
Supplementary Material for Attribute Based Interpretable Evaluation Metrics for Generative Models

Anonymous Author(s)

Affiliation

Address

email

1 A Implementation Details

2 A.1 Additional experimental setup

3 **Details of generated images** We generate samples using official checkpoints provided by
4 StyleGANs[3][5][4][6] and iDDPMs[8][1]. For LDM[9], we generate images from pesser’s un-
5 official checkpoint¹. We use 50k of training images and generated images for both FFHQ[3] and
6 LSUN Cat[10] experiment. For the Metfaces[5] experiment, we use 1,336 of training images and 50k
7 of generated images.

8 **Details of GPT queries** Table S5 provides the questions we used for preparing GPT attributes. We
9 accumulated GPT attributes by iteratively asking GPT to answer ‘Give me 50 words of useful, and
10 specific adjective visual attributes for {*question*}’. Then, we selected the top N attributes based on
11 their frequency of occurrence, ensuring that the most frequently mentioned attributes were prioritized.
12 We suppose that the extracted attributes might be biased due to the inherent randomness in GPT’s
13 answering process. This potential problem is out of our scope. We anticipate future research will
14 address it to extract attributes in an more fair and unbiased manner with large language models. For
15 smooth flow of contents, the table is placed at the end of this material.

16 **Details of extracted attribute** Table S6 describes selected attributes by each extractor. We used "A
17 photo of {attribute}" as prompt engineering for all attributes. For the sake of clarity and readability,
18 the table is presented at the end of the paper.

19 **Miscellaneous** We use `scipy.stats.gaussian_kde(dataset, 'scott', None)` to estimate the distri-
20 bution of Directional CLIPScore for given attributes. We use `spacy.load("en_core_web_sm")` to
21 extract attributes from BLIP[7] captions. We resize all images to 224x224. We used "ViT-B/32"[2]
22 as a CLIP encoder. We used a single NVIDIA RTX 3090 GPU (24GB) for the experiments.

23 A.2 Details of CelebA mean accuracy experiment

24 Table S7 shows attributes used in Table S1. We construct a refined set of attributes to ensure a more
25 objective evaluation by excluding subjective judgments such as “attractive”. For convenience and
26 organization, it is positioned at the end of this material.

¹<https://github.com/pesser/stable-diffusion>

27 B Additional Ablation Study

28 B.1 Necessity of seperating image mean and text mean

29 In the main paper, we defined Directional CLIPScore as computing angles between vectors V_x and
 30 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
 31 of captions to an attribute in CLIP space. Table S1 quantitatively validates the effectiveness of setting
 32 the origin of V_a as the center of captions (C_T) compared to the center of images (C_X).

Table S1: **Mean accuracy from CelebA ground truth labels.** Directional CLIPScore with origin at the center of images (C_X) is seriously inferior to the one with origin at the center of images (C_T). It validates the definition of V_a .

	All attributes		Refined attributes	
	From image's mean (C_X)	From text's mean (C_T)	From image's mean (C_X)	From text's mean (C_T)
mean accuracy	0.288	0.409	0.324	0.530

33 B.2 Erratum

34 StyleGAN2-ADA results on FFHQ in Table 4 and Figure 5 in the main paper were produced with
 35 the checkpoint in the official StyleGAN2-ADA repository but it was a wrong one². The corrected
 36 versions are in Table S2 and Figure S1, respectively. Our deepest apologies for the mistake. We
 37 confirm that there is no other mistakes.

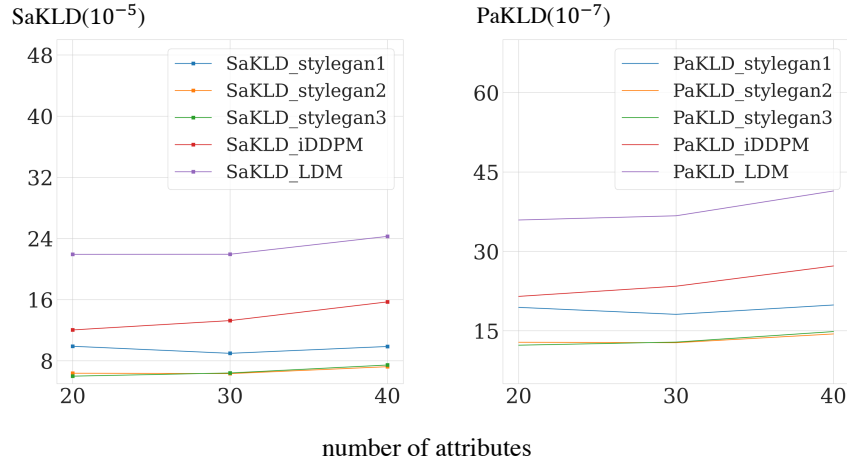


Figure S1: **The effect of the attribute counts on our metric.** The the rank of the models remained consistent regardless of the number of attributes.

38 C More detailed results and analysis

39 In this section, we provide analysis of various generative models using our metric's explicit inter-
 40 pretability. In Section C.1, we analyze the results on FFHQ, and in Section C.2, we analyze the results
 41 on LSUN Cat regarding color and shape. We summarize the key results of PaKLD in a table. The
 42 complete results, including each score from PaKLD, can be found in the appendix at the end.

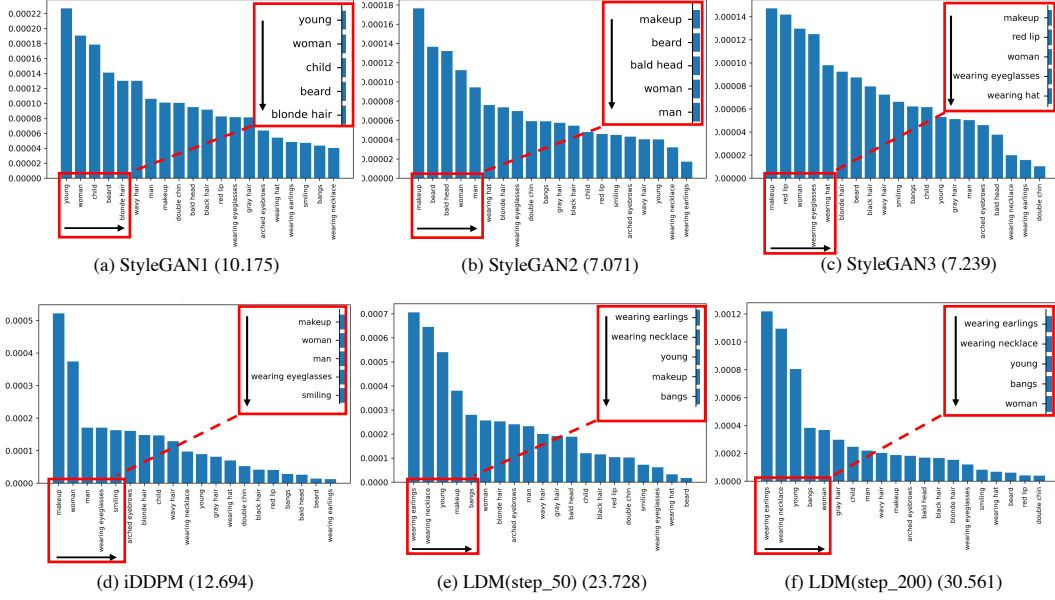
43 C.1 Challenging attributes for generative models in FFHQ with USER attributes

44 In this subsection, we provide analysis of SaKLD and PaKLD scores measured using USER attributes
 45 on FFHQ. USER attributes were used for ease of interpretation, and the trends were consistent for
 46 both BLIP and GPT. The results for both cases can be seen in Figur S10, S11, S12 and S13.

²It was the baseline StyleGAN2 trained on FFHQ 1024². There is no pretrained checkpoint of StyleGAN2-ADA on FFHQ in the official repository.

Table S2: Results with BLIP attribute on FFHQ

	FFHQ				LSUN Cat				Metfaces			
	SaKLD (10^{-5})	PaKLD (10^{-7})	FID	FID_CLIP	SaKLD (10^{-5})	PaKLD (10^{-7})	FID	FID_CLIP	SaKLD (10^{-5})	PaKLD (10^{-7})	FID	FID_CLIP
StyleGAN1	9.902	19.431	4.744	1.869	74.626	119.456	10.670	6.113	-	-	-	-
StyleGAN2	6.377	12.838	3.176	1.472	63.601	100.896	9.406	5.083	40.769	87.118	24.210	2.409
StyleGAN2 ada	-	-	-	-	-	-	-	-	31.140	58.065	24.100	2.202
StyleGAN3	5.993	12.285	3.205	1.661	-	-	-	-	-	-	-	-
iDDPM	-	-	-	-	110.229	136.579	7.590	5.789	-	-	-	-
iDDPM(P2)	12.040	21.507	7.317	2.394	-	-	-	-	129.627	230.720	37.426	6.412
LDM(step 50)	16.580	31.057	16.381	4.650	-	-	-	-	-	-	-	-
LDM(step 200)	21.920	35.940	11.869	3.570	-	-	-	-	-	-	-	-

Figure S2: SaKLD results with USER attributes on FFHQ dataset The value beside model name denotes the SaKLD value(10^{-5}) for each respective model. Please zoom in for the best view.

47 **SaKLD** Figure S2 shows the SaKLD results for StyleGAN 1, 2, 3, iDDPM, and LDM with two
 48 different step versions. For LDM, DDIM sampling steps of 50 and 200 were used, and all numbers of
 49 the images are 5k.

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

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

62 For LDM, the top-scoring attributes are similar for both 50-step and 200-step results. However, it
 63 is observed that while FID improves with 200 steps, SaKLD gets worse. Specifically, the scores
 64 for "earrings," "necklace," and "young" significantly increase with 200-step results. Analyzing the
 65 influence of attributes as the number of steps increases, leading to more frequent gradient updates,
 66 would be a highly interesting research direction. Moreover, diffusion models are known to generate
 67 different components at each timestep. Understanding how these model characteristics affect attributes
 68 remains an open question and presents an intriguing area for exploration.

Table S3: Top 3 PaKLD pair with USER attributes on FFHQ

	StyleGAN1	StyleGAN2	StyleGAN3	iDDPM	LDM(step 50)	LDM(step 200)
1st	man&woman	arched eyebrows &makeup	red lip&makeup	arched eyebrows &makeup	bald head &wearing earlings	bald head &wearing earlings
2nd	man&beard	man&beard	man&woman	red lip&makeup	man&young	wearing necklace &bald head
3rd	beard&young	red lip&makeup	man&beard	wearing eyeglasses &makeup	bald head &young	bald head &young

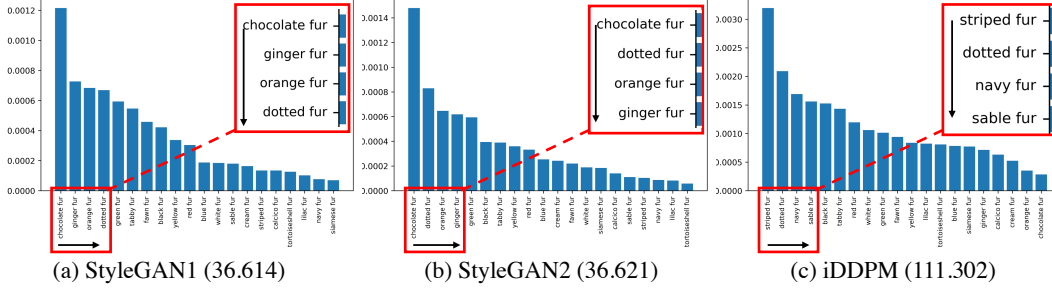


Figure S3: **SaKLD with color attributes on LSUN Cat.** The value beside model name denotes the SaKLD value (10^{-5}) for each respective model. Please zoom in for the best view.

