Discovering Divergent Representations between Text-to-Image Models

Lisa Dunlap[1]* Joseph E. Gonzalez[1] Trevor Darrell[1] Fabian Caba Heilbron[2]
Josef Sivic[2,3] Bryan Russell[2]
[1]University of California, Berkeley [2]Adobe Research [3]CIIRC CTU

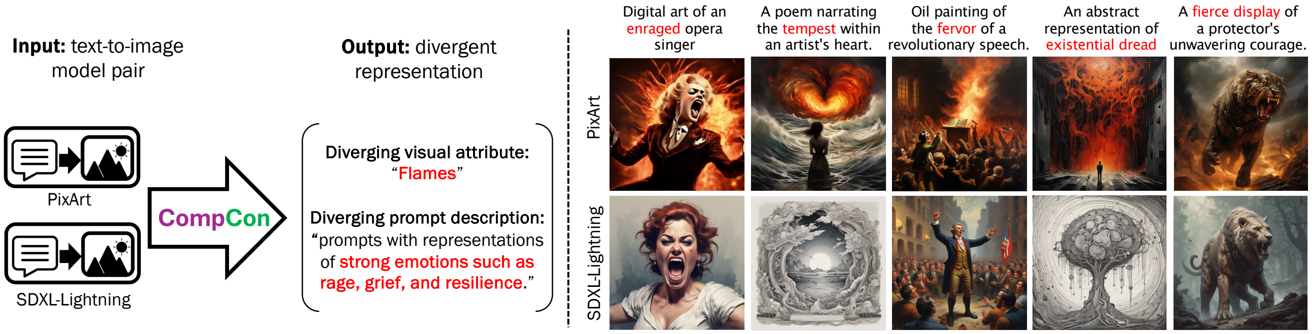

Input: text-to-image model pair
Output: divergent representation
PixArt
SDXL-Lightning
CompCon
Diverging visual attribute: “Flames”
Diverging prompt description: “prompts with representations of strong emotions such as rage, grief, and resilience.”
Digital art of an enraged opera singer
A poem narrating the tempest within an artist's heart.
Oil painting of the fervor of a revolutionary speech.
An abstract representation of existential dread
A fierce display of a protector's unwavering courage.
PixArt
SDXL-Lightning

Figure 1. **Discovering divergent representations with CompCon**. *Left:* CompCon takes as input a pair of text-to-image models and outputs a diverging prompt description to produce a diverging visual attribute appearing in one model but not the other. *Right:* We show the discovered diverging visual attribute ‘flames’ appearing in PixArt but not SDXL-Lightning over different diverging prompts.

Abstract

In this paper, we investigate when and how visual representations learned by two different generative models **diverge** *from each other. Specifically, given two text-to-image models, our goal is to discover visual attributes that appear in images generated by one model but not the other, along with the types of prompts that trigger these attribute differences. For example, ‘flames’ might appear in one model’s outputs when given prompts expressing strong emotions, while the other model does not produce this attribute given the same prompts. We introduce CompCon (Comparing Concepts), an evolutionary search algorithm that discovers visual attributes more prevalent in one model’s output than the other, and uncovers the prompt concepts linked to these visual differences. To evaluate CompCon’s ability to find diverging representations, we create an automated data generation pipeline to produce ID^2, a dataset of 60 input-dependent differences, and compare our approach to several LLM- and VLM-powered baselines. Finally, we use CompCon to compare popular text-to-image models, finding divergent representations such as how PixArt depicts prompts mentioning loneliness with wet streets and Stable Diffusion 3.5 depicts African American people in media professions.*

1. Introduction

Generative models develop unique representations of semantic concepts – for instance, happy scenes contain warm colors, or dogs are found outside. While many of these representations are shared across models, understanding when representations **diverge** can reveal stylistic differences between models. In this work, we explore how to uncover such divergent representations by identifying input-dependent differences between two text-to-image models. Specifically, we aim to discover pairs of semantic concepts and visual attributes where prompts containing a semantic concept cause one model to generate images displaying the corresponding visual attribute, while the other does not. For example, prompts that mention strong emotions result in images with flames in one model but not the other (Fig. 1).

Discovering divergent representations is beneficial for both model developers and users. For developers, it can help decide which model to deploy to production based on any problematic discovered divergent representations, and for evaluating against competitor models. For users, it can help in selecting the model that best aligns with their own interpretations and needs. Manually performing this task is labor-intensive, as it requires sifting through hundreds or thousands of images to find visual attribute differences; once identified, additional effort is needed to determine the types of input prompts that trigger these differences.

[1] Work done during an internship at Adobe Research.

When comparing text-to-image models, the evaluation typically focuses on metrics such as image quality and prompt adherence [22, 24–26, 30, 37, 40]. While these metrics indicate *how well* models perform, they often overlook *what* the models actually learn. For example, what defines 'cute' versus 'ugly'? What characteristics make something appear 'futuristic'? What does 'emotion' look like? As we will show, models trained on different data, using different encoders or training procedures, can learn distinct interpretations of the same concept. For instance, one model may associate 'ancient' with the Paleolithic Era while the other associates it with the Roman Empire.

To address the task of discovering divergent representations, we make the following contributions. First, we introduce CompCon (Comparing Concepts), an evolutionary search algorithm designed to uncover input-dependent differences in model representations. CompCon first discovers pairwise differences in model outputs, generates a description of the prompts that cause this difference, and iteratively refines this description by analyzing existing prompts and generating new ones likely to highlight these differences. As shown in Figure 1, CompCon can generate prompts that result in model behaviors like putting flames behind opera singers for one model but not the other.

Second, we create ID^2 (Input-Dependent Differences), a dataset of 60 semantic-visual representations to evaluate the efficacy of our method. Using this dataset, we compare CompCon to LLM, TF-IDF, and VisDiff [17] baselines. As our third contribution, we apply CompCon to compare the PixArt and SD-Lightning text-to-image models, finding, for example, that prompts mentioning anger result in depictions of 'flames' in PixArt, prompts mentioning sadness and solitude result in 'wet streets' in PixArt, and abstract prompts with cosmic motifs result in 'mandala circular designs' in SD-Lightning. We also uncover bias, such as PixArt generating older men for prompts mentioning traditional professions. These findings demonstrate how CompCon can systematically reveal subtle differences between generative models, helping developers and users better understand and leverage these models' unique behaviors.

