Rebuttal for " Revisiting the Evaluation of Image Synthesis with GANs "

Anonymous Author(s)
Affiliation
Address
email

To Reviewer cWrw

Q1: Include state-of-the-art generative models like diffusion models.

Reply: Thanks. Following your valuable suggestion, we further include more state-of-the-art generative models on the ImageNet dataset for synthesis evaluation, namely GigaGAN (CVPR'2023) [7], MDT (ICCV'2023) [4], and DG-Diffusion (ICML'2023) [8]. Specifically, we either gather their official models for inference or download the pre-generated images released by the authors for evaluation. Similarly, $50K$ generated images and the entire training set (*i.e.,* $1.28M$ images) are used as the synthesized and real distributions, respectively. All details are consistent with the experiments conducted in our main paper. Tab. 1 presents the quantitative comparison results. Akin to the results in our main paper, our evaluation system provides consistent ranks with FID and human visual evaluation, demonstrating the reliability of our metric. These results will be added in the next version of our paper.

Table 1: **Quantitative comparison results of Centered Kernel Alignment (CKA$_\uparrow$) on ImageNet dataset**. † scores are quoted from the original paper and others are tested three times.

Model	FID†	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall	User study
GigaGAN [7]	3.45	68.01	79.93	90.15	98.34	82.40	96.52	85.89	65%
DG-Diffusion [8]	3.18	68.22	80.06	90.56	98.46	82.51	96.88	86.12	66%
MDT [4]	1.79	69.64	81.68	91.78	99.43	83.43	98.19	87.36	69%

Q2: Provide the evaluation on more MLP-based models like mlp-mixer.

Reply: Thanks. As suggested, two MLP-based models are leveraged as the feature extractor for synthesis evaluation, namely gMLP [12] and MLP-mixer [20]. Following the experimental settings in our main paper, we identify the reliability and robustness of these MLP-based models via 1) visualizing the highlighted regions that contribute most significantly to the measurement results, and 2) attacking the feature extractor with histogram matching attack. Fig. 1 and Tab. 2 respectively present the qualitative and quantitative results. On one hand, the heatmap visualization results indicate that both gMLP and mixer-MLP capture limited semantics. Considering that more visual semantics should be considered for a more comprehensive evaluation, gMLP and MLP-mixer might not be adequate for synthesis comparison. On the other hand, the quantitative results demonstrate that their FD scores could be altered by the histogram matching attack, without actually improving the synthesis quality. That is, gMLP and MLP-mixer are susceptible to the histogram attack. Together with the finding that the FD scores of ResMLP could be manipulated without any improvement to the generative models in Tab.2 of our main paper, we do not integrate MLP-based feature extractors into our measurement system. These results will be added in the next version of our paper.

Q3: 100 Human Judgment may not enough to fully capture the complexities of evaluating generative models objectively.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

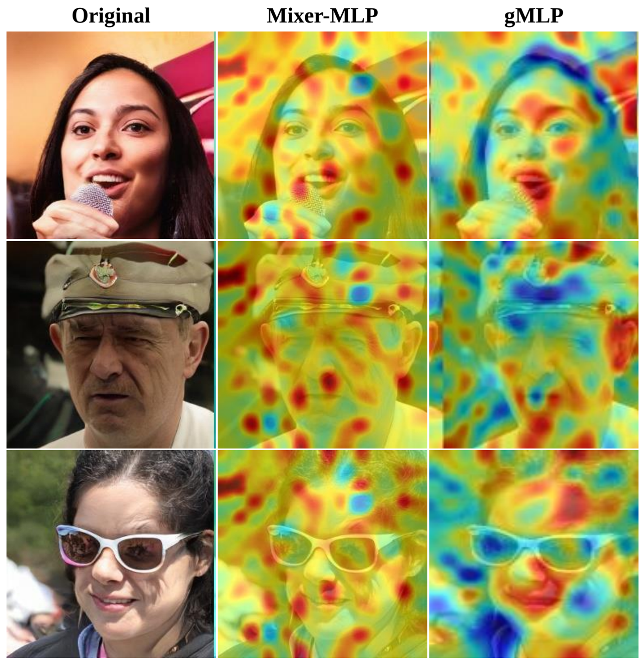


Figure 1: **Heatmaps from MLP-based extractors, namely Mixer-MLP [20] and gMLP [12].**

Table 2: **Quantitative comparison results of MLP-based extractors' Fréchet Distance ($\mathbf{FD}_{\downarrow}$) on the ImageNet dataset**. † scores are quoted from the original paper and others are tested three times.

Extractor	Random	Chosen$_I$
gMLP	$2.93_{\pm 0.004}$	$2.89_{\pm 0.004}\downarrow$
mixer-MLP	$5.51_{\pm 0.01}$	$5.35_{\pm 0.01}\downarrow$

Reply: Thanks. We agree that involving thousands of persons for human visual evaluation can provide more consistent and reliable results. However, this is too expensive for us as including thousands of participants requires massive human and time resources. Therefore, two strategies of human perceptual judgment are designed for different investigations in our main experiments, namely benchmaking the synthesis quality of one specific generative model and comparing two paired models. In particular, 100 participants are asked to vote the synthesis quality and their final scores are averaged to avoid overly subjective individual outcomes. Moreover, in order to ensure that our human evaluation is reliable and consistent, we repeat the same images several times (*i.e.,* 4) randomly for human visual comparison. In this way, if one user vote photorealistic and unrealistic two times each for the same images, the results would be considered as indistinguishable. This operation further filters overly subjective individual judgment and ensure the rationality of our user study. Additionally, we notice that common choice for human evaluation in the community is to include about 50 participants for perceptual comparison [18, 23, 14, 22, 16, 11, 10]. For instance, [23], [18] and [10] asked 50 workers to pick the unrealistic images, ProjectedGAN [16] conducted a human preference study with only 28 participants, and the most recent work [14] included only 15 graders to compare the synthesis quality. By contrast, 100 persons are involved in our human judgment, thus we believe that our perceptual comparison results are reliable. Furthermore, we view large-scale human evaluation as our future work to perform more extensive investigations.

Hope that the above discussions could address your concerns, please let us know if you have any further questions. Thanks for your effort and constructive suggestions again.

To Reviewer Z9sB

Q1: Only the mean values of metrics are reported, no stds.