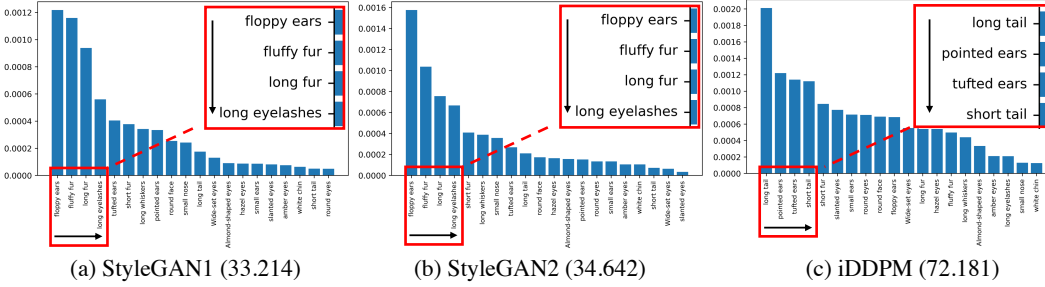


Figure S4: **SaKLD results with shape attributes on LSUN Cat.** The value beside model name denotes the SaKLD value (10^{-5}) for each respective model. Please zoom in for the best view.

69 **PaKLD** Table S3 presents the top three attributes with the highest PaKLD scores, and their
70 individual scores and overall values can be found in Figure S5.

71 PaKLD provides a quantitative measure of the appropriateness of relationships between attributes.
72 Thus, if a model generates an excessive or insufficient number of specific attributes, it affects not
73 only SaKLD but also PaKLD. Therefore, it is natural to expect that attribute pairs with high PaKLD
74 scores will often include top-ranking attributes in SaKLD.

75 Nevertheless, PaKLD reveals interesting findings. Firstly, it is noteworthy that attributes related to
76 ‘beard’ consistently receive high scores across all StyleGAN 1, 2, and 3 models. Figure S5 confirms
77 that ‘beard’ significantly contributes to the overall PaKLD scores. This indicates that GANs generally
78 fail to learn the relationship between beards and other attributes, making it an intriguing research
79 topic to explore the extent of this mislearning and its underlying reasons.

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

82 LDM also exhibits interesting attributes, particularly ‘bald head’. Despite not having a high score in
83 SaKLD, it consistently receives high scores in PaKLD. This implies that individuals with bald heads
84 have a scarcity of many attributes that are commonly found, and analyzing this phenomenon would
85 be a promising avenue for future research.

Table S4: Top 3 PaKLD pair with shape/color attributes on LSUN Cat

		StyleGAN1	StyleGAN2	iDDPM
color attributes	1st	fawn fur	fawn fur	chocolate fur
		&chocolate fur	&chocolate fur	&striped fur
	2nd	chocolate fur	chocolate fur	tabby fur
		&sable fur	&sable fur	&striped fur
	3rd	lilac fur	lilac fur	orange fur
		&chocolate fur	&chocolate fur	&striped fur
shape attributes	1st	small ears	tufted ears	point ears
		&floppy ears	&floppy ears	&long tail
	2nd	tufted ears	small ears	slanted eyes
		&floppy ears	&floppy ears	&long tail
	3rd	pointed ears	pointed ears	long tail
		floppy ears	&floppy ears	&hazel eyes

86 C.2 Comparing generative models with specific attribute types

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

90 **Color-related attributes** Figure S3 illustrates the color-related result of SaKLD that iDDPM fails
91 to preserve attributes with patterns such as striped fur and dotted fur. Considering that the color in the
92 diffusion model is largely determined by the initial noise, we suppose that creating texture patterns
93 such as stripes or dot patterns would be challenging. This characteristic is also observed in PaKLD.
94 Unlike GANs, we can observe that relationships between solid colors without patterns or textures are
95 not among the top 3 attributes.(Table S4) To provide insights into the relationship scores between
96 various attributes, we have included the PaKLD results as a separate attachment at the end of the
97 paper. (Figure S6 and Figure S7)

98 **Shape-related attributes** Unlike color-related attributes, the SaKLD scores for shape-related
99 attributes show similarities to StyleGANs. However, the attributes that have a negative impact on
100 the scores are different as shown in Figure S4. Interestingly, among the attributes that DMs struggle
101 with, the top two attributes, 'long tail' and 'pointed ears', share the commonality of being thin and
102 long.(Table S4) We speculate that this is similar to the difficulty in creating stripes, indicating a
103 similar characteristic. A similar tendency is observed in PaKLD results. (Figure S8 and Figure S9)

104 These conjectures also explain why 'arched eyebrows' in FFHQ has a high PaKLD score. Arched
105 eyebrows have a thin and elongated shape that differs from the typical eyebrow appearance. Consid-
106 ering the characteristics of diffusion models that struggle to create stripes effectively, we can gain
107 insights into the reasons behind this observation.

Table S5: **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	<i>question</i>
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’
LSUN Cat	‘spotting differences in facial appearance for identifying individuals’
	‘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’
Metfaces	‘discerning variations in eye color and shape to identify individual cats in images’
	‘spotting unique markings to distinguish between cats in photos’
	‘distinguishing visual attributes for identifying individuals in painted portraits’
	‘unique facial features for differentiating people in portrait art’
	‘notable facial characteristics in painted portraiture for individual identification’
	‘recognizable facial markers for distinguishing people in portrait art’
	‘peculiar facial distinctions in painted portraiture for person identification’
	‘distinctive facial qualities in portraits for recognizing individuals’
	‘identifying facial properties in portrait art for distinguishing people’
	‘divergent facial characteristics in painted portraiture for person differentiation’
	‘unmistakable facial features in portrait art for individual recognition’
	‘discriminating facial traits in painted portraiture for distinguishing persons’

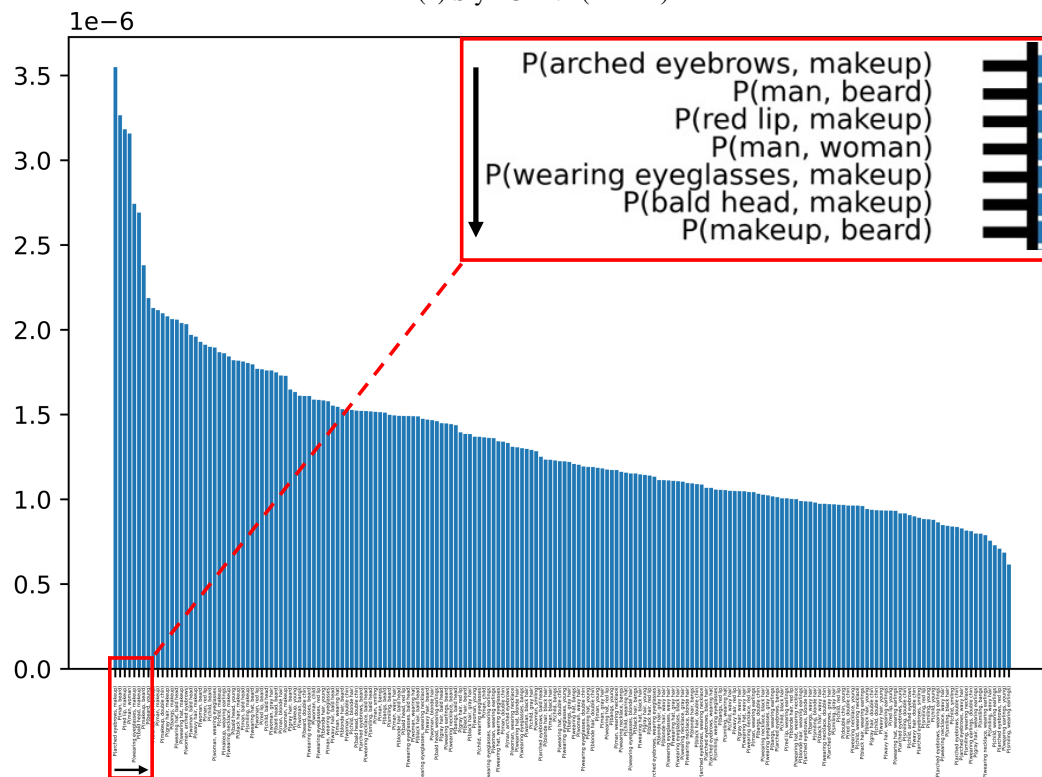
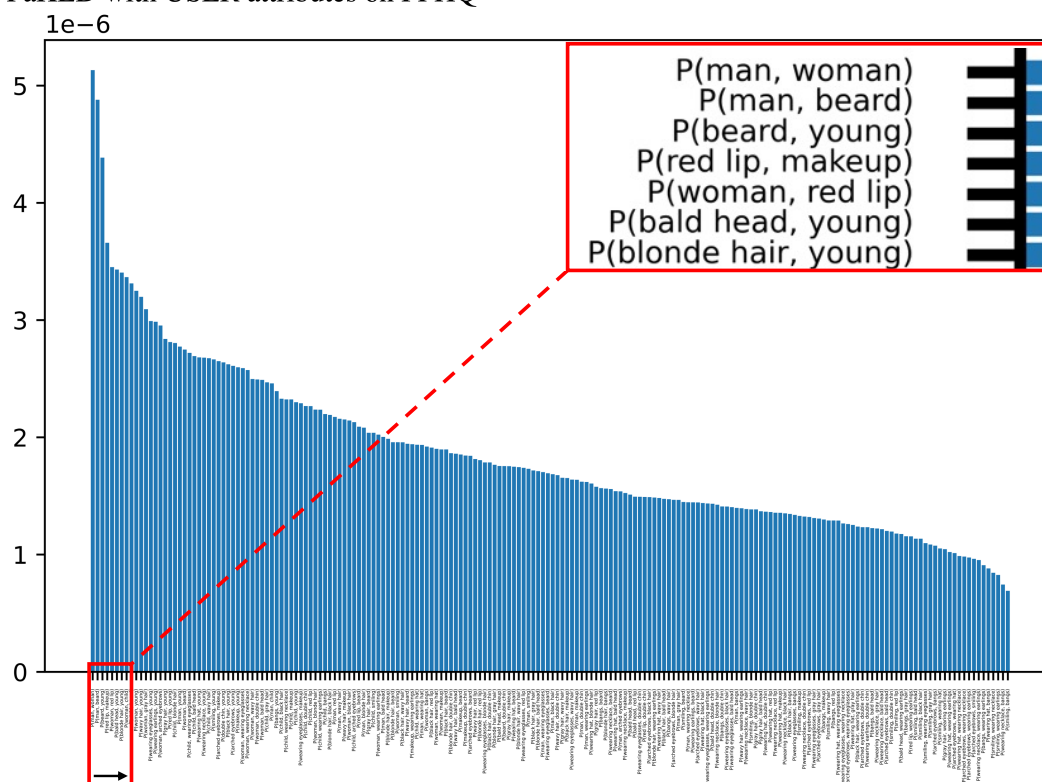
Table S6: Examples of attributes for each attribute extractor.

