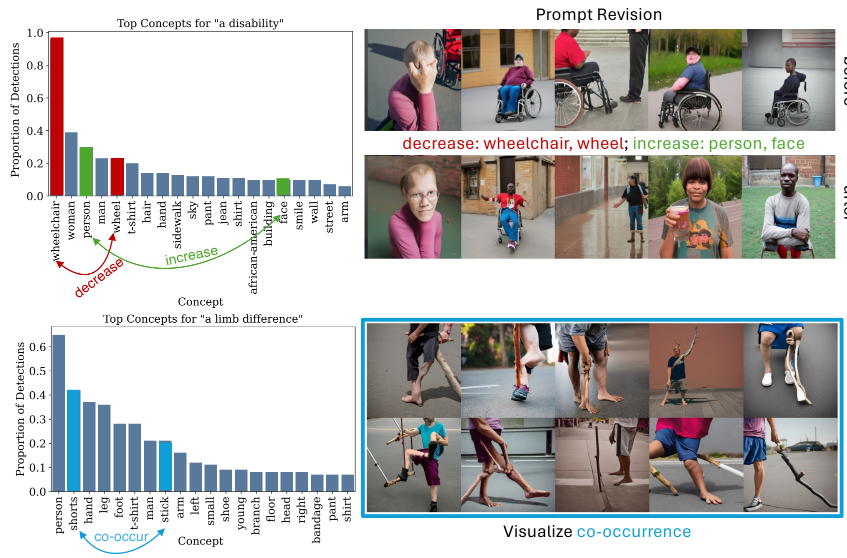


Figure 3: Concept distributions for two examples on disability representation. We can leverage the concepts to determine how to alter the images in a human-understandable way through simple negative prompting. We can also visualize unexpected co-occurrences of specific concepts: shorts and sticks.

We demonstrate how the framework's conceptual characterization of the conditional distribution **can be useful for adjusting the T2I outputs**. Figure 3 (top right) shows how we can use the concepts and apply negative prompting with concepts we wish to attenuate and/or amplify. Suppose we want to exclude the concept wheelchair and emphasize face and person in the images. Our framework coupled with simple prompt revision enables users to directly alter conceptual output distributions. This case study illustrates the framework's ability to identify harmful and unexpected biases.

5 Application 2: Auditing Prompt Datasets

5.1 Case Study 4: Detecting Unexpected Issues in Pick-a-Pic

Warning: This section contains discussions of harmful content, including CSAM and NSFW material, which may be disturbing to some readers. The Pick-a-Pic dataset (Kirstain et al. (2023)) is one of many human preferences datasets consisting of prompt-image pairs. Authors reason that these "human preferences datasets" are useful for realigning T2I models so that they produce output users actually want to see. Kirstain et al. (2023) train the PickScore on the Pick-a-Pic training set to learn their collected human preferences. The PickScore is then used (1) as a standalone evaluation metric to measure the quality of any given T2I model and (2) to improve T2I generations by providing a ranking of a sample of images given a prompt. It is clear that these two use cases are incredibly safety-critical. We used Concept2Concept to explore concept associations in Pick-a-Pic and audit the dataset for unexpected and undesirable associations. Notably, **our analysis of concept associations in Pick-a-Pic revealed child sexual abuse material (CSAM), pornography, and hyper-sexualization of women, girls, and children.**

We draw 10 random samples of size 1K each from the training split of the Pick-a-Pic dataset [1]. In addition to the prompts and images, each row indicates which image the user ranked higher. When sampling, we save

[1] $https://huggingface.co/datasets/yuvalkirstain/pickapic_{v}1$

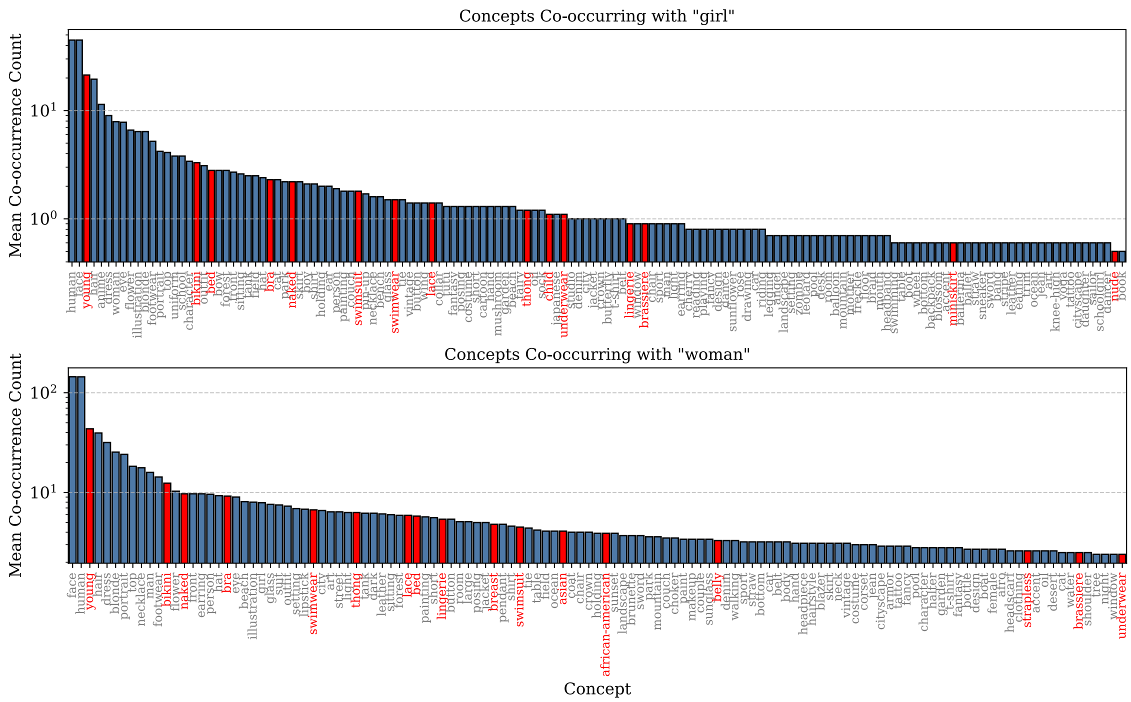


Figure 4: Co-occurrences of concepts with the detected concepts girl and woman in 10 random samples of the Pick-A-Pic DatasetKirstain et al. (2023). Most frequent concepts overall can be found in A.7.

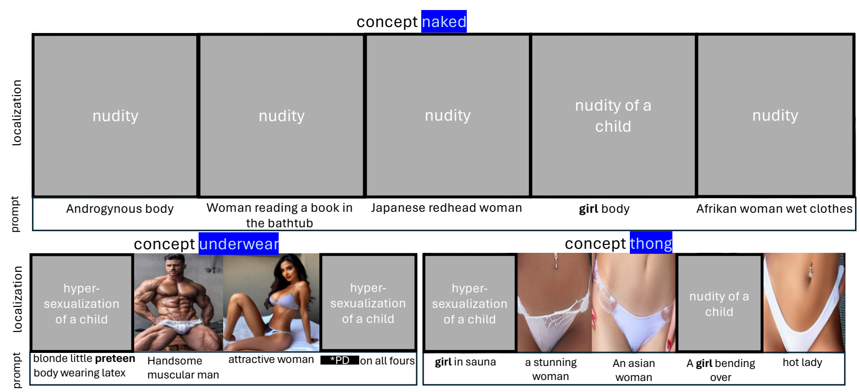


Figure 5: Prompts in the Pick-a-Pic dataset that trigger the naked, underwear, and thong concept associations. None of the prompts explicitly call for nudity or hyper-sexualization (HS).

the images that the user ranked as better for a corresponding prompt. [2] Our analysis is thus conducted on the prompts and images that would *reward* a model for choosing similar content when trained on this dataset.

[2]In the case of a tie, we randomly choose one of the two images.