Reply: Thanks. As suggested, we add the std values of our experiments to better illustrate the numerical fluctuation of various extractors towards the histogram attack. 3 presents the quantitative

Table 3: **Quantitative comparison results of Fréchet Distance (FD$_\downarrow$) on FFHQ dataset**. "Random, Chosen$_I$" respectively represent the synthesized distribution of randomly generated and matching the class prediction of Inception-V3. Moreover, "$_v$" and "$_v$" respectively denote the architecture of ResNet and ViT. ($\downarrow$) indicates the results are hacked by the histogram matching mechanism. Notably, the values across different rows are not comparable and the results are tested three times.

Model	Inception	ConvNeXt	SWAV	MoCo$_r$	RepVGG	CLIP$_r$	Swin	ViT	DeiT	CLIP$_v$	MoCo$_v$	ResMLP
Random	2.81± 0.01	78.03± 0.10	0.13± 0.002	0.24± 0.003	129.61± 0.41	10.34± 0.06	142.87± 0.12	15.11± 0.09	437.80± 0.14	1.06± 0.01	7.32± 0.03	99.11± 0.06
Chosen$_I$	2.65$\pm 0.01$$\downarrow$	78.19± 0.11	0.13± 0.002	0.24± 0.003	129.67± 0.39	10.36± 0.08	140.01$\pm 0.12$$\downarrow$	15.11± 0.10	430.81$\pm 0.16$$\downarrow$	1.06± 0.01	7.40± 0.03	95.36$\pm 0.06$$\downarrow$

results. We could tell that the FD scores of extractors that are vulnerable to the attack can be improved by matching the histogram, and the improvement of FD scores is greater than stds. For instance, the improvement of FD scores from the Inception model is 0.16 and the computation std is only 0.01, there is an order of magnitude difference between them. Moreover, the improvement of FD scores from the Swin-Transformer model is 2.86 and the computation std is only 0.12. That is, the improvement is actually caused by the histogram attack rather than the variance of attempts. Note that the generator is unchanged but the FD scores are improved by the attack, which is unacceptable for synthesis evaluation. Accordingly, extractors that are vulnerable to the histogram matching attack are not reliable for evaluation.

Q2: The authors provide many tables with the results but it is not trivial to parse them. Specifically, checking whether this or that metric correlates with the human evaluation ahowld be done manually. It would be great if this could be somehow quantified or visualized (e.g., FID/other metrics as functions of the user score, 2D plots).

Reply: Thanks. Following your valuable suggestion, we visualize the correlation between different metrics and the human evaluation results. Specifically, we plot the correlation of the averaged ranks of various models given by human judgment, CKA, and FID. Fig. 2 and Fig. 3 respectively present the visualization results of the ImageNet, FFHQ, and LSUN-Church datasets. Obviously, the averaged ranks given by CKA are more consistent with that of the human evaluation, demonstrating the accuracy of CKA. Moreover, we plot the comparison between the stds and the improvements obtained by the histogram attack for better illustration. Fig. 4 presents the results. Similarly, we could observe that the improvement is actually caused by the histogram attack rather than the variance of attempts.

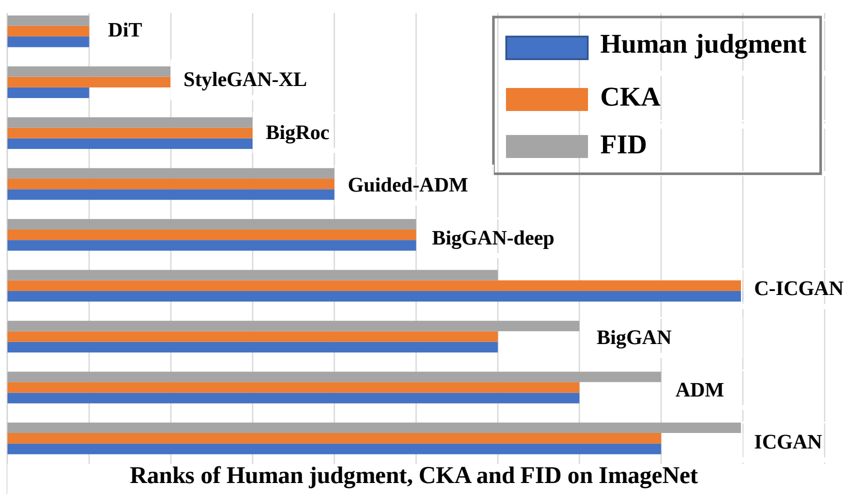


Figure 2: **The correlation of the averaged ranks of various models on ImageNet given by human judgment, CKA, and FID.**

Q3: May be it is more fair to emphasize other advantages of CKA (such as the sample efficiency) rather than consistency and reliability.

Reply: Thanks. On one hand, our results demonstrate that CKA provides a consistent ranking with the FID scores in most cases, demonstrating that CKA can deliver the similarity between different data distributions. One the other hand, CKA agrees with human visual judgment whereas FID fails in

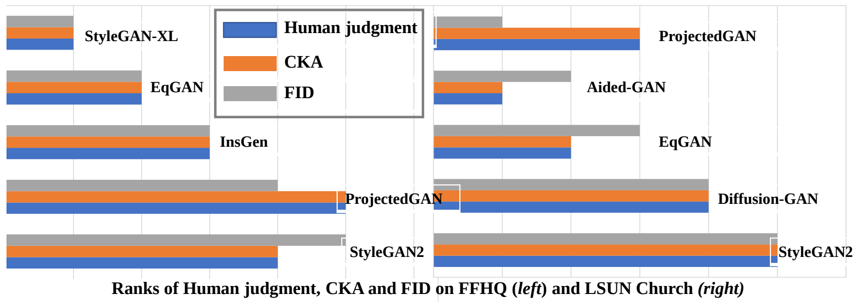


Figure 3: **The correlation of the averaged ranks of various models on FFHQ and LSUN-Church given by human judgment, CKA, and FID.**