Extractor	N	Attribute
BLIP	20	'a woman', 'a man', 'a person', 'glasses', 'a suit', 'a little girl', 'a young boy', 'a cell phone', 'a microphone', 'a necklace', 'a hat', 'tie', 'a young girl', 'blonde hair', 'long hair', 'a blue shirt', 'a beard', 'a white shirt', 'her head', 'a tie'
	30	'a woman', 'a man', 'a person', 'glasses', 'a suit', 'a little girl', 'a young boy', 'a cell phone', 'a microphone', 'a necklace', 'a hat', 'tie', 'a young girl', 'blonde hair', 'long hair', 'a blue shirt', 'a beard', 'a white shirt', 'her head', 'a tie', 'her face', 'a couple', 'a baby', 'her hair', 'a black shirt', 'a smile', 'a young man', 'a toothbrush', 'his face', 'a red shirt'
	40	'a woman', 'a man', 'a person', 'glasses', 'a suit', 'a little girl', 'a young boy', 'a cell phone', 'a microphone', 'a necklace', 'a hat', 'tie', 'a young girl', 'blonde hair', 'long hair', 'a blue shirt', 'a beard', 'a white shirt', 'her head', 'a tie', 'her face', 'a couple', 'a baby', 'her hair', 'a black shirt', 'a smile', 'a young man', 'a toothbrush', 'his face', 'a red shirt', 'a scarf', 'a little boy', 'a child', 'red hair', 'a flower', 'her hand', 'his mouth', 'blue eyes', 'women', 'her mouth'
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'
USER	20	'makeup', 'bangs', 'wearing eyeglasses', 'wearing earlings', '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'

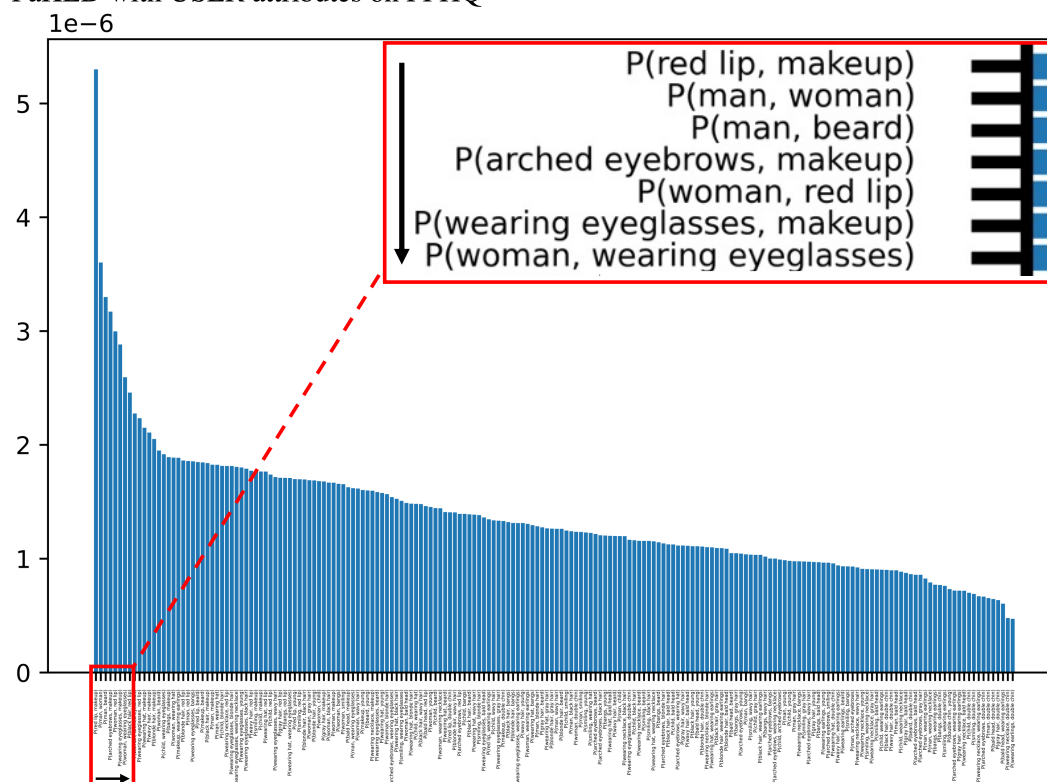
Table S7: Attributes used for CelebA accuracy experiment

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'

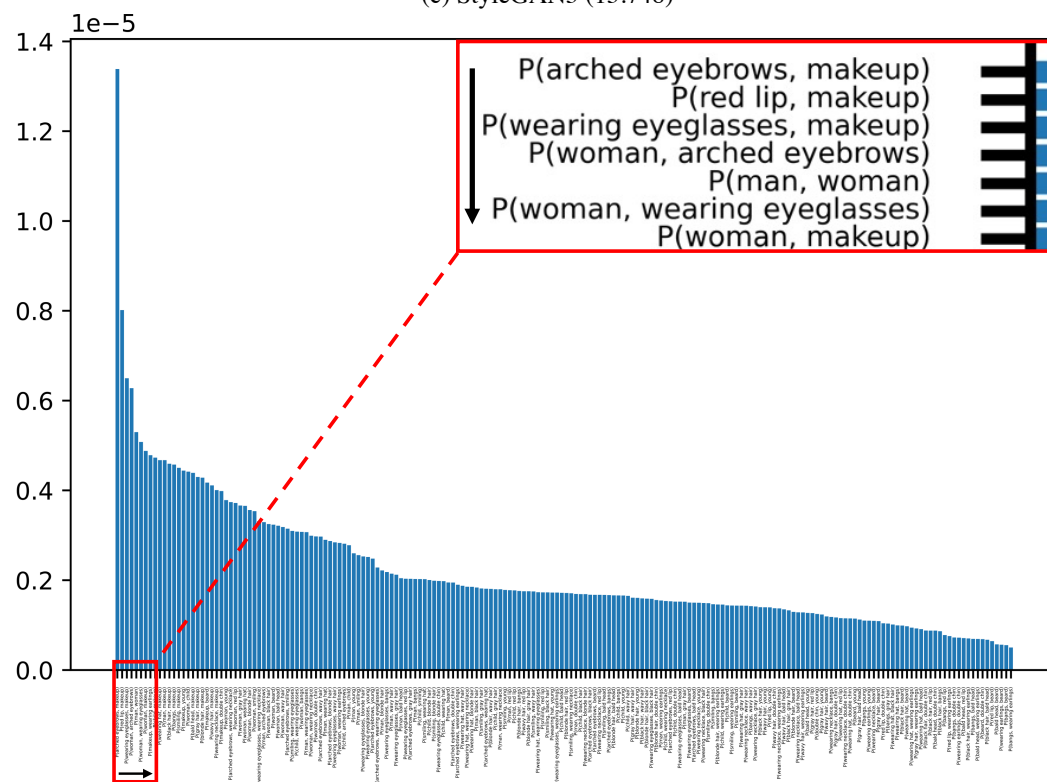
PaKLD with USER attributes on FFHQ



PaKLD with USER attributes on FFHQ



(c) StyleGAN3 (13.748)



(d) iDDPM (21.787)

