Attribute Based Interpretable Evaluation Metrics for Generative Models

Dongkyun Kim [*] [1 2] Mingi Kwon [*] [1] Youngjung Uh [1]

Abstract

When the training dataset comprises a 1:1 proportion of dogs to cats, a generative model that produces 1:1 dogs and cats better resembles the training species distribution than another model with 3:1 dogs and cats. Can we capture this phenomenon using existing metrics? Unfortunately, we cannot, because these metrics do not provide any interpretability beyond "diversity". In this context, we propose a new evaluation protocol that measures the divergence of a set of generated images from the training set regarding the *distribution of attribute strengths* as follows. Single-attribute Divergence (SaD) reveals the attributes that are generated excessively or insufficiently by measuring the divergence of PDFs of individual attributes. Paired-attribute Divergence (PaD) reveals such pairs of attributes by measuring the divergence of *joint* PDFs of pairs of attributes. For measuring the attribute strengths of an image, we propose Heterogeneous CLIPScore (HCS) which measures the cosine similarity between image and text vectors with *heterogeneous initial points*. With SaD and PaD, we reveal the following about existing generative models. ProjectedGAN generates implausible attribute relationships such as `baby` with `beard` even though it has competitive scores of existing metrics. Diffusion models struggle to capture diverse colors in the datasets. The larger sampling timesteps of the latent diffusion model generate the more minor objects including `earrings` and `necklace`. Stable Diffusion v1.5 better captures the attributes than v2.1. Our metrics lay a foundation for explainable evaluations of generative models. Code: github.com/notou10/sadpad .

1. Introduction

The advancement of deep generative models, including VAEs (Kingma & Welling, 2013), GANs (Karras et al., 2019; 2020b; 2021; Sauer et al., 2021), and Diffusion Models (DMs) (Song et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2022), has led to generated images that are nearly indistinguishable from real ones. Evaluation metrics, especially those assessing fidelity and diversity, play a pivotal role in this progress. One standout metric is Fréchet Inception Distance (FID) (Heusel et al., 2017), measuring the disparity between training and generated image distributions in embedding space. Coupled with other metrics like precision, recall, density, and coverage, the difference between generated and real image distributions is effectively gauged.

Figure 1 illustrates the evaluation metrics for two models with distinct properties. While Model 1's generated images align closely with the training dataset, Model 2 exhibits a lack of diversity. Notably, in Figure 1a gray box, Model 1 consistently outperforms Model 2 across all metrics. Yet, these metrics fall short in explicability; for example, they don't highlight the overrepresentation of `long hair` and `makeup` in Model 2.

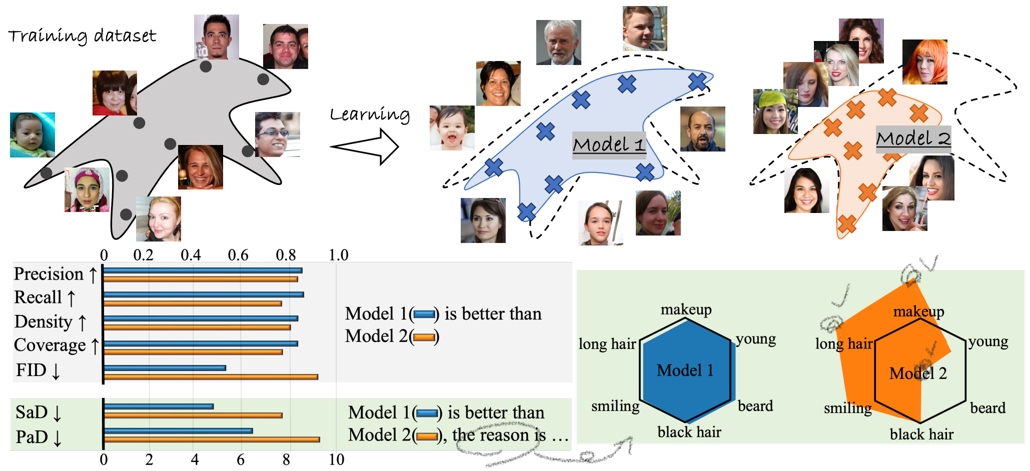
Addressing this gap, our paper proposes a methodology to quantify discrepancies between generated and training images, focusing on specific attributes. Figure 1b shows the concept of our alternative approach that measures the distribution of attribute strengths compared to the training set: while Model 1 offers a balanced attribute distribution akin to the training dataset, Model 2 overemphasizes `long hair` and underrepresents `beard`.

To build metrics that quantify the difference between two image sets in an interpretable manner, we introduce Heterogeneous CLIPScore (HCS), an enhanced variant of CLIPScore (Radford et al., 2021). Compared to CLIPScore, Heterogeneous CLIPScore captures the similarity between modalities—image and text—by establishing distinct origins for text and image vectors.

Utilizing HCS, we introduce new evaluation protocols to assess the attribute distribution alignment between generated images and training data as follows. 1) Single-attribute Divergence (SaD) measures how much a generative model

[*]Equal contribution [1]Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea [2]AI Lab, CTO Division, LG Electronics, Seoul, Republic of Korea. Correspondence to: Youngjung Uh <yj.uh@yonsei.ac.kr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) The scores of metrics

(b) Interpretable score

Figure 1. **Conceptual illustration of our metric.** We design the scenario, Model 2 lacks diversity. (a) Although existing metrics (gray box) capture the inferiority of Model 2, they do not provide an explanation for the judgments. (b) Our attribute-based proposed metric (green box) has an interpretation: Model 2 is biased regarding `long hair`, `makeup`, `smiling`, and `beard`.

deviates from the distribution of each attribute in the training data. 2) Paired-attribute Divergence (PaD) measures how much a generative model breaks the relationship between attributes in the training data, such as "babies do not have beards." With the proposed metrics, users can now realize which specific attributes (or pairs of attributes) in generated images differ from those in training images.

Figure 1b shows the concept of SaD with 6 attributes, where `long hair, makeup, beard` are the most influential attributes to SaD. This allows us to explain why Model 2 is not good. We note elaborate quantification of attribute preservation could be one of the meaningful tasks since the generative model can be utilized for diverse purposes such as text-to-image generation not only for generating a plausible image.

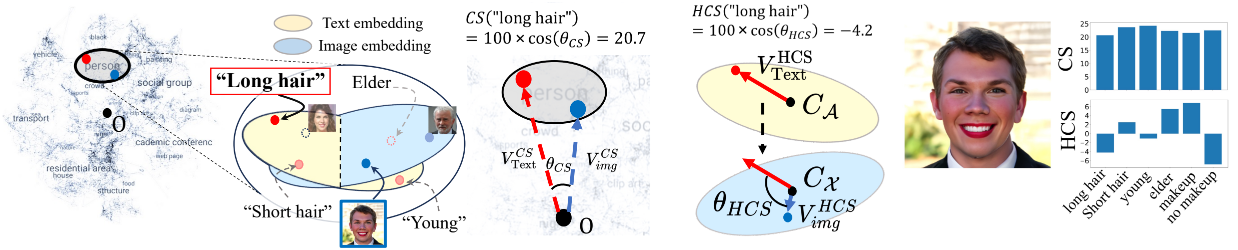
We conduct a series of carefully controlled experiments with varying configurations of attributes to validate our metrics in Section 5.2 and 5.3. Then we provide different characteristics of state-of-the-art generative models (Karras et al., 2019; 2020b; 2021; Sauer et al., 2021; Nichol & Dhariwal, 2021; Rombach et al., 2022; Yang et al., 2023) which could not be seen in the existing metrics. For instance, GANs better synthesize color-/texture-related attributes such as `striped fur` which DMs hardly preserve in LSUN-Cat (Section 5.4). When we increase the sampling steps of DMs, tiny objects such as `necklaces` and `earrings` tend to appear more frequently. Even though Stable diffusion v2.1 is reported that have a better FID score than Stable diffusion v1.5, the attribute-aspect score is worse than v1.5 (Section 5.5). Our approach is versatile, and applicable wherever image comparisons are needed. The code will be publicly available.

2. Related Work

Fréchet Inception Distance Fréchet Inception Distance (FID) (Heusel et al., 2017) calculates the distance between the estimated Gaussian distributions of two datasets using a pre-trained Inception-v3 (Szegedy et al., 2016). However, Kynkäänniemi et al. (2022) pointed out that patterns resembling ImageNet classes significantly influence FID. They proposed using embeddings from the CLIP encoder to make FID less susceptible to intentional or accidental distortions. Additionally, Stein et al. (2023) suggested using embeddings from DINO-v2. Despite these suggestions, both approaches merely changed the embedding space for measuring FID, relying on the raw embeddings as they are. In contrast, we design a new representation for this purpose.

Fidelity and diversity Sajjadi et al. (2018) devised precision and recall for generative model evaluation. Further refinements were provided by Kynkäänniemi et al. (2019) and Naeem et al. (2020). Generally, these metrics use a pretrained network to evaluate how embeddings of generated images match with those of real images and vice-versa.

Other metrics Beyond these, metrics such as Kernel Inception Distance (KID) (Bińkowski et al., 2018), Perceptual path length (Karras et al., 2019), Fréchet segmentation distance (Bau et al., 2019), and Rarity score (Han et al., 2022) have been introduced. The first calculates squared Maximum Mean Discrepancy (MMD) between inception representations, the second indicates latent space smoothness, the third measures pixel segmentation differences, and the latter assesses the rarity of generated images. However, these metrics predominantly rely on raw embeddings from pretrained



Visualization of CLIP embedding space (a) CLIPScore (b) Heterogeneous CLIPScore (c) CS and HCS for an image

Figure 2. **Illustration of CLIPScore and Heterogeneous CLIPScore.** We visualized the CLIP embedding space obtained from multiple texts. The yellow ellipse represents the embedding space of CelebA's text attributes, while the blue ellipse visualizes the embedding space of images. (a) CLIPScore (CS) evaluates the similarity between V_{img}^{CS} and V_{Text}^{CS} from the coordinate origin, where the angle between the two vectors is bounded, resulting in a limited similarity value. (b) Heterogeneous CLIPScore (HCS) gauges the similarity between V_{img}^{HCS} and V_{Text}^{HCS} using the defined means of images $C_{\mathcal{X}}$ and texts $C_{\mathcal{A}}$ as the origin, the range of similarity is unrestricted. (c) shows flexible values of HCS compared to CS.

classifiers, yielding scores with limited interpretability. As Figure 1a indicates, while some metrics highlight poor image generation performance, they lack in-depth explanatory insights. We aim to fill this gap with our novel, detailed, and insightful evaluation metrics.

TIFA (Hu et al., 2023) uses visual question answering to validate if text-to-image results correspond to the input texts. On the other hand, our metrics evaluate the distribution of attribute strengths in a set of images.

3. Toward Explainable Metrics

Existing metrics for evaluating generated images often use embeddings from Inception-V3 (Szegedy et al., 2016) or CLIP image encoder (Dosovitskiy et al., 2020). Yet, these embeddings lack clarity in interpreting each channel in the embedding. Instead, we opt to measure attribute strengths in images for a predefined set of attributes. We first explain CLIPScore as our starting point (Section 3.1), introduce Heterogeneous CLIPScore (Section 3.2), and describe ways of specifying the target attributes (Section 3.3.)