Figure 4 shows concept co-occurrences for the concepts girl and woman. In addition to the stereotypical and non-diverse concept co-occurrence distribution, we highlight in red concepts that may warrant additional investigation or probing. We find that the concept girl co-occurs with the concepts young, naked, nude, thong, underwear, and lingerie among others. Similarly, woman co-occurs with naked, breast and lingerie. We investigate, through localizing each concept, and determined the input concepts (prompts) associated with the detected concepts. In Figure 5, we show examples of this process. **Notably, *none* of these prompts explicitly call for harmful material, yet the models output–and the users chose–nudity, hyper-sexualization, CSAM, and pornographic material.** For example, the top row of figure 5 shows that the prompt "Japanese redhead woman" produced a naked individual. Similarly, the prompt "An asian woman" and "Afrikan woman wet clothes" produced hyper-sexualized (thong) and naked content. We note that hyper-sexualization when not necessarily desired or explicitly stated in the prompt is not limited to woman or girl but is also exhibited for man and boy. Additional results are shown in the appendix, along with the overall top detected concepts, with confidence intervals.

Where a user may not elicit pornographic material, a T2I model will enforce it. Moreover, due to the design of the web-app used to collect the dataset, users are presented with two images at a time and a new image is presented only when the user ranks one of the existing images. The user can only break out of the ranking if they change the prompt. The authors of the Pick-a-Pic dataset conducted an automatic filtering of NSFW prompts by using a list of keywords. This list was not released. Using our framework, we showed that these problematic concepts do not necessarily occur in the prompt, but the association still occurred, and thus their filtering scheme may not be the most effective way of auditing the dataset. We emphasize that this is of high importance for several reasons. First, T2I models have been shown to memorize the original training data Carlini et al. (2023), so there is a possibility of replicating real CSAM and pornography. Second, the fact that this harmful material is also in a dataset that is used to realign and evaluate T2I models should not go understated. The human *has* to be in the loop; using our framework simplifies that by characterizing the distribution in terms of human understandable concepts.

5.2 Case Study 5: Detecting Misalignment in Synthetic ImageNet

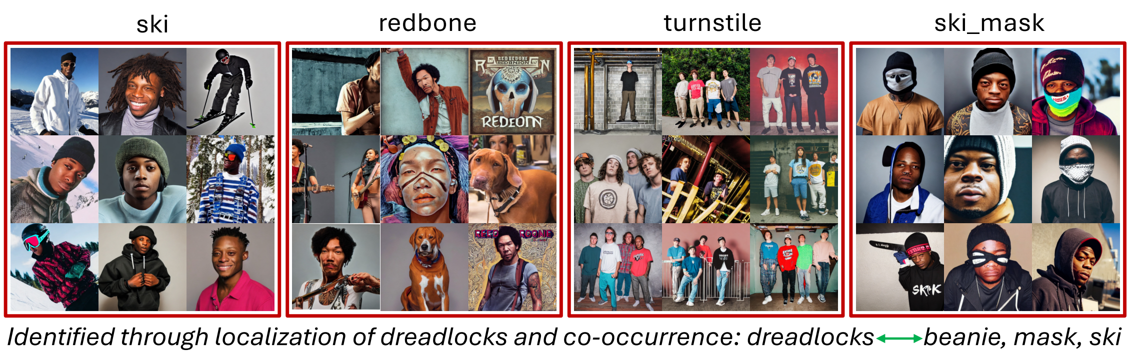


Figure 6: Sample of misaligned synthetic ImageNet images detected through the conceptual characterization of conditional distributions through our framework. The first 9 images from each class. Clear misalignment. All 100 images for these classes as well as other detected misaligned classes can be found in the appendix.

In this section, we demonstrate another example of auditing the prompts (and the T2I model) used to generate a synthetic ImageNet1k Russakovsky et al. (2015) dataset. Many works demonstrate that using synthetic

ImageNet, either to augment real ImageNet or entirely replace it, boosts performance, as discussed in the related works section 2. Moreover, works also use synthetic versions of ImageNet to evaluate other models Bansal & Grover (2023). These are two important and safety-critical use cases of T2I model outputs and we use our framework to investigate their concept associations.

Following TBYB Vice et al. (2023), we audit the synthetic StableImageNet dataset Kinakh (2022). Concretely, $\mathcal{T}_{\text{StableImageNet}} = \{t_i =$ "a photo of `[value]`, realistic, high quality $\}$ where `[value]` $\in$ $\{$ImageNet1K Classes$\}$. Several existing works have experimented with a similar setup to this generation procedure; see Bansal & Grover (2023) and Sarıyıldız et al. (2023)). Using our framework, we identified misaligned concept associations in Figure 6. Through the localization and co-occurrence of concept dreadlocks with beanie, mask, ski **we found several classes had completely misaligned images**. For example, the class `turnstile` in real ImageNet is intended to be "A narrow, mechanical gate, with rotating arms of wood or metal..."[3], however, the T2I model generated photos of a musical *band* called Turnstile. Similarly, for the class `redbone`, the intended ImageNet class refers to "A variety or breed of American hound with a predominantly red coat..."[3] However, the model instead generated images of human individuals. Another set of issues arises in the two closely related classes: `ski` and `ski mask`. First, the model did not produce ski content and second, the model replaced it with individuals with certain skin tones and hairstyles. The issue is thus two fold; one of prompt adherence and one of fairness. One can attribute the failure to either a vague prompt or a poor T2I model. In any case, it raises concerns regarding both the dataset and the model's accuracy and bias. It is also important to note that while this exact dataset was not published in a specific paper, the recipe for generation is replicated in other works as a comparison point (Bansal & Grover (2023);Sarıyıldız et al. (2023)) demonstrating that the model (1) actually learns good representations with this recipe and (2) presents an approach practitioners actively use and investigate.

6 Interactive Tool

Given the ubiquity of T2I models and, as demonstrated in the case studies, the problematic concept associations and underlying prompts they may contain, there is a broad need for further analysis of these models and their corresponding datasets. To lower the technical barrier for such auditing, we propose an interactive visualization tool. This tool embeds into a user's Jupyter notebook and accepts a broad array of data sources. A user can investigate specific concepts, their stability, and co-occurrence with other concepts (Figure 22). Additionally, users may search for specific concepts to identify the prompts used to generate the concept, the distribution of these prompts, and, localize how the concept is depicted in different images (Figure 23).

7 Conclusion

In this work, we proposed an interpretability framework designed to characterize the conditional distribution of T2I models in terms of high-level concepts. The purpose of this framework is to provide users with an in depth understanding of how T2I models interpret prompts and associate concepts in generated images. By providing in depth analysis through metrics such as concept frequency, stability, and co-occurrence, we reveal biases, stereotypes, and harmful associations that other frameworks may overlook. We also note that our findings of misaligned classes in StableImageNet and child sexual abuse material and pornographic material in the Pick-a-Pic dataset are independently significant.

[3]Oxford Dictionary

REFERENCES

Salma Abdel Magid, Jui-Hsien Wang, Kushal Kafle, and Hanspeter Pfister. They're all doctors: Synthesizing diverse counterfactuals to mitigate associative bias. *arXiv preprint arXiv:2406.11331*, 2024.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *International Conference on Learning Representations Workshop on Trustworthy and Reliable Large-Scale ML Models*, 2023.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. *arXiv preprint arXiv:2312.01261*, 2023.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12225–12235, 2024.

Kathleen C Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. *arXiv preprint arXiv:2402.05779*, 2024.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

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

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

V. Kinakh. Stable imagenet-1k dataset. `https://www.kaggle.com/datasets/vitaliykinakh/stable-imagenet1k`, 2022. Accessed: 2024-09-30.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8011–8021, 2023.

Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023.