PaKLD with USER attributes on FFHQ

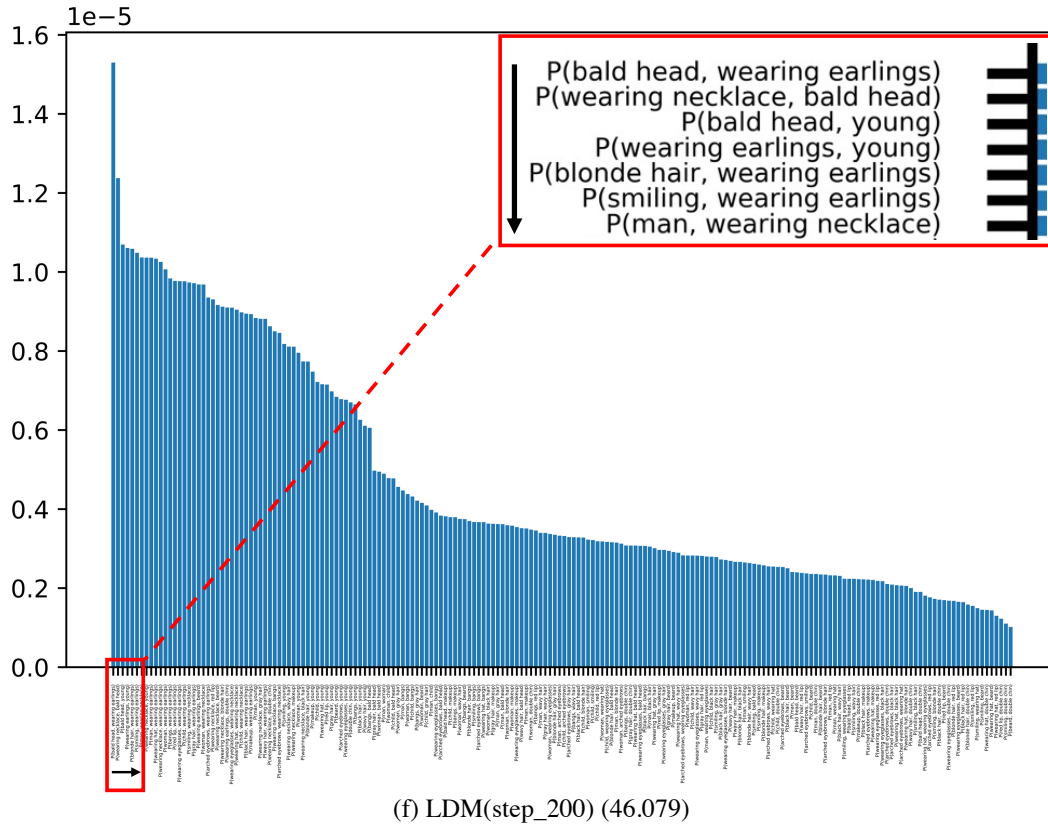
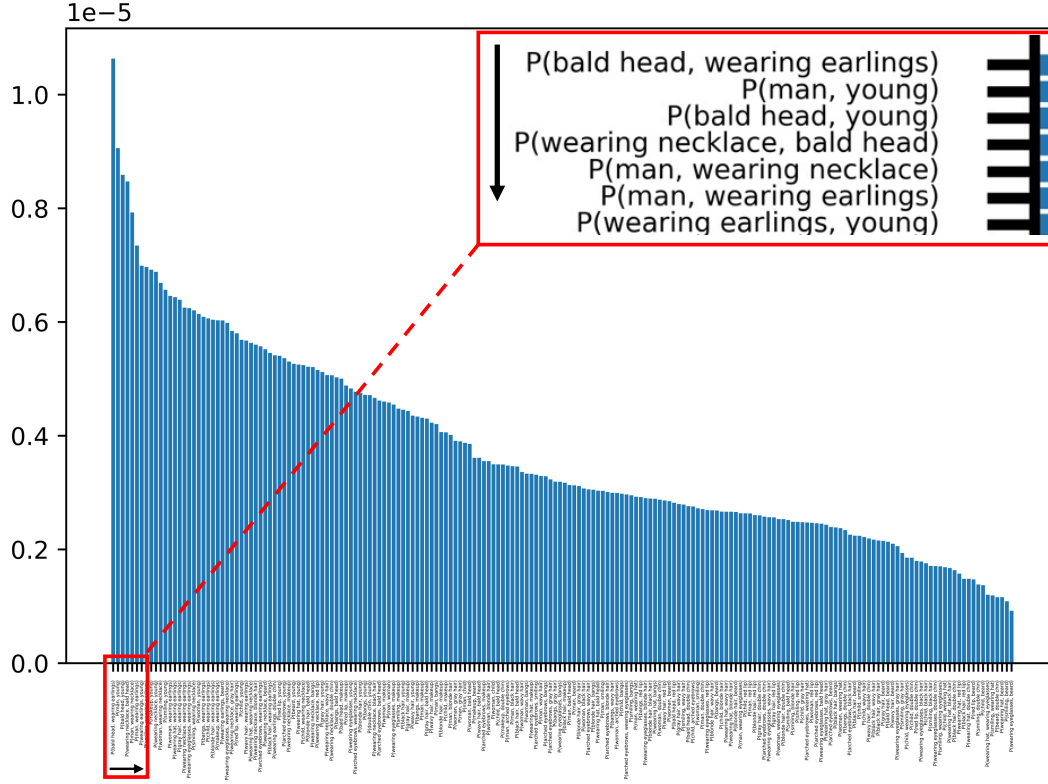
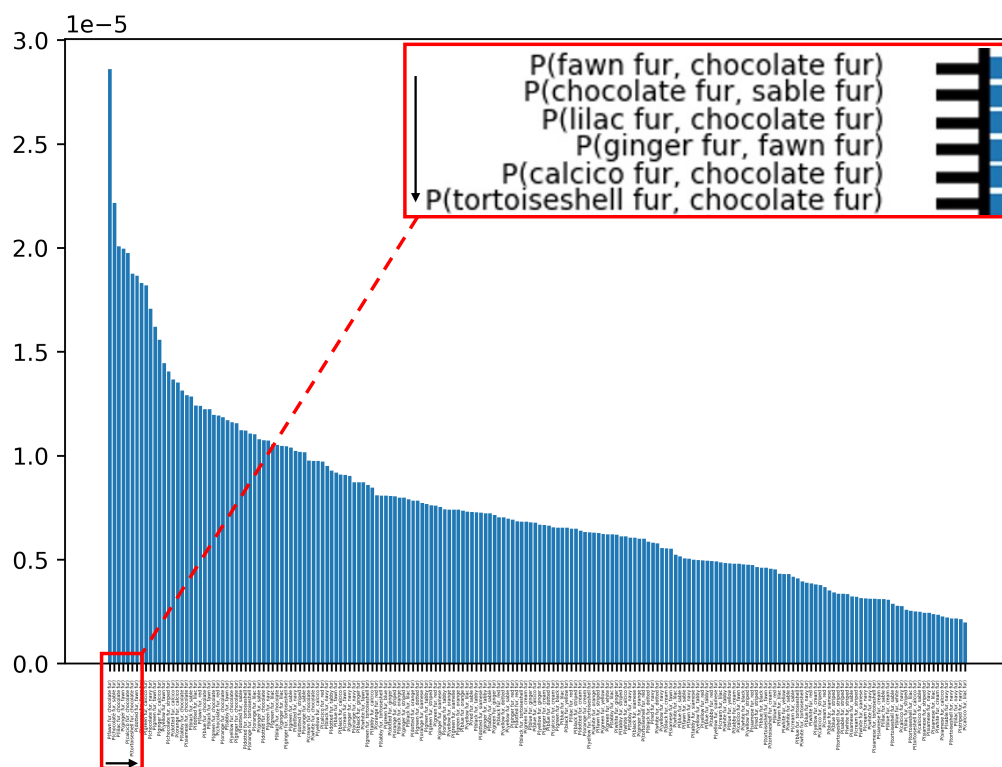
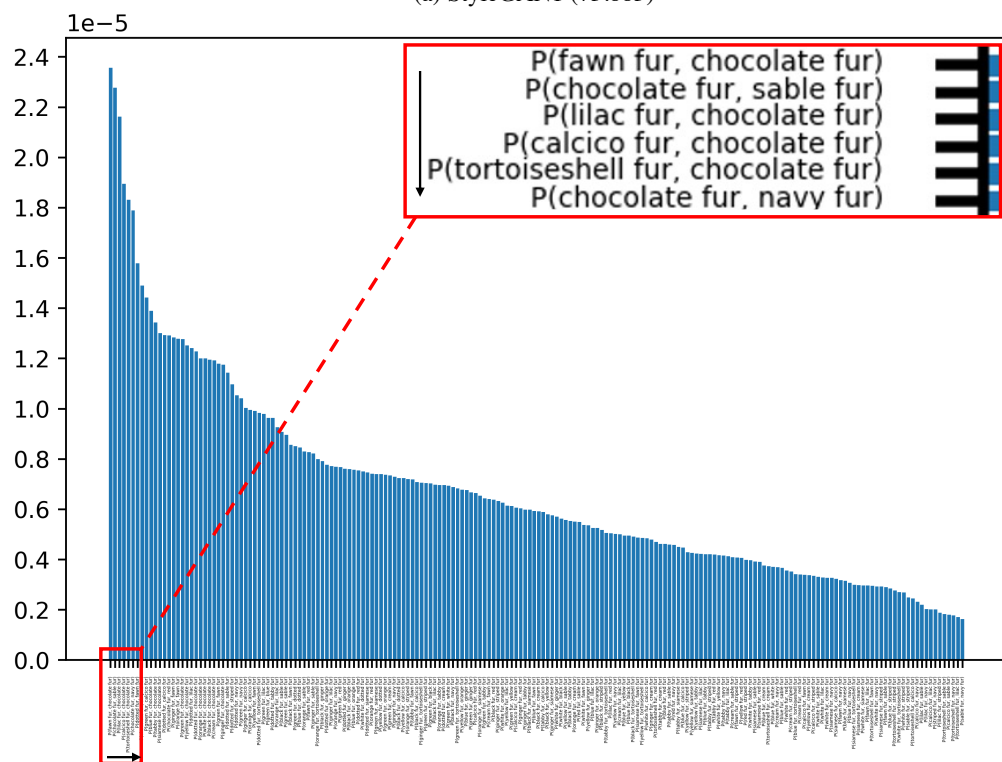


Figure S5: **PaKLD results with USER attributes on FFHQ.** The value beside model name denotes the PaKLD value(10^{-7}) for each respective model. Please zoom in for the best view.



(a) StyleGAN1 (75.885)



(b) StyleGAN1 (67.518)

Figure S6: **PaKLD results with color attributes on LSUN Cat in StyleGANs.** The value beside model name denotes the PaKLD value(10^{-7}) for each respective model.

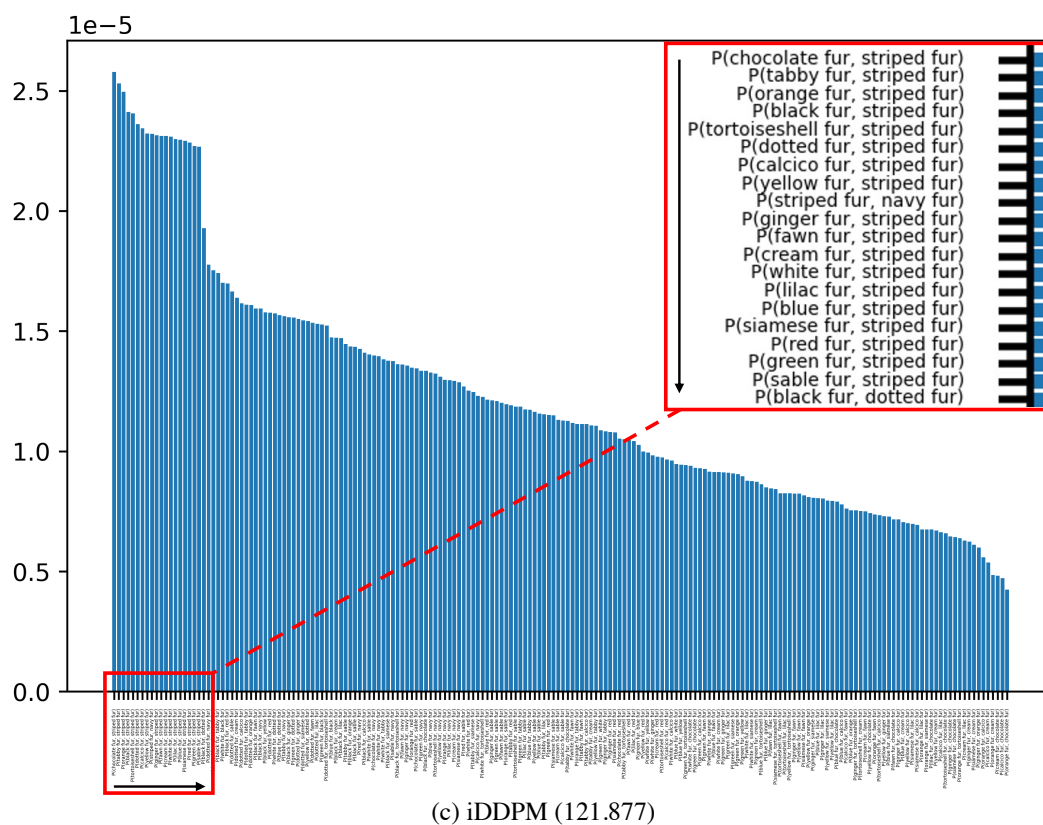


Figure S7: **PaKLD results with color attributes on LSUN Cat.** The value beside model name denotes the PaKLD value(10^{-7}) for each respective model. Please zoom in for the best view.