3.1. Measuring attribute strengths with CLIP

For a set of attributes, we start by measuring the attribute strengths of images. The typical approach is computing CLIPScore:

$$\text{CLIPScore}(x, a) = 100 \times \text{sim}(\mathbf{E_I}(x), \mathbf{E_T}(a)), \quad (1)$$

where x is an image, a is a given text of attribute, $\text{sim}(*, *)$ is cosine similarity, and $\mathbf{E_I}$ and $\mathbf{E_T}$ are CLIP image encoder and text encoder, respectively. Figure 2c shows an example CLIPScores of an image regarding a set of attributes. Yet, CLIPScores themselves do not provide a clear notion of attribute strengths as we observe ambiguous similarities between opposite attributes. The research community is already aware of such a problem. To overcome this, we introduce Heterogeneous CLIPScore in the subsequent subsections, showcased in Figure 2c, ensuring more accurate attribute strengths.

3.2. Heterogeneous CLIPScore

In the earlier section, we noted that CLIPScore tends to have a narrow value range, as visualized in Figure 2a. To remedy this, we introduce Heterogeneous CLIPScore (HCS). It uses heterogeneous initial points for image and text embedding vectors as follows.

Given training images denoted as $\{x_1, x_2, ..., x_{N_{\mathcal{X}}}\} \in \mathcal{X}$, and a set of attributes defined as $\{a_1, a_2, ..., a_{N_{\mathcal{A}}}\} \in \mathcal{A}$, we define $C_{\mathcal{X}}$ as the center of images and $C_{\mathcal{A}}$ as another center of text attributes on CLIP embedding, respectively as

$$C_{\mathcal{X}} = \frac{1}{N_{\mathcal{X}}} \sum_{i=1}^{N_{\mathcal{X}}} \mathbf{E_I}(x_i), \quad C_{\mathcal{A}} = \frac{1}{N_{\mathcal{A}}} \sum_{i=1}^{N_{\mathcal{A}}} \mathbf{E_T}(a_i). \quad (2)$$

These centers act as initial points of the embedding vectors. HCS is defined by the similarity between the two vectors, V_x and V_a. The former connects the image center to a specific image, while the latter connects the attribute center to a particular attribute. Then we define

$$V_x = \mathbf{E_I}(x) - C_{\mathcal{X}}, \quad V_a = \mathbf{E_T}(a) - C_{\mathcal{A}}, \quad (3)$$

$$\text{HCS}(x, a) = 100 \times \text{sim}(V_x, V_a), \quad (4)$$

where $\text{sim}(*, *)$ computes cosine similarity. For extending HCS from a single sample to all samples, we denote the probability density function (PDF) of $\text{HCS}(x_i, a_i)$ for all $x_i \in \mathcal{X}$ as $\text{HCS}_{\mathcal{X}}(a_i)$.

Figure 2 illustrates the difference between HCS (Heterogeneous CLIPScore) and CS (CLIPScore). HCS uses the respective centers as initial points, allowing for clearer determination of attribute magnitudes, whereas CS lacks this clarity.

3.3. Attribute selection

The effectiveness of our evaluation metric is contingent upon the target attributes we opt to measure. Inspired by recent work (Hu et al., 2023) evaluate generative models utilizing large language model (LLM), we employ a vision-language model (VLM) to determine the best attributes that truly capture generator performance. We pinpoint and assess the attributes evident in the training data via image descriptions. By analyzing the frequency of these attributes in image captions, we can identify which ones are most prevalent. To achieve this for captionless datasets, we employ the image captioning model, BLIP (Li et al., 2022), to extract words related to attributes from the training data. We then adopt N frequently mentioned ones as our target attributes, denoted as $\mathcal{A}$, for the metric. Given that these attributes are derived automatically, utilizing BLIP for this extraction could serve as a foundational method. In our Table S15, we demonstrate that using BLIP to extract attributes exhibits similar tendencies to both using CelebA GT labels and defining attributes through LLM.

4. Evaluation Metrics with Attribute Strengths

In this section, we harness the understanding of attribute strengths to devise two comprehensible metrics. Section 4.1 introduces Single-attribute Divergence (SaD), quantifying the discrepancy in attribute distributions between training data and generated images. Section 4.2 brings forth Paired-attribute Divergence (PaD), evaluating the relationship between attribute strengths.

4.1. Single-attribute Divergence

If we have a dataset with dogs and cats, and a generative model only makes dog images, it is not an ideal model because it does not produce cats at all (Goodfellow et al., 2016). With this idea, we say one generative model is better than another if it makes a balanced number of images for each attribute similar to the training dataset. Since we do not know the true distribution of real and fake images, we came up with a new metric, Single-attribute Divergence (SaD). This metric checks how much of each attribute is in the dataset by utilizing interpretable representation. Our metric, SaD, quantifies the difference in density for each attribute between the training dataset ($\mathcal{X}$) and the set of generated images ($\mathcal{Y}$). We define SaD as

$$\text{SaD}(\mathcal{X},\mathcal{Y}) = \frac{1}{M}\sum_{i}^{M}\text{KL}(\text{HCS}_{\mathcal{X}}(a_i),\text{HCS}_{\mathcal{Y}}(a_i)), \quad (5)$$

where i denotes an index for each attribute, M is the number of attributes, KL(*) is Kullback-Leibler divergence, and $\text{HCS}_{\mathcal{X}}(a_i)$ denotes PDF of $\text{HCS}(x_i, a_i)$ for all $x_i \in \mathcal{X}$.

We first estimate PDFs of Heterogeneous CLIPScore for each attribute present in $\mathcal{X}$ and $\mathcal{Y}$ by applying Gaussian Kernel Density Estimation (KDE) on the entire sample for each dataset. Subsequently, we compare these HCS PDFs which reflect the distribution of attribute strengths within the datasets. If an attribute's distribution in $\mathcal{X}$ closely mirrors that in $\mathcal{Y}$, their respective HCS distributions will align, leading to similar PDFs. To measure discrepancies between these distributions, we employ Kullback-Leibler Divergence (KLD). This quantifies how much the generated images either over-represent or under-represent specific attributes compared to the original data. Subsequently, we determine the average divergence across all attributes between $\mathcal{X}$ and $\mathcal{Y}$ to derive the aggregated metric for SaD.

In addition, we define the mean difference of attribute strength to further examine whether poor SaD comes from excessive or insufficient strength of an attribute a:

$$\text{mean difference} = \frac{1}{N_x}\sum_{i}^{N_x}\text{HCS}(x_i, a) - \frac{1}{N_y}\sum_{i}^{N_y}\text{HCS}(y_i, a). \quad (6)$$

where N_x and N_y are the number of training images and generated images, respectively. Intuitively, a high magnitude of mean difference indicates the mean strength of $\mathcal{Y}$ differs significantly from $\mathcal{X}$ for attribute a. A positive value indicates $\mathcal{Y}$ has images with stronger a than $\mathcal{X}$, and vice versa for a negative value. While this does not conclusively reveal the exact trend due to a's complex distribution, it provides an intuitive benchmark.

4.2. Paired-attribute Divergence

We introduce another metric, Paired-attribute Divergence (PaD), aimed at evaluating whether generated images maintain the inter-attribute relationships observed in the training data. Essentially, if specific attribute combinations consistently appear in the training data, generated images should also reflect these combinations. To illustrate, if every male image in the training dataset is depicted wearing glasses, the generated images should similarly represent males with glasses. We assess this by examining the divergence in the joint probability density distribution of attribute pairs between the training data and generated images. This metric, termed Paired-attribute Divergence (PaD), leverages joint probability density functions as detailed below:

$$\text{PaD}(\mathcal{X},\mathcal{Y}) = \frac{1}{|P|}\sum_{(i,j)}^{P}\text{KL}(\text{HCS}_{\mathcal{X}}(a_{i,j}),\text{HCS}_{\mathcal{Y}}(a_{i,j})), \quad (7)$$

where M is the number of attributes, $P = \binom{M}{2}$, (i, j) denotes an index pair of attributes selected out of M, and the joint PDF of the pair of attributes is denoted as $\text{HCS}_{\mathcal{X}}(a_{i,j})$.

When utilized together with SaD, PaD will offer a compre-

Table 1. **CLIPScore and Heterogeneous CLIPScore's accuracy on CelebA dataset.**

	accuracy	f1 score
Heterogeneous CLIPScore	**0.817**	**0.616**
CLIPScore	0.798	0.575

hensive analysis of the model's performance. For instance, if the probability density function of the generator for the attribute pair (`baby`, `beard`) diverges notably from the training data's distribution while SaD for `baby` and `beard` are comparatively low, it suggests that the generator may not be effectively preserving the (`baby`, `beard`) relationship. Consequently, PaD enables us to quantify how well attribute relationships are maintained in generated data. To the best of our knowledge, we are the first to propose a metric for interdependencies relationship. Moreover, this becomes interpretable.

5. Experiments

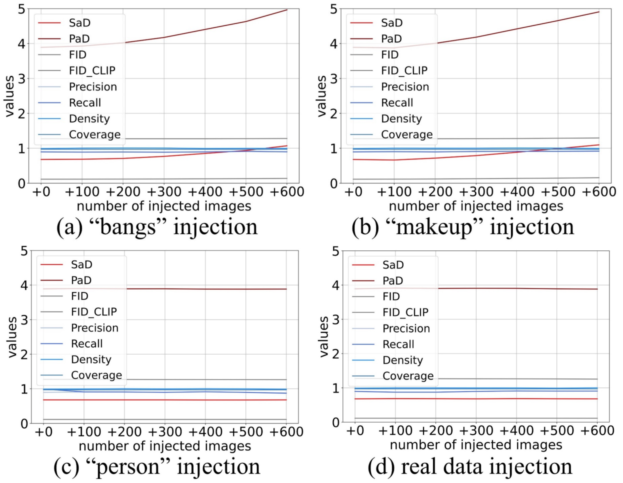
Experiment details For estimating the probability density function (PDF) of Heterogeneous CLIPScore (HCS) in both the training data and generated images, we use Gaussian Kernel Density Estimation (KDE). In this process, we extract 10,000 samples from generated and real images to obtain PDFs of attribute strengths. These PDFs are then used to compute SaD and PaD. In every experiment, we use a set of $N_{\mathcal{A}} = 20$ attributes. In the case of the toy experiments on FFHQ, we use attributes from CelebA ground truth label (Table S17) for the convenience of interpretation.

5.1. Heterogeneous CLIPScore outperforms CLIPScore

Heterogeneous CLIPScore (HCS) outshines CLIPScore (CS) in binary classification of the attributes in CelebA. Table 1 reports accuracy and F1 score computed as follows. For each attribute, we sort the scores of test images and classify the images with top k scores as positives to compute accuracy and F1 score where k denotes the number of positive images in the test set. Then we compute their mean over all attributes. This superiority persists even for refined attributes which excludes subjective attributes such as `attractive` or `blurry` as shown in Table S8. Table S7 provides the full list of attributes and the refined attributes. More details are available in the Appendix A.2.