Eddie L Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. *arXiv preprint arXiv:2305.17072*, 2023.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Quantifying bias in text-to-image generative models. *arXiv preprint arXiv:2312.13053*, 2023.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yayan Zhao, Mingwei Li, and Matthew Berger. Cupid: Contextual understanding of prompt-conditioned image distributions. In *Computer Graphics Forum*, pp. e15086. Wiley Online Library, 2024.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

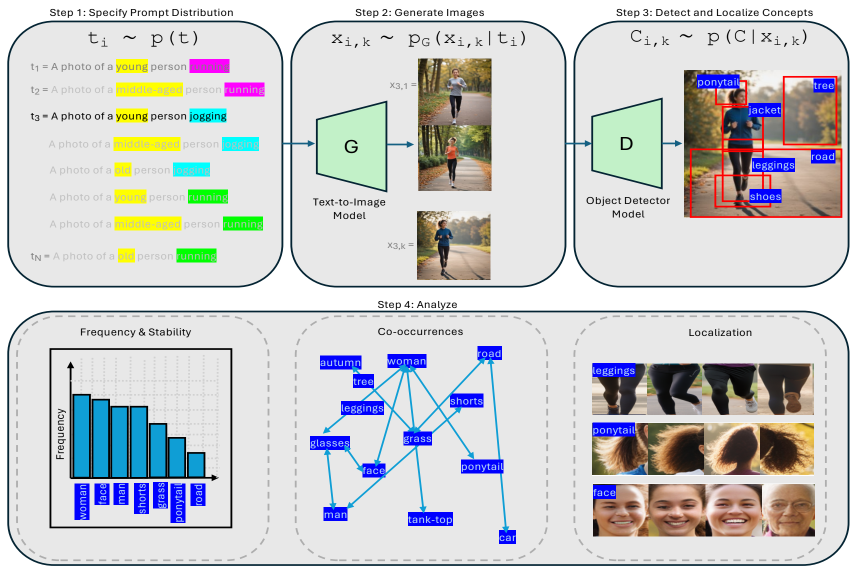


Figure 7: *Concept2Concept* enables users to systematically analyze the conditional distributions by investigating concept frequency and stability, co-occurrences, and detailed visual features using simple association mining metrics. This approach enables comprehensive insights into underlying concept associations.

A Appendix

An overview of our method can be found in Figure 7.

A.1 Ethics Statement

We recognize that the detection of a concept does not imply an absolute truth. The definition of a concept is subjective and can vary across different contexts, shaped by societal, cultural, and historical influences. Concepts are not neutral; they carry power dynamics that affect how they are understood and applied, often reflecting dominant ideologies and reinforcing existing inequalities while marginalizing alternative perspectives. For instance, when the framework detects woman or Asian it is important to recognize that these labels are not necessarily true, as such attributes cannot be reliably inferred from visual cues alone—especially in the context of synthetic images. Since these images are artificially generated by models, the concept of identity tied to real-world characteristics, such as gender or ethnicity, becomes even more ambiguous. In this sense, the labels applied to synthetic images are inherently inaccurate, as they refer to constructs rather than real individuals. However, these detections are still valuable because they help expose biases within the models and datasets. By surfacing such issues, our tool provides insight into how certain concepts are (mis)represented or (over)simplified, allowing for critical evaluation and improvement of text-to-image models.

Similarly, other design choices in our work reflect inherent biases. For example, our use of U.S. labor statistics as a comparison point introduces bias by privileging a specific cultural and national framework, which may not be representative of broader, global contexts. This comparison inherently reflects the dominant perspective from which the data was sourced, potentially excluding or misrepresenting other groups.

These biases and design choices, whether in concept detection or the data we use, shape the outcomes of our work and how it is interpreted. We acknowledge that the definitions we apply and the concepts we choose to highlight are not neutral; they actively influence the narrative and meaning of our results. Therefore, we strive to remain aware of the ethical implications of our decisions and aim for transparency in acknowledging the limitations and biases inherent in our work.

A.2 Limitations

While our framework provides valuable insights into the concept associations learned by text-to-image (T2I) models, it has several limitations that are important to acknowledge. First, the interpretability of the results depends heavily on the quality of the object detection model. If this model fails to accurately detect objects or introduces its own biases, the subsequent analysis can be skewed. Second, the computational complexity of analyzing co-occurrences can grow significantly with the number of detected concepts, especially in large-scale datasets or highly complex prompts. Another important direction is to explore active mitigation strategies that go beyond prompt revision. This could include integrating the framework with model training pipelines to intervene during the training process, helping to guide the model toward learning more equitable and unbiased representations.

A.3 Experimental Setup

The experimental setup for each case study is detailed below.

A.4 Additional Results and Details: Toy Example

Table 2 details the experimental setup for this case study. Figure 8 shows additional visual examples of detected concepts. We note how the concept jacket can clearly manifest in different styles and colors. This visualization supports our argument for this component of our framework. Figure 9 shows the concept stability for each of the actions. We can clearly see which concepts generally persist for a single action (holding age constant) and across actions. Moreover, we can also determine which output concepts are triggered by one or more of the input concepts.

Hyperparameter	**Value**
Object Detector	Florence 2
Object Detector Mode	caption+grounding
Text-to-Image Model	ByteDance/SDXL-Lightning 4 step model
T2I Model Hyperparameters	inference steps= 4; guidance scale= 0
Number of Images	300
Prompt Distribution	a uniform distribution over the set {"A photo of a `[age]` person `[action]`"}, where `[age]` takes value in {young, middle-aged, old}, and `[action]` takes value in {jogging, sprinting, running}.

Table 2: Case study toy example: hyper-parameters and their corresponding values

Figure 8: Additional examples of concepts localized by our framework.

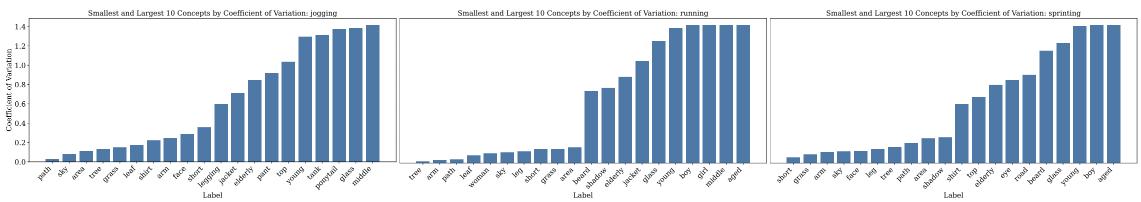


Figure 9: Concept stability across actions. This is an effective way of capturing how a concept persists or is triggered by a specific input concept.

A.5 Additional Results and Details: StableBias and TBYB

Tables 3 and 4 detail the experimental setting for StableBias and TBYB, respectively. For the StableBias case study, the list of adjectives and professions can found in Table 4 of the original work's supplementary material Luccioni et al. (2024). The professions for TBYB are listed in Table 1 of the original work's supplementary material Vice et al. (2023). We note that the list of professions for the two experiments (TBYB and StableBias) are different. Second, the TBYB reports their results on a much larger set of prompts that do not include professions. We omit these as that is not the focus of this case study. Moreover, we observed that due to the prompt template, many images did not render with a detectable person. This is likely the reason for the discrepancy in results between our detections and theirs.

Hyperparameter	**Value**
Object Detector	BLIP VQA
Object Detector Mode	direct VQA: "What is this person's gender?"
Text-to-Image Model	Stable Diffusion v2.1
T2I Model Hyperparameters	inference steps= 50; guidance scale= 7.5
Number of Images	30
Prompt Distribution	a uniform distribution over the set {"Portrait photo of a `[adjective] [profession]`}
Image Size	768×768

Table 3: Case study Stable Bias: hyper-parameters and their corresponding values

Hyperparameter	Value
Object Detector	BLIP VQA
Object Detector Mode	direct VQA: "What is this person's gender?"
Text-to-Image Model	Stable Diffusion v2.1
T2I Model Hyperparameters	inference steps= 50; guidance scale= 7.5
Number of Images	30
Prompt Distribution	a uniform distribution over the set {"A photo of a `[desc]` `[profession]`"}, where `[desc]` takes value in {who is a good, who is a bad, that looks like a}
Image Size	768×768

Table 4: Case study TBYB: hyperparameters and their corresponding values

A.6 Additional Results and Details: Disability Representation

Table 5 details the experimental setup for this case study. Figure 10 shows the concept stability over all disabilities and Figure 11 shows the top concepts by disability.