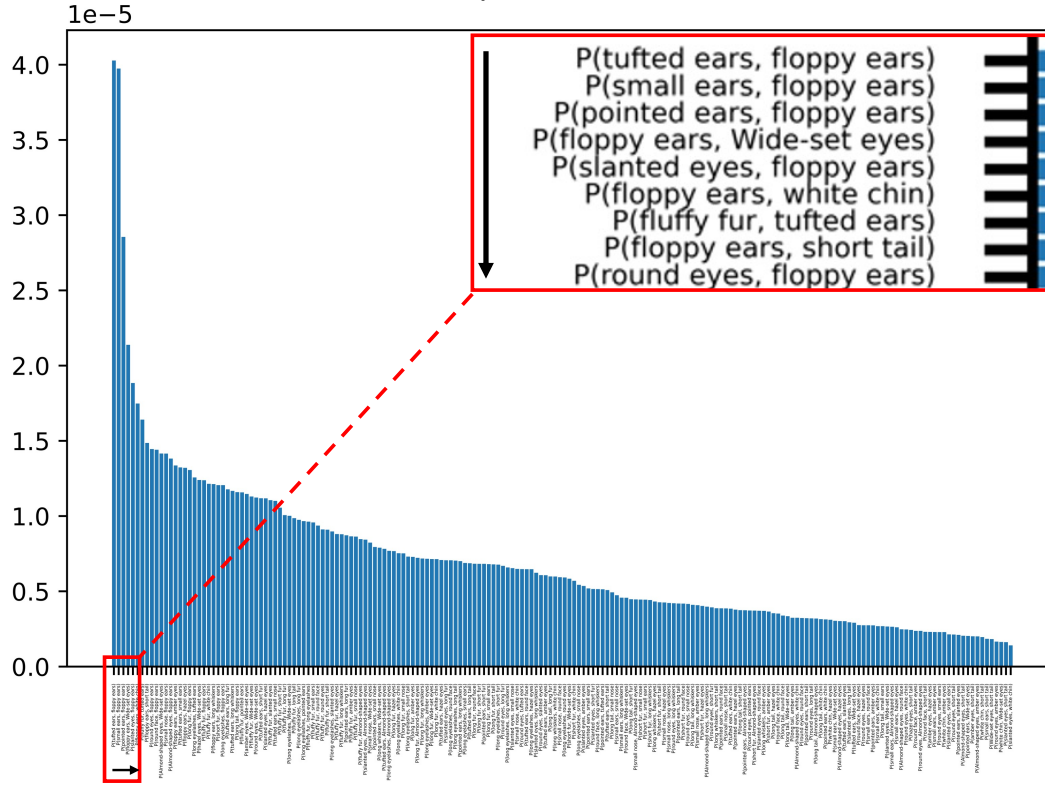
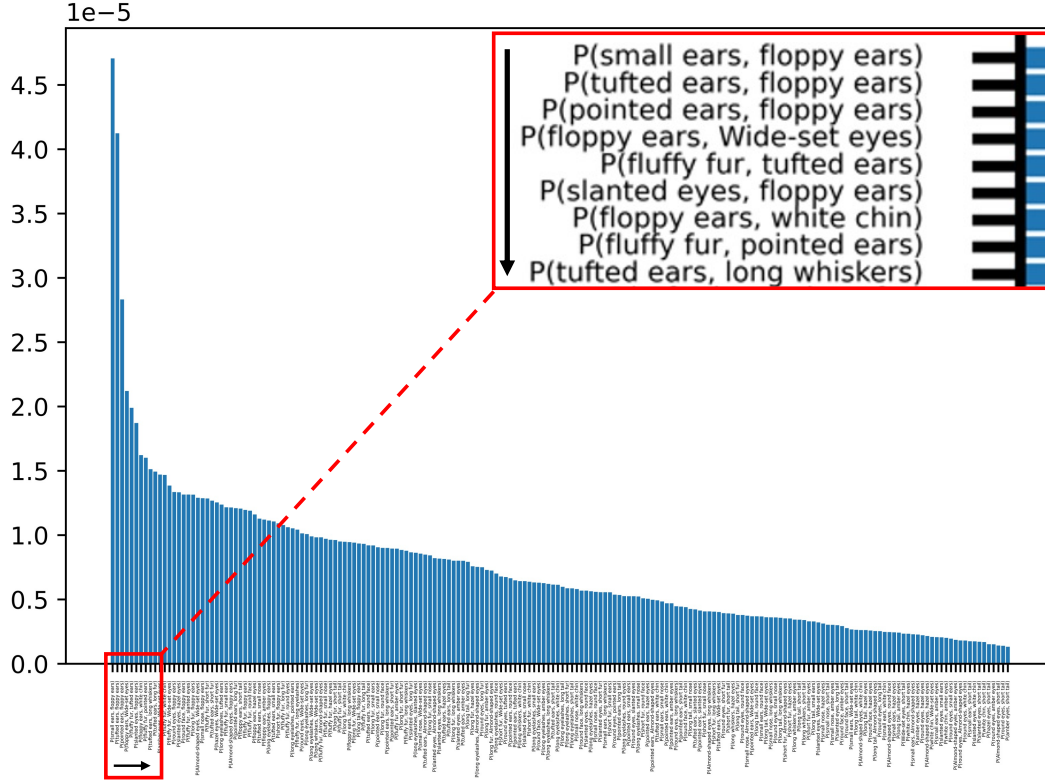


Figure S8: **PaKLD results with shape attributes on LSUN Cat in StyleGAN.** The value beside model name denotes the PaKLD value(10^{-7}) for each respective model. Please zoom in for the best view.

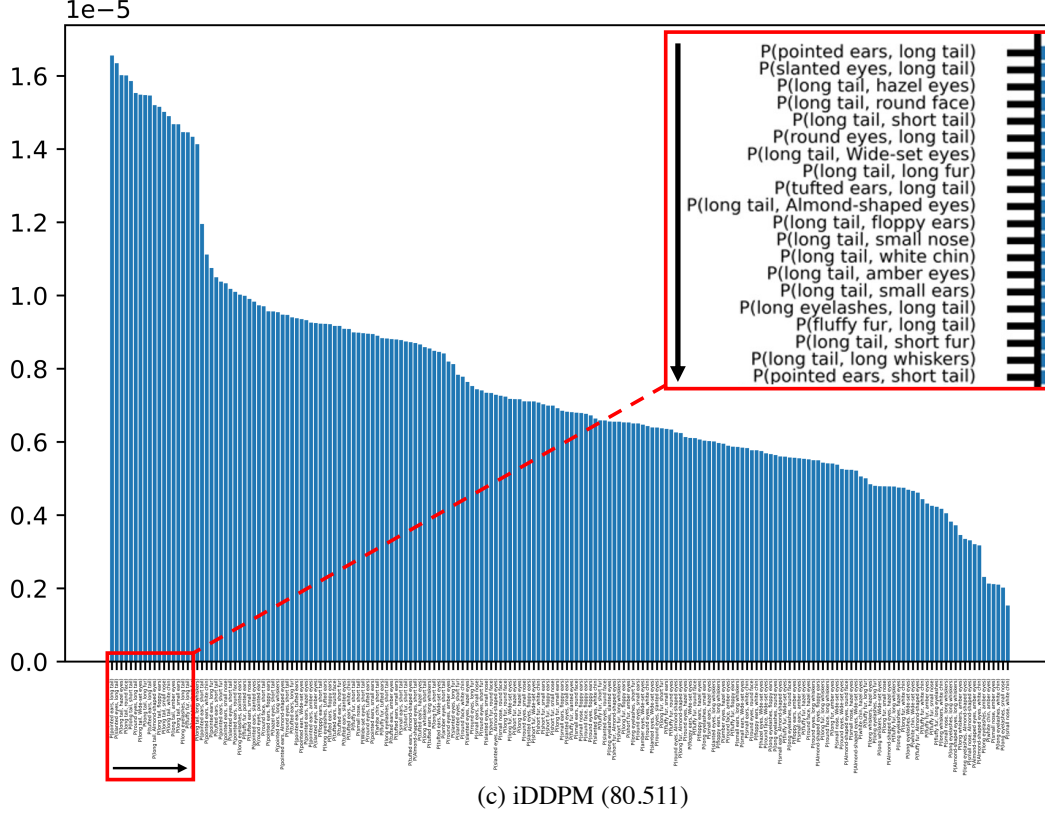


Figure S9: **PaKLD results with shape attributes on LSUN Cat.** While the PaKLD shape of (a) and (b) appears sharp due to the specific attribute "floppy ears" combined with other attributes, the PaKLD shape of (c) is more rounded despite "long tail" being a primary factor affecting PaKLD. We analyze StyleGAN models generally capture attribute relationships more accurately than the DMs across the majority of attributes. The value beside model name denotes the PaKLDvalue(10^{-7}) for each respective model. Please zoom in for the best view.