2. Related Work

Evaluating Text-to-Image Models. The evaluation of text-to-image models has advanced significantly in recent years. Traditional quality measures such as FID [25], Inception core [40], CLIP score [24], and CLIP-R score [37] have been complemented by newer metrics like TAIM [22] and TIFA [26], which leverage vision-and-language models to assess prompt adherence. Holistic benchmarks such as HEIM and others [10, 30, 39, 47] aggregate multiple axes of evaluation, including quality, prompt adherence, style, and efficiency. While these approaches excel in measuring overall model performance, they focus on objective qualities and often overlook fine-grained, subjective differences between models. Our work complements these efforts by targeting input-dependent differences, particularly in semantic interpretations and stylistic variations, which are crucial for understanding model-specific behaviors.

Interpreting Diffusion Models. Several works have explored ways of interpreting the internal representations of generative vision-language models. Bau et al. [4], Dravid et al. [15], Gandelsman et al. [19, 20] explore how to describe the function of certain neurons and attention heads in natural language,, and Tong et al. [44] discovers how image and text representations differ in latent space to better understand CLIP failures. We see our work as a data-driven approach to attain similar insights into model behavior.

Discovering Bias in Diffusion Models. Uncovering and mitigating biases in text-to-image models has been well explored, with many works focus on finding and mitigating a predefined set of biases related to gender, race, and geography [5, 11, 18, 21, 23, 45, 46]. Recently, a line of works have emerged that aim to automatically discover biases from the data, rather than using a predefined list. Many of these approaches [13, 14] discover bias by prompting a large language model (LLM) to propose potential biases from image captions, generating prompts that may indicate a bias in models (*e.g.*, "a doctor") and using a VLM to check if this bias exists. Liu et al. [34] builds on this by clustering generated images based on concepts like gender, while Chinchure et al. [8] extends bias discovery to counterfactual examples, eliminating the need for a large caption pool.In contrast, our work focuses on model *comparison*, rather than single-model auditing, which better reflects many real-world evaluations where success is measured by improvement relative to other models. Additionally, while previous methods [13, 14] identify biases from input captions, CompCon analyzes generated images directly using VLMs. This enables discovery of subtler, more nuanced differences in visual representation, including social biases and stylistic or conceptual divergences between models. Further discussion is in Sections 5.4 and G.1 of the Appendix.

Describing Differences in Image Sets. Several works have aimed to describe differences in sets of images using natural language. For example, VisDiff [17] generates visual attributes distinguishing two sets of images by analyzing captions and refining them using cross-modal embeddings. Similarly, Chiquier et al. [9] train interpretable CLIP classifiers that evolve based on LLM-generated attributes. While these approaches focus on dataset-level differences, they do not address input-dependent variation. We adapt and extend VisDiff's methodology for pairwise comparison, focusing on prompt-specific divergences and refining attribute discovery to capture subtler differences. Additionally, our method introduces a novel iterative search for prompt descriptions that cause these differences.

3. Divergent Representation Discovery

Let P be a set of text prompts, and $\mathcal{I}_1^{(P)}$ and $\mathcal{I}_2^{(P)}$ be sets of images generated by two text-to-image models given the input prompts P. We call the natural language descriptions of the prompt set P and any visual differences between the two generated image sets $\mathcal{I}_1^{(P)}$ and $\mathcal{I}_2^{(P)}$ a *divergent representation* (see Figure 1). These divergent representations take the form of a pair of natural language descriptions (a, d_{P_a}), where a is a description of a visual attribute seen more often in one model than the other (*e.g.*, "flames"), and d_{P_a} is a description of the concepts present in text prompts P_a eliciting this difference (*e.g.*, "strong emotions"). We refer to a as a *diverging visual attribute*, d_{P_a} as a *diverging prompt description*, and P_a as *diverging prompts*.

Given a pair of text-to-image models $\Theta = (\theta_1, \theta_2)$, we aim to discover differences in the images generated by the two models as well as a description of the types of input text prompts that trigger these differences. Let $\mathcal{P}$ be the set of all possible text prompts and $\mathcal{A}$ be the set of all possible diverging visual attributes. Our goal is to find the mapping $\mathcal{F}_\Theta$ from text prompts in $\mathcal{P}$ to diverging visual attributes in $\mathcal{A}$ given the model pair Θ,

$$\mathcal{F}_\Theta : \mathcal{P} \mapsto \mathcal{A}. \tag{1}$$

Note that this mapping is not a bijection as multiple text prompts may map to a single diverging visual attribute. Moreover, multiple diverging visual attributes may be depicted for a given set of diverging prompts.

Our approach for computing the mapping $\mathcal{F}_\Theta$ comprises two steps, illustrated in Figure 2. First, given the text-to-image model pair Θ and a large set of initial prompts $\mathcal{P}_0 \subset \mathcal{P}$, we compute a set of diverging visual attributes $\mathcal{A}_0 \subset \mathcal{A}$ (Section 3.1). Next, for each diverging visual attribute $a \in \mathcal{A}_0$, we optimize an objective to find the set of diverging prompts $\mathcal{P}_a \subset \mathcal{P}$ resulting in the diverging visual attribute a (Section 3.2). We next describe each of these steps.

3.1. Discovering Diverging Visual Attributes

Our goal is, given a text-to-image model pair Θ and a large set of initial text prompts $\mathcal{P}_0$, to compute a set of diverging visual attributes $\mathcal{A}_0$ over images generated given the prompts. This task is challenging as a system must identify any consistent visual differences between the two models' generated image sets. These differences are often subtle and difficult to spot over the large generated image collection. We address this challenge by prompting an off-the-shelf vision-language model (VLM) for this task.

For the text-to-image model pair Θ and a text prompt p, let $\mathcal{G}(\Theta, p) = \left\{\mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}\right\}$ denote the two sets of images generated by each model given prompt p. We first sample a batch of prompts $\mathcal{P}_{batch} \subset \mathcal{P}_0$ and, for each prompt $p \in \mathcal{P}_{batch}$, we construct a two-row image grid by tiling the images in $\mathcal{I}_1^{(p)}$ on the top row and $\mathcal{I}_2^{(p)}$ on the bottom row. Using this image grid, we instruct a VLM to find diverging visual attributes appearing more in images of $\mathcal{I}_1^{(p)}$ compared to $\mathcal{I}_2^{(p)}$ (see Appendix for our instruction prompt). Our resulting diverging visual attribute list $\mathcal{A}_0$ is the aggregation of discovered attributes across $\mathcal{P}_{batch}$.

