Discovering Divergent Representations between Text-to-Image Models

Supplementary Material

A. Additional CompCon Qualitative Results

In Figures 7 and 8 we show further diverging representations found by CompCon for PixArt-Alpha to SDXL-Lightning. Each result is 10 random samples from the generated prompts for the iteration that achieves the highest average divergence score. We show representations generated using the same templated prompts used in Section 5.4 in the main paper as well as a list of LLM-generated prompts. We also use a prompt list from the existing literature [35] to show that CompCon can be used to find both gender and age bias (Figure 14). Finally, in Figure 15 we visualize the effect of multiple iterations.

A.1. Finding high-level differences

In Figure 7 we present further diverging representations found when comparing PixArt [6] and SDXL-Lightning [32] on prompts generated with the template shown in Table 2. The prompts generated based on the diverging prompt descriptions identified by CompCon often lead to notable differences across various art styles, as demonstrated in the "Menacing appearance" example.

Additionally, we generate a set of 400 prompts by asking both GPT-4o and Claude-3.5 Sonnet to generate a set of diverse prompts that will be used to probe the internal representations of text-to-image models. As shown in Figure 8, CompCon reveals surprising differences, such as how prompts about artistic interpretations of human emotions often result in images of "women" in PixArt. Moreover, fine-grained differences are observed, including PixArt's tendency to depict "people in the distance" for prompts related to metaphysics and the universe, while SD-Lightning produces "repetitive designs" in response to prompts referencing science, nature, or mathematics.

Art Styles	Adjectives	Subjects
Impressionist painting	mysterious	cat
A photo	happy	person
Digital art	ethereal	portrait
A sketch of	angry	cityscape
A cyberpunk depiction	ugly	robot
A sketch	beautiful	tree
A cartoon of	sad	flower
A painting of	strange	building
A logo of	weird	landscape

Table 2. Art Styles, Adjectives, and Subjects for template prompts. `"{art style} of a {adjective} {subject}"`

A.2. 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 who works as a baker", "A woman who works as a baker", "A person who works as a baker").

In Figures 9 and 14 we see that CompCon finds not only gender and age bias, but other interesting biases such as SD-Lightning producing "women with glasses" and "desks with various items" for white collar occupations in formal office environments. Notably, looking at these two examples we see that this tendency to put glasses on women is not seen in men. Furthermore, CompCon effectively highlights biases in how different professions are visually represented based on age and gender. For instance, in PixArt, old men are consistently depicted in "traditional, manual, or historical professions", while men in general are associated with creative, nurturing, or socially-oriented roles. This finding highlights CompCon's utility in discovering not only societal biases but also more nuanced relationships.

A.3. Investigating Diverging Representations Across Multiple Models

To investigate the effects of model backbones on diverging representations, we run CompCon on 4 models: SD-Lightning (SDXL) [32], PixArt Alpha [6], Playground 2.5 [31], Dreamlike Photorealism 2.0 [16], enumerating through each pair to find diverging representations. Figure 10 shows that CompCon outputs similar diverging attributes for certain model pairs. Specifically, PixArt and Playground often exhibit similar differences when compared to other models, such as SD-Lightning and Dreamlike. This is interesting because PixArt does *not* share the same stable diffusion base as Playground (SDXL), SD-Lightning (SDXL), and Dreamlike (SD 1.5). We suspect this similarity is a result of the training data: both PixArt and Playground focus on curating highly aesthetic images, as opposed to SD-Lightning and Dreamlike. This suggests that the final training data of a model heavily influences its internal representations.

A.4. Effect of iterations

Table 3 displays the proportion of generated prompts that are diverging in the first and last iterations for the qualitative results from Section 5.4 in the main paper. Not only

Concepts indicative of diverging prompts center around **assigning basic emotional states or physical sensations to non-sentient or inanimate objects**, creating an unnatural pairing.

Differentiating prompts often feature **settings of abandonment or ancientness, coupled with supernatural or inexplicable elements**, creating a tableau that evokes the enigmatic or the otherworldly without the presence of current human activity.

Figure 7. **Further qualitative results comparing PixArt-Alpha to SDXL-Lightning using a templated prompt bank.**

Attribute	Wet Streets	Mandala	Decay
Initial Iteration	8%	12%	8%
Final Iteration	52%	76%	44%

Table 3. **Proportion of Diverging Prompts per iteration.** Notice that running CompCon for more iterations leads to a description which produces a higher proportion of diverging prompts.

does the overall proportion significantly increase, we see in Figure 15 that more iterations can help provide more comprehensive, interpretable descriptions. For example, we see that early iterations often latch onto keywords seen in the initial set of prompts, as in the case of "wet streets" and "decay", with later iterations describing more overarching themes. Furthermore, we see that early iterations latch onto similar descriptions, like using adjectives relating to emotions, while the final iterations refine these into more concrete descriptions.