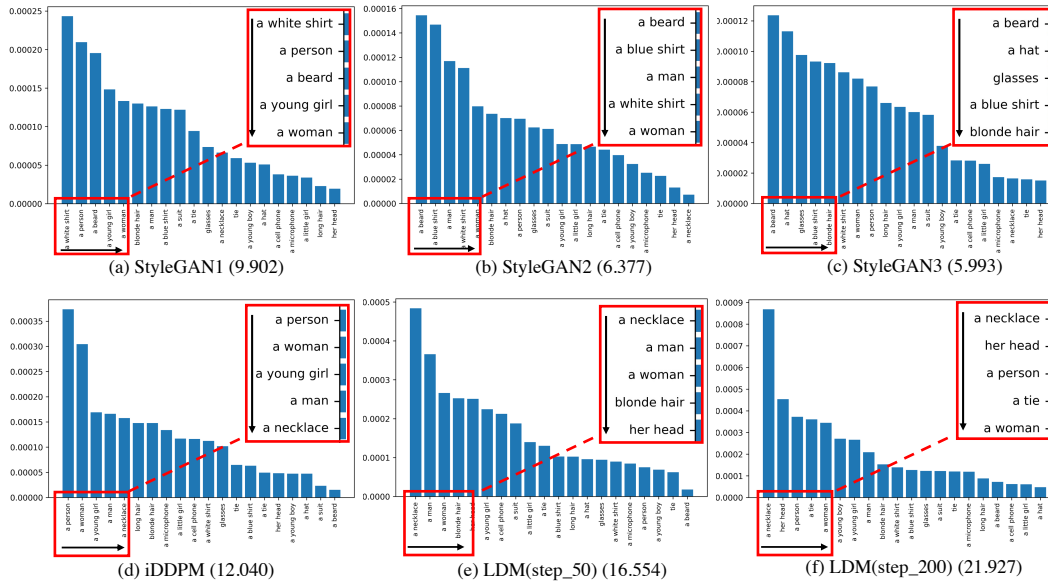
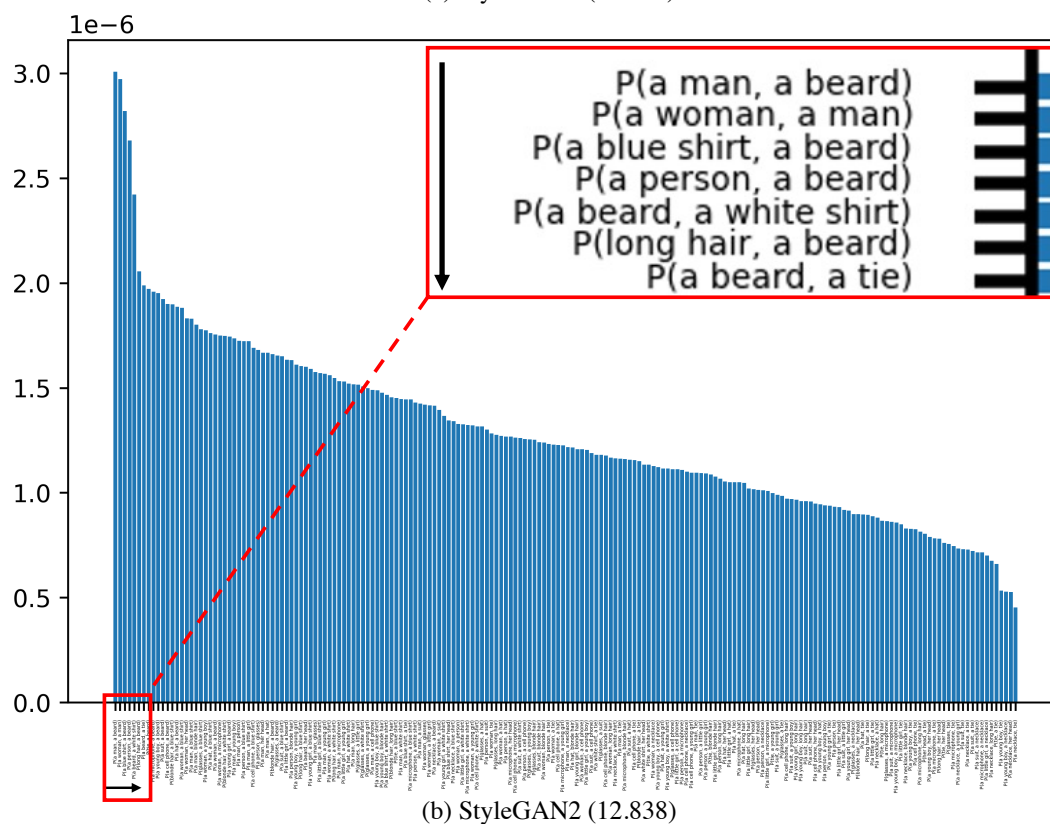
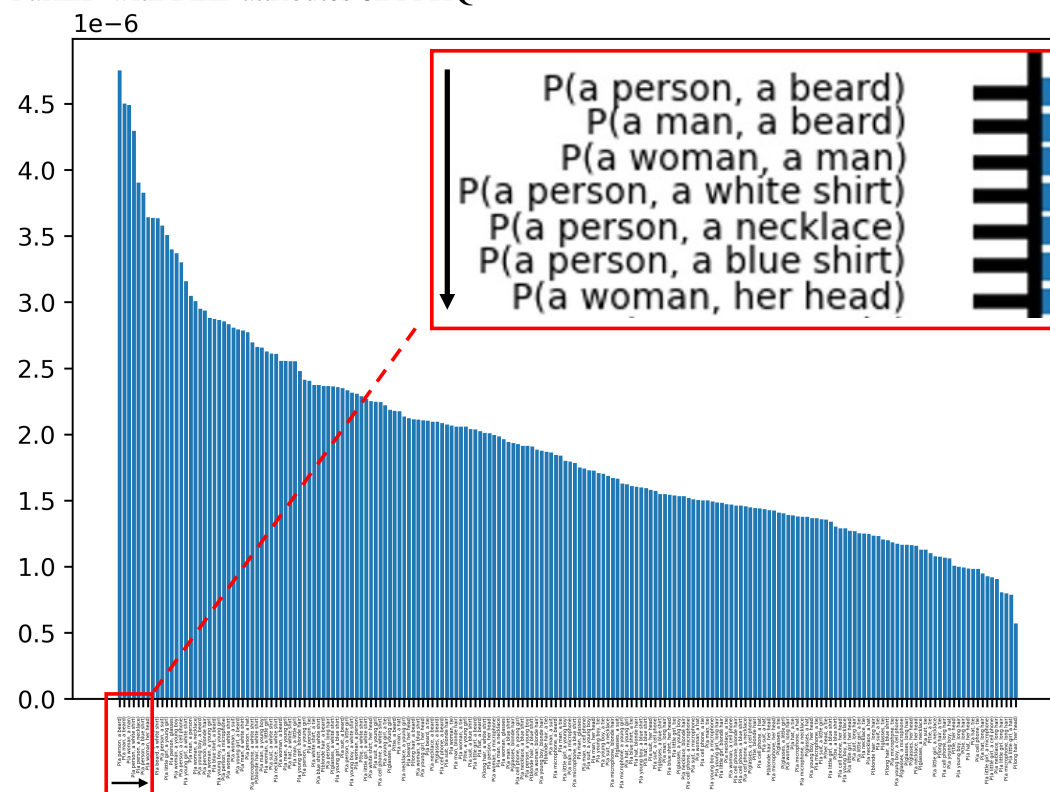
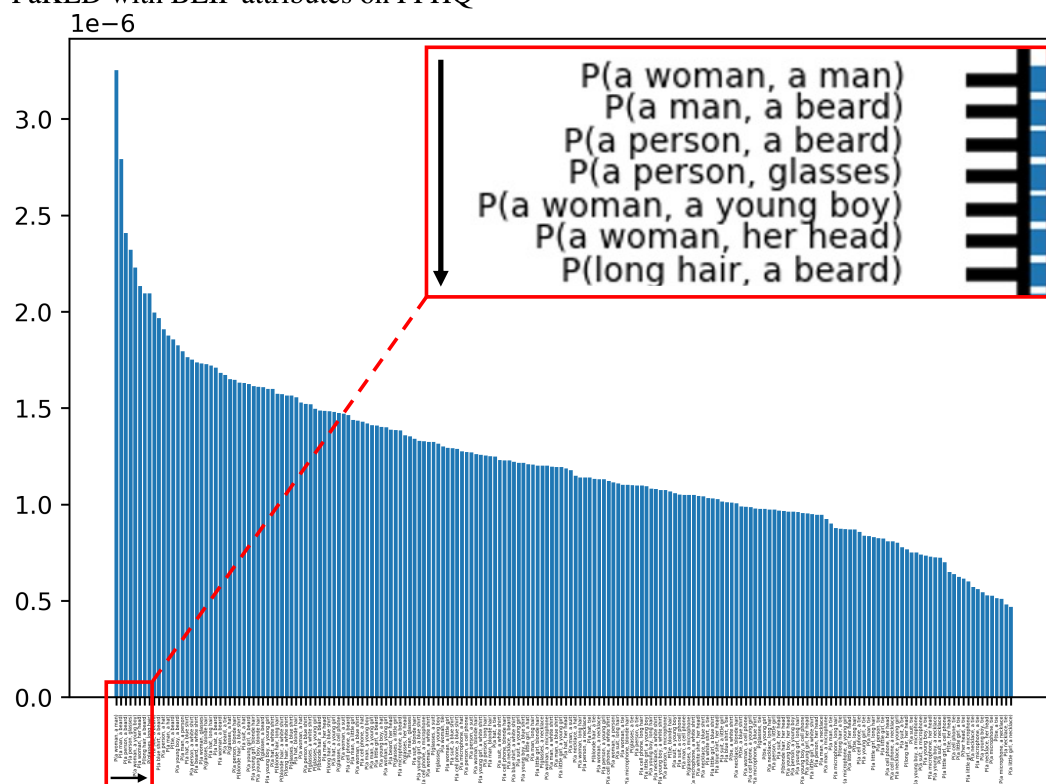


Figure S10: **SaKLD results with BLIP attributes on FFHQ dataset.** The value beside model name denotes the SaKLD value (10^{-5}) for each respective model. Please zoom in for the best view.

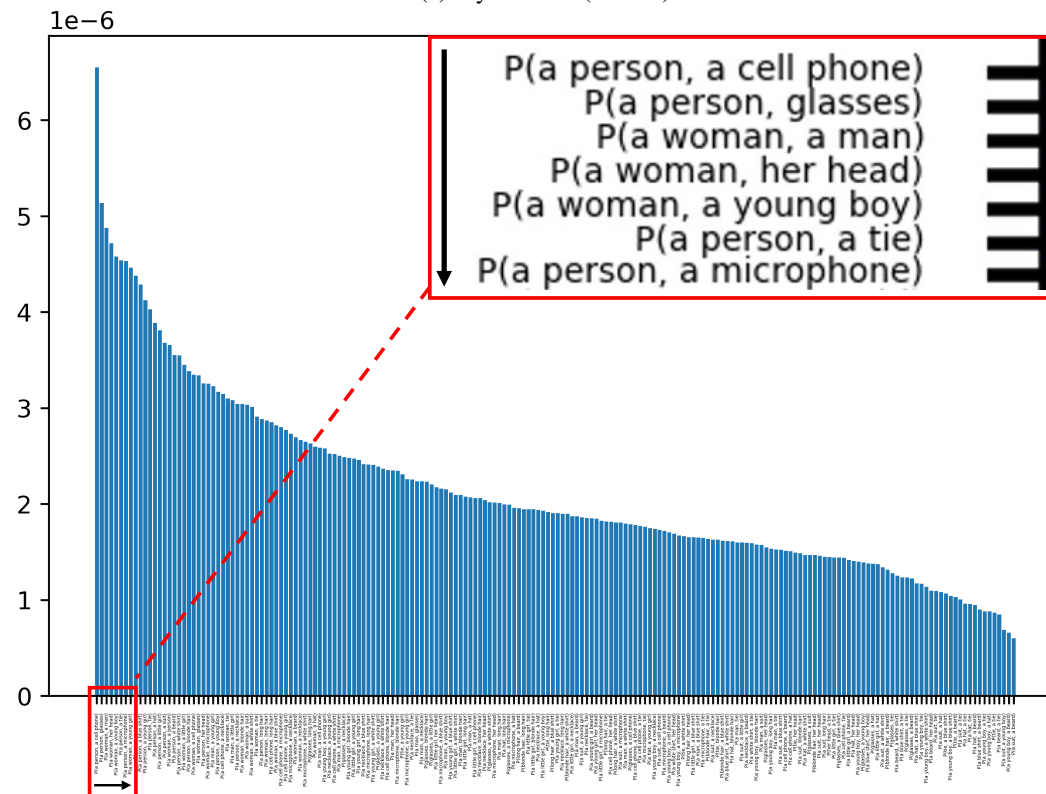
PaKLD with BLIP attributes on FFHQ



PaKLD with BLIP attributes on FFHQ



(c) StyleGAN3 (12.285)



(d) iDDPM (21.507)