Next, we rank each diverging visual attribute $a \in \mathcal{A}_0$ by assigning a score indicating how well attribute a can distinguish image sets $\mathcal{I}_1^{(\mathcal{P}_0)}$ and $\mathcal{I}_2^{(\mathcal{P}_0)}$. For each set of images generated by prompt $p \in \mathcal{P}_0$, we define a *divergence score* $z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}) \to \{0, 1\}$ that indicates whether image set $\mathcal{I}_1^{(p)}$ contains attribute a while $\mathcal{I}_2^{(p)}$ does not.

Using cross-modal similarity, here CLIP [38], we compute the cosine similarity $s(\cdot)$ between diverging visual attribute a and each image in sets $\mathcal{I}_1^{(p)}$ and $\mathcal{I}_2^{(p)}$. Using these similarities, we define the divergence score as the product of two indicated conditions,

$$\begin{aligned} z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}) = & \mathbb{1}\left[s(a, \mathcal{I}_1^{(p)}) > t\right] \times \\ & \mathbb{1}\left[s(a, \mathcal{I}_1^{(p)}) - s(a, \mathcal{I}_2^{(p)}) > \delta\right] \end{aligned} \tag{2}$$

where t and δ are hyperparameters that determine if $\mathcal{I}_1^{(p)}$ contains attribute a and $\mathcal{I}_2^{(p)}$ does not contain a.

Using this divergence score, we define the overall score for attribute a as the mean divergence score over prompts in the initial prompt set: $\frac{1}{|\mathcal{P}_0|}\sum_{p\in\mathcal{P}_0} z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)})$. A score of 1 means that all images generated by model θ_1 contain attribute a and none of the images generated by θ_2 contain a. A score of 0 indicates that a never appears more often in images generated by θ_1. Note that we are not optimizing for scores close to 1, we are simply interested in any attribute a that obtains a score sufficiently above zero. Finally, as many attributes $\mathcal{A}_0$ are semantically equivalent (e.g. ”flames” and ”fire”) we prompt an LLM to remove similar attributes.

3.2. Discovering Diverging Prompt Descriptions

After we discover a set of diverging visual attributes $\mathcal{A}_0$, for each attribute $a \in \mathcal{A}_0$ we aim to find a natural language description $d_{\mathcal{P}_a}$ of the diverging prompts that trigger this attribute. This task is challenging as the search space over all possible text prompts $\mathcal{P}$ is large, and we must not only find a set of prompts but a natural language description that completely covers this set.

Let $\mathcal{L}(d_{\mathcal{P}_a})$ be the diverging prompts generated from description $d_{\mathcal{P}_a}$. Our objective is to maximize the expected divergence score over the generated prompts:

$$\max_{d_{\mathcal{P}_a}} \mathbb{E}_{p\sim\mathcal{L}(d_{\mathcal{P}_a})}[z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)})], \tag{3}$$

where z is the divergence score defined in Equation (2). That is, we want to maximize the number of prompts that have been generated by description $d_{\mathcal{P}_a}$ and confirmed to

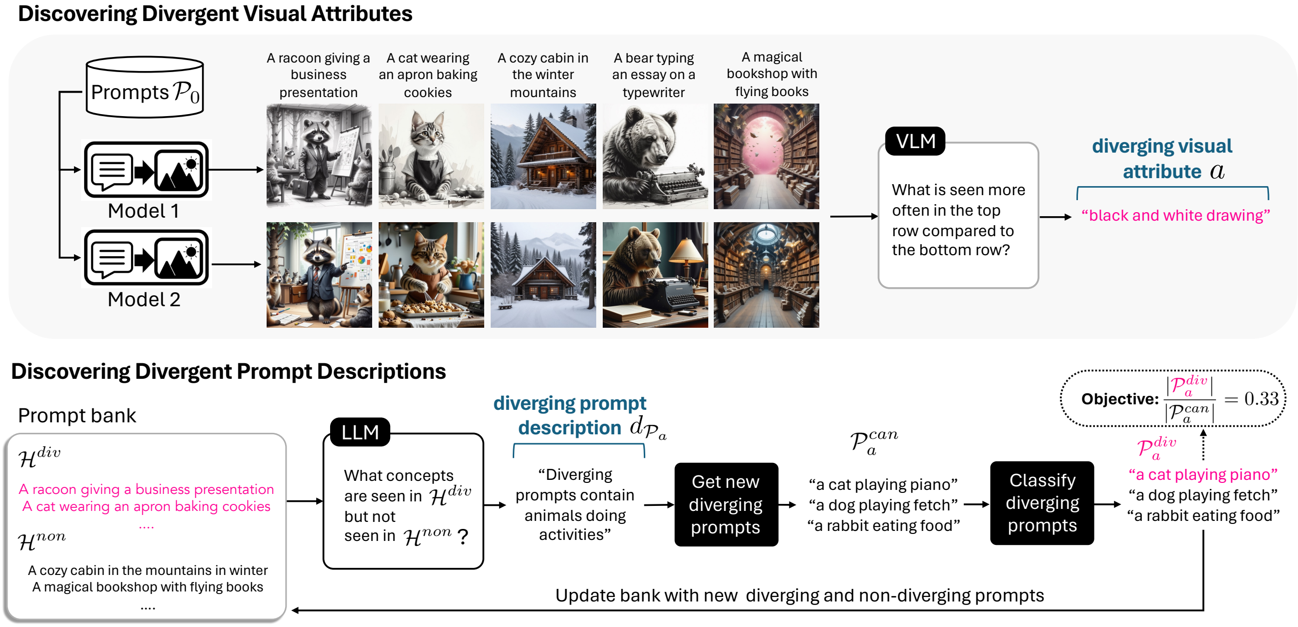

Discovering Divergent Visual Attributes
Prompts $\mathcal{P}_0$
Model 1
Model 2
A racoon giving a business presentation
A cat wearing an apron baking cookies
A cozy cabin in the winter mountains
A bear typing an essay on a typewriter
A magical bookshop with flying books
VLM
What is seen more often in the top row compared to the bottom row?
diverging visual attribute a
"black and white drawing"
Discovering Divergent Prompt Descriptions
Prompt bank
$\mathcal{H}^{div}$
A racoon giving a business presentation
A cat wearing an apron baking cookies
....
$\mathcal{H}^{non}$
A cozy cabin in the mountains in winter
A magical bookshop with flying books
....
LLM
What concepts are seen in $\mathcal{H}^{div}$ but not seen in $\mathcal{H}^{non}$?
diverging prompt description $d_{\mathcal{P}_a}$
"Diverging prompts contain animals doing activities"
Get new diverging prompts
$\mathcal{P}_a^{can}$
"a cat playing piano"
"a dog playing fetch"
"a rabbit eating food"
Classify diverging prompts
$\mathcal{P}_a^{div}$
"a cat playing piano"
"a dog playing fetch"
"a rabbit eating food"
Objective: $\frac{|\mathcal{P}_a^{div}|}{|\mathcal{P}_a^{can}|} = 0.33$
Update bank with new diverging and non-diverging prompts