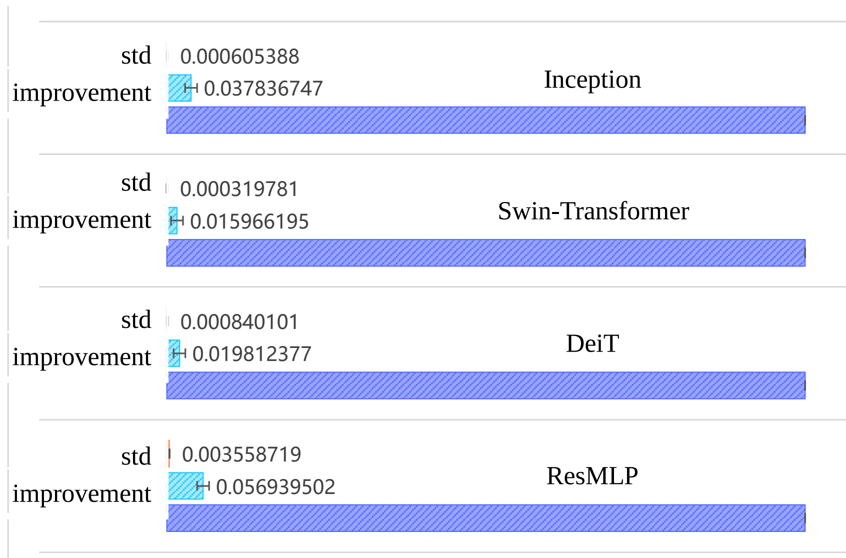


Figure 4: **The quantitative comparison between the stds and the improvements obtained by the histogram attack.**

some circumstances. That is, CKA can measure the synthesis performance more reliable than FID. Additionally, CKA shows better sample efficiency than both FID and KID. Thus we integrate CKA as the distributional distance to evaluate the synthesis performance in our system. Together with several robust feature extractors, our new measurement system is more consistent and reliable than exiting alternatives. In the main paper, we emphasize the reliability and consistency of our entire system rather than only the distributional distance (*i.e.,* CKA) as both the extractors and distances are important. We will proofread our presentation and emphasize the advantages of our overall system following your valuable suggestions.

Q4: Include at least some evaluation/comparison/comment with this KID (both in terms of correlation with human evaluation and sample efficiency).

Reply: Thanks. Following your valuable suggestion, we further involve Kernel Inception Distance (KID) [1], precision, and recall [15] into our comparison. Note that the original KID employs Inception-V3 as the feature extractor, and there is a large "perceptual null space" in Inception-V3. Therefore, we first investigate whether KID scores can be altered by attacking the feature extractor with the histogram matching mechanism. The experimental details are consistent with computing Fréchet Distance ($FD_{\downarrow}$) in Tab.2 of the main paper. Tab. 6 presents the quantitative results. Still, some extractors, such as Inception, Swin-Transformer, and ResMLP, are susceptible to the histogram matching attack. For instance, the KID score of Swin-Transformer is improved by 5.31% when the chosen set is used. These observations agree with our findings in our main paper, suggesting that certain extractors can be hacked when KID is employed as the distributional distance. Then, we investigate the sample efficiency of KID, Precision, and Recall to probe the impacts of the amount of generated samples. Fig. 5 presents the curves of KID, Precision, and Recall scores computed under different data regimes. Similarly, we could observe that the KID scores can be improved by synthesizing more images. Interestingly, the recall scores decrease as the generated sample size increases whereas the precision is stable. This is caused by the definition of recall: recall measures the proportion of the real distribution that is covered by the synthesized distribution. In practical computation, the denominator increases as the synthesized samples increases, while the numerator (*i.e.,* images from the real distribution) remain unchanged. In this way, the recall scores decrease as the generated sample size increases and vice versa. By contrast, CKA scores are stable under different data regimes, (please see Fig. 2 in the main paper). Moreover, CKA can provide reliable synthesis evaluation that agrees with human visual judgment. Accordingly, CKA is a proper choice for building a consistent and reliable measurement system.

Q5: The authors should better explain CKA metric in the main text.
Reply: Thanks. Following your valuable suggestion, we add more details of the CKA metric as follows:

Centered Kernel Alignment (CKA) as a widely used similarity index for quantifying neural network representations [2, 9, 3], could also serve as a metric of similarity between two given distributions. To be specific, CKA is normalized from Hilbert-Schmidt Independence Criterion (HSIC) [5] to ensure invariant to isotropic scaling and is calculated by

$$\mathrm{CKA}(\mathrm{X}, \mathrm{Y}) = \frac{\mathrm{HSIC}(x, y)}{\sqrt{\mathrm{HSIC}(x, x)\mathrm{HSIC}(y, y)}}. \tag{1}$$

Here, HSIC determines whether two distributions are independent. Formally, let $K_{ij} = k\left(\mathrm{x}_i, \mathrm{x}_j\right)$ and $L_{ij} = l\left(\mathrm{y}_i, \mathrm{y}_j\right)$, where k and l are two kernels. HSIC is defined as

$$\mathrm{HSIC}(K, L) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(KHLH), \tag{2}$$

where H denotes the centering matrix (*i.e.,* $H_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$). For kernel selections of k and l, we find that different kernels (RBF, polynomial, and linear) give similar results and rankings, and the RBF kernel contributes to the distinguishability of quantitative results. Therefore, RBF kernel is used for all experiments, and the bandwidth is set as a fraction of the median distance between examples [9]. These metrics are compared in a consistent setting for fair comparison, more implementation details are given in *Supplementary Material*.

Q6: Provide a more explanatory discussion of what is CKA (beside the formulas) and some intuition what it measure and why it is a good metric?
Reply: Thanks. As a widely used similarity index for measuring the correspondence between representations in neural networks, CKA has been identified to have several advantages: 1) CKA is invariant to orthogonal transformation and isotropic scaling, making is stable under various image transformations; 2) CKA can capture the non-linear correspondence between representations due to its kernel mapping; and 3) CKA can determine the correspondence across different features and with different widths, whereas previous metrics fail [9]. Additionally, through extensive experiments, we demonstrate that CKA can provide a accurate evaluation for synthesis comparison and is sample-efficient. Accordingly, CKA is a good metric for delivering the distributional discrepancy.

Q7: What is the "null space" of this metric?
Reply: Thanks. In fact, it is the feature extractor that might have a “perceptual null space". For instance, the Inception model has been identified to have a large “perceptual null space", leading it vulnerable to the histogram matching attack. Moreover, CKA measures the similarity between different distributions, and $\mathrm{CKA}(\mathrm{X}, \mathrm{Y}) = 1$ if and only if the two sets coincide.