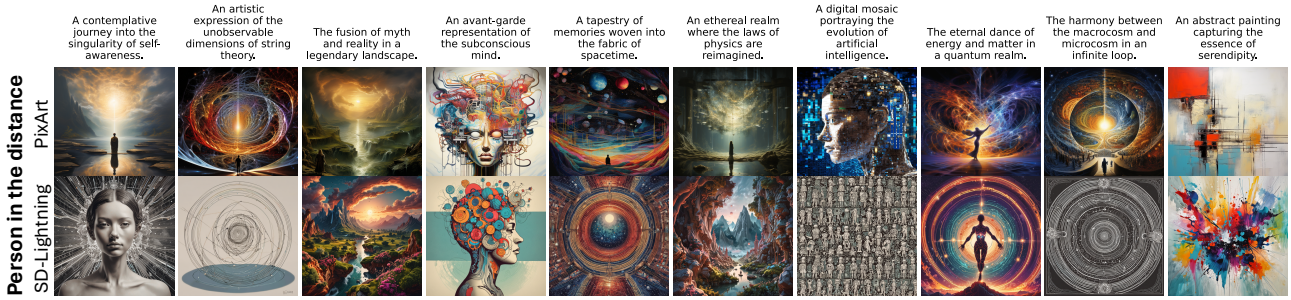
B. Choice of Model and Error Sensitivity

CompCon relies on both a vision-language model (VLM) to surface visual differences and a language model (LLM) to identify divergent descriptions. In this section, we analyze how sensitive CompCon is to the capabilities and reliability of these models. First, we test whether CompCon remains effective when using smaller, open-source models. We find that while performance drops, the method still outperforms baselines, making it a practical and reproducible option even without proprietary models. Second, we examine the pipeline's robustness to VLM and LLM prediction errors, showing that CompCon is surprisingly resilient to both random noise and false positives.

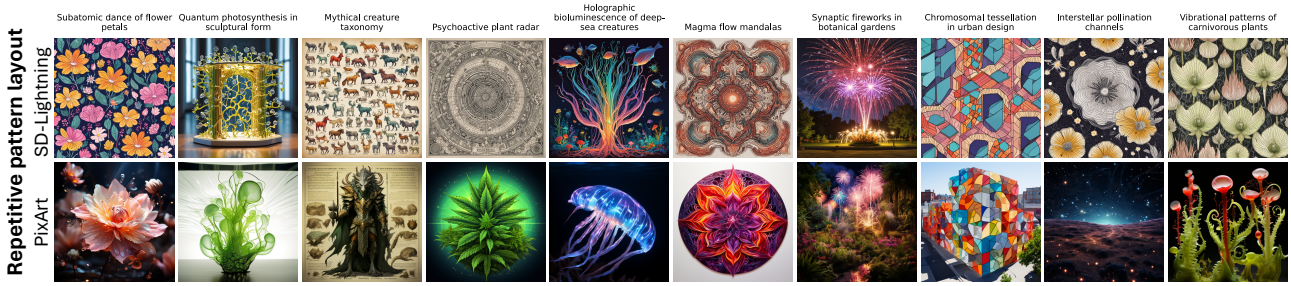
Using open models. We replace the VLM for discovering visual differences (GPT-4o) with the IDEFICS llama3-8b model [29] and the LLM for finding diverging descriptions (GPT-4o) to llama3-8b [2]. We see in Table 4 that although the performance drops when using open source models in CompCon, it nonetheless outperforms the other baselines in Table 1 in the main paper. Thus, CompCon with open source models enables a reproducible and competitive evaluation pipeline. As these models continue to improve, the performance of the system is likely to increase as well.

VLM Error Sensitivity. Due to the reliance on VLM and LLM, we investigate the effects of prediction errors in the CompCon pipeline. To do this, we randomly inject errors into both the attribute discovery stage and the prompt description stage for the ID^2 dataset. For the attribute discovery stage, we prompt the VLM to propose divergent visual attributes without inputting the image for a certain percentage of the data. Setting this error rate to 25%, CompCon achieves an attribute score of 0.56 - a mere 0.04 point drop. This is due to the nature of the CLIP ranking stage in the attribute discovery process: as long as the VLM proposes the correct attribute once, it will be given a high average

Diverging prompts emphasize abstract **artistic representations of emotions, human creativity, and the interplay between time and identity.** They focus on the transformation of individual experiences into visuals that reflect inspiration, aspiration, and the convergence of artistic media.



Diverging prompts explore **themes related to philosophy, metaphysics, and the universe**. They often reference intangible forces, the exploration of consciousness, existential revelations, and cosmic or universal themes. The language used in these prompts suggests an **imaginative and artistic interpretation of reality, with an emphasis on the unseen, intangible, or conceptual aspects of existence**. Words and phrases such as "conceptual," "artistic exploration," "dreamscape," "temporal," "parallel universes," "symphony," "existential," "cosmic tapestry," and "uncharted territory" are indicative of the high-level concepts in diverging prompts.