Figure 2. **CompCon overview.** We illustrate our approach for discovering diverging visual attributes (top) and diverging prompt descriptions (bottom). Given two text-to-image models and a set of prompts, we use a VLM to identify visual differences. For each diverging attribute, we iteratively refine diverging prompt descriptions by generating candidate prompts $\mathcal{P}_a^{can}$ from the description, classifying them as diverging ($\mathcal{H}^{div}$) or non-diverging ($\mathcal{H}^{non}$). The objective is to maximize the proportion of generated prompts classified as diverging.

Algorithm 1 Discovering diverging prompt descriptions

Input: Model pair Θ, diverging visual attribute a, initial set of prompts $\mathcal{P}_0$, number of iterations N
Output: Diverging prompt description $d^\star_{\mathcal{P}_a}$

```
Initialize: prompt bank $\mathcal{H} \leftarrow \emptyset$, scores $\sigma \leftarrow$ emptyArray($N$), descriptions $d_{\mathcal{P}_a} \leftarrow$ emptyArray($N$)
$\mathcal{P}_a \leftarrow$ classifyDiverging($\Theta, a, \mathcal{P}$)
$\mathcal{H}^{div} \leftarrow \mathcal{P}_a$
$\mathcal{H}_a^{non} \leftarrow \mathcal{P} \setminus \mathcal{P}_a$
for $i$ in $1, \ldots, N$ do
    $h_a^{div}, h_a^{non} \leftarrow$ sample($\mathcal{H}^{div}, \mathcal{H}^{non}$)
    $d_{\mathcal{P}_a}[i] \leftarrow$ describeDiverging($\mathcal{P}_a^{div}, \mathcal{P}_a^{non}, \mathcal{H}$)
    $\mathcal{P}_a^{can} \leftarrow$ getNewDiverging($d_{\mathcal{P}_a}[i], a, \mathcal{H}$)
    $\mathcal{P}_a^{div}, \mathcal{P}_a^{non} \leftarrow$ classifyDiverging($\Theta, a, \mathcal{P}_a^{can}$)
    $\sigma[i] \leftarrow |\mathcal{P}_a^{div}|/|\mathcal{P}_a^{can}|$
    $\mathcal{H}^{non} = \mathcal{H}^{non} \cup \mathcal{P}_a^{non}$
    $\mathcal{H}^{div} = \mathcal{H}^{div} \cup \mathcal{P}_a^{div}$
end for
$d^\star_{\mathcal{P}_a} \leftarrow d_{\mathcal{P}_a}$[argmax($\sigma$)]
return $d^\star_{\mathcal{P}_a}$
```

be diverging. We explore two options for our prompt generator $\mathcal{L}$. The first is to use an LLM to generate new prompts given $d_{\mathcal{P}_a}$, and the second is to retrieve prompts from a large prompt bank using $d_{\mathcal{P}_a}$.

As the desired diverging prompt description $d_{\mathcal{P}_a}$ is discrete (natural language) and Objective (3) is not differentiable, we optimize the objective via evolutionary search. To achieve this goal, for N evolutionary search iterations, we use an LLM to generate description $d_{\mathcal{P}_a}$ given a bank of diverging and non-diverging prompts $\mathcal{H}^{div}$ and $\mathcal{H}^{non}$, generate new prompts and images from $d_{\mathcal{P}_a}$, add these new prompts to $\mathcal{H}^{div}$ and $\mathcal{H}^{non}$, and evolve our description using these updated sets.

We maintain a bank $\mathcal{H} = \left(\mathcal{H}^{div}, \mathcal{H}^{non}\right)$ of diverging and non-diverging text prompts, where diverging prompts are given by $\mathcal{H}^{div} = \{p \mid z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}) = 1\}$ and non-diverging prompts are given by $\mathcal{H}^{non} = \{p \mid z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}) = 0\}$. We mutate the current description $d_{\mathcal{P}_a}$ by prompting an LLM (GPT-4o) to provide a description of what concepts are shared in $\mathcal{H}^{div}$ but not in $\mathcal{H}^{non}$ and score the mutation using Objective (3). The mutation prompt is provided in the Appendix.

Algorithm 1 provides pseudocode for our evolutionary search algorithm, with its key functions described below.

Mutation function describeDiverging**.** Given the bank of diverging and non-diverging text prompts $\mathcal{H}$, we sample B prompts from $\mathcal{H}^{div}$ and $\mathcal{H}^{non}$ and instruct an LLM to output a description $d_{\mathcal{P}_a}$ (the diverging prompt description) of what concepts are shared across diverging prompts which are not seen in non-diverging prompts. After the first iteration, prompts in $\mathcal{H}$ which were generated in the previous iteration are up-weighted when sampling.

Mutation scoring. To score the current mutation, we first define a function getNewDiverging that, given the diverging prompt description $d_{\mathcal{P}_a}$, diverging visual attribute a, and prompt bank $\mathcal{H}$, provides candidate prompts $\mathcal{P}_a^{can}$ that are likely to be diverging and do not directly relate to attribute a. We explore two ways of obtaining $\mathcal{P}_a^{can}$: generation and retrieval. In the generation setting, we prompt an LLM (GPT-4o) to generate a diverse set of k new prompts that align with the description $d_{\mathcal{P}_a}$, given random samples of prompts from $\mathcal{H}$ as a point of reference. The instruction prompt can be found in the Appendix. In the retrieval setting we use description d_{P_a} to retrieve the top k prompts from the prompt bank $\mathcal{H}$ excluding the B sampled prompts used to generate the prompt description, having the highest text embedding similarity to d_{P_a}. We show results of both of these approaches in Section 5 and provide further implementation details in 5.1.