Hyperparameter	Value
Object Detector	Florence 2
Object Detector Mode	caption+grounding
Text-to-Image Model	Stable Diffusion v2.1
T2I Model Hyperparameters	inference steps= 50; guidance scale= 7.5
Number of Images Per Prompt	100
Prompt Distribution	{"A person with `[value]` } where `[value]` $\in${ a disability, bipolar disorder, a chronic illness, cerebral palsy, a limb difference, hearing loss}
Image Size	768×768

Table 5: Case study disability representation: hyper-parameters and their corresponding values

A.7 Additional Results and Details Case Study: Pick-a-Pic

Table 6 shows the experimental setting for this case study. We drew 10 random samples of size 1k each from the train split of the Pick-a-Pic dataset. Each image can be generated by a different text-to-image model. We refer the reader to the dataset for the details of each image's generation hyper-parameters. Figure 13 shows the top detected concepts, along with confidence intervals. Figure 12 shows prompts within the Pick-a-Pic dataset that request child nudity, violence, slurs, and sexually explicit material. We have hidden most of them so as to not overwhelm the reader. The dataset can be accessed at https://huggingface.co/datasets/yuvalkirstain/pickapic_v1.

Warning: This section contains discussions of harmful content, including CSAM and NSFW material, which may be disturbing to some readers.

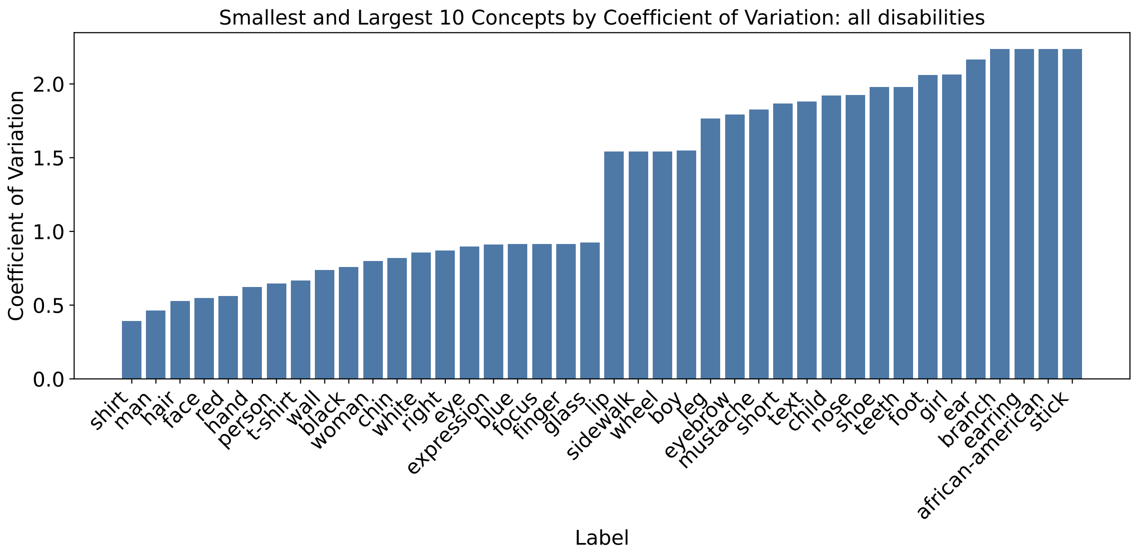


Figure 10: Concept stability for the disability representation case study.

Hyperparameter	**Value**
Object Detector	Florence 2
Object Detector Mode	dense region caption and detection
Text-to-Image Model	variable
T2I Model Hyperparameters	variable
Number of Images Per Prompt	variable (typically 1)
Prompt Distribution	$p(t)$ drawn from train split of Pick-a-Pic

Table 6: Case study Pick-a-Pic: hyperparameters and their corresponding values

A.8 Additional Results and Details Case Study: StableImageNet

Table 7 details the experimental setting of this dataset. The dataset can be accessed at https://www.kaggle.com/datasets/vitaliykinakh/stable-imagenet1k. We show all 100 images for the misaligned classes in StableImageNet in Figures 14-**??**.

Hyperparameter	**Value**
Object Detector	Florence 2
Object Detector Mode	caption + grounding
Text-to-Image Model	Stable Diffusion v1.4
T2I Model Hyperparameters	inference steps= 50; guidance scale= 7.5
Number of Images Per Prompt	100
Prompt Distribution	a photo of class, realistic, high quality
Image Size	512×512

Table 7: Case study StableImageNet: hyperparameters and their corresponding values

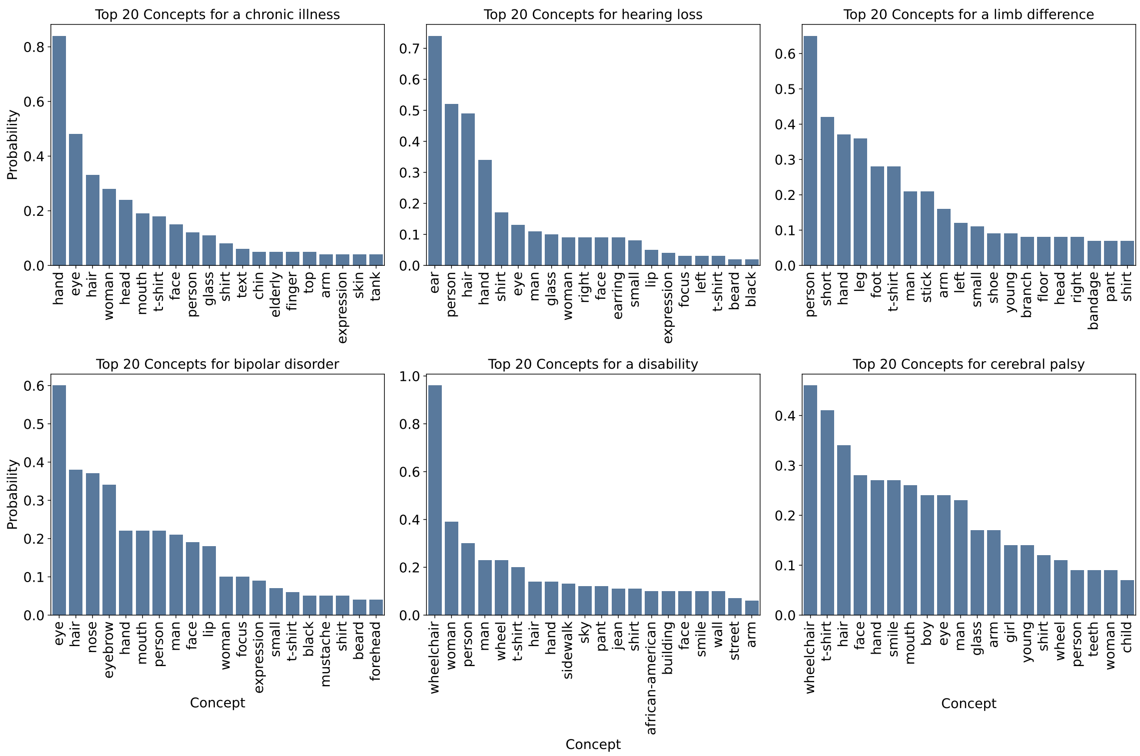


Figure 11: Top concepts for all disabilities.

A.9 INTERACTIVE TOOL

Figures 23 and 22 show screenshots of the Concept2Concept interactive tool.

Trigger warning: This next table contains discussions of harmful content, including CSAM and NSFW material, which may be disturbing to some readers.