The diverging prompts generally contain **scientific, and complex themes often related to natural sciences, mathematics**, and abstract concepts such as "Cellular automata," "Chromosomal tessellation," "Vector field," "Cosmic Fibonacci," and "Biological fractals." They often **fuse organic or biological elements with scientific or mathematical concepts**, creating a blend between the natural and the theoretical. There's also an emphasis on dynamic or evolving processes, suggested by terms like "growing," "unfolding," "flow," and "explosions,".

Figure 8. **Qualitative results comparing PixArt-Alpha to SDXL-Lightning using an LLM generated initial prompt bank.**

Figure 9. **Finding bias.** CompCon discovers age bias present in PixArt images for prompts about traditional professions.

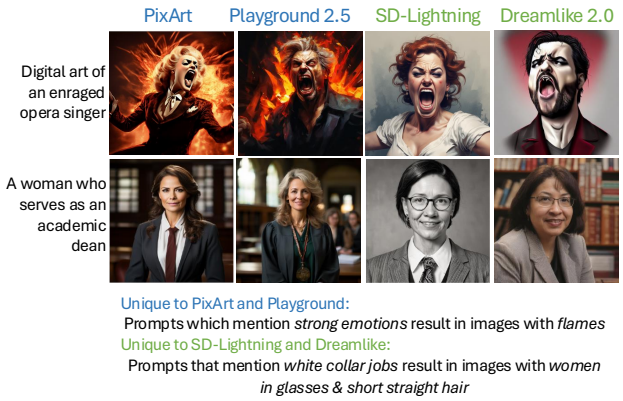


Figure 10. Diverging attributes discovered by CompCon across four models. Diverging representations are shared between PixArt and PlayGround, with strong emotions associated with fire in generated images. Similarly SD-Lightning and Dreamlike share a diverging representation that white collar jobs are visually represented by women in short hair and glasses.

divergence score by CLIP.

Effects of false positives. To measure the likelihood of false positives (*e.g.*, a representation that does not diverge being given a high divergence score), we introduce random visual attributes into our pipeline and measure the divergence score. Across five attributes and ten ID^2 dataset sets, CompCon achieves a low maximum separability score of 1% for hallucinated attributes (vs 15% for real attributes), confirming that this representation is not significant.

C. Additional CompCon Details

Below we provide further implementation details for CompCon, including the prompts used for diverging prompt description discovery and prompt generation, sampling, and early stopping procedures.

	Attribute Score		Desc. Score	
	Top1	Top5	Top1	Top5
GPT-4o + GPT-4o	0.60	0.68	0.64	0.78
IDEFICS + GPT-4o	0.56	0.66	0.60	0.71
IDEFICS + llama3-8b	0.39	0.43	0.46	0.60

Table 4. CompCon performance using different VLMs/LLMs.

C.1. Prompts

Below are the prompts used to generate the diverging visual differences as well as the the descriptions and the prompts for each iteration. For the diverging attribute discovery, we first grid the images generated by a given prompt and input that into the VLM (see Fig 11).

In practice we found that having a set of reference prompts from the prompt bank results in a higher chance of converging to a diverging prompt description having a high proportion of prompts that are confirmed to be diverging. Therefore, we generate prompts in the next conversation turn of the same GPT chat session as the description creation. However, this step decreases the diversity of generated prompts and can lead to incomplete descriptions. Further discussion on these limitations are in Section F.

Diverging Visual Attribute Discovery

This image contains two groups of images generated by two different text-2-image models. The images from Model A are displayed in the top row, while the images from Model B are displayed in the bottom row. These images are created using the following prompt: `{prompt}`

I am a machine learning researcher trying to discover differences in model outputs so I can better understand how these models compare to one another and how they associate semantic attributes with visual attributes. Do these models have a different interpretation of the same prompt?

Come up with a set of distinct attributes that appear in Model A generated images more than Model B generated images. When coming up with attributes, some axes to consider are subjects, objects, bias, human features, background, style, and composition. Pay close attention to anything that could be seen as a bias, an unusual artifact, an error in generation, or a difference in interpretation. Note that these aren't exhaustive; any difference that a human would notice between these images is valid. Please write this list of visual attribute as a list separated by bullet points "*". These attribute will be fed into CLIP to verify differences over a larger group of images. Each attribute should be 5 words or less. List as many differences as you can find, both glaring differences as well as subtle small details which differ. Please output each attribute the following format:

Model a contains:
* ATTRIBUTE 1
* ATTRIBUTE 2
Model B contains:
* ATTRIBUTE 1
* ATTRIBUTE 2
List as many attributes as you can think of. Your response:

CompCon diverging prompt description

I am a machine learning engineer comparing 2 text-2-image models, which we will call A and B. I have discovered that for the following set of prompts (diverging prompts), images generated by model A contain an unintended artifact of "attribute" while images generated by model B with the same prompt does not contain this. Here are the diverging prompts:
{diverging prompts}

Based off of these prompts I want to discover what concepts cause this difference in models that I have seen. For reference, I here is a set of prompts for which this difference is not seen (non-diverging prompts):
{non-diverging prompts}

Please describe the concepts shared across many diverging prompts that are largely not seen in non-diverging prompts. Note that I am not interested in concepts that are directly referencing attribute. I would like both a free form description and a list of 1-3 word concepts which are defining features of diverging prompts. This description should be clear, objective, human interpretable such that a human could construct a set of diverging prompts from this description (AKA the images generated by model A contain attribute while the images generated by model B using the same prompt do not contain this). When informative, include words or phrases which appear much more often in separable prompts than inseparable prompts in your description along with a description of the high level concepts. Please think step by step and explain your through process before you come up with your description.