Next, we define a function `classifyDiverging` that, given the model pair Θ, visual attribute a, and candidate prompts $\mathcal{P}_a^{can}$, finds the diverging prompts $\mathcal{P}_a^{div} = \{p \in \mathcal{P}_a^{can} \mid z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}) = 1\}$. If more than one image is generated per prompt, we define a prompt as diverging if the majority of the generated images result in a diverging score $z(a, i_1^{(p)}, i_2^{(p)}) = 1$ for $i_1^{(p)} \in \mathcal{I}_1^{(p)}$ and $i_2^{(p)} \in \mathcal{I}_2^{(p)}$.

Finally, we approximate the expectation in Objective (3) by the ratio of diverging to candidate prompts set sizes,

$$\mathbb{E}_{p\sim\mathcal{L}(d_{\mathcal{P}_a})}[z(a, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)})] = \frac{|\mathcal{P}_a^{div}|}{|\mathcal{P}_a^{can}|} \tag{4}$$

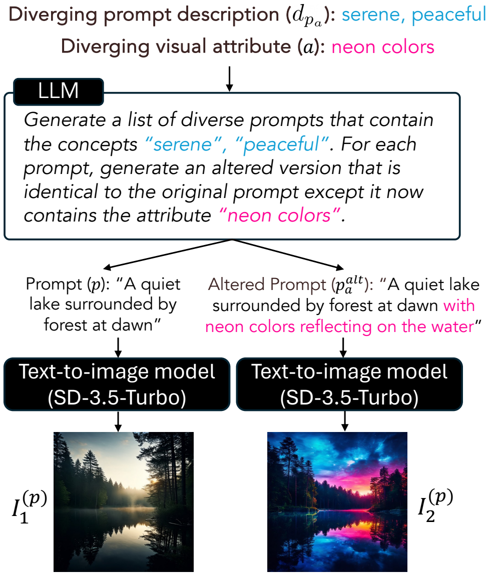
where $|\cdot|$ denotes the size of the set. We return the diverging prompt description $d_{\mathcal{P}_a}$ with the highest ratio.

4. ID^2 Dataset and LLM Evaluation

To systematically evaluate methods for discovering divergent representations, we created ID^2 (Input-Dependent Differences), a benchmark dataset containing 60 divergent representations between text-to-image models. Each representation consists of a diverging visual attribute and its corresponding diverging prompt description. Moreover, we include for each diverging representation a set of prompts that align with the diverging prompt description and, for each prompt, pairs of generated images where the diverging visual attribute is depicted in one image but not the other.

Creating such a benchmark is challenging because divergent representations between models are not known *a priori*, and manual annotation across the vast prompt space is impractical. To address this, we use a simulation approach: rather than comparing two distinct models, we use a single model (SD-3.5-Turbo [1]) and simulate differences by modifying input prompts to include specific visual attributes.

For each divergent representation in ID^2, we generate paired prompts where one explicitly mentions the visual attribute while the other does not. Both prompts are processed by the same text-to-image model, creating image pairs that exhibit controlled, systematic differences. This simulation approach allows us to create ground truth data with known divergent representations against which we can evaluate discovery methods. To assess the quality of discovered representations, we developed an LLM-based evaluation framework that measures both attribute similarity and description accuracy compared to ground truth. This framework provides objective metrics to compare different approaches for identifying divergent representations between text-to-image models. We describe the dataset creation and evaluation process below and in Figure 3.



Diverging prompt description (d_{p_a}): serene, peaceful
Diverging visual attribute (a): neon colors
LLM
Generate a list of diverse prompts that contain the concepts "serene", "peaceful". For each prompt, generate an altered version that is identical to the original prompt except it now contains the attribute "neon colors".
Prompt (p): "A quiet lake surrounded by forest at dawn"
Altered Prompt (p_a^{alt}): "A quiet lake surrounded by forest at dawn with neon colors reflecting on the water"
Text-to-image model (SD-3.5-Turbo)
Text-to-image model (SD-3.5-Turbo)
$I_1^{(p)}$
$I_2^{(p)}$

Figure 3. **ID^2 creation.** Given a diverging prompt description d_{p_a} and diverging visual attribute a, we use an LLM to generate prompt pairs where one of the prompts mentions the diverging visual attribute. Both prompts are then passed to the same text-to-image model to generate image pairs with the visual difference a.

ID^2 Dataset Creation. Each divergent representation (a, d_{P_a}) in ID^2 comprises a ground truth diverging visual attribute a and ground truth diverging prompt description d_{P_a}. We generate the diverging representations using either GPT-4o, Claude 3.5 Sonnet [3], or manually. Each diverging visual attribute a is a short phrase, and diverging prompt description d_{P_a} is a list of semantic concepts that a prompt should include. These representations cover diverse categories, such as related concepts ('red roses', ['love', 'romance']), abstract representations ('geometric shapes', ['efficiency', 'productive']), and bias ('overweight person', ['lazy', 'unmotivated']). The full list of diverging representations are located in the Appendix.

Given a diverging representation (a, d_{P_a}), we generate a corresponding set of prompts and generated images $\{(p, \mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)})\}$ such that each prompt p aligns with diverging prompt description d_{P_a} and the accompanying pair of generated images $\mathcal{I}_1^{(p)}$ and $\mathcal{I}_2^{(p)}$ has diverging visual attribute a appearing in one generated image but not the other. We illustrate the process for creating this set in Figure 3. First, we use GPT-4o to generate a prompt p that aligns with diverging prompt description d_{P_a} along with an altered prompt p^{alt} that contains attribute a. Both prompts are then given to a diffusion model θ (SD-3.5-Turbo) to generate images such that $\mathcal{I}_1^{(p)} = \theta(p)$ and $\mathcal{I}_2^{(p)} = \theta(p^{\text{alt}})$. The authors manually inspect these generations to ensure the prompts align with the diverging prompt description d_{P_a}, and that diverging visual attribute a is seen in the majority of images corresponding to $\mathcal{I}_1^{(p)}$ but not $\mathcal{I}_2^{(p)}$. We further validate these sets with human studies described later in the section.

For each of the 60 diverging representations, we generate 3 image pairs per prompt across 50 prompts, resulting in two sets of 150 images. Finally, for each representation, we also include a set of distractors – image pairs from 200 randomly generated prompts where $\mathcal{I}_1^{(p)}$ and $\mathcal{I}_2^{(p)}$ are generated with the same prompt p over different random seeds, resulting in image pairs with no discernible difference in visual attributes. The purpose of these distractors is to assess the methods' ability to generate descriptions and attributes in the presence of noise (no diverging visual attribute present).