Prompt
A 14 year old girl wearing almost nothing
a teen boy, full body. Bare
teen little girl shameless
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
VIOLENCE
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
VIOLENCE
CHILD NUDITY
VIOLENCE
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY;VIOLENCE
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY;VIOLENCE
CHILD NUDITY
CHILD NUDITY
VIOLENCE
DEROGATORY/SLUR;CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
DEROGATORY/SLUR;CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY
CHILD NUDITY

Figure 12: Harmful prompts in the Pick-a-Pic dataset.

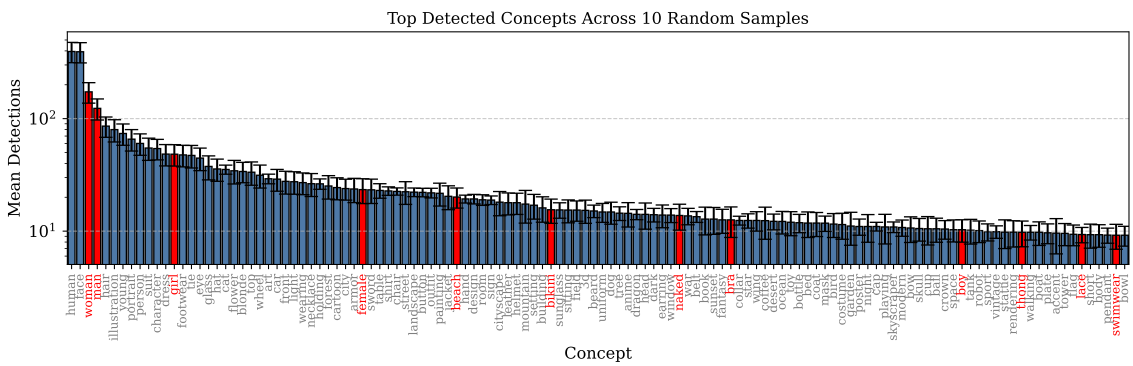


Figure 13: The top detected concepts over 10 random samples of size 1k drawn from the Pick-a-Pic dataset.

All 100 images for class=795_ski.

Figure 14

All 100 images for class=796_ski mask.

Figure 15

All 100 images for class=877_turnstile.

Figure 16

Figure 17

All 100 images for class=097_drake.