Q8: What is the computational complexity of CKA compared to FID? How it scales with the feature dimension and sample size?
Reply: Thanks. Assume N samples from the evaluated distributions are used for calculating CKA, the main computational complexity of CKA comes from: 1) centering the kernel matrix with the pre-defined centering matrix with the complexity of $\mathcal{O}(N^2)$; and 2) computing HSIC scores with

the complexity of $\mathcal{O}(N^3)$. Therefore, the overall computational complexity of CKA is $\mathcal{O}(N^3)$. By contrast, the computational complexity of FID mainly comes from calculating the mean and variance of the sample features ($N \times d$, where d denotes the feature dimension). The computational complexity is $\mathcal{O}(N \times d)^3$. The computational complexity of both CKA and FID increases linearly with the cubic power of sample size. Moreover, as suggested, we provide the clock time of FID and CKA in the following table. Concretely, we use the full FFHQ dataset ($70K$ images) as the reference distribution and generate $50K$ images for evaluation, the clock time is tested on a single 3090 24G GPU. We could tell that CKA takes shorter time than FID when the same amount of samples are calculated.

Extractor	Inception	ViT
FID	3426 (s)	3630 (s)
CKA	3225 (s)	3328 (s)

Q9: What is the theoretical sample complexity of CKA? Are there any known results here?
Reply: Thanks. To the best of our knowledge, there are no known results of the theoretical sample complexity of CKA. CKA measures the distributional discrepancies between different distributions with a considerable samples from each distribution. Accordingly, involving sufficient samples for evaluation ensures more accurate results in practice. However, through evaluating the CKA scores under various data regimes, we observe that CKA shows satisfactory sample-efficiency and stability under different number of samples. Therefore, we can synthesize subsets with fixed number of images (*e.g.,* 50 K) for evaluation. By contrast, the FID and KID scores could be improved by producing more samples, which is unacceptable for a reliable evaluation.

Q10: Is centered kernel alignment somehow related to the kernel maximum mean discrepancy (KID/MMD)?
Reply: Thanks. Centered Kernel Alignment (CKA) is normalized from Hilbert-Schmidt Independence Criterion (HSIC) [5] to ensure it is invariant to isotropic scaling and is formally defined by

$$\mathrm{CKA}(\mathrm{X}, \mathrm{Y}) = \frac{\mathrm{HSIC}(x, y)}{\sqrt{\mathrm{HSIC}(x, x)\mathrm{HSIC}(y, y)}}. \tag{3}$$

HSIC is equivalent to maximum mean discrepancy (MMD) between the joint distribution and the product of the marginal distributions, and HSIC with a specific kernel family is equivalent to distance covariance [17]. HSIC determines whether two distributions are independent, and HSIC = 0 implies independence. However, HSIC is not invariant to isotropic scaling, making it sensitive to isotropic transformation of images when used for synthesis evaluation.

Hope that the above discussions could address your concerns, please let us know if you have any further questions. Thanks for your effort and constructive suggestions again.

To Reviewer 95Dj

Q1: There are no ablation studies to separately prove the effectiveness of six extractors and CKA.
Reply: Thanks. In this work, we seek to develop a new measurement system that could provide reliable and consistent synthesis comparisons. In particular, two key components are crucial for the measurement system, *i.e.,* the feature extractor that defines representation space and the distributional distance that deliver similarities. Accordingly, we make in-depth analyses on the reliability and robustness of various feature extractors and different distributional distances. For the feature extractors, we gather multiple models that are pre-trained with different objectives (fully-supervised/self-supervised) and various architectures (CNN/ViT/MLP). Notably, these models are chosen for a systematic investigation to comprehensively understand the intrinsic properties of various extractors, rather than based on existing findings. Then, we testify their performance on 1) how many semantic features they can capture for evaluation, 2) how robust they are when being attacked by the histogram matching mechanism, and 3) how distinct the representation space they can define. These investigations provide several new findings to the community, including 1) one specific extractor can only capture limited semantics and provide ons-side results, 2) extractors that are vulnerable

Table 4: **Quantitative comparison results of Fréchet Distance (FD$_\downarrow$) on ImageNet dataset**. "Random, Chosen$_I$" respectively represent the synthesized distribution of randomly generated and matching the class prediction of Inception-V3. Moreover, "$_v$" and "$_v$" respectively denote the architecture of ResNet and ViT. ($_\downarrow$) indicates the results are hacked by the histogram matching mechanism. Notably, the values across different rows are not comparable and the results are tested three times.

Model	Inception	ConvNeXt	SWAV	MoCo$_r$	RepVGG	CLIP$_r$	Swin	ViT	DeiT	CLIP$_v$	MoCo$_v$	ResMLP
Random	34.29$\pm$0.09	78.02$\pm$0.16	0.13$\pm$0.003	0.32$\pm$0.002	54.98$\pm$0.22	27.64$\pm$0.15	323.12$\pm$0.88	50.97$\pm$0.20	621.98$\pm$1.02	5.46$\pm$0.09	50.01$\pm$0.21	145.32$\pm$1.02
Chosen$_I$	33.05$\pm$0.08$\downarrow$	78.10$\pm$0.14	0.13$\pm$0.002	0.32$\pm$0.002	54.30$\pm$0.24	27.66$\pm$0.17	301.91$\pm$0.92$\downarrow$	50.96$\pm$0.18	597.32$\pm$1.11$\downarrow$	5.46$\pm$0.07	50.00$\pm$0.19	133.06$\pm$1.09$\downarrow$

to the histogram matching attack are not reliable, and 3) different feature extractors might define similar representation spaces. For the distributional distances, we investigate the numerical stability of different distances across various representation spaces and the sample efficiency of different distances. Through extensive comparisons, we find that Centered Kernel Alignment (CKA) provides a better comparison across various extractors and hierarchical layers with its bounded score. Moreover, CKA is more sample-efficiency and exhibits better agreement with human visual judgment. Together with these findings, we build a new measurement system that can accurately reflect the synthesis performance. Following this line, the effectiveness of each feature extractors and CKA is identified in our main experiments. In particular, Fig. 1 of the main paper indicates that the chosen six feature extractors can incorporate more visual semantics for evaluation in a complementary manner. And Tab. 2 of the main paper demonstrates that each of the chosen extractors is robust towards the histogram attack. Furthermore, in Tab.4 of the main paper and Tab.4, 5, 6, 7, 8 of the supplementary material, we provide qualitative and quantitative results of each extractor from various semantic levels. These results also demonstrate the reliability of each extractor when used for synthesis evaluation. In addition to evaluating the robustness of these extractors on the FFHQ dataset, we further perform the same experiment on the ImageNet dataset. Tab. 4 presents the quantitative results. We can tell from these results that the chosen feature extractors are robust to the attack, further demonstrating their reliability.