5.2. Biased data injection experiment: the effectiveness of our metric

In this subsection, we conduct a toy experiment to validate our metrics against existing methods. Initially, two non-overlapping subsets, each with 30K images from FFHQ, are defined as training data $\mathcal{X}$ and generated images $\mathcal{Y}$. Starting with these subsets that share a similar distribution, we gradually infuse biased data into $\mathcal{Y}$. The biased data is generated using DiffuseIT (Kwon & Ye, 2022). We translate samples from the training data, without overlap to the initial 60K images, into `makeup` (Figure 3a) and `bangs` (Figure 3b). We also provide controlled counterpart where injected samples are unbiased data translated into the `person` (Figure 3c), or injected samples remain untranslated (Figure 3d).



(a) "bangs" injection (b) "makeup" injection

(c) "person" injection (d) real data injection

Figure 3. **Validation of metrics through biased injection.** We design one set: typical 30K of FFHQ images, and another set: 30K FFHQ + injected images. Biased data injection, illustrated in (a) with `makeup` and (b) with `bangs` leads to an increase in both SaD and PaD rise. In contrast, unbiased data injection (c) `person` and (d) real data, injecting the same distribution as the training set results in no SaD and PaD rise. Our metrics effectively capture changes in attribute distribution, while existing metrics cannot.

As depicted in Figure 3, our metrics display a consistent trend: SaD and PaD rise with the inclusion of more edited images in $\mathcal{Y}$, whereas other metrics are static. Thanks to the attribute-based design, our metric suggests that `makeup` or `bangs` is the dominant factor for SaD, and relationships that are rarely seen in training data such as (`man`, `makeup`) and (`man`, `bangs`) for PaD. The impact on SaD and PaD scales linearly with the number of images from different attribute distributions. For an expanded discussion and additional experiments, refer to Figure S8 and Appendix B.4. These results underscore that SaD adeptly discerns the attribute distribution variation, and PaD identifies the joint distribution shift between attribute pairs, outperforming other metrics.

5.3. Discernment of PaD

In another toy experiment, we designed a scenario where SaD metric struggled to detect specific attribute relation-

Table 2. **Discernment of PaD.** Table 2 shows SaD to be consistent, while PaD to be different because the correlation between gender and eyeglasses is corrupted. We use 10,000 images for each set, and marginals (the number of images for `men`, `women` and `eyeglasses`) are the same across set A, B, and C.

	SaD	PaD	PaD(`men` & `eyeglasses`)
between set A & set B	1.52	6.14	5.68
between set A & set C	1.51	**9.60**	**251.34**

ships, while PaD metric successfully pinpointed them. We curate three sets of images sampled from CelebA, with the identical marginal distribution of individual attributes. **set A**: randomly sampled with a restriction: only `men` wear `eyeglasses`, **set B**: randomly sampled with a restriction: only `men` wear `eyeglasses` (identical to set A), and **set C**: randomly sampled with a restriction: only `women` wear `eyeglasses` (corrupted correlation).

We then observe whether SaD and PaD capture difference due to the corrupted correlation between gender (`men` or `women`) and `eyeglasses`, measuring SaD and PaD of **set B** and **set C** from **set A**. Table 2 shows PaD correctly reveals the most influential pair of attributes (`men` & `eyeglasses`) by 26 times higher PaD than the mean PaD, while SaD struggles to capture corruption of correlation between gender (`men` or `women`) and `eyeglasses`. This highlights the effectiveness of PaD in capturing errors in pairwise relations, and the necessity of employing PaD for comprehensive model analysis. Full SaD and PaD is available in Figure S12.

5.4. Comparing generative models with our metrics

Leveraging the superior sensitivity and discernment of our proposed metrics, we evaluate the performance of GANs and Diffusion Models (DMs) in Table 3. Generally, the tendency of SaD and PaD align with other existing metrics. However three notable points emerge; 1) ProjectedGAN (Sauer et al., 2021) lags in performance, 2) As sampling timesteps in DM increase, FIDs improve, while SaD and PaD decline. 3) GANs and Diffusion models vary in their strengths and weaknesses concerning specific attributes.

1) ProjectedGAN (Sauer et al., 2021) prioritizes matching the training set's embedding statistics for improving FID rather than improving actual fidelity (Kynkäänniemi et al., 2022). While it performs well in existing metrics, it notably underperforms in SaD and particularly in PaD. This implies that directly mimicking the training set's embedding stats does not necessarily imply correct attribute correlations. Figure 4 provides failure cases generated by ProjectedGAN.

2) Diffusion models typically yield better quality with higher number of sampling timesteps. Yet, SaD and PaD scores for LDM with 200 steps surpass those of LDM with 50 steps.

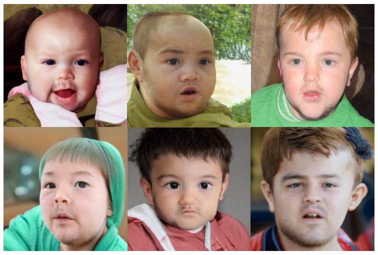


Figure 4. **Failure cases by ProjectedGAN.** ProjectedGAN disregards attribute relationships, such as generating babies with beards.

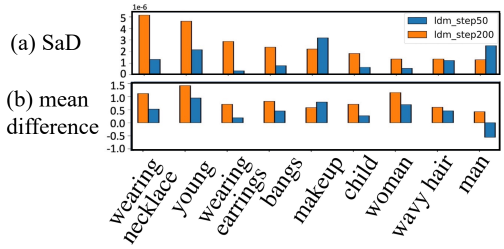


Figure 5. **LDM with 50 steps v.s. LDM with 200 timesteps.** With increased sampling timesteps, (a) SaD of LDM gets worse, (b) since making too many fine objects such as `earrings` or `necklace`.

As illustrated in Figure 5, higher sampling timesteps in the LDM model produce more high-frequency elements such as `necklaces` and `earrings`. This could explain the dominance of attributes such as `young, makeup, woman, wavy hair` naturally. We suppose that dense sampling trajectory generates more high-frequency objects. The scores and mean differences of each attribute are depicted in Figure 5a and Figure 5b respectively.

3) Diffusion models fall short on modeling color-related attributes than shape-related attributes. As our metrics provide flexible customization, we report SaD and PaD of color attributes (e.g., `yellow fur`, `black fur`) and shape attributes (e.g., `pointy ears`, `long tail`) within LSUN Cat dataset. Table 4 shows that iDDPM excels in matching shape attributes compared to color attributes.

This aligns with the hypothesis by Khrulkov et al. (2022) suggesting that DMs learn the Monge optimal transport map, the shortest trajectory, from Gaussian noise distribution to image distribution regardless of training data. This implies that when the initial latent noise x_T is determined, the image color is also roughly determined because the diffused trajectory tends to align with the optimal transport map.

In addition, iDDPM shows notable scores, with the attribute `arched eyebrows` showing scores over two times higher than GANs in SaD, and attributes related to `makeup` consistently receive high scores across all StyleGAN 1, 2, and 3 models in PaD. Investigating how the gen-

Table 3. **Comparing the performance of generative models.** We computed each generative model's performance on our metric with their official pretrained checkpoints on FFHQ (Karras et al., 2019). We used 50,000 images for both GT and the generated set.

	StyleGAN1	StyleGAN2	StyleGAN3	iDDPM	LDM (50)	LDM (200)	StyleSwin	ProjectedGAN
SaD (10^{-7})↓	11.35	**7.52**	7.79	14.78	10.42	14.04	10.76	17.61
PaD (10^{-7})↓	27.25	**19.22**	19.73	34.04	25.36	30.71	26.56	41.53
FID↓	4.74	**3.17**	3.20	7.31	12.18	11.86	4.45	5.45
FID_{CLIP}↓	3.17	**1.47**	1.66	2.39	3.89	3.57	2.45	3.63
Precision↑	0.90	0.92	0.92	0.93	**0.94**	0.91	0.92	0.92
Recall↑	0.86	0.89	0.90	0.84	0.82	0.88	0.91	**0.92**
Density↑	1.05	1.03	1.03	**1.09**	1.09	1.07	1.01	1.05
Coverage↑	0.97	0.97	0.97	0.95	0.94	0.97	0.97	0.97

Table 4. **SaD and PaD of models with different attributes for LSUN Cat.** Analyzing the weakness of iDDPM for specific attribute types, such as color or shape.

	color attributes		shape attributes	
	SaD (10^{-7})↓	PaD (10^{-7})↓	SaD (10^{-7})↓	PaD (10^{-7})↓
StyleGAN1 (Karras et al., 2019)	**139.03**	**248.96**	169.76	318.46
StyleGAN2 (Karras et al., 2020b)	**112.06**	**195.75**	132.41	246.44
iDDPM (Nichol & Dhariwal, 2021)	46.93	85.99	**32.48**	**62.69**

eration process of GANs or DMs affects attributes such as attributes would be an intriguing avenue for future research. See Appendix D for details.

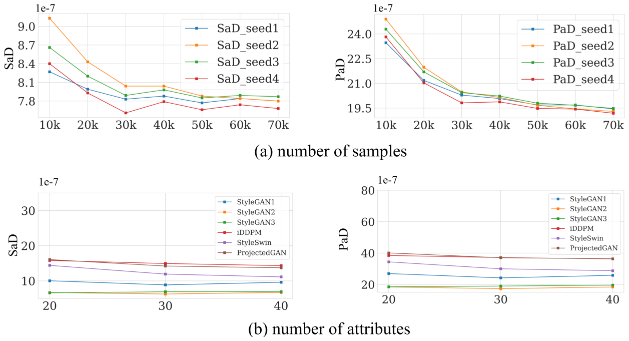
5.5. Evaluating text-to-image models

Recently, there has been a huge evolution of text-to-image generative models (Nichol et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Balaji et al., 2022). To evaluate text-to-image models, zero-shot FID score on COCO (Lin et al., 2014) is widely used including Stable Diffusion (SD). Instead, we use our metrics to examine text-to-image models regarding excessively or insufficiently generated attributes. We generate 30K images with captions from COCO using SDv1.5 and SDv2.1 to calculate SaD and PaD with attributes extracted from the captions. We use $N_{\mathcal{A}} = 30$.

Table 5 shows SDv1.5 has twice better SaD and PaD than SDv2.1. Interestingly, the mean difference of attribute strengths is below zero. It implies that SDs tend to omit some concepts such as `group`[1] or `plate`[2]. In particular, SDv2.1 struggles to generate scenes with multiple people. It aligns with common claims[3] about SDv2.1 even though it achieves low FID. We also conduct similar experiments on the LAION-2B dataset, and SDv1.5 shows continuously better SaD and PaD. We provide more details in Appendix B.5.

Table 5. **SaD and PaD of different versions of Stable Diffusion.** Stable Diffusion v1.5 is almost twice better than v2.1. We generate 30,000 images using the captions from COCO. We use $N_{\mathcal{A}} = 30$.

$N_{\mathcal{A}} = 30$	SaD (10^{-7})↓	PaD (10^{-7})↓	SaD worst-rank attribute (mean difference)		
			1st	2nd	3rd
SDv1.5	**24.37**	**60.71**	plate (-1.9)	group (-1.6)	building (-1.6)
SDv2.1	48.23	106.86	group (-3.7)	plate (-2.5)	person (-2.7)



(a) number of samples

(b) number of attributes

Figure 6. **SaD and PaD over a different number of samples and attributes.** (a) SaD and PaD are stable with more than 50,000 images. (b) The ranking of models mostly remains consistent regardless of the number of attributes.