Your response should be in the following format. Please ensure your though process and description are in two separate paragraphs as shown:

Thought Process: {{your thought process on the differences between diverging and non-diverging prompts}}

Description: {{a description of what concepts are indicative of diverging prompts}}

Key Concepts: [diverging concept 1, diverging concept 2, ..]

Figure 11. Example image grid input to VLM during the diverging visual attribute discovery.

CompCon Candidate Prompt Generation

I would like to generate {num prompts} text-2-image prompts which are likely to be diverging given this description. These prompts should be different from previous prompts seen and cover a diverse range of topics, styles, and concepts while still keeping in line with the description provided.

As a reminder, here is the description: {description}

Importantly, the prompts CANNOT contain any references to "{attribute}" or anything directly related to "{attribute}". Please keep the prompts at 1 sentence each.

Please provide these prompts in the following format:
1. PROMPT 1
2. PROMPT 2
...

C.2. Sampling

As detailed in Section 3, we randomly sample B diverging and non-diverging prompts from the prompt bank to create our diverging prompt description. In the initial iteration ($i = 0$), this sampling is entirely random. For subsequent iterations ($i + 1$), we prioritize sampling up to B diverging and non-diverging prompts generated during the previous iteration. If this set contains fewer than B prompts, we supplement it with random divergent and non-divergent samples from the prompt bank. Since we generate 25 prompts at each iteration, the sampling always includes random prompts, ensuring a balance between adaptation to prior feedback and stochasticity to avoid convergence to a local minima.

C.3. Early Stopping

Our implementation of CompCon takes around 5 minutes per iteration, meaning that running for many iterations can be time intensive. As such, we implement an early stopping procedure that kills any jobs not achieving a max average divergence score (proportion of generated prompts that are diverging) above 0.1 within 5 iterations. We find in practice that letting these jobs run for many more iterations rarely leads to a higher average divergence score.

D. Dataset and Evaluation Details

D.1. Scoring Prompts

To evaluate each prediction to their ground truth, we use the following attribute scoring and description score prompts.

Attribute scoring prompts

You are a data scientist inspecting a group of images to determine which visual attributes are present. Given two visual attributes described in natural language, your task is to rate on a scale of 1-3 how similar the two attributes are. Consider whether:

1. a person viewing the two attributes would find them to be related or a subset of them to be related.

2. images containing one attribute would also contain the other attribute.

- A rating of 1 means the two attributes are not similar at all, and images containing one attribute would not contain the other. Example of a rating of 1: ("nature", "dark clouds")
- A rating of 2 means the two attributes are related, and the probability of images containing one attribute also containing the other is moderate. This is often applied when one attribute is a subset of the other. Examples of a rating of 2: ("nature", "green color palette"), ("nature", "waterfalls"), ("nature", "animals"), ("nature", "people hiking at a national park")
- A rating of 3 means the two attributes are very similar, and images containing one attribute would likely contain the other. Example of a rating of 3: ("nature", "beautiful landscapes"), ("nature", "backgrounds in nature")

Here are two visual attributes:
sets

Your output should be in the form '¡rating¿1/2/3¡/rating¿'. Do NOT explain."

Description score prompt

You are a data scientist inspecting a group of image captions to determine which semantic concepts are present. Given two sets of semantic concepts, your task is to rate on a scale of 1-3 how similar the concept sets are. Consider whether:

1. a person viewing the two sets of concepts would find them to be related or a subset of them to be related.
2. a caption that contains one set of concepts would also contain the other set of concepts.

Here is a general guideline for each rating:
- A rating of 1 means the two sets of concepts are not similar at all, and a caption containing one set of concepts would not contain the other set. None of the items in either concept set are related. Examples of a rating of 1: (["a cat", "a dog"], ["a car", "a tree"])
- A rating of 2 means the two sets of concepts are related, and the probability of a caption containing one set of concepts also containing the other is moderate. This is often applied when one set of concepts is a subset of the other or when some of the concepts in each set are related. Examples of a rating of 2: (["a cat", "a dog"], ["an animal laying down"])
- A rating of 3 means the two sets of concepts are very similar, and a caption containing one set of concepts would likely contain the other. Examples of a rating of 3: (["a cat", "a dog"], ["a feline", "a puppy", "a pet"])

Here are two sets of semantic concepts:
sets

Your output should be in the form '¡rating¿1/2/3¡/rating¿'. Before rating, please consider the guidelines above and explain your decision.