PaKLD with BLIP attributes on FFHQ

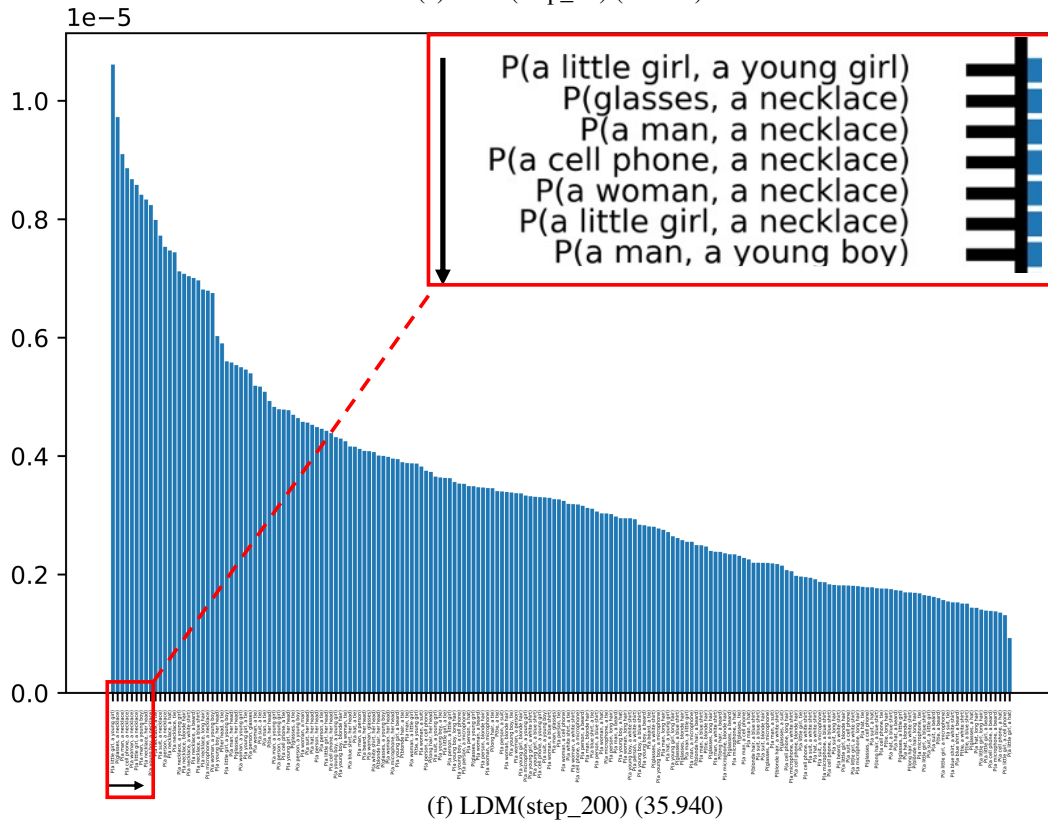
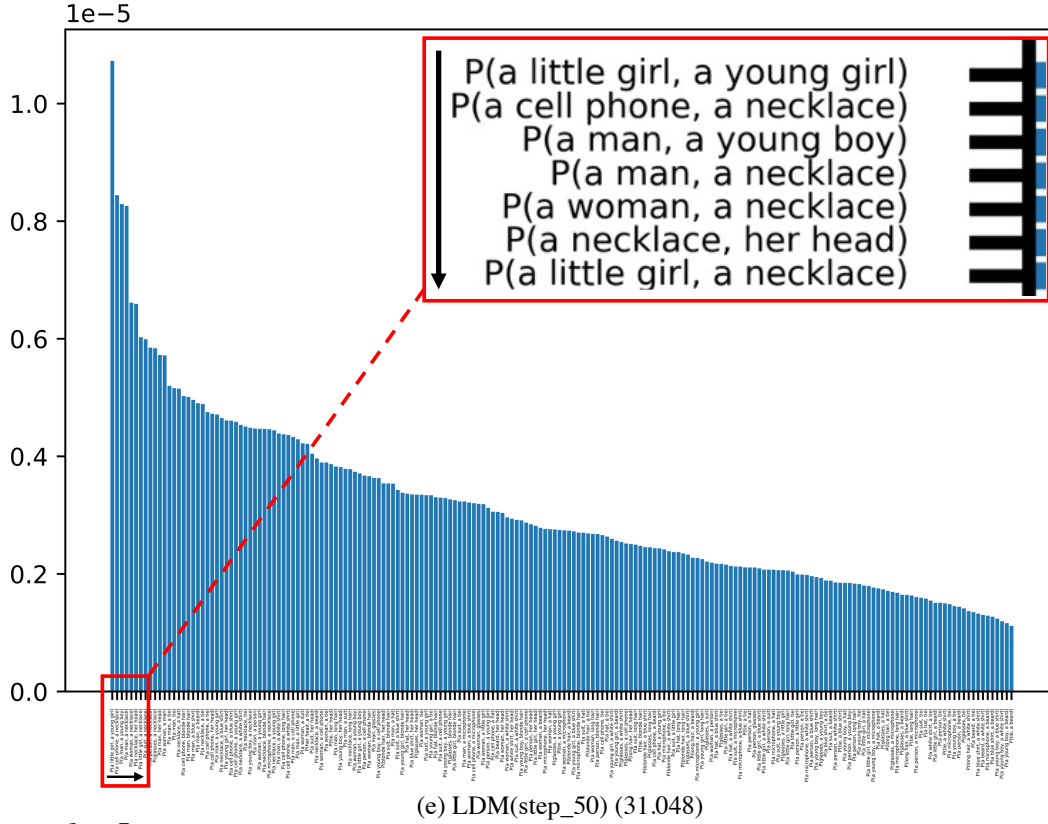


Figure S11: **PaKLD results with BLIP attributes on FFHQ**. The value beside model name denotes the PaKLD value(10^{-7}) for each respective model. Please zoom in for the best view.

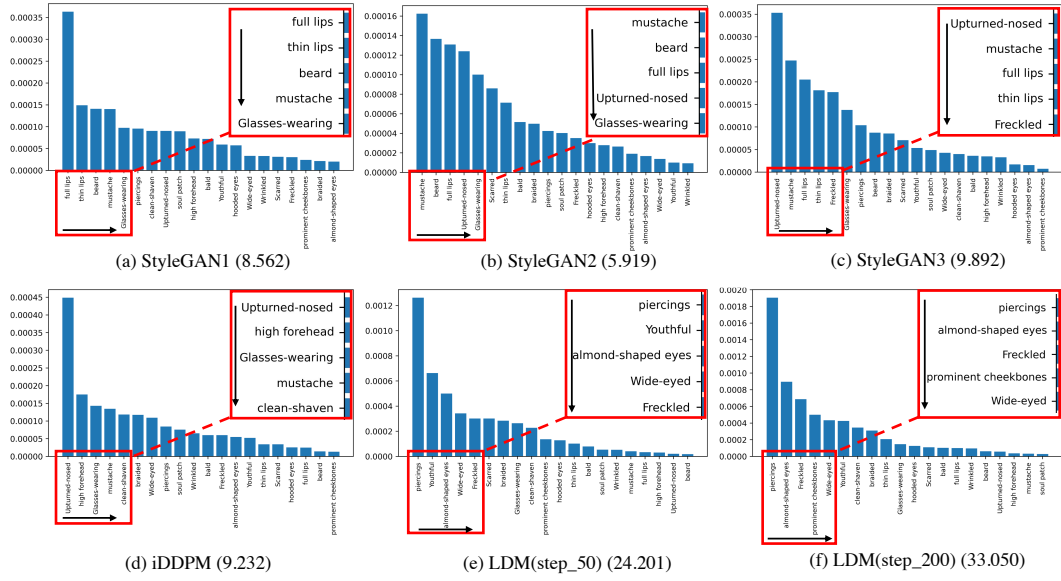
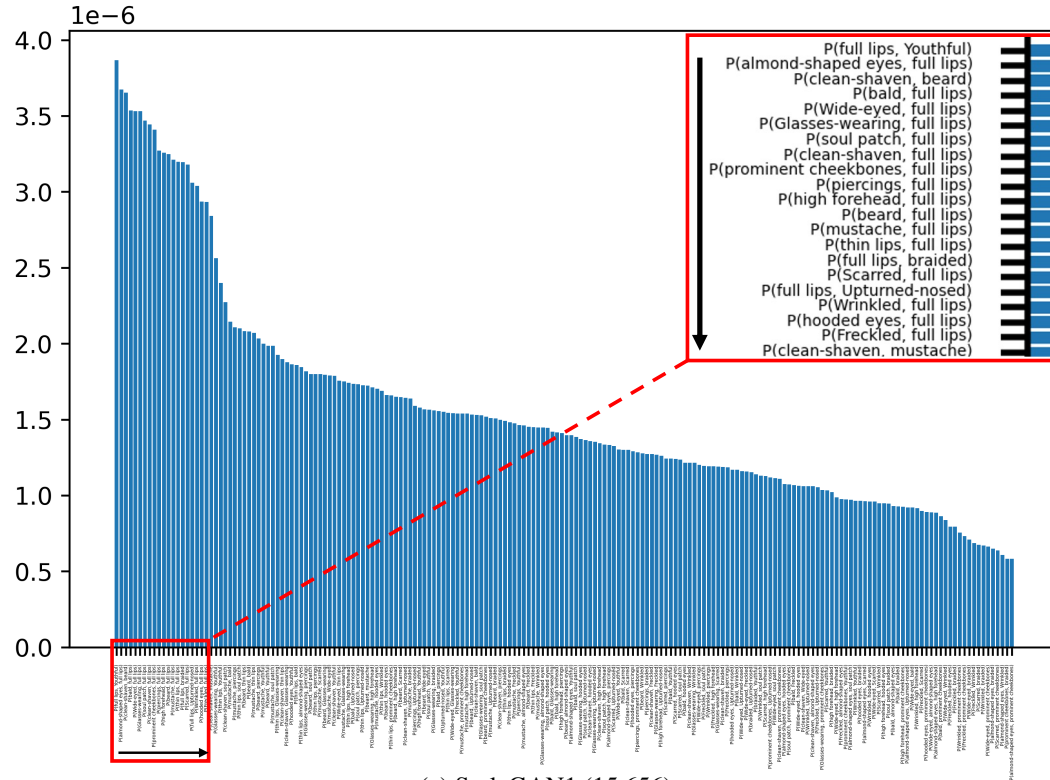
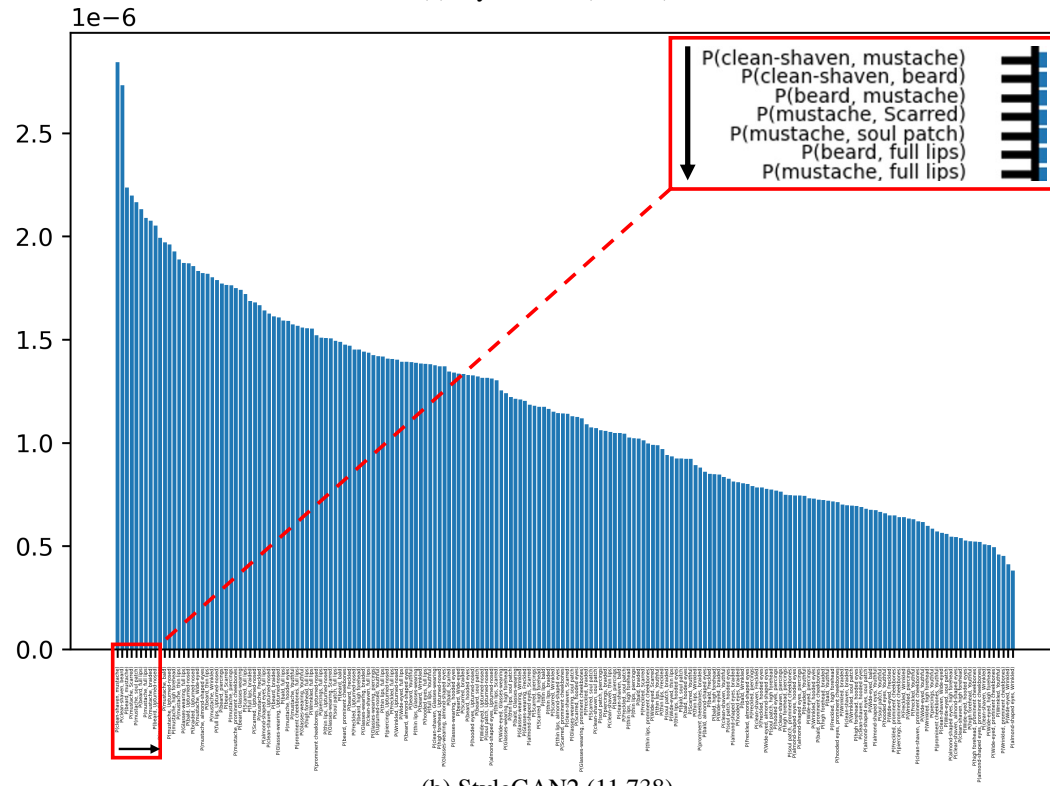


Figure S12: **SaKLD** We analyze results with **GPT** attributes on **FFHQ** dataset. While (a), (d), or (f) noticeably falls short in preserving certain specific attributes such as "full lips", "upturned-nose" or "piercings", (b) and (c) preserves all attributes in a balanced manner. Please observe whether the shape of SaKLD is sharp or smooth, and note that scale of y axis is different for each model. The value beside model name denotes the SaKLD value(10^{-5}) for each respective model.