5.6. Impact of sample size and attribute count on proposed metric

In Figure 6, we conduct ablation experiments to study the impact of the number of samples and attributes. Using four random seeds, we generate images with StyleGAN3 from FFHQ. We posit that SaD and PaD begin to standardize with 30,000 images and become more stable with over 50,000 images. Figure 6b provides SaD and PaD of various models over different numbers of attributes where the attributes from BLIP are sorted by their number of occurrences in the dataset. The ranking of the models largely stays stable irrespective of the number of attributes. We suggest that 20 attributes are sufficient for typical evaluation, but leveraging a broader range offers richer insights.

[1] e.g., A group of people is standing around a large clock.

[2] e.g., A table is set with two plates of food and a candle.

[3] https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/

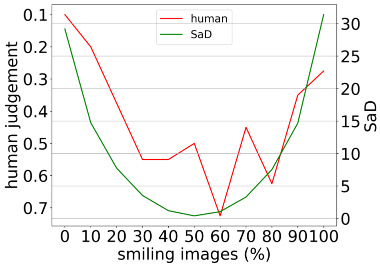


Figure 7. **Correlation between human judgements and SaD.**

5.7. Alignment with human judgment

Prior metrics (Heusel et al., 2017; Sajjadi et al., 2018; Naeem et al., 2020) faced challenges in evaluating the correspondence with human judgment between large image sets, given the impracticality of storing visual features of thousands of images in human memory. In response, we conduct an alternative approach, assessing the alignment between our metric and human judgment, particularly when there are shifts in attribute distribution(s) within one set.

We show SaD and PaD are consistent with human judgment on the CelebA dataset. 40 participants participated in these surveys.

SaD Figure 7 shows a correlation between SaD and human judgments. We asked the participants to mark if two sets have different distribution of smile. One set is fixed as a training set with 50% of them `smile`. Another set varies from 0% of them `smile` to 100% of them `smile`. We used smiling and non-smiling images from CelebA ground truth labels. Meanwhile, we measure SaD between the two sets and for comparison. Notably, both SaD and human judgement rapidly increase with increasing and decreasing smile in >80% and <30% range, respectively. Likewise, both SaD and human judgement have gentle change with same sign of slope in 30% < smile < 80% range.

PaD Table 6 shows a correlation between PaD and human judgments. Based on the given ground-truth set A, participants ranked three sets; 1) a set with strong positive correlation (r=1) 2) a set with zero-correlation (r=0) and 3) a set with strong negative correlation (r=-1).

We opt to use the correlations between `man` and `smile` and we gave five triplets to the participants to rank within the triplets. Most (about 94%) of the participants identified the rank of correlation between `man` and `smile` correctly and it aligns with PaD.

6. Conclusion and Discussion

We have introduced novel metrics that evaluate the distribution of attribute strengths. Single-attribute Divergence reveals which attributes are correctly or incorrectly modeled. Paired-attribute Divergence considers the joint occurrence of attributes in individual images. The explicit interpretability of these metrics allows us to know which generative model suits the user's necessity. Furthermore, Heterogeneous CLIPScore more accurately captures the attribute strengths than CLIPScore.

Table 6. **Correlation between human judgements and PaD.**

set A	set B	PaD	Human 1st (%)	Human 2nd (%)	Human 3rd (%)
r=1	r=1	**4.57**	**94.36**	3.59	2.05
	r=0	38.43	2.56	**93.85**	3.59
	r=-1	117.58	3.08	2.56	**94.36**

Our metrics have the advantage of revealing the distribution of attributes from a *set* of generated images where human judgment faces difficulty in observing attributes in excessively many images. Furthermore, our research establishes a solid foundation for the development of explainable evaluation metrics for generative models and contributes to the advancement of the field.

Discussion 1) Estimating probability density functions (PDFs) with Kernel Density Estimation (KDE) requires a sufficient (>50K) number of samples. With a sufficient number of samples, KDE can effectively approximate the PDF of a dataset, capturing the underlying distribution of the data. This is particularly important in complex data sets where the distribution of attributes might be intricate or not immediately obvious. 2) Our metrics can be influenced by quality of vision-language model (VLM). I.e., a biased or limited extraction of attributes may bring misrepresentation of our metrics. For instance, a VLM that disproportionately emphasizes certain attributes or overlooks others can skew the analysis, leading to an inaccurate assessment of the data. 3) Exploring strengths of other aspects such as texture (Caron et al., 2021; Oquab et al., 2023; Kirillov et al., 2023) or other modalities (Girdhar et al., 2023) may provide valuable insights and enhance the robustness.

Furthermore, we wish to highlight the flexibility of our approach. While we have conducted evaluations using attributes defined by BLIP, our method allows for the customization of attributes to suit the specific needs of the task and the user's objectives. For instance, in pursuing fairness with a focus on equitable generation of features such as race and gender, these can be directly employed as attributes. Alternatively, for specific tasks like image translation, desired attributes can be selectively chosen to tailor the evaluation process. However, it's crucial to exercise caution in this selection process of attributes to avoid introducing bias.

We believe our new evaluation metrics, with interpretability and the ability to encapsulate user intention, will have a healthy impact on the research community.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, such as if the training set contains biases, our metrics favor biased generation, which may introduce negative societal impact.

References

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022.

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. Seeing what a gan cannot generate. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4502–4511, 2019.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. arXiv preprint arXiv:1801.01401, 2018.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11472–11481, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190, 2023.

Goodfellow, I., Bengio, Y., and Courville, A. Deep learning. book in preparation for mit press. URL¡ http://www. deeplearningbook. org, 1, 2016.

Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. Rarity score: A new metric to evaluate the uncommonness of synthesized images. arXiv preprint arXiv:2206.08549, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897, 2023.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401–4410, 2019.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. Advances in neural information processing systems, 33:12104–12114, 2020a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110–8119, 2020b.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. Advances in Neural Information Processing Systems, 34:852–863, 2021.

Khrulkov, V., Ryzhakov, G., Chertkov, A., and Oseledets, I. Understanding ddpm latent codes through optimal transport. arXiv preprint arXiv:2202.07477, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.

Kwon, G. and Ye, J. C. Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264, 2022.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems, 32, 2019.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in frechet inception distance. arXiv preprint arXiv:2203.06026, 2022.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International

Conference on Machine Learning, pp. 12888–12900. PMLR, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755. Springer, 2014.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In International Conference on Machine Learning, pp. 7176–7185. PMLR, 2020.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pp. 8162–8171. PMLR, 2021.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. Advances in neural information processing systems, 31, 2018.

Sauer, A., Chitta, K., Müller, J., and Geiger, A. Projected gans converge faster. Advances in Neural Information Processing Systems, 34:17480–17492, 2021.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

Stein, G., Cresswell, J. C., Hosseinzadeh, R., Sui, Y., Ross, B. L., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=08zf7kTOoh.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826, 2016.

Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., and Wen, F. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18381–18391, 2023.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., and Guo, B. Styleswin: Transformer-based gan for high-resolution image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11304–11314, 2022.

A. Appendix

A. Implementation Details

A.1. Additional experiment setup

Details of generated images We generate samples using official checkpoints provided by StyleGANs (Karras et al., 2019; 2020b;a; 2021), ProjectedGAN (Sauer et al., 2021), Styleswin (Zhang et al., 2022), iDDPMs (Nichol & Dhariwal, 2021; Choi et al., 2022), and LDM (Rombach et al., 2022). We use 50K of training images and generated images for both FFHQ (Karras et al., 2019) and LSUN Cat (Yu et al., 2015) experiment.

Miscellaneous We use `scipy.stats.gaussian_kde`(dataset, '*scott*', None) to estimate the distribution of Heterogeneous CLIPScore for given attributes. We observe HCS values mainly feature unimodal and bimodal distributions, ensuring that the sample sizes are large enough for effective KDE application. We've chosen Scott's Rule for bandwidth selection, as recommended by the default settings in `scipy.stats.gaussian_kde`. This recommendation is due to its balanced approach to managing bias and variance in our data estimation, adjusting the bandwidth based on data size and dimensionality. In KDE, the bandwidth directly influences the standard deviation of the Gaussian kernels; a larger bandwidth leads to a smoother density estimate, while a smaller bandwidth results in a more detailed density estimate. This directly affects the std of the kernels used in our analysis. This rule scales the bandwidth with $n^{-1/(d+4)}$, where n is the number of data points, and d is the number of dimensions.

We use `spacy.load("en_core_web_sm")` to extract attributes from BLIP(Li et al., 2022) captions. We resize all images to 224x224. We used `"ViT-B/32"` (Dosovitskiy et al., 2020) as a CLIP encoder. We used a single NVIDIA RTX 3090 GPU (24GB) for the experiments.

A.2. Details of CelebA accuracy experiment

Table S8 displays binary classification results for all attributes in CelebA using both CS and HCS, comparing them to the ground truth attribute labels. By setting the threshold based on the number of positive labels for each CelebA attribute, we found that the accuracy and F1 score of HCS are superior to CS, regardless of whether we use micro or macro averaging. Additionally, we conducted experiments by setting the origin of HCS as the overall mean of both image and text means, validating that using separate text and image means is essential.

Table S7. **Attributes used for CelebA accuracy experiment.**

Attribute type	Attribute
Refined attributes	Arched_Eyebrows, Bags_Under_Eyes, Bald, Bangs, Big_Nose, Black_Hair, Blond_Hair,Brown_Hair, Chubby, Double_Chin, Eyeglasses, Goatee, Gray_Hair, Heavy_Makeup, Male, Mouth_Slightly_Open, Mustache, No_Beard, Sideburns, Smiling, Straight_Hair, Wavy_Hair, Wearing_Earrings, Wearing_Hat,Wearing_Lipstick, Wearing_Necklace, Wearing_Necktie, Young
All attributes	5_o_Clock_Shadow, Arched_Eyebrows, Attractive, Bags_Under_Eyes, Bald, Bangs, Big_Lips,Big_Nose, Black_Hair, Blond_Hair, Blurry, Brown_Hair, Chubby, Double_Chin, Eyeglasses, Goatee, Gray_Hair, Heavy_Makeup, High_Cheekbones, Male, Mouth_Slightly_Open, Mustache, Narrow_Eyes, No_Beard, Oval_Face, Pale_Skin, Pointy_Nose, Receding_Hairline, Rosy_Cheeks, Sideburns, Smiling, Straight_Hair, Wavy_Hair, Wearing_Earrings, Wearing_Hat,Wearing_Lipstick, Wearing_Necklace, Wearing_Necktie, Young

B. Additional Ablation Study

B.1. Necessity of separating image mean and text mean

In the main paper, we defined Heterogeneous CLIPScore as computing angles between vectors V_x and V_a. V_x is a vector from the center of images to an image in CLIP space. V_a is a vector from the center of captions to an attribute in CLIP space. Table S8 quantitatively validates the effectiveness of setting the origin of V_a as the center of captions ($C_{\mathcal{A}}$) compared to the

Table S8. **Accuracy from CelebA ground truth labels.** Heterogeneous CLIPScore with origin at the entire center of images and texts is seriously inferior to the one with origin at the separate center of images ($C_{\mathcal{X}}$) and texts ($C_{\mathcal{A}}$). It validates the definition of V_a.

		accuracy	f1 score(macro)	f1 score(micro)
All attributes	HCS	**0.794**	**0.442**	**0.545**
	CS	0.781	0.392	0.515
Refined attributes	HCS	**0.817**	**0.519**	**0.616**
	CS	0.798	0.450	0.575

Table S9. **Top 40 appeared attribute in COCO validation captions.** The first and third rows represent the attributes in COCO validation captions, while the second and fourth rows represent the corresponding number of appearances of these attributes in the captions.

man	woman	he	people	person	table	group	street	water	plate
20262	9352	8212	8164	7196	6584	6401	4382	3741	3717
cat	field	couple	dog	side	food	beach	bed	bathroom	road
3476	3385	3301	3071	2981	2973	2731	2687	2477	2377
grass	kitchen	skateboard	picture	road	train	building	snow	surfboard	toilet
2346	2286	2259	2209	2165	2140	2108	2097	1968	1879
giraffe	room	men	bunch	ball	air	bench	clock	boy	sign
1874	1827	1819	1809	1807	1710	1630	1607	1573	1569

Table S10. **SaD and PaD for text-to-image models.** Stable diffusion v1.5 outperforms Stable Diffusion v2.1 in SaD and PaD regardless of the number of attributes despite being known as inferior in FID.