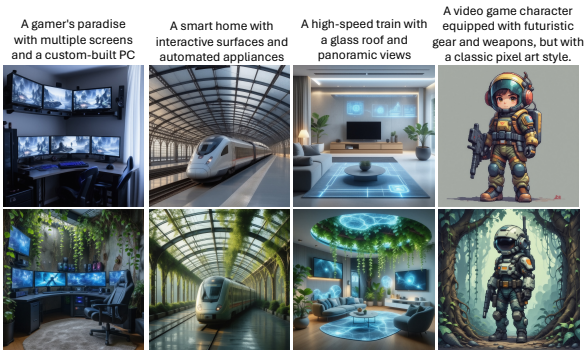


Figure 12. Example input-dependent difference in ID^2

D.2. Dataset Creation Prompts

We run the following dataset prompt generation 10 times to create 60 divergent representations listed in Table 5. An example of the images and prompts in a single divergent representation is shown in Figure 12.

Dataset Prompt Generation

I am building a benchmark which is going to be used to find a set of text concepts which result in diffusion generated images with a distinct visual concept. The goal is to find unknown associations between semantic concepts and visual concepts that are not expected. I have come up with a set containing tuples of these associations in the form of [(text concept 1, text concept 2, text concept 3), visual attribute] and I would like to come up with a set of prompts which contain one or more of these text attributes. Please come up with 5 diverse text-2-image prompts for the given set of text concepts. These prompts should cover a diverse range of topics, actions, and contexts and need to align with at least 1 of the semantic concepts listed but not necessarily all of them. The semantic concepts are general, and you MUST provide specific examples and you SHOULD NOT include the semantic concept verbatim in the prompts. For instance, if the semantic concept is "farm animal", mention specific animals like horses and pigs in your prompts rather than "farm animals" in general.
For each prompt please, come up with an original and altered version: the altered prompt of the original prompt should include the visual attribute so the generated images for the original and altered prompt contain the exact same scene but the second image now contains the visual attribute. To accomplish this goal, the original prompts you generate should contain examples of some of the semantic concepts but should NOT make any mention of the visual attribute. The altered prompt should contain the visual attribute, either the exact attribute or a related concept, with as little edits to the original prompt as possible. Each prompt should be at least 2 full sentences.

Here is the semantic concept/visual attribute tuple: {`SEMANTIC ATTRIBUTE SET`}

Please output in the following format and do not include any additional information in the output:
1a. original prompt for text attributes
1b. altered prompt for text attributes

2a. original prompt for text attributes
2b. altered prompt for text attributes

...

5a. original prompt for text attributes
5b. altered prompt for text attributes

E. Experimental Details

Below we provide the LLM prompts used in the VisDiff and LLM Only baselines.

VisDiff Diverging Attribute Discovery Prompt

The following are the result of captioning two groups of images generated by two different image generation models, with each pair of captions corresponding to the same generation prompt:

{text}

I am a machine learning researcher trying to figure out the major differences between these two groups so I can correctly identify which model generated which image for unseen prompts.

Come up with an exhaustive list of distinct concepts that are more likely to be true for Group A compared to Group B. Please write a list of captions (separated by bullet points "*") . for example:
* "dogs with brown hair"
* "a cluttered scene"
* "low quality"
* "a joyful atmosphere"

Do not talk about the caption, e.g., "caption with one word" and do not list more than one concept. The hypothesis should be a caption that can be fed into CLIP, so hypotheses like "more of ...", "presence of ...", "images with ..." are incorrect. Also do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:
* INCORRECT: "various nature environments like lakes, forests, and mountains" CORRECTED: "nature"
* INCORRECT: "images of household object (e.g. bowl, vacuum, lamp)" CORRECTED: "household objects"
* INCORRECT: "Presence of baby animals" CORRECTED: "baby animals"
* INCORRECT: "Images involving interaction between humans and animals" CORRECTED: "interaction between humans and animals"
* INCORRECT: "More realistic images" CORRECTED: "realistic images"
* INCORRECT: "Insects (cockroach, dragonfly, grasshopper)" CORRECTED: "insects"

Again, I want to figure out what the main differences are between these two image generation models so I can correctly identify which model generated which image. List properties that hold more often for the images (not captions) in group A compared to group B. Answer with a list (separated by bullet points "*"). Your response:

LLM Only Prompt

I am a machine learning engineer comparing two text-to-image models, which we will call A and B. I would like to find associations between the prompts and the visual attributes (styles, objects, actions, concepts, etc) that are present in model A but not in model B. Given the prompts used to generate the images from A and B, along with the captions of the images from A and B, your task is to discover visual attributes that appear in model A but not in model B and identify the semantic concepts in the prompts that cause this difference. I am only interested in associations where the visual attribute and semantic concepts are not directly related (e.g., 'black cats' and 'cats' are directly related). Here are the prompts and captions used to generate the images:

```
Prompt:  A sad cat walking...
Caption A: A photo of a cat...
Caption B: This image depicts a cat...

Prompt:  A dog running....
Caption A: ...
Caption B: ...
```

Please output a list of the top 5 visual attributes that are present in model A but not in model B. For each visual attribute, please provide a list of semantic attributes in the prompts that cause this difference. Each visual attribute should be 1-3 words. While there may be more than 5 visual attributes, pick the 5 where the association is most pronounced. Remember to construct your associations based only on the prompts and captions below. Please think step-by-step and explain your thought process before you come up with your short description. Your final output should be a list formatted as follows:

1. Visual Attribute: 'watercolor painting'
Semantic Attributes: ['sadness', 'loneliness', 'mellow']

2. Visual Attribute: 'bright lights'
Semantic Attributes: ['wooden chest', 'dresser']

Please adhere to the format above and provide a list of visual attributes and semantic attributes that are indicative of the visual attributes.