Q2: It is important to research how to improve the speed of evaluation without affecting the evaluation accuracy.

Reply: Thanks. We agree. Both evaluation speed and accuracy are very important in practice. This work focuses on developing a measurement system that could reliably and consistently reflect the synthesis performance. Based on the findings that one certain feature extractor might capture only limited semantics for evaluation, we integrate multiple extractors to alleviate this. Therefore, the evaluation time is relatively longer than using one extractor for evaluation. However, the inference time of these feature extractors is much shorter than the inference time of diffusion models. For instance, the evaluation time of our measurement system on $50K$ images is about 5 hours on a single 3090 24G GPU, but it takes about several days to generate $50K$ images with diffusion models (about 4 days for MDT and 2.5 days for DG-Diffusion). Consequently, improving both the speed of our evaluation and the inference speed of diffusion models is also important. In the future, we plan to integrate various accelerate techniques to improve our evaluation speed without compromising the evaluation accuracy, such as optimizing the model architecture, model pruning and distillation, *etc.*

Q3: The layers of Section 2.3, 3.1 and 3.2 are not prominent and the organization of them is not clear. Specifically, the summary sentences are not emphasized and paragraph are not strictly parallel.

Reply: Thanks. Our presentation is organized for following reasons: In Section 2.3, we present the details of generative models, evaluated datasets, and analysis approaches (including our visualization tool, histogram matching attack, and human evaluation). They are independent of each other, thus we discuss them in parallel in the main paper. In Section 3.1, we investigate the feature extractors by first identifying their attention on visual semantics, followed by investigating their robustness to the histogram matching attack. Finally, we filter extractors that define similar representation spaces. These studies are gradually deepening, thus they are organized in a progressive manner. In Section 3.2, we first study the numerical scales of CKA and FID across various extractors and hierarchical layers of one certain extractor. After that, we investigate the sample efficiency of CKA and KID. In the last paragraph of Section 3.2, we summarize our findings about the feature extractors and the distributional distances. Moreover, the summary sentences of each paragraph provide our primary findings of this paragraph. Following your valuable suggestions, we will carefully proofread and revise the corresponding presentation to make our paper more logical.

Hope that the above discussions could address your concerns, please let us know if you have any further questions. Thanks for your effort and constructive suggestions again.

To Reviewer y8MJ

Q1: The novelty is limited. CKA is a well-known metric for evaluating the similarity between distributions.

Reply: Thanks. In this work, we seek to develop a new measurement system that could provide reliable and consistent synthesis comparisons. In particular, two key components are crucial for the measurement system, *i.e.,* the feature extractor that defines representation space and the distributional distance that deliver similarities. Accordingly, we make in-depth analyses on the reliability and robustness of various feature extractors and different distributional distances. For the feature extractors, we gather multiple models that are pre-trained with different objectives (fully-supervised/self-supervised) and various architectures (CNN/ViT/MLP). Then, we testify their performance on 1) how many semantic features they can capture for evaluation, 2) how robust they are when being attacked by the histogram matching mechanism, and 3) how distinct the representation space they can define. These investigations provide several new findings to the community, including 1) one specific extractor can only capture limited semantics and provide ons-side results, 2) extractors that are vulnerable to the histogram matching attack are not reliable, and 3) different feature extractors might define similar representation spaces. For the distributional distances, we investigate the numerical stability of different distances across various representation spaces and the sample efficiency of different distances. Through extensive comparisons, we find that Centered Kernel Alignment (CKA) provides a better comparison across various extractors and hierarchical layers with its bounded score. Moreover, CKA is more sample-efficiency and exhibits better agreement with human visual judgment. Together with these findings, we build a new measurement system that can accurately reflect the synthesis performance. To the best of our knowledge, this paper is the first work to present these findings about feature extractors and to incorporate CKA for synthesis measurement in the community. We believe that these findings can provide potential insights to further works that develop new evaluation protocols.

Q2: Lacks discussion of some state-of-the-art methods, such as stable diffusion and midjourney.

Reply: Thanks. Following your valuable suggestion, we further include more state-of-the-art generative models on the ImageNet dataset for synthesis evaluation, namely GigaGAN (CVPR'2023) [7], MDT (ICCV'2023) [4], and DG-Diffusion (ICML'2023) [8]. Specifically, we either gather their official models for inference or download the pre-generated images released by the authors for evaluation. Similarly, $50K$ generated images and the entire training set (*i.e.,* $1.28M$ images) are used as the synthesized and real distributions, respectively. All details are consistent with the experiments conducted in our main paper. Tab. 5 presents the quantitative comparison results. Akin to the results in our main paper, our evaluation system provides consistent ranks with FID and human visual evaluation, demonstrating the reliability of our metric. Notably, this paper focuses on evaluating the performance of various generative models trained on single modality (*i.e.,* images). Therefore, evaluating generative models that are trained on multiple modality synthesis tasks (*e.g.,* text-to-image generation) is slightly out of our scope. However, multiple modality tasks such as text-to-image/video have made remarkable progress recently, and evaluating their performance accurately is a very important and promising topic. Accordingly, we plan to investigate the performance of our measurement system under multiple modality synthesis tasks in our future work.

Q3: This paper only compares CKA with FID and lacks a comparison with the other metrics. A discussion of these related metrics is needed.

Reply: Thanks. Following your valuable suggestion, we further involve Kernel Inception Distance (KID) [1], precision, and recall [15] into our comparison. Note that the original KID employs Inception-V3 as the feature extractor, and there is a large "perceptual null space" in Inception-V3.

Table 5: **Quantitative comparison results of Centered Kernel Alignment (CKA$_\uparrow$) on ImageNet dataset**. † scores are quoted from the original paper and others are tested three times.

Model	FID†	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall	User study
GigaGAN [7]	3.45	68.01	79.93	90.15	98.34	82.40	96.52	85.89	65%
DG-Diffusion [8]	3.18	68.22	80.06	90.56	98.46	82.51	96.88	86.12	66%
MDT [4]	1.79	69.64	81.68	91.78	99.43	83.43	98.19	87.36	69%

Table 6: **Quantitative comparison results of Kernel Inception Distance (KID$_\downarrow$, $\times e^{-3}$) on FFHQ dataset**. "Random, Chosen$_I$" respectively represent the synthesized distribution of randomly generated and matching the class prediction of Inception-V3. Moreover, "$_v$" and "$_v$" respectively denote the architecture of ResNet and ViT. ($_\downarrow$) indicates the results are hacked by the histogram matching mechanism. Notably, the values across different rows are not comparable and the results are tested three times.