	SaD			PaD		
	$N_{\mathcal{A}} = 20$	$N_{\mathcal{A}} = 30$	$N_{\mathcal{A}} = 40$	$N_{\mathcal{A}} = 20$	$N_{\mathcal{A}} = 30$	$N_{\mathcal{A}} = 40$
SDv1.5	**37.91**	**24.37**	**25.44**	**87.53**	**60.71**	**62.47**
SDv2.1	69.49	48.23	47.53	146.12	106.86	105.03

center of images ($C_{\mathcal{X}}$).

B.2. Replacing Heterogeneous CLIPScore with CLIPScore

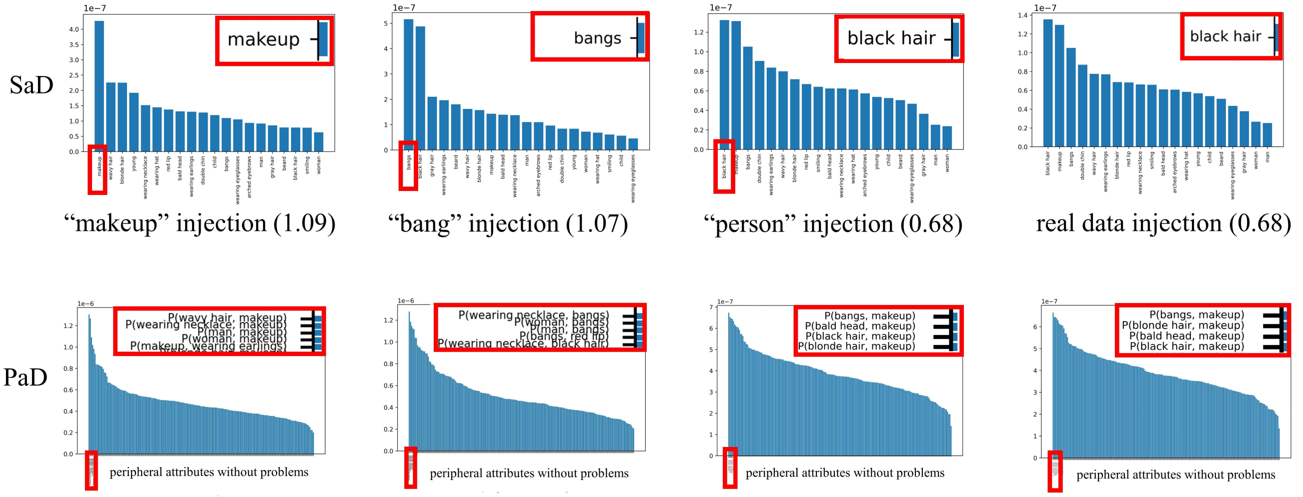
We also include additional comparisons of SaD and PaD across different image injection settings with CLIPScore rather than Heterogeneous CLIPScore in Figure 3. Compared to the validation result with Heterogeneous CLIPScore, both results reflect a corresponding tendency: the more correlated image injected, the worse performance in the proposed metric. However, considering the quantitative effectiveness we demonstrated for Heterogeneous CLIPScore in Table S8, we highly recommend using Heterogeneous CLIPScore with proposed metrics: SaD and PaD.

B.3. Attribute combination involving more than two attributes

Proposed metric is capable of examining relationships involving any number of attributes (N-way relationships), as it fundamentally relies on measuring the joint probability among involved attributes. The methodology proposed in our paper remains applicable for considering the joint probability of three or more attributes. This is particularly true for exceeding 30k images, where we did not observe statistical instability.

While it's technically feasible to extend the analysis to encompass interactions among three or more attributes (N-way relationships), we focus on the two-attribute relationships, which provide the finest level of granularity. Furthermore, our experiments have shown that the model rankings remains consistent regardless of the complexity of attribute relationships considered.

Table S11 supports the above statement: the left two columns report the worst three triplets of attributes and the rightmost column reports the worst pairs with their ranks connected to each triplet. The pairs being included in the triplets in the same row indicates that the worst triplets can be identified by the worst pairs.



"makeup" injection (1.09) "bang" injection (1.07) "person" injection (0.68) real data injection (0.68)

"makeup" injection (4.90) "bang" injection (4.96) "person" injection (3.88) real data injection (3.88)

Figure S8. **Correlated images injection experiment.**

Table S11. **Results with attribute triplets**

triplet ranks	triplet attributes	similar PaD ranks (PaD attributes)
1	man & woman & wearing necklace	R1(man& woman)
2	child & red lip & makeup	R2(red lip& makeup), R3(child & makeup)
3	red lip & makeup & young	R2(red lip& makeup), R4(makeup & young)

B.4. Can SaD and PaD also capture skips of attribute?

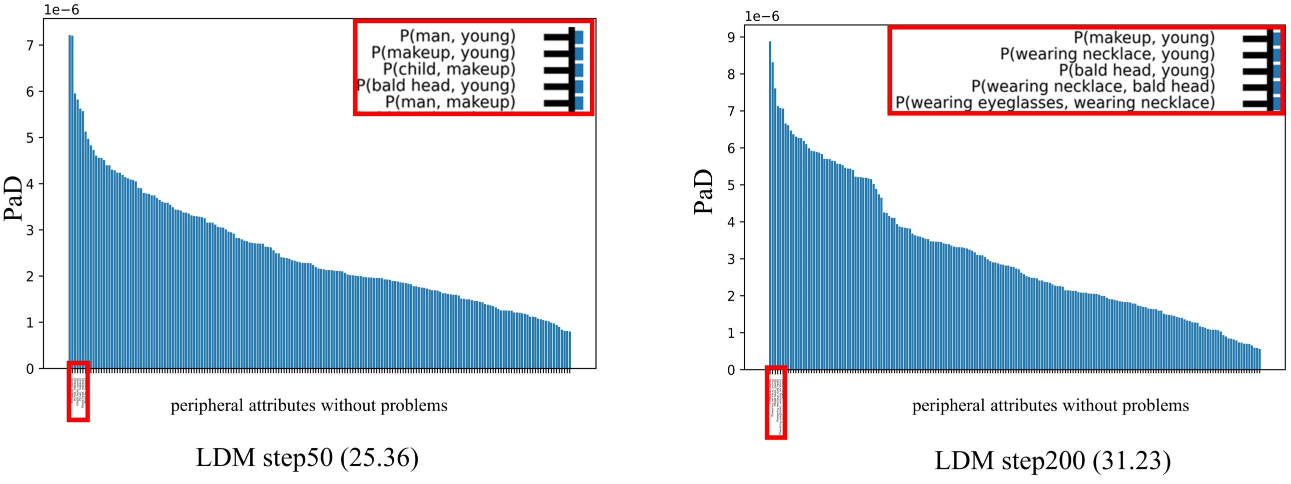
We validate that SaD and PaD accurately capture the skipness of certain attributes in Table S12. Using CelebA annotation labels, we construct sets A and B with 50k images, each naturally containing 3,325 and 3,260 images with eyeglasses, respectively. As we intentionally replace images with eyeglasses in set B with images without eyeglasses, SaD and PaD deteriorated linearly with an increasing number of replaced images, with the `eyeglasses` attribute making a more significant contribution to SaD and PaD. It demonstrates proposed metric effectively catches the skipness of some attributes, and accurately captures the distribution change of the attribute HCS probability density function.

B.5. More details: text-to-image model evaluation

The number of attributes We compare Stable Diffusion v1.5 and Stable Diffusion v2.1 on the COCO dataset using top-N appeared attributes in COCO validation captions (Table S9). Regardless of number of attributes, SDv1.5 outperforms SDv2.1 in SaD and PaD (Figure S14, Table S10).

Table S12. **Validation result of skips experiment.**

	SaD	PaD	most influencing attribute for SaD
$\frac{\text{eyeglasses 3325}}{\text{total 50000}}$ v.s. $\frac{\text{eyeglasses 3260}}{\text{total 50000}}$	0.63	3.42	beard
$\frac{\text{eyeglasses 3325}}{\text{total 50000}}$ v.s. $\frac{\text{eyeglasses 2000}}{\text{total 50000}}$	0.89	4.05	**eyeglasses**
$\frac{\text{eyeglasses 3325}}{\text{total 50000}}$ v.s. $\frac{\text{eyeglasses 1000}}{\text{total 50000}}$	1.54	5.66	**eyeglasses**
$\frac{\text{eyeglasses 3325}}{\text{total 50000}}$ v.s. $\frac{\text{eyeglasses 3325}}{\text{total 50000}}$	3.25	11.59	**eyeglasses**



LDM step50 (25.36)

LDM step200 (31.23)

Figure S9. **PaD for LDM with different sampling timesteps.**

Table S13. **SaD and PaD for text-to-image models: LAION-2B.** Stable diffusion v1.5 outperforms Stable Diffusion v2.1 in SaD and PaD regardless of dataset type despite being known as inferior in FID.

	SaD	PaD
SDv1.5	**14.26**	**40.27**
SDv2.1	32.30	62.54

Table S14. **Correlation between human judgements and SaD.**

set A	set B	SaD	Human 1st (%)	Human 2nd (%)	Human 3rd (%)
	strong smile	**0.89**	**99.48**	0	0.51
strong smile	medium smile	19.39	0	**99.48**	0.51
	no smile	92.46	0.51	0.51	**98.97**

Table S15. **Alignment of result from different attribute selection methologies.** Regardless of attribute selection methologies (BLIP, CelebA GT labels), tendency of SaD and PaD remains same.

		StyleGAN1	StyleGAN2	StyleGAN3	iDDPM	LDM (50)	LDM (200)	StyleSwin	ProjectedGAN
BLIP	SaD (10^{-7})↓	10.04	6.70	**6.60**	15.81	12.97	10.36	14.41	16.06
	PaD (10^{-7})↓	26.96	**18.44**	18.56	38.48	31.87	26.03	34.41	40.09
CelebA GT label	SaD (10^{-7})↓	11.35	**7.52**	7.79	14.78	10.42	14.04	10.76	17.61
	PaD (10^{-7})↓	27.25	**19.22**	19.73	34.04	25.36	30.71	26.56	41.53

Different dataset We compare Stable Diffusion v1.5 and Stable Diffusion v2.1 with a 30k subset of the LAION-2B dataset as shown in Table S13. The outcomes align closely with the values reported in the paper.