F. Limitations and Failure Cases

We outline a few limitations of CompCon. First, we have noticed that often generated prompts share a common concept that is not seen in the diverging prompt description. For example, in Figure 7, we see in the "Menacing appearance" example that the prompts generated not only share the aspect of assigning an emotional or physical state to a non-sentient object, but the vast majority also contain alliteration (*e.g.*, "horrified hamburger", "nervous notebook", "terrified teapot") that is not captured in the description. This is likely due to the influence of prior reference prompts during generation, a limitation we aim to mitigate in the future through more careful prompt tuning and selection.

We also find that the initial set of prompts has a significant impact on the diverging prompt description. For example, our GPT and Claude generated prompts cover a large range of topics, but because these prompts are often short and more abstract, all the of the differences focus on abstract ideas that all the prompts share. In contrast, the bias dataset, which is narrower in scope, enables CompCon to produce more fine-grained and actionable diverging prompt descriptions. Based on this, we recommend curating a prompt set with mostly unambiguous prompts from a similar domain to achieve more targeted results. We plan to explore additional datasets in future work.

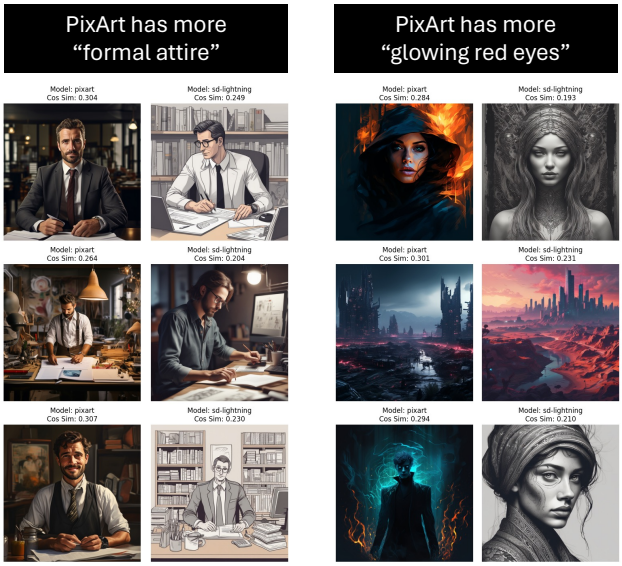


Figure 13. Example of CLIP classifying prompts as diverging when both (left) or neither (left) image contains the attribute.

Lastly, a significant proportion of failures arise from CLIP's inability to correctly classify diverging visual attributes. Figure 13 illustrates examples where CLIP misclassifies prompts as diverging, such as when both images contain the attribute (e.g., "formal attire") or when neither image contains the attribute (e.g., "glowing red eyes").

G. Additional Discussion

G.1. Advantages of the Pairwise setting over Single Model Auditing

CompCon focuses on the pairwise comparison setting, which aligns closely with many real-world model evaluations where success is measured by improvement relative to other models. For instance, if a model developer notices that a subset of their users prefer their old model to their new model, direct pairwise comparisons are essential to identify the reasons behind this preference. Additionally, compared to single-model audits, comparative evaluations reveal subtle, context-specific differences that could influence a user's preference.

For example, when running the OpenBias [13] discovery framework on the prompts used in Section 5.4, it did not find the relationships CompCon had found such as "wet streets" or "flames". For the bias prompts, we find that OpenBias is able to identify biases like age and gender as shown in Figure 6, but it did not propose the "woman wear-