PaKLD with GPT attributes on FFHQ

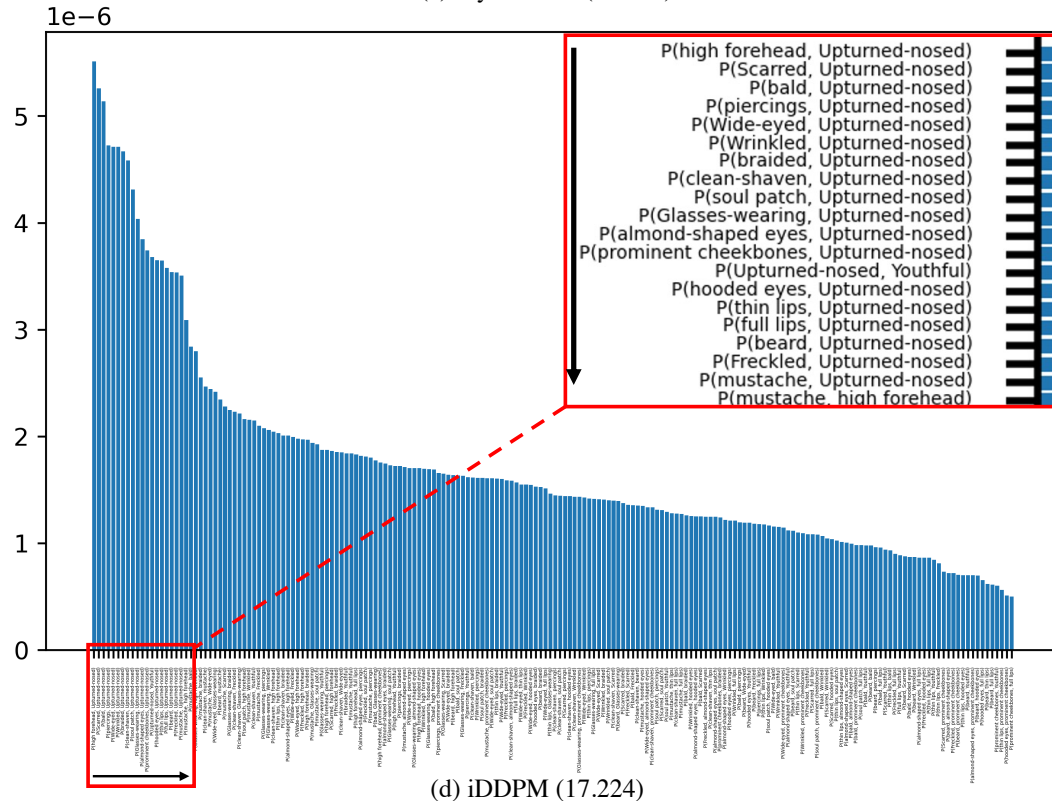
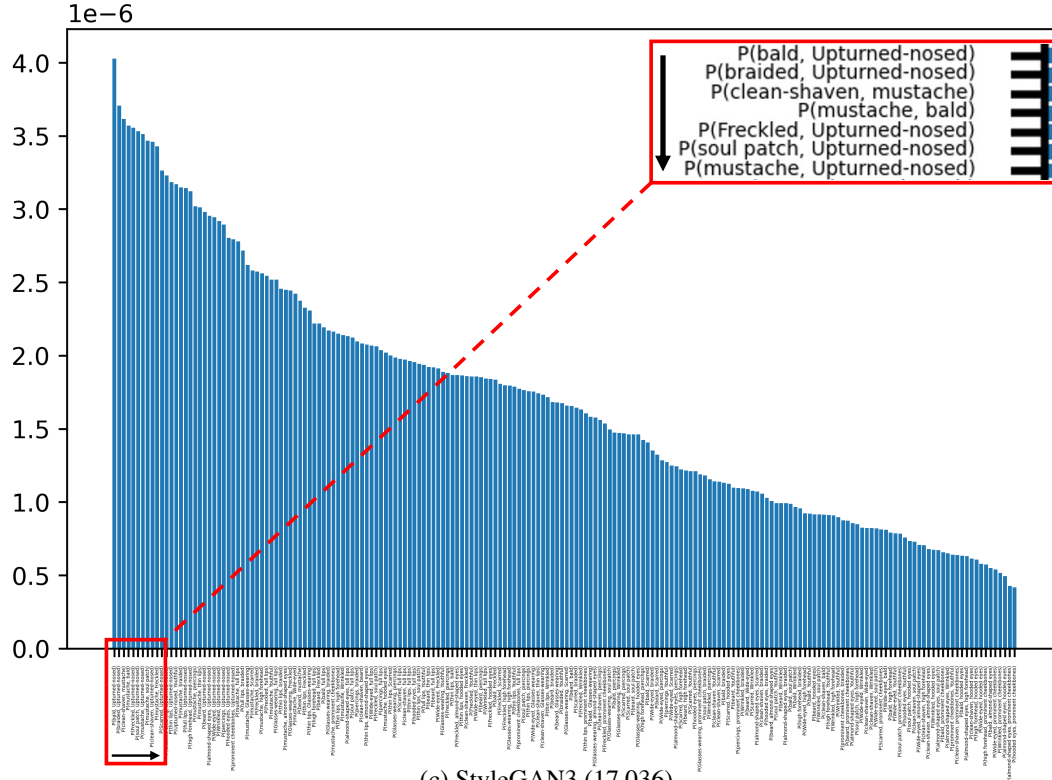


(a) StyleGAN1 (15.656)



(b) StyleGAN2 (11.738)

PaKLD with GPT attributes on FFHQ



PaKLD with GPT attributes on FFHQ

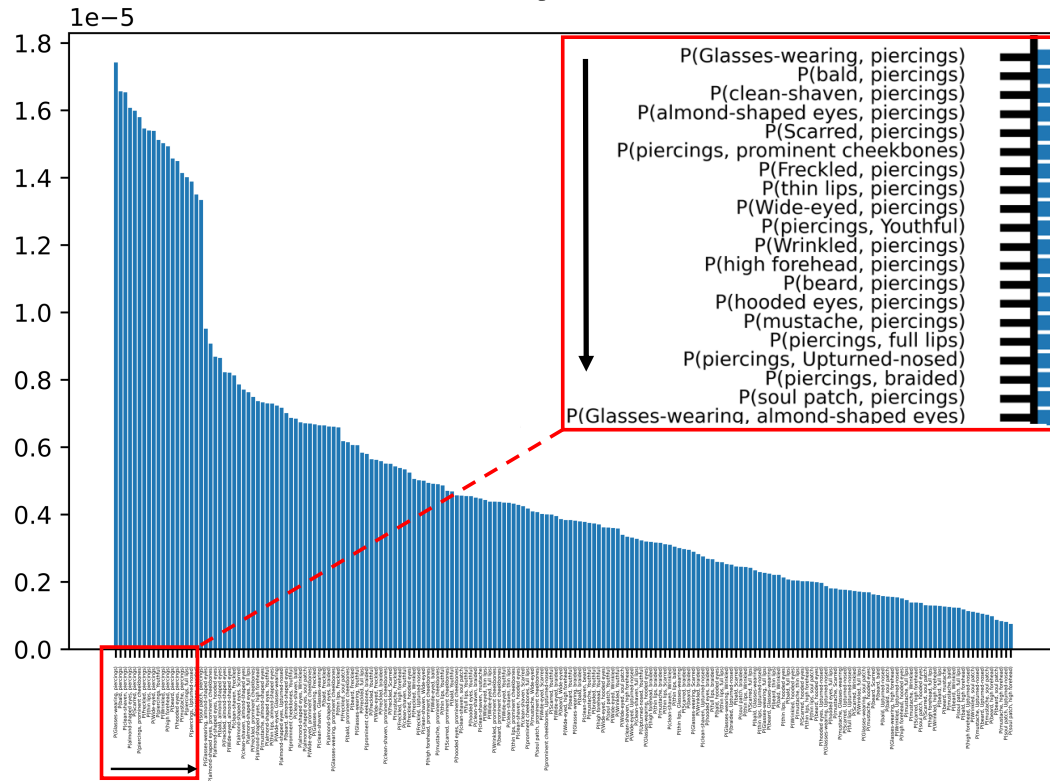
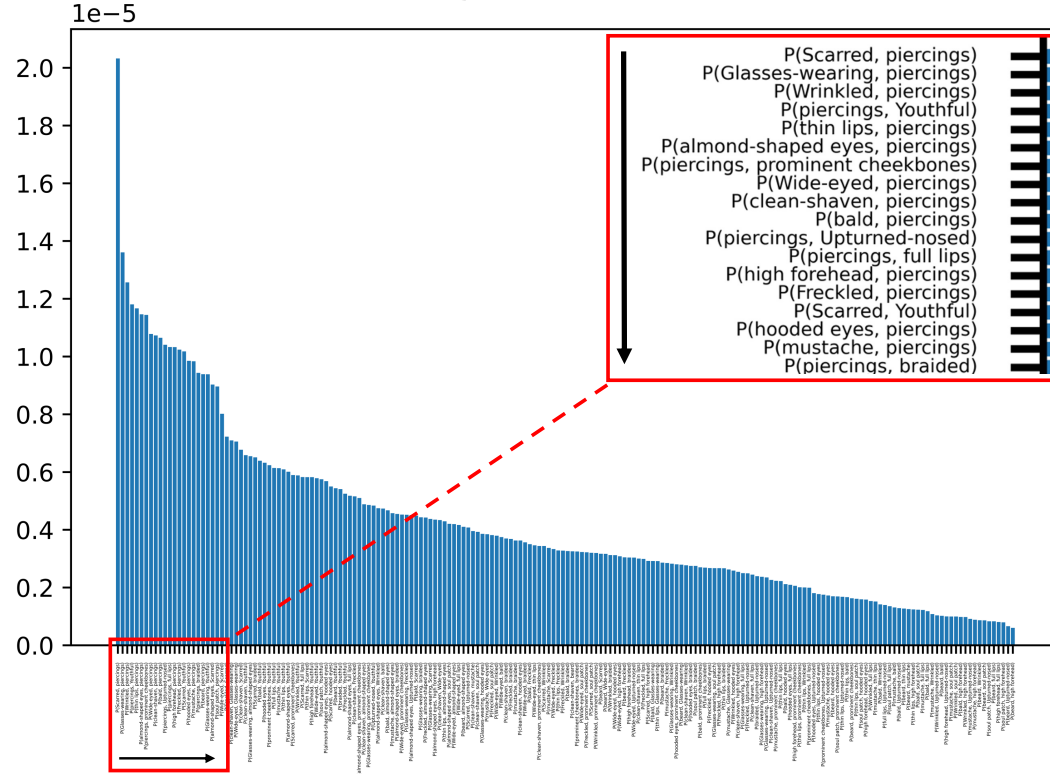


Figure S13: **PaKLD results with gpt attributes on FFHQ.** (f) LDM captures wrong relation when "piercings" combined with other attributes, resulting in a significant PaKLD value. The value beside model name denotes the PaKLD value(10^{-7}) for each respective model. Please zoom in for the best view.

References

- [1] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [4] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.
- [5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.
- [6] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.
- [7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [8] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [10] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.