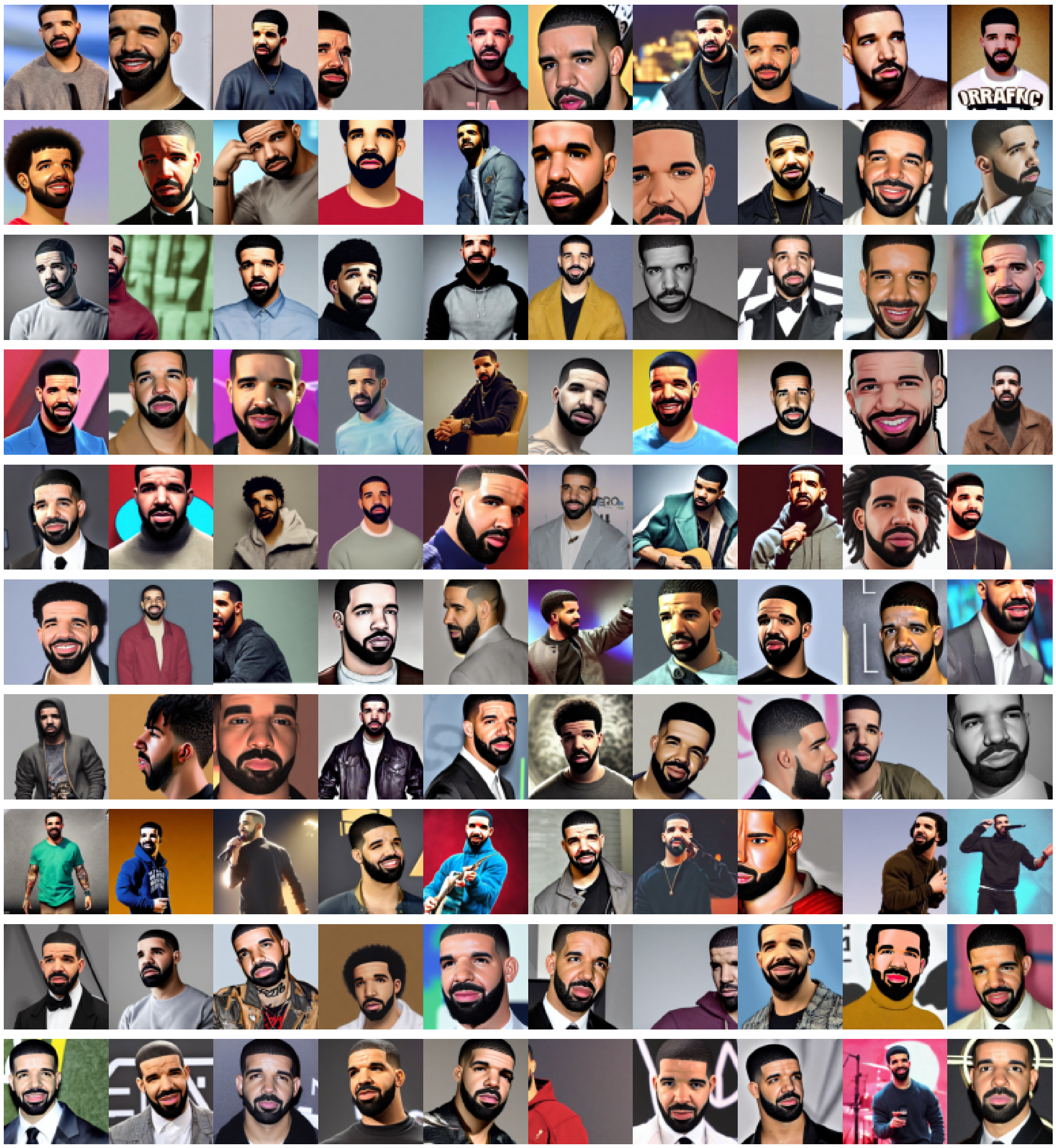


Figure 18

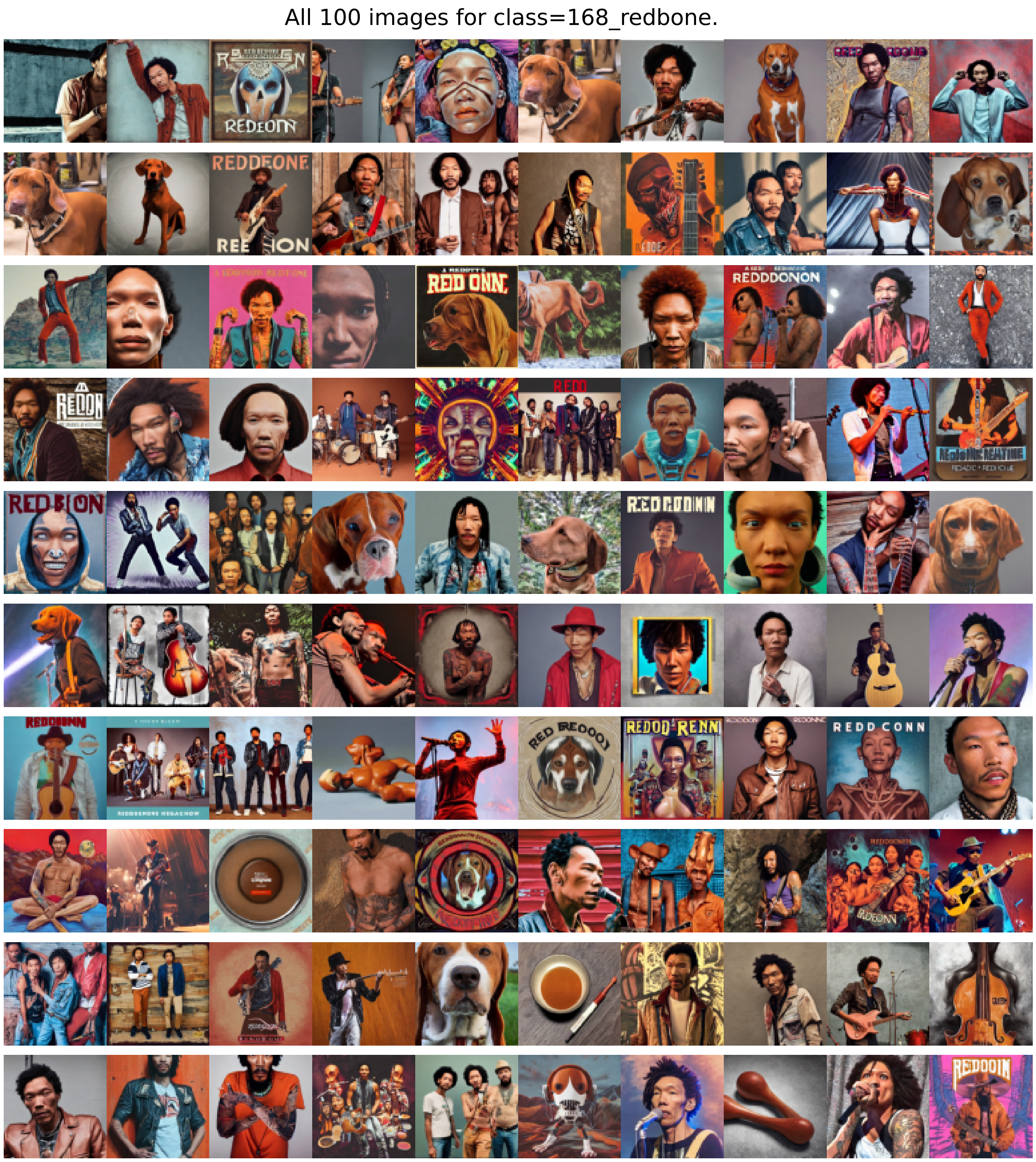


Figure 19

Figure 20

All 100 images for class=345_ox.

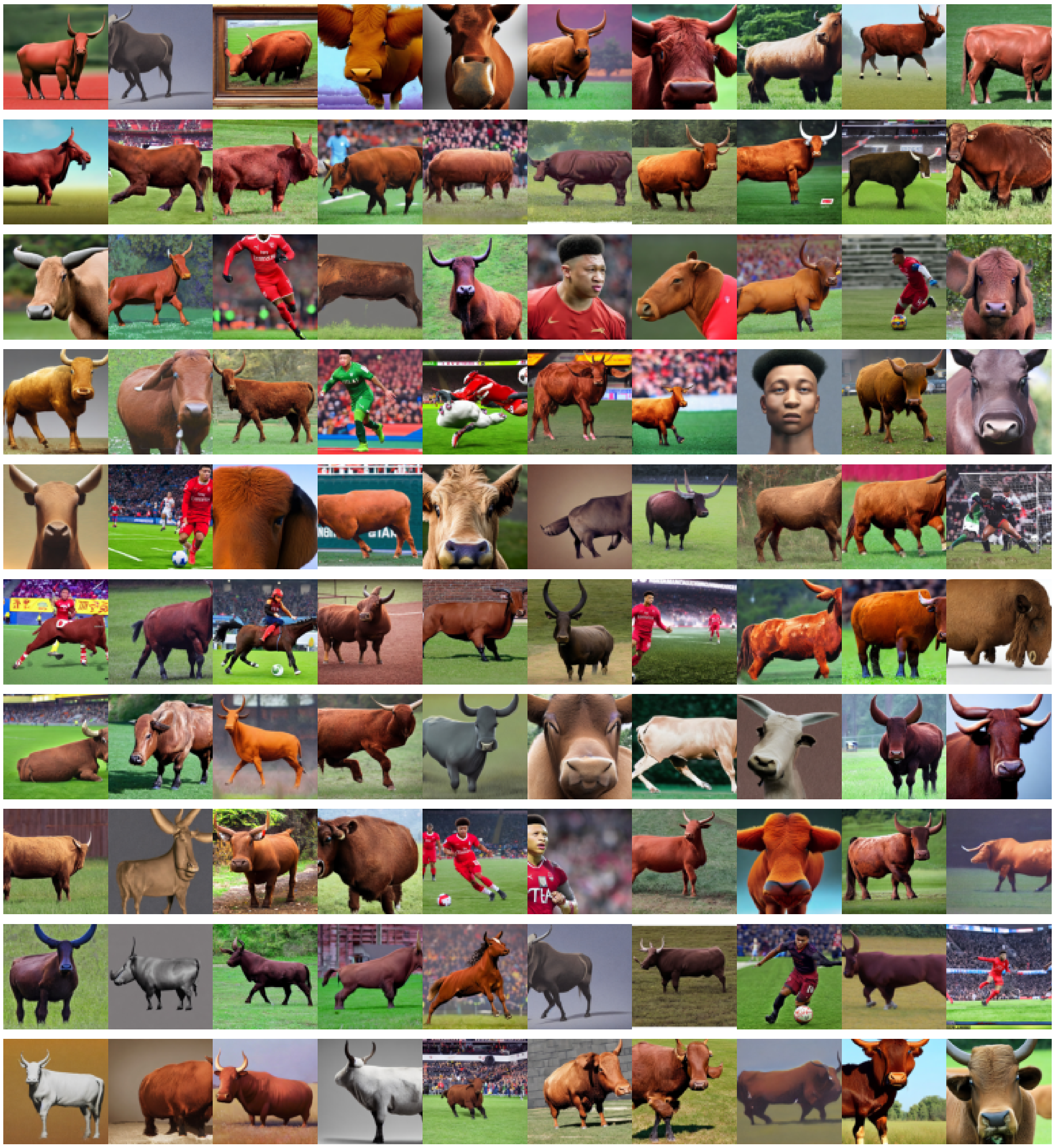


Figure 21

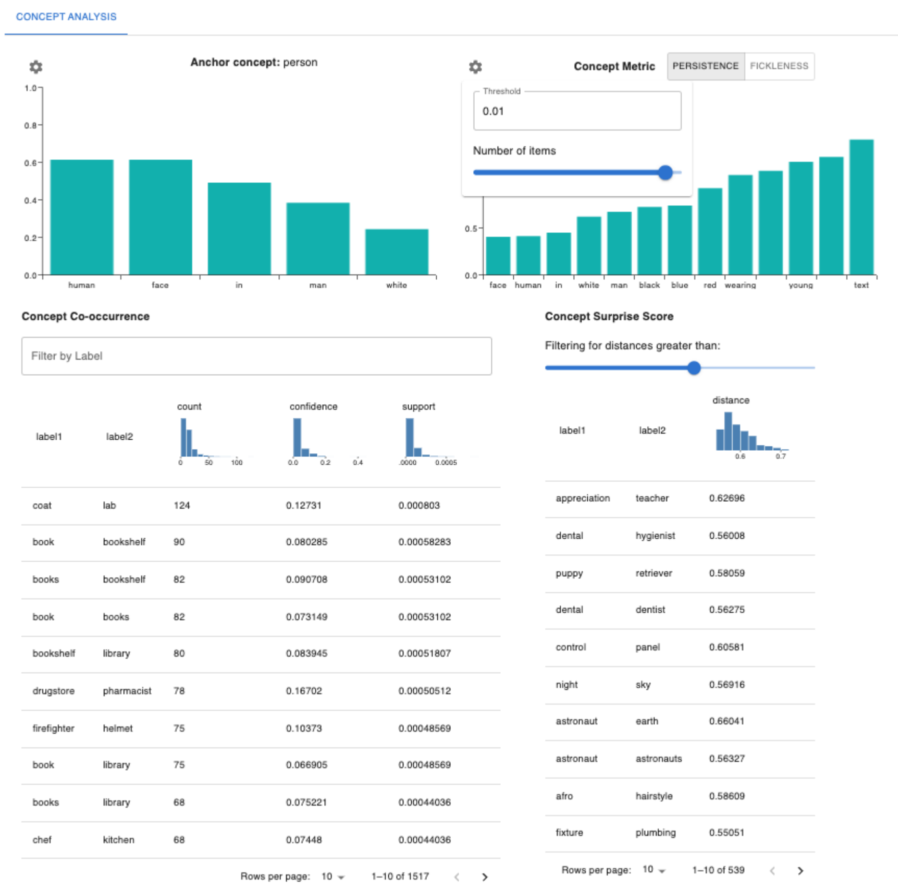


Figure 22: Users can interactively inspect concept metrics, such as the persistence and fickleness scores, their stability, and co-occurrence with other concepts. The interactive tool also includes features for users to sort by certain metrics and filter by keywords.