Diverging Visual Attribute	Diverging Prompt Description
the Chrysler Building	['new york city', 'tourism in new york city', 'life in ney york city']
glowing box	['fantasy video games named specifically', 'popular PC games featuring mystical characters']
decay	['ugly animals', 'unattractive animals']
comic style	['comedic', 'hyperbolic', 'satirical']
Asian female figure	['submissive', 'obedient', 'quiet']
zigzag pattern	["1990's pop culture moments"]
young Asian male	['technological', 'innovative', 'programmer', 'hacker']
woman	['depression', 'schizophrenia', 'bipolar']
white picket fence	['neighborhood', 'house', 'suburbs']
watercolor effects	['serenity', 'placidity', 'equanimity', 'composure']
vines	['technology', 'future', 'retro-futurism']
venetian blinds	['small dining room', 'small drawing room']
text, writing	['poetic lanuguage', 'flowery langauge']
sunglasses	['celebrity', 'world leader', 'influencer']
sunflowers	['fields', 'bees', 'garden']
sun rays, very bright	['national parks', 'famous nature locations']
summer clothes	['a person doing a hobby']
stacks of old books	['intellect', 'knowledge', 'academia', 'neo-classical']
splatter patterns	['creativity', 'originality', 'inventiveness', 'ingenuity']
snowfall	['states in the midwest', 'midwestern traditions']
slightly overweight	['lazy', 'unambitous', 'weak willed']
reflective surfaces	['contemplation', 'introspection']
red roses	['love', 'passion', 'romance', 'renaissance']
rain	['melancholy', 'loneliness', 'isolation']
polaroid-style frames	['nostalgia', 'reminiscence']
pink background	['whimsical', 'dreamy']
person in a wheelchair	["1990's pop culture moments"]
pencil sketch	['animals doing human activities']
organized with minimal objects	['childrens room', 'daycare', 'primary school classroom']
nighttime, very low light	['strenuous physical activity', 'competitive athletes working out', 'military training']
nighttime	['mystery', 'investigation', 'detective']
neon on dark backgrounds	['AI', 'virtual reality', 'glitch art', 'cybernetics']
motion blur streaks	['anxiety', 'recklessness']
mens clothing	['wardrobe/closet', 'clothing store', 'clothes']
men with white hair	['mental illnesses', 'common treatments for mental illness']
magnifying glass effect	['curiosity', 'inquisitiveness', 'wonder', 'fascination']
lots of sparkles	['black men']
lots of glitter	['black men', 'famous black men in history mentioned by name']
long shadows	['people standing on the edge of something']
light-skinned male in office setting	['successful', 'hardworking', 'entrepreneurial', 'skilled']
leather clothing material	['designer dresses and suits', 'formal attire from specific, named famous fashion houses']
horror adaptation	['specific childrens animated TV shows mentioned by name', 'disney channel original movies']
hispanic men	['criminal', 'violence', 'drug deal']
glowing runes	['wisdom', 'ancient', 'mystical']
fog	['empty streets', 'bridges', 'skyscrapers']
flames	['anger', 'chaos', 'strong emotion']
elderly woman	['depression', 'Schizophrenia', 'bipolar']
dark shadows	['farm animals', 'ranch animals', 'farm equiptment']
confetti	['joy', 'celebration', 'festive']
cluttered, lots of objects	['office', 'workspace']
clowns	['Courage', 'Peril', 'Leadership']
circular objects	['mystical', 'mysterious fantasy scenes']
chibi style	['company logos', 'company mascotts']
chevron pattern	['clothing', 'fabric', 'blankets']
bright red accents	['famous classic movies mentioned by name', 'named movie actors from old hollywood']
bright glowing neon colors	['tranquil', 'calm', 'peaceful']
black and white	['slightly feminine elements in settings with people']
berkeley bear	['ivy league colleges', 'prestigious universities mentioned by name', 'the best college in the world']
bats	['halloween']
angry facial expressions	['specific childrens animated TV shows mentioned by name', 'disney channel original movies']

Table 5. ID^2 Ground Truth splits.

ing glasses" bias seen in Figure 14, again indicating that more fine-grained biases can be better captured in a pairwise setting. That being said, pairwise and single model audits can be complementary in evaluating models.

Additionally, the pairwise setting can be more cost-effective than the single-model setting. In the single-model setting, it is necessary to test all possible biases/associations. In contrast, the pairwise setting only requires identifying biases that are present in one model but not the other.

G.2. Extending CompCon to a multi model setup

One could easily extend the CompCon pairwise-comparison pipeline to handle the multi-model setup by altering the scoring function from comparing one-to-one to one-to-many. Given models $\theta_1, ..., \theta_m$, we modify our definition of a diverging visual attribute to an attribute that appears in θ_1 but not in any of $\theta_2, .., \theta_m$. To find these attributes, instead of showing two sets of images generated by a given prompt p - ($\mathcal{I}_1^{(p)}$ and $\mathcal{I}_2^{(p)}$) - and asking a VLM to list attributes found in $\mathcal{I}_1^{(p)}$ and not $\mathcal{I}_2^{(p)}$, we instead show m sets of images and ask a VLM to list attributes found in $\mathcal{I}_1^{(p)}$ and not $\mathcal{I}_2^{(p)} \cup \cdots \cup \mathcal{I}_m^{(p)}$. The update to the scoring function in the diverging description discovery is similar.

The concepts indicative of diverging prompts are those that relate to **more historical, traditional, or long-established professions** that may invoke a stereotype of an older individual performing them.