Model	Inception	ConvNeXt	SWAV	MoCo$_r$	RepVGG	CLIP$_r$	Swin	ViT	DeiT	CLIP$_v$	MoCo$_v$	ResMLP
Random	1.88$\pm$0.02	34.81$\pm$0.11	9.61$\pm$0.06	5.31$\pm$0.06	33.88$\pm$0.29	2.85$\pm$0.05	21.64$\pm$0.10	16.74$\pm$0.10	18.01$\pm$0.19	38.06$\pm$0.20	15.41$\pm$0.09	4.86$\pm$0.02
Chosen$_I$	1.71$\pm$0.02$\downarrow$	34.82$\pm$0.10	9.61$\pm$0.06	5.31$\pm$0.05	33.89$\pm$0.27	2.85$\pm$0.05	20.49$\pm$0.09$\downarrow$	16.74$\pm$0.12	19.39$\pm$0.22	38.09$\pm$0.19	15.40$\pm$0.07	4.70$\pm$0.02$\downarrow$

Therefore, we first investigate whether KID scores can be altered by attacking the feature extractor with the histogram matching mechanism. The experimental details are consistent with computing Fréchet Distance (FD$_\downarrow$) in Tab.2 of the main paper. Tab. 6 presents the quantitative results. Still, some extractors, such as Inception, Swin-Transformer, and ResMLP, are susceptible to the histogram matching attack. For instance, the KID score of Swin-Transformer is improved by 5.31% when the chosen set is used. These observations agree with our findings in our main paper, suggesting that certain extractors can be hacked when KID is employed as the distributional distance. Then, we investigate the sample efficiency of KID, Precision, and Recall to probe the impacts of the amount of generated samples. Fig. 5 presents the curves of KID, Precision, and Recall scores computed under different data regimes. Similarly, we could observe that the KID scores can be improved by synthesizing more images. Interestingly, the recall scores decrease as the generated sample size increases whereas the precision is stable. This is caused by the definition of recall: recall measures the proportion of the real distribution that is covered by the synthesized distribution. In practical computation, the denominator increases as the synthesized samples increases, while the numerator (*i.e.,* images from the real distribution) remain unchanged. In this way, the recall scores decrease as the generated sample size increases and vice versa. By contrast, CKA scores are stable under different data regimes, (please see Fig. 2 in the main paper). Moreover, CKA can provide reliable synthesis evaluation that agrees with human visual judgment. Accordingly, CKA is a proper choice for building a consistent and reliable measurement system. These results will be added in the next version of our paper.

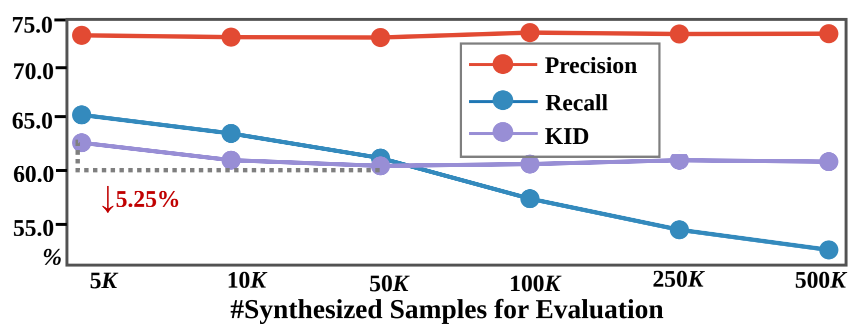


Figure 5: **Kernel Inception Distance (KID), Precision, and Recall scores evaluated under various data regimes on FFHQ dataset.** The scores are scaled for better visualization. $\downarrow$ denotes the results fluctuate downward. The percentages represent the magnitude of the numerical variation.

Q4: It will be good if the reviewer can see the dataset during the review process.

Reply: Thanks. All evaluated datasets and generative models are publicly available thanks to the original authors' generous release. For synthesized images, we either gather the pre-computed datasets from the official repositories or use public models with the official settings to generate new images for evaluation. We will make our code and evaluation scripts publicly available, making it easier to evaluate synthesis performance.

Hope that the above discussions could address your concerns, please let us know if you have any further questions. Thanks for your effort and constructive suggestions again.

To Reviewer xEcW

Q1: It would be beneficial to have a proposed metric for evaluating the performance of image translation.

Reply: Thanks. Following your valuable suggestion, we employ our measurement system to evaluate the performance of image-to-image translation. We collect publicly available image-to-image translation models that are officially released to translate images from one domain to another domain for evaluation. Specifically, three translation benchmarks are involved here, namely Horse-to-Zebra [19, 23, 13], Cat-to-Dog [21, 13], and Dog-to-Cat [21, 6]. For each benchmark, we translate the tested images to the target domain following the original experimental settings. Then we compute the distributional discrepancies between the translated images and the real target images. Tab. 7, 8, and Tab. 9 respectively present the quantitative results of the evaluated three image-to-image translation benchmarks. It can be seen from these results that CKA provides consistent ranks with FID among various extractors, and the averaged score can reflect the performance of different image translation models. For instance, the performance of CUT [13] on Horse-to-Zebra is identified better than that of CycleGAN [23] by both FID and our proposed metric. And the qualitative results in the original paper of CUT [13] also suggest that the performance of CUT surpasses CycleGAN. That is, our measurement system can provide a reliable evaluation under such settings. This indicates that our measurement system can also be used for evaluating the performance of image translation tasks. These results will be added in the next version of our paper, and we plan involve more state-of-the-art image translation models for evaluation for future work.

Table 7: **Quantitative comparison results of Centered Kernel Alignment (CKA$_\uparrow$) on Horse-to-Zebra dataset**.

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
CycleGAN [23]	83.32	73.55	88.67	85.82	83.96	74.72	73.74	80.08
AttentionGAN [19]	76.05	75.59	91.73	86.37	85.16	76.65	75.49	81.83
CUT [13]	51.29	78.48	93.22	88.83	87.84	78.75	77.36	84.08

Table 8: **Quantitative comparison results of Centered Kernel Alignment (CKA$_\uparrow$) on Cat-to-Dog dataset**.