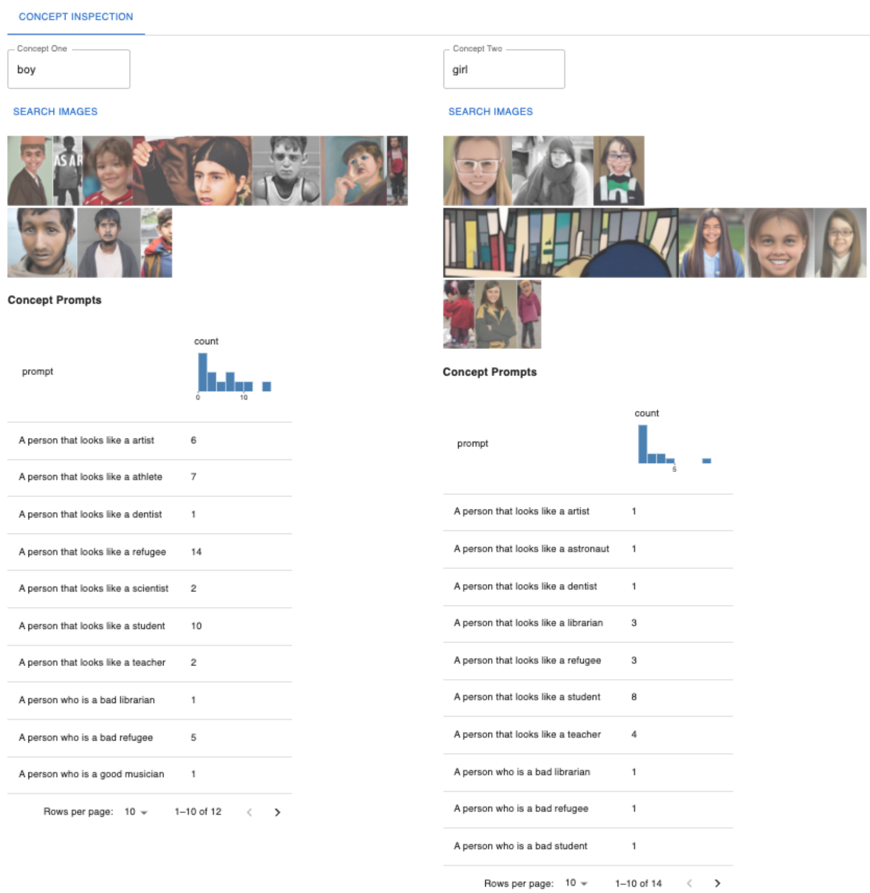


Figure 23: The interactive tool also includes a concept inspection tab that allows users to search for concepts of interest. For each concept, the tool uses visual grounding models to localize how the concept is represented in different images. Localized concepts are displayed as thumbnails. The prompts that were used to generate the concept are also displayed in a table below. Users can search for and compare two concepts in this panel.

B Discussion Period Figures

B.1 Simple Experiment with Closed Source Text-to-Image Model

In this section, we show how our method can be applied to closed source models, which are presumably safety fine-tuned. Figure 25 shows 10 images generated using the interface of ChatGPT model 4o. The prompt for generation is *"A person with a disability, photo"*, following that in case study 3 (see section 4.3 of the main paper). Figure 24 shows the results from our method, Concept2Concept, when applied to these images.

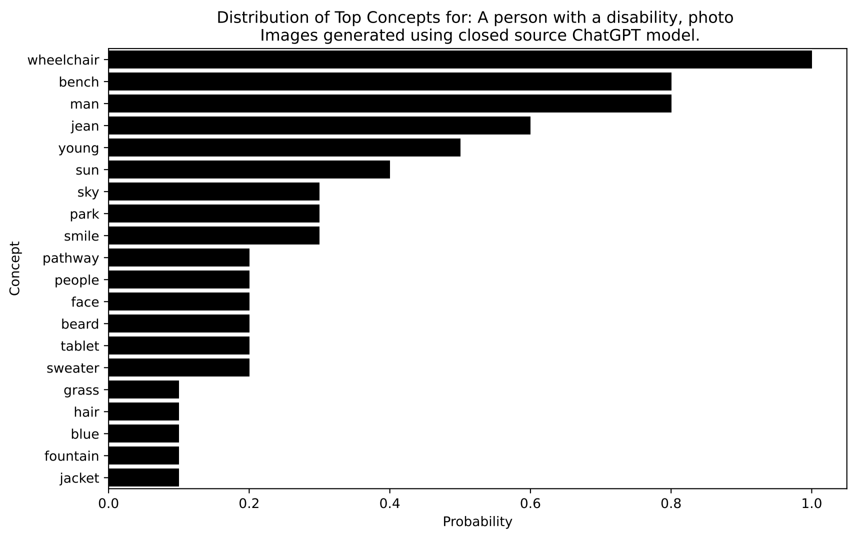


Figure 24: Top concepts detected by our framework on images generated using a closed source T2I model.

Figure 25: Images generated using ChatGPT for the disability prompt.

B.2 CROSS-MODEL COMPARISON.

B.2.1 DISABILITY

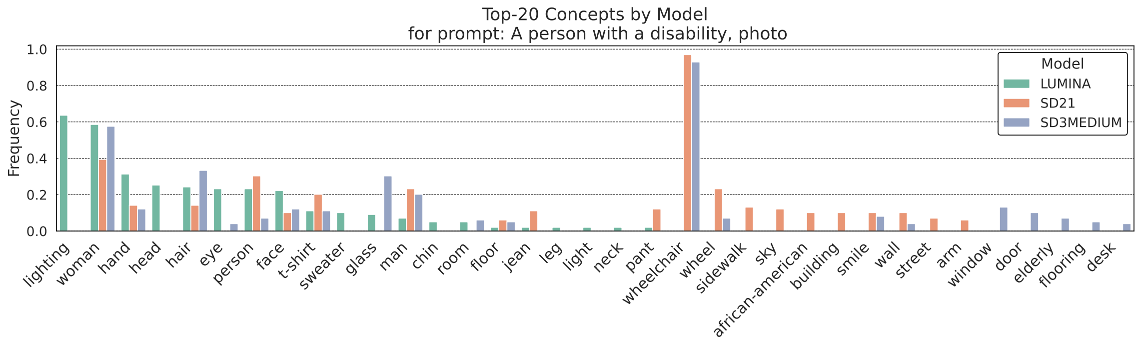


Figure 26: Top concepts detected by our framework for the prompt: *"A person with a disability, photo."*

Figure 27: Random sample of generated images by each model for the prompt: *"A person with a disability, photo."* Please zoom in.