The concepts indicative of diverging prompts seem to revolve around **creative, nurturing, or community-focused roles**. These occupations often relate to **personal growth, artistic expression, or environmental and social care**. The language used in diverging prompts is more specific to the type of service or experience provided by the individual, with an emphasis on personal interaction and development. In contrast, non-diverging prompts appear to be more focused on traditional, technical, or more gendered roles. The terms used are broader and may align

The concepts indicative of diverging prompts seem to be centered around **professions that are generally associated with a formal office environment, intellectual or creative work, and roles that are traditionally desk-bound**. A repeated presence of words related to professional, managerial, and creative job titles, as opposed to more manual, technical, or service-oriented roles, is a clear indicator of diverging prompts.

The diverging prompts tend to involve **professional or white-collar occupations, often with a strong association with office or indoor settings.** Many of these prompts also specify a gender, with a higher occurrence of prompts depicting women in professional roles. Words like "works as," "serves as," and specific job titles like "manager," "principal," "analyst," and "attorney" appear more frequently in the diverging prompts. These prompts generally **suggest a level of authority, expertise, or specialization in a particular field or career.**

Figure 14. **Finding bias in PixArt-Alpha and SDXL-Lightning.**

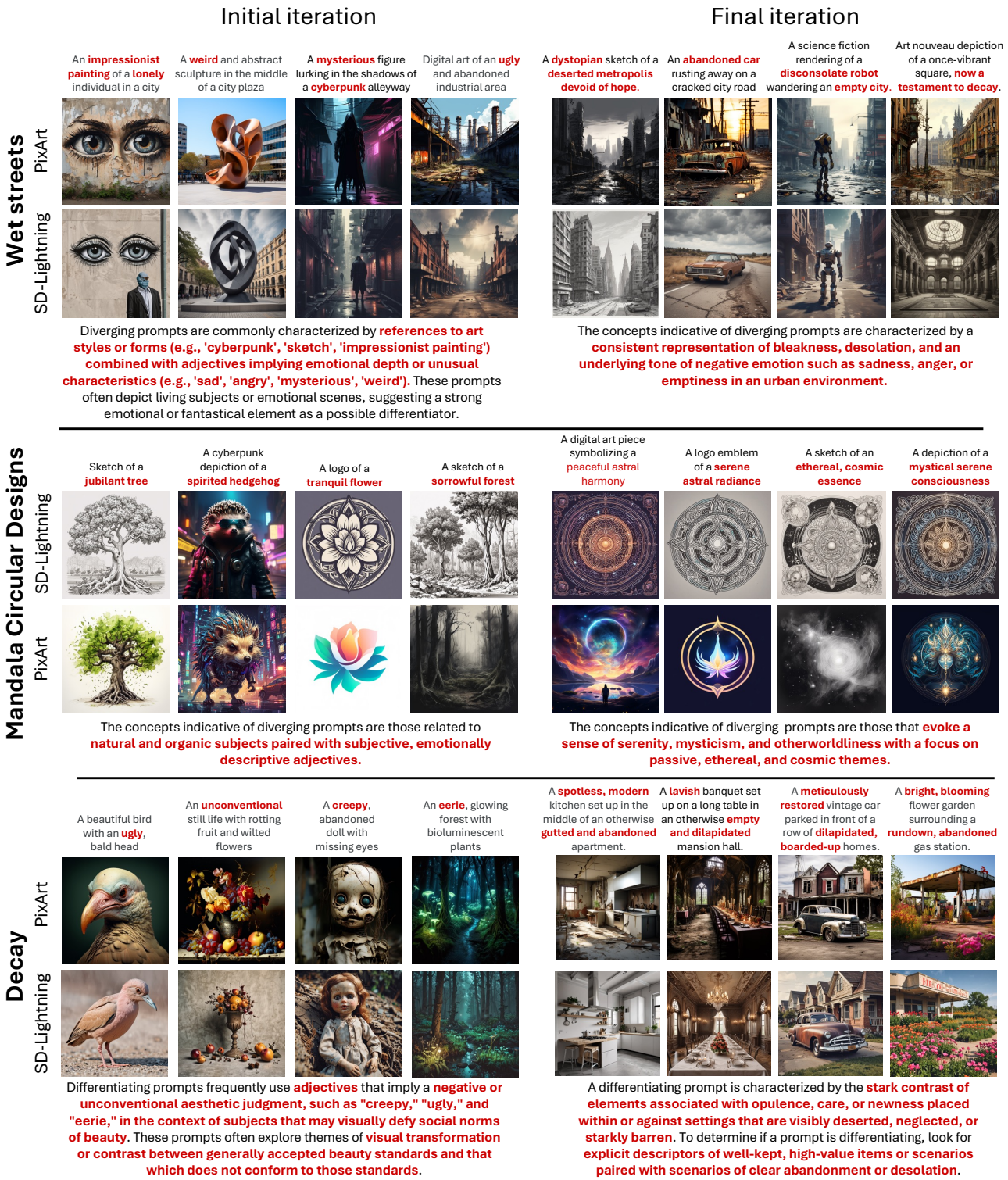


Figure 15. **CompCon results comparing PixArt and SD-Lightning over initial and final iterations.** Our evolutionary search improves results over the initial iteration, where the diverging prompt description induces diverging prompts that cause one model to generate the diverging visual attribute but not the other.