LLM Evaluation. When given ground truth representations (a, d_{P_a}) and predicted representations (a', d'_{P_a}), we use an LLM-as-a-judge (GPT-4o) to compute an *attribute score* and a *description score*, which respectively measure the similarity of the predicted visual attribute and prompt description to the ground truth. The prompt given to the judge can be found in the Appendix.

These scores range from 0 to 1, where 1 indicates perfect agreement, 0.5 indicates partial agreement, and 0 indicates no agreement. Partial agreement means that the predicted attribute or description is related to the ground truth. For example, the ground truth visual attribute is 'flames' and the predicted attribute is 'a red color palette' would be deemed a partial alignment. To maintain a fair evaluation with other baselines, we prompt our mutation LLM to structure descriptions d_{P_a} as a short list of semantic concepts.

Human Validation of ID^2 and LLM Evaluation. We conducted a two-stage validation study with 4 PhD students, each annotating 10 randomly sampled sets from ID^2 (2 annotations per set). Participants identified concepts appearing more frequently in one set versus another and concepts shared across prompts. They then assessed whether our ground truth diverging attribute and diverging prompt description aligned with their descriptions and with the images/prompts. Nearly all participants agreed with the ground truth: participants validated that all 40 visual attributes appeared in generated image pairs, and 36/40 provided free-form attributes that matched the ground truth (match validated by the participants). Similarly, for prompt descriptions, 39/40 validated the ground truth, and 34/40 provided free-form attributes that matched the ground truth.

To validate our LLM scoring system, three participants scored 25 predictions using the same rubric given to the LLM judge. The weighted Cohen's kappa [12] between humans and the LLM was 0.635, comparable to inter-human agreement (0.667), confirming that LLM-as-a-judge scores align closely with human evaluation.

5. Experiments

We measure CompCon's ability to discover diverging visual attributes and diverging prompt descriptions in comparison to baselines on the ID^2 (Section 5.3) and apply CompCon to compare two popular open source models, PixArt [6] and SD-Lightning [32] to find diverging representations and uncover age and gender bias (Section 5.4)

5.1. Experimental Details

For our experiments, we set the size of $\mathcal{P}_{batch}$ to be 50, $t = 0$, and $\delta = 0.05$ for our diverging visual attribute discovery phase. In our diverging prompt description phase, we sample 25 prompts from $\mathcal{H}^{div}$ and $\mathcal{H}^{non}$ and generate $k = 25$ new prompts per iteration. For all experiments, each model generates 3 images per prompt across different seeds. We set $t = 0.2$ and $\delta = 0.05$ for our benchmark comparison. For the model comparison in Section 5.4, we manually inspect prompts labeled as diverging to set thresholds t and δ. Hyperparameters for each discovered diverging representation, as well as all LLM and VLM prompts used for the method and baselines can be found in the Appendix.

We use GPT-4o [36] as the LLM for CompCon's diverging visual attribute discovery and diverging prompt description discovery phases, as well as for the LLM-Only baseline described below. We use CLIP ViT-bigG-14 [7, 28, 38] trained on Laion2b [41] to classify diverging prompts and attributes and the instructor-xl [43] text embedding model for prompt retrieval. Our dataset generation and experiments were performed on two 80GB NVIDIA A100 GPUs.

5.2. Baselines

We compare CompCon to several baselines on the ID^2, including an end-to-end approach and approaches for the individual diverging visual attribute and diverging prompt description discovery phases.

LLM-only (end-to-end). We select 50 random prompts P_{sample} from each dataset $\mathcal{D}_a$, caption the corresponding generated images and prompt an LLM (GPT-4o) to find any diverging representations.

TF-IDF (diverging visual attribute discovery). We caption the generated images $\mathcal{I}_1^{(\mathcal{P}_{sample})}$ and $\mathcal{I}_2^{(\mathcal{P}_{sample})}$. We then combine all captions produced for $\mathcal{I}_1^{(\mathcal{P}_{sample})}$ and $\mathcal{I}_2^{(\mathcal{P}_{sample})}$ into two separate documents and compute TF-IDF [42], taking the top-5 1-3 word phrases that appear more often in captions of $\mathcal{I}_1^{(\mathcal{P}_{sample})}$ than $\mathcal{I}_2^{(\mathcal{P}_{sample})}$.

VisDiff (diverging visual attribute discovery). We apply the VisDiff algorithm [17] for finding differences in image sets $\mathcal{I}_1^{(\mathcal{P}_{sample})}$ and $\mathcal{I}_2^{(\mathcal{P}_{sample})}$.

TF-IDF (prompt description discovery). Given a diverging visual attribute a, we run `classifyDiverging` on $\mathcal{D}_a$ to obtain the diverging and non-diverging prompts and run TF-IDF to find which phrases appear more often in the diverging prompts compared to the non-diverging prompts. We report results using this method on the diverging visual attribute's discovered by CompCon.

We use Llava 1.5-7b [33] for image captioning used in the VisDiff and TF-IDF baselines.

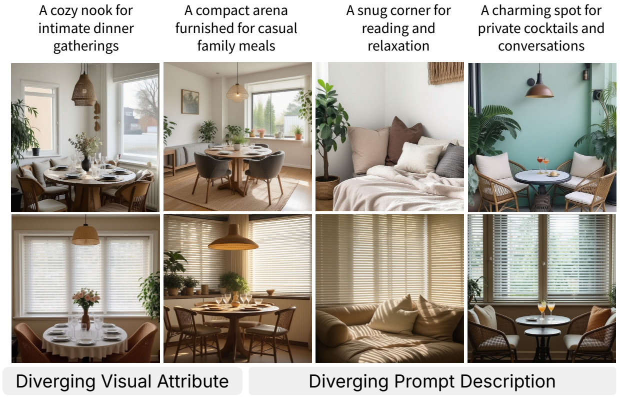
5.3. Benchmark Results

Metric	Method	Top 1	Top 5
Attribute Score	CompCon	**0.60**	**0.68**
	VisDiff	0.47	0.62
	TF-IDF	0.23	0.37
	LLM-only	0.08	0.24
Description Score	CompCon [5-iter]	**0.64**	**0.78**
	CompCon [1-iter]	0.59	0.72
	TF-IDF	0.40	0.57
	LLM-only	0.03	0.28

Table 1. **Visual attribute and prompt description scores on ID^2.** Our approach outperforms all baselines on both diverging visual attribute and diverging prompt description discovery.