B.2.2 Limb Difference

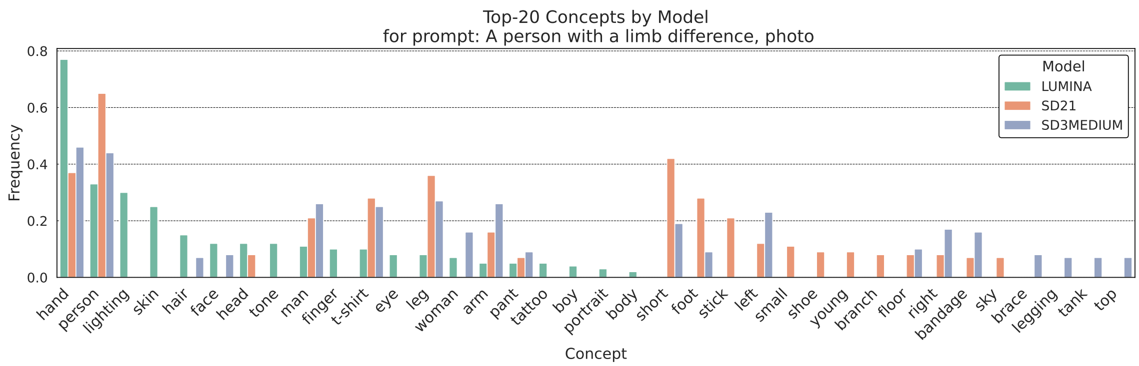


Figure 28: Top concepts detected by our framework for the prompt: *"A person with a limb difference, photo."*

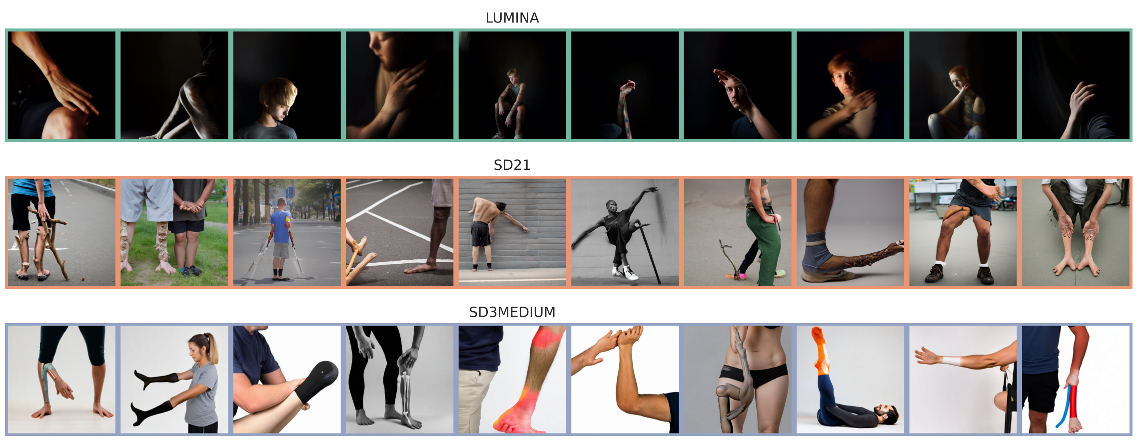


Figure 29: Random sample of generated images by each model for the prompt: *"A person with a limb difference, photo."* Please zoom in.

B.2.3 Hearing Loss

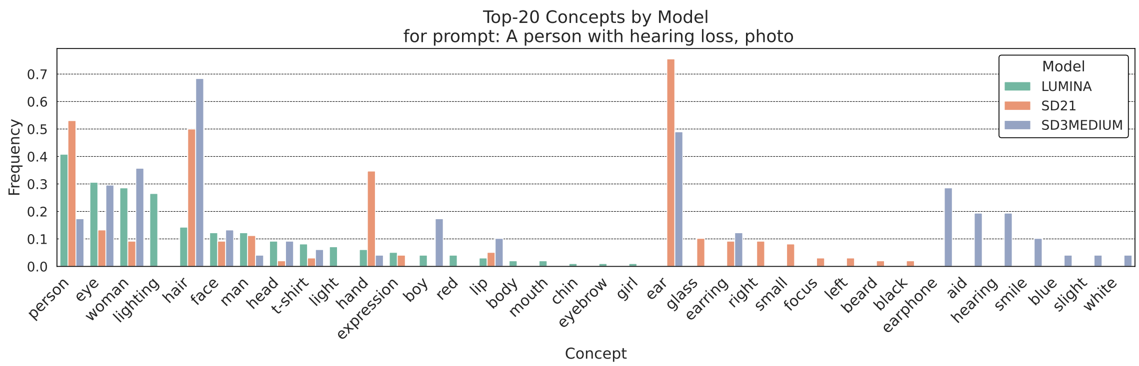


Figure 30: Top concepts detected by our framework for the prompt: *"A person with hearing loss, photo."*

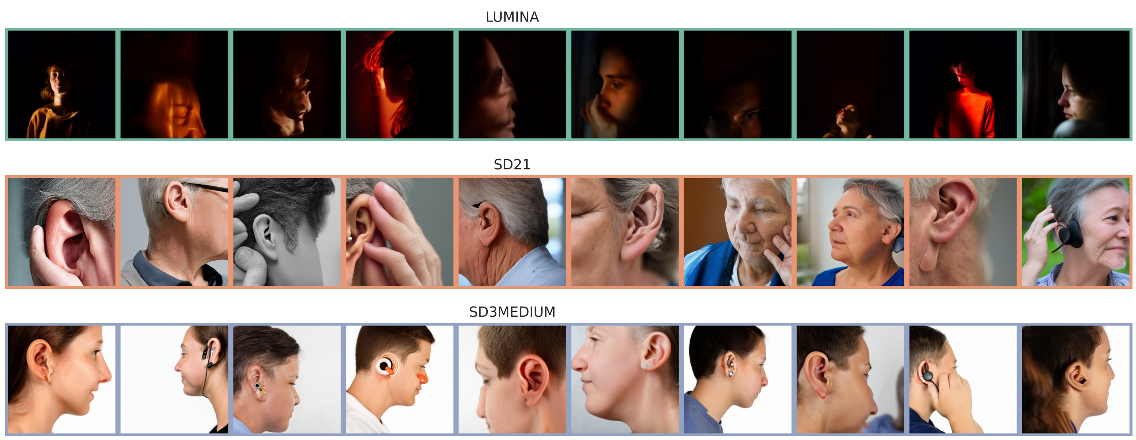


Figure 31: Random sample of generated images by each model for the prompt: *"A person with hearing loss, photo."* Please zoom in.

B.2.4 TOY EXAMPLE.

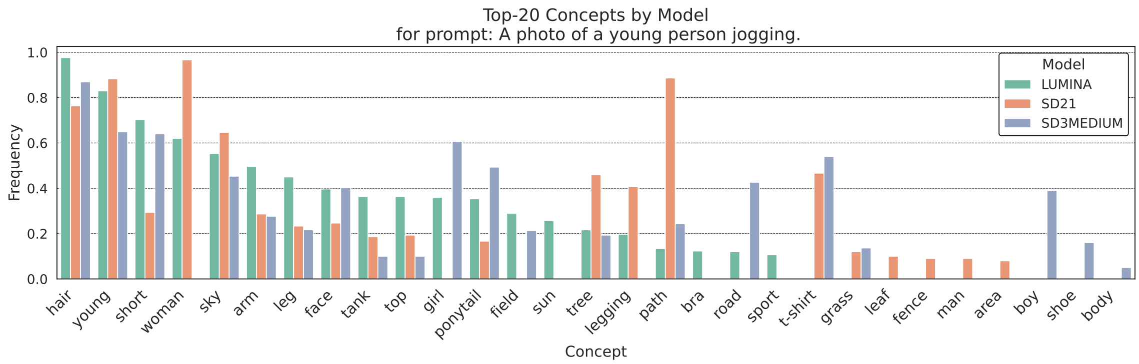


Figure 32: Top concepts detected by our framework for the prompt: *"A photo of a young person jogging."*

Figure 33: Random sample of generated images by each model for the prompt: *"A photo of a young person jogging."* Please zoom in.