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
CUT [13]	74.95	84.93	78.75	88.83	84.31	93.56	70.91	83.55
GP-UNIT [21]	60.96	90.45	87.79	94.05	90.12	95.91	75.32	88.94

Table 9: **Quantitative comparison results of Centered Kernel Alignment (CKA$_\uparrow$) on Dog-to-Cat dataset**.

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
GP-UNIT [21]	31.66	79.58	78.18	96.79	86.93	93.92	77.42	85.47
MUNIT [6]	18.88	84.87	84.11	98.51	88.11	95.95	86.10	89.61

Q2: Although a large amount of experiments has been conducted, this work seems to simply exploited existing findings such as advantages of CNN-transformer networks.

Reply: Thanks. In this work, we seek to develop a new measurement system that could provide reliable and consistent synthesis comparisons. In particular, two key components are crucial for the measurement system, *i.e.,* the feature extractor that defines representation space and the distributional distance that deliver similarities. Accordingly, we make in-depth analyses on the reliability and robustness of various feature extractors and different distributional distances. For the feature extractors, we gather multiple models that are pre-trained with different objectives (fully-supervised/self-supervised) and various architectures (CNN/ViT/MLP). Notably, these models are chosen for a systematic investigation to comprehensively understand the intrinsic properties of various extractors, rather than based on existing findings. Then, we testify their performance on 1) how many semantic features they can capture for evaluation, 2) how robust they are when being attacked by the histogram matching mechanism, and 3) how distinct the representation space they can define. These investigations provide several new findings to the community, including 1) one specific extractor can only capture limited semantics and provide ons-side results, 2) extractors that are vulnerable to the histogram matching attack are not reliable, and 3) different feature extractors might define similar representation spaces. For the distributional distances, we investigate the numerical stability of different distances across various representation spaces and the sample efficiency of different distances. Through extensive comparisons, we find that Centered Kernel Alignment (CKA) provides a better comparison across various extractors and hierarchical layers with its bounded score. Moreover, CKA is more sample-efficiency and exhibits better agreement with human visual judgment. Together with these findings, we build a new measurement system that can accurately reflect the synthesis performance. To the best of our knowledge, this paper is the first work to present these findings in the community of generative models. We believe that these findings can provide potential insights to further works that develop new evaluation protocols.

Q3: In addition to quality, this study should extend the metric to assess the diversity and novelty of generated samples.
Reply: Thanks. The target of generative models is to reproduce the observed data distribution, thus a good metric should accurately deliver the distributional discrepancy between the synthesized distribution and the real distribution to reflect the synthesis performance. Accordingly, our proposed evaluation system focuses on capturing the similarity between different data distributions instead of one certain aspect of the synthesized images, *e.g.,* quality and fidelity. By comparing the distributional distances between the original distribution and the synthesized distribution produced by various generative models, we can capture their actual improvement . We agree that assessing the diversity and novelty of generated samples is crucial to understand the intrinsic properties of the synthesized distributions, but this is slightly out of scope in this paper. We plan to investigate the performance of our evaluation system in assessing the synthesis diversity and novelty in our future studies as suggested.

Hope that the above discussions could address your concerns, please let us know if you have any further questions. Thanks for your effort and constructive suggestions again.

References

[1] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *Int. Conf. Learn. Represent.*, 2018.

[2] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. *Adv. Neural Inform. Process. Syst.*, 2001.

[3] M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022.

[4] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan. Masked diffusion transformer is a strong image synthesizer. *Int. Conf. Comput. Vis.*, 2023.

[5] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77, 2005.

[6] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Eur. Conf. Comput. Vis.*, pages 172–189, 2018.

[7] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up gans for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10124–10134, 2023.

[8] D. Kim, Y. Kim, W. Kang, and I.-C. Moon. Refining generative process with discriminator guidance in score-based diffusion models. *Int. Conf. Mach. Learn.*, 2023.

[9] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *Int. Conf. Mach. Learn.*, pages 3519–3529, 2019.

[10] N. Kumari, R. Zhang, E. Shechtman, and J.-Y. Zhu. Ensembling off-the-shelf models for gan training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10651–10662, 2022.

[11] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen. The role of imagenet classes in fr\'echet inception distance. In *arXiv preprint arXiv:2203.06026*, 2022.

[12] H. Liu, Z. Dai, D. So, and Q. V. Le. Pay attention to mlps. *Adv. Neural Inform. Process. Syst.*, 34: 9204–9215, 2021.

[13] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu. Contrastive learning for unpaired image-to-image translation. In *Eur. Conf. Comput. Vis.*, pages 319–345, 2020.

[14] S. Qian, H. Chang, Y. Li, Z. Zhang, J. Jia, and H. Zhang. Strait: Non-autoregressive generation with stratified image transformer. *arXiv preprint arXiv:2303.00750*, 2023.

[15] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Adv. Neural Inform. Process. Syst.*, 31, 2018.

[16] A. Sauer, K. Chitta, J. Müller, and A. Geiger. Projected gans converge faster. *Adv. Neural Inform. Process. Syst.*, pages 17480–17492, 2021.

[17] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013.

[18] T. R. Shaham, T. Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural image. In *Int. Conf. Comput. Vis.*, pages 4570–4580, 2019.

[19] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Trans. Neural Networks Learn. Syst.*, 2021.

[20] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inform. Process. Syst.*, 34:24261–24272, 2021.

[21] S. Yang, L. Jiang, Z. Liu, and C. C. Loy. Unsupervised image-to-image translation with generative prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18332–18341, 2022.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2223–2232, 2017.

Revisiting the Evaluation of Image Synthesis with GANs — *Supplementary Material*

Anonymous Author(s)
Affiliation
Address
email

This *Supplementary Material* is organized as follows: appendix A discusses the limitations of our paper and appendix B provides the implementation details of our experiments, appendix C demonstrates how human visual judgment is performed, and appendix D presents more quantitative and qualitative results.

A Limitations

Despite a comprehensive investigation, our study could still be extended in several aspects. For instance, the impacts of different low-level image processing techniques (*e.g.,* resizing) could be identified since they also play an important role in synthesis evaluation [11]. Besides, comparing datasets with various resolutions could be further studied. Nonetheless, our study could be considered an empirical revisiting towards the paradigm of evaluating generative models. We hope this work could inspire more fascinating works of synthesis evaluation and provide potential insight to develop more comprehensive evaluation protocols. We will also conduct more investigation on the unexplored factors and compare more generative models with our system.

B Implementation Details

B.1 Datasets

FFHQ [14] contains unique $70,000$ human-face images with large variations in terms of age, ethnicity, and facial expressions. We employ the resolution of $256 \times 256 \times 3$ for our experiments.