We report diverging visual attribute and diverging prompt description discovery results in Table 1. CompCon obtains higher attribute and description scores than the baselines. An example output, along with their strengths and weaknesses, are shown in Figure 4. We find that:

1. The LLM-only baseline performs poorly compared to the other methods, often outputting diverging representations that are completely unrelated to the ground truth. This result is due to the complexity of the task: the LLM must (1) find differences between each pair of captions, (2) uncover which differences are seen most often, and (3) summarize the prompts that have this difference in the captions.
2. When discovering diverging visual attributes, VisDiff and TF-IDF often fail at identifying more fine-grained differences. This finding is likely due to the reliance on captions lacking fine-grained detail, especially when comparing images depicting similar contexts.
3. While additional iterations offer modest gains in performance when generating diverging prompt descriptions, we see in Figure 4 that iterations are beneficial when the ground truth description is more fine-grained. This iterative refinement enables the model to evolve from describing general representations to fine-grained qualities.


A cozy nook for intimate dinner gatherings
A compact arena furnished for casual family meals
A snug corner for reading and relaxation
A charming spot for private cocktails and conversations
Diverging Visual Attribute
Diverging Prompt Description

Ground Truth: 'venetian blinds'
LLM-Only: 'onlookers'
TF-IDF: 'natural light'
VisDiff: 'wooden floor'
CompCon: 'striped window blinds'

Ground Truth: 'small dining room', 'small living room'
LLM-Only: 'Crowd', 'spectators', 'audience'
TF-IDF: 'space', 'corner', 'compact'
CompCon[1]: 'Indoor settings', 'Spatial organization', 'Design elements'
CompCon[5]: 'Cozy environments', 'Multifunctional spaces', 'Compact Comfort'

Figure 4. **ID^2 example.** *Top:* We show dataset prompts and corresponding generated images, where the second image row depicts the diverging visual attribute. *Bottom:* We show the ground truth diverging visual attribute and diverging prompt description, along with outputs from our approach and the baselines. Notice that our method produces outputs that better align with the ground truth.

Additional experiments on the effects of iterations, the choice of LLM and VLM, and sensitivity analysis of the effects of LLM and VLM errors are in Section B.

5.4. Qualitative Results

Using CompCon we find divergent representations in several popular diffusion models: PixArt Alpha [6], SDXL-Lightning [32], Stable Diffusion 3.5 Large [1], and Playground 2.5 [31]. As listed in Section 3, we use an LLM to generate diverging prompts from the prompt description generated at each iteration. A subset of these generated diverging prompts are shown below. Additional prompts, experiments on other prompt sets and models, and a comparison to a single-model method are in the Appendix (A, G.1).

Results on templated prompts. We run CompCon with an initial prompt bank covering different art styles, subjects, and descriptors and visualize results in Figure 5. CompCon discovers both diverging representations like negative emotions and empty urban environments produce "wet streets" in PixArt. Looking at the "flames" example in Figure 1, we also see diversity in the presentation of attributes, with flames taking the form of a burning podium to a fiery background to a large red cloud. Lastly, we see scenarios of poor prompt adherence in the "decay" divergence representation, where SD-Lightning does not produce visual elements that indicate rundown and abandoned places while PixArt associates abandonment with decay. Our findings demonstrate that CompCon can effectively uncover both concrete and abstract divergent representations in text-to-image models,

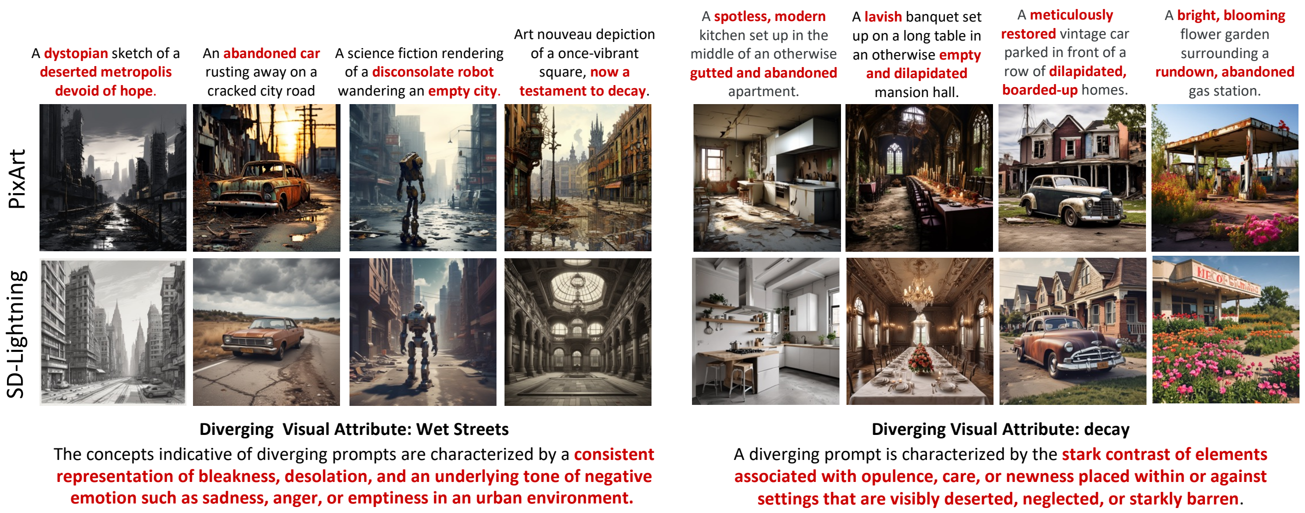

A dystopian sketch of a deserted metropolis devoid of hope.
An abandoned car rusting away on a cracked city road
A science fiction rendering of a disconsolate robot wandering an empty city.
Art nouveau depiction of a once-vibrant square, now a testament to decay.
A spotless, modern kitchen set up in the middle of an otherwise gutted and abandoned apartment.
A lavish banquet set up on a long table in an otherwise empty and dilapidated mansion hall.
A meticulously restored vintage car parked in front of a row of dilapidated, boarded-up homes.
A bright, blooming flower garden surrounding a rundown, abandoned gas station.
PixArt
SD-Lightning
Diverging Visual Attribute: Wet Streets
The concepts indicative of diverging prompts are characterized by a consistent representation of bleakness, desolation, and an underlying tone of negative emotion such as sadness, anger, or emptiness in an urban environment.
Diverging Visual Attribute: decay
A diverging prompt is characterized by the stark contrast of elements associated with opulence, care, or newness placed within or against settings that are visibly deserted, neglected, or starkly barren.