C. Is BLIP enough to represent attributes of images?

Table S15 shows the alignment between the results extracted using BLIP and the CelebA GT labels used as attributes. Through this, we argue that the attributes extracted by BLIP can adequately represent the significant attributes of an image dataset. Additionally, we can select specific domain attributes using LLMs. In the case of LSUN, we used LLMs to extract attributes related to shape and texture.

Details of GPT queries Table S16 provides the questions we used for preparing GPT attributes. We accumulated GPT attributes by iteratively asking GPT to answer 'Give me 50 words of useful, and specific adjective visual attributes for $\{question\}$'. Then, we selected the top N attributes based on their frequency of occurrence, ensuring that the most frequently mentioned attributes were prioritized. We suppose that the extracted attributes might be biased due to the inherent randomness

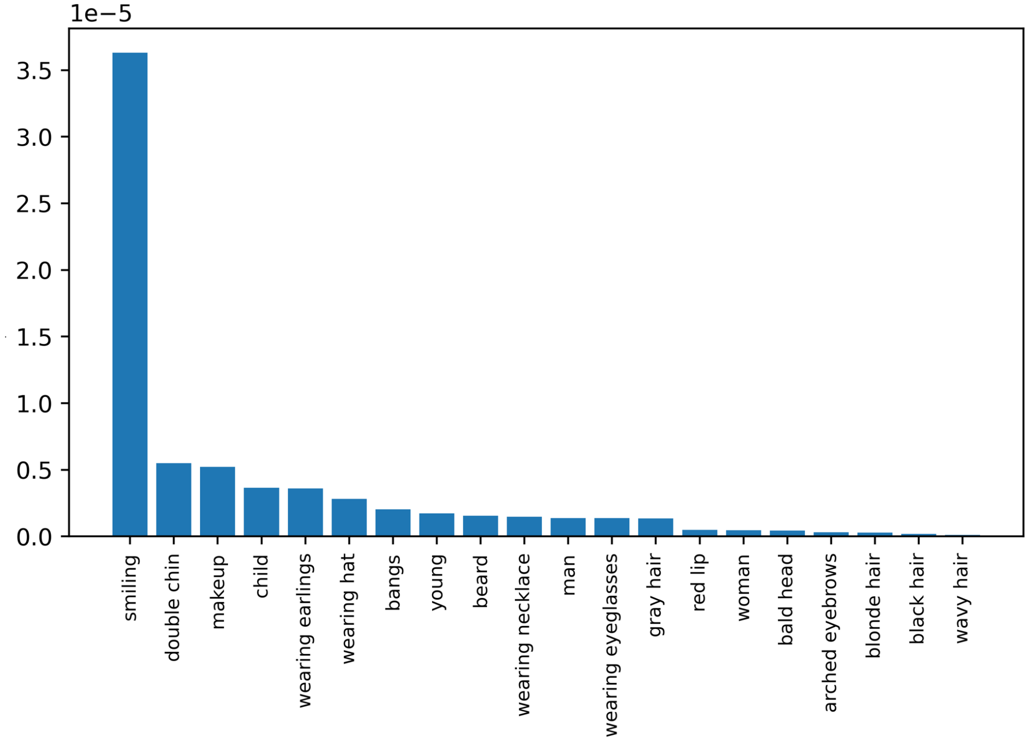


Figure S10. **SaD result for Section 5.3 (Discernment of PaD).**

in GPT's answering process. This potential problem is out of our scope. We anticipate future research will address it to extract attributes in a more fair and unbiased manner with large language models. For a smooth flow of contents, the table is placed at the end of this material.

Details of extracted attribute Table S17 describes selected attributes by each extractor. We used "A photo of {attribute}" as prompt engineering for all attributes.

D. More detailed results and analysis

In this section, we provide analysis of various generative models using our metric's explicit interpretability.

SaD Figure S13 shows the SaD results for StyleGAN 1, 2, 3, iDDPM, and LDM with two different step versions, StyleSwin, and ProjectedGAN. For LDM, DDIM sampling steps of 50 and 200 were used, and all numbers of the images are 50k.

SaD directly measures the differences in attribute distributions, indicating the challenge for models to match the density of the highest-scoring attributes to that of the training dataset. Examining the top-scoring attributes, all three StyleGAN models have similar high scores in terms of scale. However, there are slight differences, particularly in StyleGAN3, where the distribution of larger accessories such as `eyeglasses` or `earrings` differs. Exploring the training approach of alias-free modeling and its relationship with such accessories would be an interesting research direction.

In contrast, iDDPM demonstrates notable scores, with attributes `makeup` and `woman` showing scores over two times higher than GANs. Particularly, apart from these two attributes, the remaining attributes are similar to GANs, highlighting significant differences in the density of `woman` and `makeup`. Investigating how the generation process of diffusion models, which involves computing gradients for each pixel, affects attributes such as `makeup` and `woman` would be an intriguing avenue for future research.

For LDM, while FID improves with more timesteps, SaD gets worse. Specifically, the scores for `earrings`, `necklace`, and `young` significantly increase with 200-step results. Analyzing the influence of attributes as the number of steps increases,

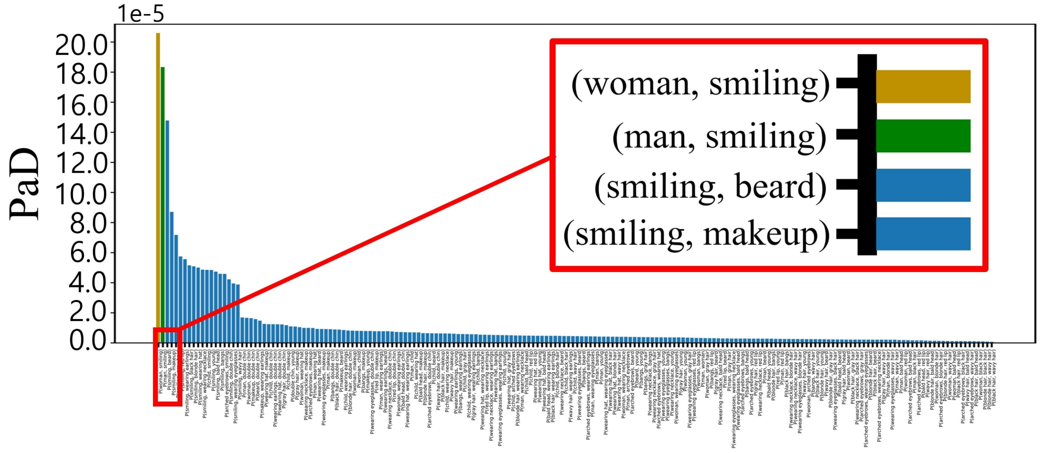


Figure S11. **PaD for Section 5.3: Discernment of PaD.**

leading to more frequent gradient updates, would be a highly interesting research direction. Moreover, diffusion models are known to generate different components at each timestep. Understanding how these model characteristics affect attributes remains an open question and presents an intriguing area for exploration.

PaD PaD provides a quantitative measure of the appropriateness of relationships between attributes. Thus, if a model generates an excessive or insufficient number of specific attributes, it affects not only SaD but also PaD. Therefore, it is natural to expect that attribute pairs with high PaD scores will often include worst-ranking attributes in SaD. Table S18 presents the worst three attributes with the highest PaD scores, and their overall values can be found in Table 3.

PaD reveals interesting findings. Firstly, it is noteworthy that attributes related to `makeup` consistently receive high scores across all StyleGAN 1, 2, and 3 models. (Table S18) This indicates that GANs generally fail to learn the relationship between `makeup` and other attributes, making it an intriguing research topic to explore the extent of this mislearning and its underlying reasons.

In the case of iDDPM, the values for `arched eyebrows` and `makeup` are overwhelmingly higher compared to other attributes. The reasons behind this will be discussed in the following subsection.

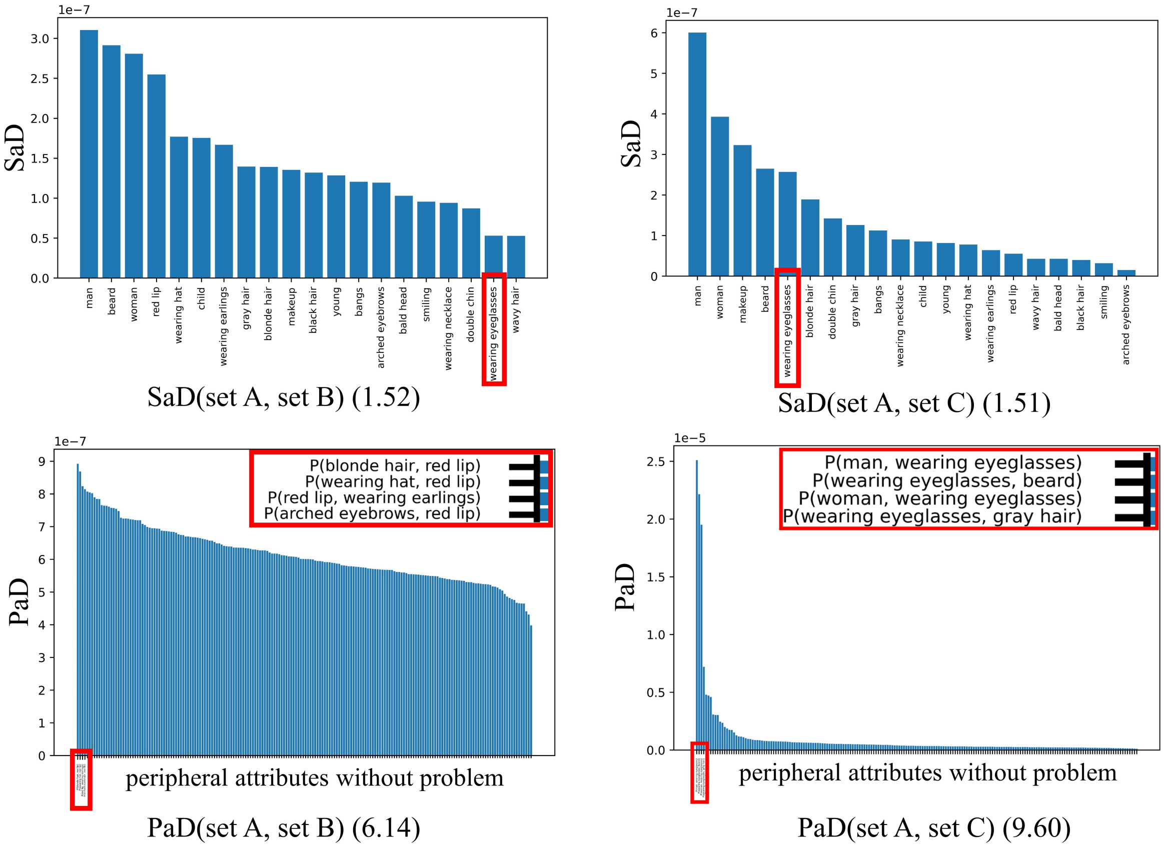
Comparing generative models with specific attribute types In the main paper, we suppose that the distribution of color-related attributes has a harmful effect on the DMs' performances compared to shape-related attributes on the proposed metric. In this section, we analyze which specific attribute DMs are hard to generate compared to StyleGAN models.

Color-related attributes Figure S15 illustrates the color-related result of SaD that iDDPM fails to preserve attributes with patterns such as `striped fur` and `dotted fur`. Considering that the color in the diffusion model is largely determined by the initial noise, we suppose that creating texture patterns such as stripes or dot patterns would be challenging. This characteristic is also observed in PaD. Unlike GANs, we can observe that relationships between solid colors without patterns or textures are not among the worst 3 attributes. (Table S19)

Shape-related attributes SaD and PaD of Shape-related attributes were relatively lower than color-related attributes. However, the attributes that have a negative impact on the scores are different in StyleGANs and iDDPM as shown in Figure S16.

Interestingly, among the attributes that DMs struggle with, the worst two attributes, `long tail` and `tufted ears`, share the commonality of being thin and long. We speculate that this is similar to the difficulty in creating `stripes`, indicating a similar characteristic.

These conjectures also explain why `arched eyebrows` in FFHQ have a high PaD score. Arched eyebrows have a thin and elongated shape that differs from the typical eyebrow appearance. Considering the characteristics of diffusion models that struggle to create stripes effectively, we can gain insights into the reasons behind this observation.



SaD(set A, set B) (1.52)

SaD(set A, set C) (1.51)

PaD(set A, set B) (6.14)

PaD(set A, set C) (9.60)

Figure S12. **Additional experiment for Section 5.3: Necessity of PaD over SaD.** In set A and set B only `men` wear `eyeglasses`, while only `women` wear `eyeglasses` in set C. PaD successfully captures pairwise relation errors between set A and set C, whereas SaD cannot.

E. Overfitting

We conducted an experiment to examine a model that performs exact copies of the training set and achieves the highest scores in SaD and PaD. We created two subsets of real images, Set A and Set B, each containing 30,000 images. Subsequently, we gradually replaced images in Set B with images from Set A, and we measured SaD, PaD and other metrics to understand the impact. This approach underscores the limitation that an overfitted model may achieve the best score, a challenge inherent to all evaluation metrics, including FID.

Specifically, the tendency due to overfitting reveals that among all metrics, FID, SaD, and PaD all exhibit R-squared values of over 99.5%, demonstrating a similar level of linearity. These findings indicate that the overfitting tendency prompted by our fine-grained attribute does not worsen.

F. N-way relationships statistical stability

Our metric is capable of examining relationships involving any number of attributes (N-way relationships), as it fundamentally relies on measuring the joint probability among involved attributes. The methodology proposed in our paper remains applicable for considering the joint probability of three or more attributes. This is particularly true for exceeding 30k images, where we did not observe statistical instability.

We provide an experiment, structured similarly to Figure 6 (metric value's variations across different seeds), to ascertain the minimum number of images that might lead to statistical instability. The proposed metric with attribute triplets demonstrates comparable standard deviation changes with an increased number of images, in comparison to SaD and PaD, indicating statistical stability. Particularly, when measuring with more than 40k images, it exhibits stability with a standard deviation of around 0.1. We conclude that extending our proposed metric to N metrics is feasible, provided that a sufficient number of

Table S16. **Scripts used for extracting attributes from GPT.** We stack GPT attributes by iteratively asking GPT to answer ‘Give me 50 words of useful, and specific adjective visual attributes for {*question*}’.

Dataset	*question*
FFHQ	‘distinguishing faces in a photo’
	‘distinguishing human faces in a photo’
	‘distinguishing different identities of people in photos of faces’
	‘differentiating between people’s faces by their distinctive features’
	‘people to change there styles in hairs, accessories around their faces’
	‘recognizing changes in hair and accessory styles in photographs of people’s faces’
	‘identifying distinct faces within an image
	‘recognizing facial characteristics to distinguish people in photos’
	‘discerning variations in facial features to identify people in images’
	‘spotting differences in facial appearance for identifying individuals’
LSUN Cat	‘ recognizing individuals from facial features in photographs’
	‘identifying distinct faces within an image
	‘recognizing variations in feline appearance to identify individual cats’
	‘discerning differences in fur patterns and colors to distinguish cats in photos’
	‘detecting subtle facial expressions to distinguish emotions in cat photos’
	‘differentiating between cats based on body type and size in photos’
	‘identifying distinctive facial features to distinguish between cats in images’
	‘recognizing changes in coat texture and length in photos of cats’
	‘discerning variations in eye color and shape to identify individual cats in images’
	‘spotting unique markings to distinguish between cats in photos’

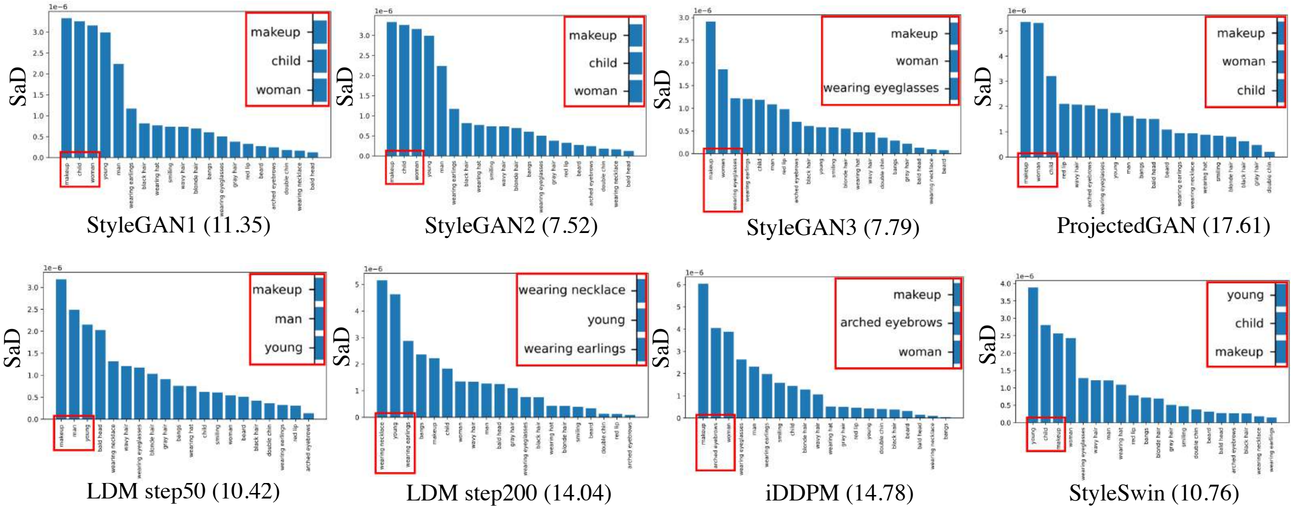
images are used.

G. Evaluating inherent biases in CLIP-like models

In this appendix, we present the detailed evaluation of inherent biases within various CLIP-like models. The Single-attribute Divergence (SaD) and Paired-attribute Divergence (PaD) metrics were used to assess the biases in ProjectedGAN with a variety of publicly available models similar to CLIP. The results indicated that certain attributes consistently showed high divergence across different models, highlighting underlying biases.

Impact of model variation on SaD and PaD. To further understand the impact of model variation, we analyzed the SaD and PaD metrics across different generative models. Despite some differences in the SaD and PaD values, the overall trend remained consistent, indicating similar biases across the models.

Consistent trends and future research directions Despite the variability in the values introduced by different CLIP-like models, certain attributes such as `makeup`, `woman`, `red lip`, `wearing necklace`, `child`, and `young` consistently showed high divergence. This indicates a uniform trend of bias across models. Addressing the variations in result values caused by changes in the encoder presents an interesting avenue for future research. Notably, StyleSwin demonstrated strong performance with certain CLIP models, suggesting potential pathways for mitigating these biases.



StyleGAN1 (11.35)

StyleGAN2 (7.52)

StyleGAN3 (7.79)

ProjectedGAN (17.61)

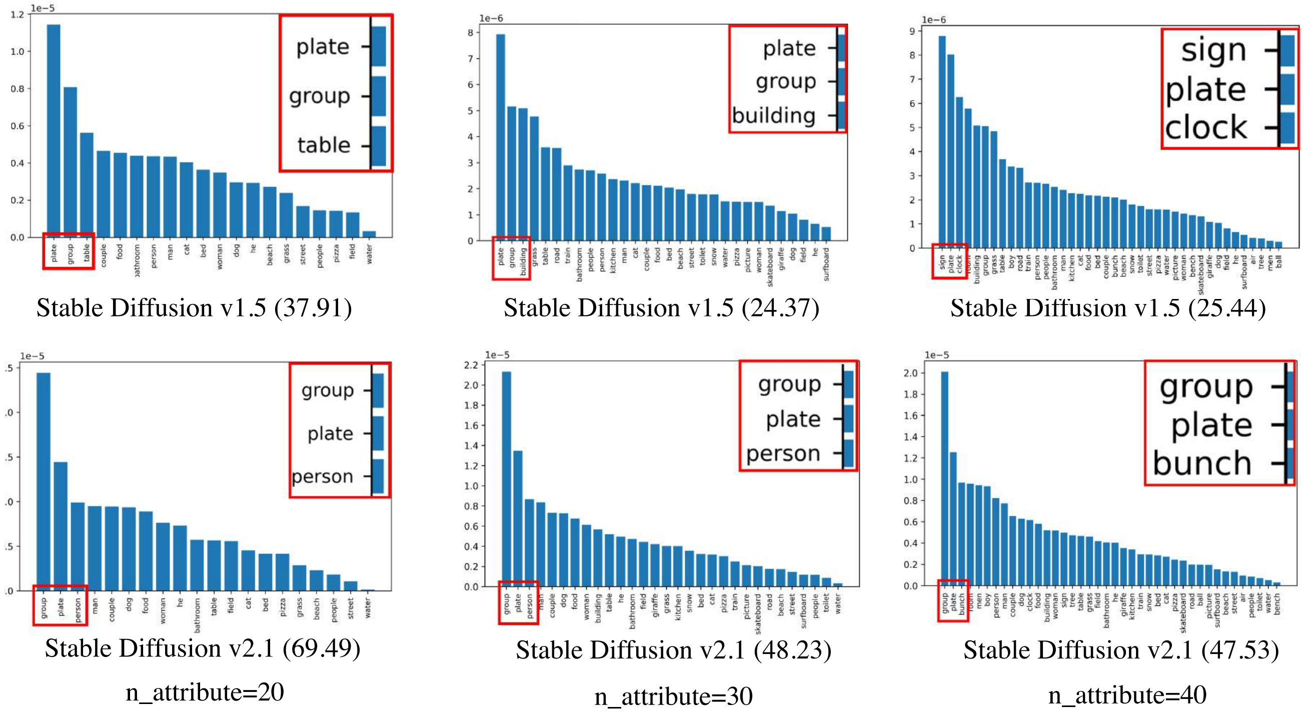
LDM step50 (10.42)

LDM step200 (14.04)

iDDPM (14.78)

StyleSwin (10.76)

Figure S13. **SaD with USER attributes on FFHQ.**



Stable Diffusion v1.5 (37.91)

Stable Diffusion v1.5 (24.37)

Stable Diffusion v1.5 (25.44)

Stable Diffusion v2.1 (69.49)

Stable Diffusion v2.1 (48.23)

Stable Diffusion v2.1 (47.53)

n_attribute=20

n_attribute=30

n_attribute=40

Figure S14. **SaD for text-to-image models.** Stable Diffusion v1.5 outperforms stable Diffusion v2.1 in SaD. regardless of the number of attributes despite being known as inferior in FID.