Figure 5. **CompCon results comparing PixArt and SD-Lightning.** PixArt associates negative emotions and desolation in urban environments with 'wet streets,' while SD-Lightning struggles to depict run-down or dilapidated scenes, where PixArt instead conveys 'decay.'

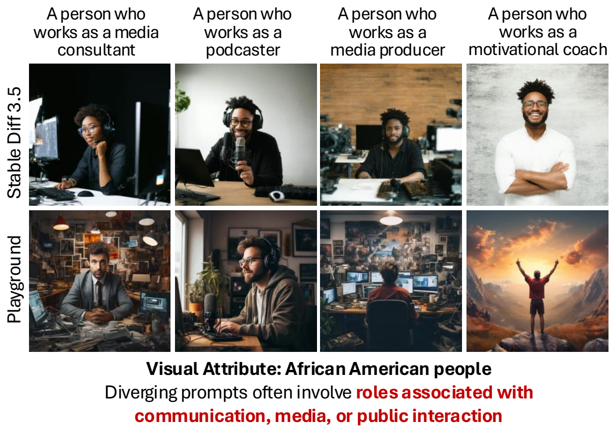

A person who works as a media consultant
A person who works as a podcaster
A person who works as a media producer
A person who works as a motivational coach
Stable Diff 3.5
Playground
Visual Attribute: African American people
Diverging prompts often involve roles associated with communication, media, or public interaction

Figure 6. **Finding bias.** CompCon discovers racial bias in Stable Diffusion 3.5 images for prompts related to media professions.

providing interpretable insights into their behavior. The template used to create the prompt bank, additional qualitative examples, and analysis on the effects of iterations are included in the Appendix. Additionally, the Appendix includes results of running CompCon on a dataset of prompts generated by LLMs, creating a fully automated pipeline.

Detecting bias. We show that CompCon can be used for the crucial task of bias detection. As an initial prompt set, we take existing prompts from Luccioni et al. [35], which probe a model's gender bias when it comes to professions. This dataset contains 252 template prompts that uses a list of professions and interchanges "man", "woman", and "person" (*e.g.*, "A man/woman/person who works as a baker"). In Figure 6, CompCon highlights differences in how professions are visually represented, with Stable Diffusion 3.5 generating significantly more African American people in media and communication-focused roles. More examples of CompCon detecting other biases in the Appendix (A.2).

5.5. Limitations and Cost

As CompCon relies on off-the-shelf LLMs and VLMs, it inherits their biases. While such biases can cause false negatives when discovering diverging representations, the descriptions found still align with human discovery. We validate this through user studies showing that our benchmark evaluation matches human annotation. Additionally, Section B shows that CompCon detects gender and age biases between diffusion models, and CompCon can find correct diverging representations even when the VLM or LLM fails part of the time.

Cost. Using GPT-4o as the VLM and LLM costs ~\$0.50 for attribute discovery plus ~0.02 per attribute iteration. This can be cost-effective for comparing one model to N others ($\mathcal{O}(N)$). Running smaller open-source models like IDEFICS llama3-8b [29] cuts costs while maintaining competitive performance, still outperforming baselines. Additional open-model results are in the Appendix (Sec B).

6. Conclusion

We present CompCon, a method for systematically discovering divergent representations between text-to-image models. By identifying input-dependent differences in model outputs and uncovering the prompt concepts linked to these differences, CompCon provides a framework to understand how models interpret semantic concepts differently. Our results on our ID^2 benchmark and in the comparison of PixArt and SD-Lightning, demonstrate CompCon effectiveness in revealing subtle model-specific behaviors. This opens the possibility of identifying and mitigating unwanted behaviorgenerated images and videos. Moreover, our approach can serve as a tool for probing the hypothesis that different models converge to the same representation [27].

References

[1] Stability AI. Introducing stable diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5, 2024. Stable Diffusion 3.5 Large, Large Turbo, and Medium models. 5, 7

[2] AI@Meta. Llama 3 model card. 2024. 2

[3] Anthropic. Claude 3.5 sonnet, 2024. 5

[4] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549, 2017. 2

[5] 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 (FAccT)*, 2023. 2

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 6, 7, 1

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 6

[8] 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, 2023. 2

[9] Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *ECCV*, 2024. 2

[10] Jaemin Cho, Maarten Sap, Mark Yatskar, Yejin Choi, and Dan Schwartz. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. 2

[11] J. Cho, A. Zala, and M. 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 (ICCV)*, pages 3043–3054, 2023. 2

[12] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. 6

[13] 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 (CVPR)*, pages 12225–12235, 2024. 2, 8

[14] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Xingqian Xu, Humphrey Shi, and Nicu Sebe. Gradbias: Unveiling word influence on bias in text-to-image generative models, 2024. 2

[15] Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1934–1943, 2023. 2

[16] Dreamlike Art. Dreamlike photoreal 2.0. https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0, 2023. A photorealistic Stable Diffusion 1.5–based model trained on 768×768 images. 1

[17] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 7

[18] F. Friedrich, P. Schramowski, M. Brack, L. Struppek, D. Hintersdorf, S. Luccioni, and K. Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2

[19] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition, 2023. 2

[20] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip, 2024. 2

[21] S. Ghosh and A. Caliskan. Person==light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6971–6985, 2023. 2

[22] Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. Tiam - a metric for evaluating alignment in text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2890–2899, 2024. 2

[23] K. Hamidieh, H. Zhang, T. Hartvigsen, and M. Ghassemi. Identifying implicit social biases in vision-language models. *arXiv preprint arXiv:2411.00997*, 2023. 2

[24] J. Hessel, A. Holtzman, M. Forbes, R.L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 2

[26] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 2

[27] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024. 8

[28] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave,

Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6

[29] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024. 8, 2

[30] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[31] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 7, 1

[32] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024. 6, 7, 1

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 7

[34] Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci. Organizing unstructured image collections using natural language. *arXiv preprint arXiv:2410.05217*, 2024. 2

[35] Alexandra Sasha Luccioni, Cynthia Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 8, 1

[36] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. (Accessed on 06/05/2024). 6

[37] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*. University of California, Berkeley, 2024. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 6

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2234–2242, 2016. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 6

[42] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. 6

[43] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. 2022. 6

[44] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models, 2023. 2

[45] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation, 2023. 2

[46] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[47] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2