Table S17. **Examples of extracted attributes by each attribute extractors.**

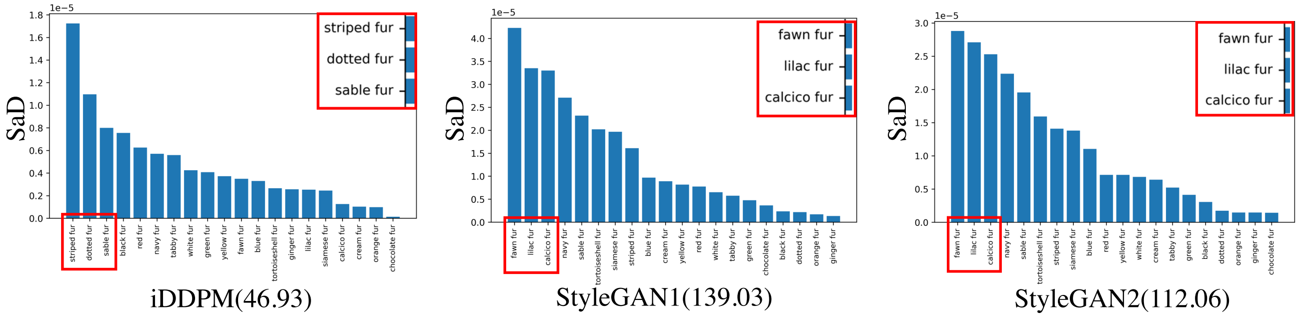
Extractor	N	Attribute
BLIP	20	woman, man, person, glasses, suit, little girl, tie, picture, sunglasses, young boy, cell phone, microphone, necklace, hat, young girl, blonde hair, long hair, blue shirt, beard, white shirt
	30	woman, man, person, glasses, suit, little girl, tie, picture, sunglasses, young boy, cell phone, microphone, necklace, hat, young girl, blonde hair, long hair, blue shirt, beard, white shirt, her head, her face, couple, baby, her hair, scarf, black shirt, smile, young man, little boy, child
	40	woman, man, person, glasses, suit, little girl, tie, picture, sunglasses, young boy, cell phone, microphone, necklace, hat, young girl, blonde hair, long hair, blue shirt, beard, white shirt, her head, her face, couple, baby, her hair, scarf, black shirt, smile, young man, little boy, child, red hair, flower, her hand, his mouth, blue eyes, women
GPT	20	clean-shaven, beard, mustache, wide-eyed, thin lips, bald, glasses-wearing, freckled, almond-shaped eyes, scarred, wrinkled, soul patch, high forehead, hooded eyes, piercings, prominent cheekbones, full lips, braided, upturned-nosed, youthful
	30	clean-shaven, beard, mustache, wide-eyed, thin lips, bald, glasses-wearing, freckled, almond-shaped eyes, scarred, wrinkled, soul patch, high forehead, hooded eyes, piercings, prominent cheekbones, full lips, braided, upturned-nosed, youthful, approachable, arched eyebrows,thin-lipped, thin-eyebrowed, birthmark, bobbed, composed, curly hair, deep-set eyes, thick-eyebrowed
	40	clean-shaven, beard, mustache, wide-eyed, thin lips, bald, glasses-wearing, freckled, almond-shaped eyes, scarred, wrinkled, soul patch, high forehead, hooded eyes, piercings, prominent cheekbones, full lips, braided, upturned-nosed, youthful, approachable, arched eyebrows,thin-lipped, thin-eyebrowed, birthmark, bobbed, composed, curly hair, deep-set eyes, thick-eyebrowed, earrings, eyebrow thickness, facial hair, goatee, heart-shaped face, long eyelashes, low forehead, monolid eyes, nasolabial folds, diamond-shaped face
CelebA GT label	20	makeup, bangs, wearing eyeglasses, wearing earrings, black hair, arched eyebrows, blonde hair, red lip, gray hair, beard, wavy hair, child, bald head, smiling, double chin, wearing hat, young, man, woman, wearing necklace

Table S18. **Top 3 PaD pair with USER attributes on FFHQ.**

	StyleGAN1	StyleGAN2	StyleGAN3	iDDPM	LDM (50)	LDM (200)	StyleSwin	ProjectedGAN
1st	man &woman	arched eyebrows &makeup	red lip &makeup	arched eyebrow &makeup	man &young	makeup &young	makeup &young	man &woman
2nd	child &makeup	child &makeup	arched eyebrow &makeup	woman &arched eyebrow	makeup &young	wearing necklace &young	woman &young	red lip &makeup
3rd	makeup &young	man &woman	child &makeup	child &makeup	child &makeup	bald head &young	wavy hair &young	child &makeup

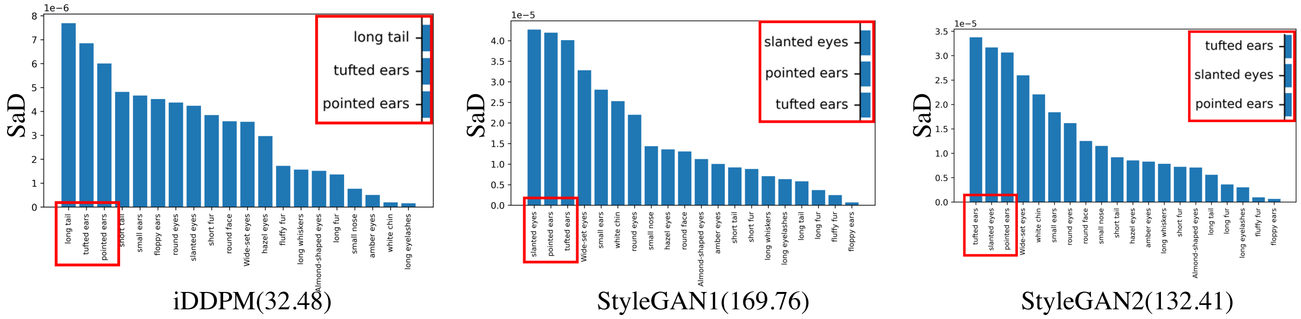
Table S19. **Worst 3 PaD pair with shape or color attributes on LSUN Cat.**

		StyleGAN1	StyleGAN2	iDDPM
color attributes	1st	fawn fur &navy fur	fawn fur &calcico fur	tabby fur &striped fur
	2nd	fawn fur &calcico fur	fawn fur &lilac fur	dotted fur &striped fur
	3rd	lilac fur &fawn fur	lilac fur &navy fur	black fur &striped fur
shape attributes	1st	tufted ears &slanted eyes	tufted ears &slanted ears	hazel eyes &long tail
	2nd	pointed ears &slanted eyes	tufted ears &white chin	Almond-shaped eyes &long tail
	3rd	slanted eyes small ears	pointed ears &white chin	long tail &wide-set eyes



iDDPM(46.93) StyleGAN1(139.03) StyleGAN2(112.06)

Figure S15. **SaD for LSUN Cat with color attributes.**



iDDPM(32.48) StyleGAN1(169.76) StyleGAN2(132.41)

Figure S16. **SaD for LSUN Cat with shape attributes.**

Table S20. **Performance of various metrics with different levels of image replacement.**

Number of replaced images	**SaD**	**PaD**	**FID**	**FID_CLIP**	**Precision**	**Recall**	**Density**	**Coverage**
100%	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
90%	0.05	0.36	0.11	0.01	0.98	0.98	0.99	0.99
80%	0.12	0.77	0.21	0.02	0.97	0.98	0.99	0.99
70%	0.19	1.19	0.32	0.03	0.96	0.96	0.99	0.99
60%	0.24	1.55	0.43	0.04	0.95	0.95	0.99	0.98
50%	0.29	1.91	0.54	0.05	0.95	0.95	0.99	0.98
40%	0.35	2.26	0.64	0.07	0.94	0.94	0.99	0.98
30%	0.42	2.67	0.74	0.08	0.92	0.93	0.99	0.97
20%	0.46	3.04	0.85	0.09	0.90	0.90	0.99	0.97
10%	0.55	3.44	0.95	0.10	0.90	0.90	0.99	0.97
0%	0.61	3.82	1.07	0.11	0.89	0.89	0.99	0.97
r^2 (%)	99.76	99.97	99.98	99.52	97.5	96.96	25.0	90.75

Table S21. **Mean and standard deviation of SaD, PaD, and proposed metric with attribute triplets.**

	10K	20K	30K	40K	50K	60K	70K
SaD	8.56 (0.35)	8.15 (0.19)	7.89 (0.19)	7.96 (0.13)	7.85 (0.16)	7.87 (0.11)	7.83 (0.12)
PaD	24.02 (0.58)	21.48 (0.38)	20.32 (0.30)	20.15 (0.20)	19.76 (0.25)	19.64 (0.20)	19.43 (0.21)
Attribute Triplets	18.67 (0.68)	16.21 (0.45)	15.29 (0.46)	14.31 (0.13)	13.93 (0.17)	13.67 (0.08)	13.44 (0.09)

Table S22. **Results for StyleGAN2 and StyleGAN3 trained on AFHQv2.**

	SaD	**PaD**
StyleGAN2	285.09	421.90
StyleGAN3	**281.18**	**413.53**

Table S23. **Comparison of different models and metrics.**

		StyleGAN1	StyleGAN2	StyleGAN3	iDDPM	LDM (200)	StyleSwin	ProjectedGAN
SigLIP	SaD	9.31	6.73	6.59	19.62	20.47	**6.41**	13.59
	PaD	23.34	17.16	**16.02**	43.63	43.62	17.26	33.61
CLIP-ViT	SaD	11.35	**7.52**	7.79	14.78	14.04	10.76	17.61
	PaD	27.25	**19.22**	19.73	34.04	30.71	26.56	41.53
CLIP-ConvNext	SaD	17.75	22.09	23.07	26.16	22.34	**9.51**	26.14
	PaD	40.22	47.79	49.81	58.45	51.50	**24.16**	61.51

Table S24. **SaD and PaD values along with the three worst-performing attributes for various CLIP-like models.** We used ProjectedGAN for this experiment.

Architecture	Model	Dataset	SaD	PaD	Worst 3 attributes
SigLIP	ViT-B-16	WebLi	13.60	33.61	red lip, child, woman
	ViT-L-16	WebLi	17.49	41.15	wearing necklace, wearing eyeglasses, woman
CLIP-ViT	ViT-B-32	OpenAI's WiT	17.61	41.53	makeup, woman, child
	ViT-B-16	LAION-2B	15.02	36.40	child, wearing necklace, wearing eyeglasses
	ViT-L-14	LAION-2B	23.55	54.34	beard, red lip, woman
	ViT-H-14	LAION-2B	17.68	43.15	red lip, gray hair, beard
	ViT-L-14	DataComp.XL	36.30	80.02	woman, child, blonde hair
	ViT-L-14	DFN-2B	16.32	40.17	child, smiling, double chin
	ViT-H-14	DFN-5B	21.24	52.03	wearing eyeglasses, woman, wearing necklace
CLIP-ConvNext	base_w	LAION-2B	13.60	33.61	bald head, wearing necklace, wearing earrings
	large_d	LAION-2B	17.49	41.15	makeup, smiling, young