ImageNet [4] includes $1,280,000$ images with $1,000$ classes of different objects such as goldfish, bow tie, etc. All experiments on ImageNet are performed with the resolution of $256 \times 256 \times 3$ unless otherwise specified.

LSUN Church [17] consists of $126,227$ images of the church, varies in the background, perspectives, etc. We employ the resolution of $256 \times 256 \times 3$ for our experiments.

B.2 Experimental Settings and Hyperparameters

Kernel selection. We consistently employ the RBF kernel

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2})$$

for calculating the CKA. The bandwidth σ is set as a fraction of the median distance between examples. In practice, three commonly used kernels could be employed for calculation, namely linear, polynomial, and RBF kernels. In order to investigate their difference, three publicly available

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

Table 1: **CKA results with different kernels.** The publicly available models are gathered for comparison. † results are quoted from the original paper.

Kernel	InsGen-2k	InsGen-10k	InsGen-140k
FID†	11.92	4.90	3.31
Linear	99.83	99.93	99.98
Poly	99.58	99.87	99.92
RBF	95.72	98.65	99.10

Table 2: **CKA results with different features for calculation.**

Metrics	InsGen-2k	InsGen-10k	InsGen-140k
Local Features	96.62	97.42	97.38
Global Token	97.46	97.88	97.93

models with clear performance margins are collected for evaluation. Concretely, we gather models of InsGen [16] trained on FFHQ with different data regimes (*i.e.,* $2K$, $10K$, $140K$), the ranking of their synthesis quality is clear and reasonable.

Tab. 1 demonstrates the quantitative results of CKA with different kernels. Obviously, these kernels give similar results and rankings. However, the RBF kernel contributes to the distinguishability of quantitative results, making the results more comparable. Consequently, the RBF kernel is employed in our experiments.

ViT features calculation. The feature maps of ViT-based extractors are three-dimensional tensors (N, W, C), where W contains the global token and local features. The global token captures the same semantic information as the local features. Thus the global taken features are used for computation in implementation. Tab. 2 shows the comparison results of using local features and the global token. Consistently, they give similar results and rankings, so we use the global token for calculation in our experiments.

Feature normalization. In practice, the activations of features play an essential role in computing the similarity index. Namely, the quantitative results would be dominated by a few activations with large peaks, neglecting other correlation patterns [15]. To investigate the activations of our self-supervised extractors, we visualize the activations of different samples and their statistics.

Fig. 1 and Fig. 2 respectively illustrate the activation of different samples and their statistics. Obviously, there are several peaks in the activations. And these peaks may dominate the similarity index as they are substantially larger than other activations. To mitigate the peaks and create a more uniform distribution, we employ the softmax transformation [15] to normalize the features. Such operation smooths the activations while maintaining the original distributional information of features. Thus the similarity index remains consistent to deliver the distribution discrepancy. Besides the softmax transformation, we also compare the behavior of different normalization techniques (*i.e.,* L_1 and L_2 normalization).

Tab. 3 demonstrate the quantitative results with different normalization techniques. They consistently provide similar results and rankings, and the softmax transformation ameliorates the peaks more significantly, providing more comparable results. Consequently, we adopt Softmax normalization in our experiments.

Histogram matching. In order to investigate the robustness of the measurement system, we employ the histogram matching [8] to attack the system. To be specific, a subset with a considerable number (*e.g.,* $50K$) of images is chosen as the referenced distribution, and the corresponding class distribution is predicted by a given classifier (*i.e.,* Inception-V3 [13]. With the guidance of the classifier, the generator is encouraged to produce a synthesis distribution that matches the predicted class distribution of real images. Recall that the generator used to produce these synthesized distributions stays unchanged, thus a robust measurement system should give consistent similarities between the randomly generated and the matched distribution.

Table 3: **CKA scores with different normalization techniques**.

Metrics	InsGen-2k	InsGen-10k	InsGen-140k
CKA_{No}	97.46	97.88	97.93
CKA_{L1}	96.62	98.91	99.33
CKA_{L2}	96.62	98.91	99.32
$\text{CKA}_{Softmax}$	95.72	98.65	99.10

Fig. 3 provides the class distribution of real and synthesized FFHQ images predicted by Inception-V3. Obviously, the class distribution of the matched distribution is well-aligned with the predicted real distribution.

Sample-efficiency. In order to investigate the impacts of the number of synthesized samples, we compute the distributional distances between the real distribution with synthesized distributions with various numbers of generated images. Concretely, FFHQ (with $70K$ images) and ImageNet (with 1.28 million images) are investigated for universal conclusions. For both datasets, we synthesis $500K$ images as candidate, and randomly choose $5K$, $10K$, $50K$, $100K$, $250K$, and $500K$ images as the synthesized distribution for computing the metrics. The entire training data is utilized as the real distribution, and the publicly accessible models on FFHQ[1] and ImageNet[2] are employed.

The curve of FD and CKA under various data regimes on ImageNet dataset is shown in Fig. 4. Consistent with the aforementioned results in the main paper, CKA could measure the distributional distances precisely with only $5K$ samples, whereas FID fails to deliver the actual measurement until sufficient samples are used. That is, CKA could give reliable results even when limited data is given, suggesting impressive sample efficiency. Equipped with the bounded quantitative results and consistency under different data regimes, as well as the robustness to the histogram matching attack, CKA outperforms FID as a reliable distance for delivering the distributional discrepancy.

C User Preference Study

Here we present more details about our human perceptual judgment. Recall that two strategies are designed for different investigations, namely benchmarking the synthesis quality of one specific generative model and comparing two paired generative models. Fig. 5 shows the user interface for benchmarking the synthesis quality of one specific generative model (*i.e.,* BigGAN on ImageNet here). To be more specific, considerable randomly generated images are shown to the user, and the user is required to determine the fidelity of synthesized images. We then obtain the final scores by averaging the judgments of the participants (*i.e.,* 100 individuals).

Fig. 6 and Fig. 7 show the human evaluation results on FFHQ and ImageNet dataset, respectively. The percentages denote how many samples of the selected images are considered photo-realistic. Together with the quantitative results in our main paper, we could tell that the proposed metric shows a better correlation with human visual comparison.

Recall that in our main paper, we find that our evaluation system gives the opposite ranking to the existing metric (*i.e.,* FID) in some circumstances. For instance, the synthesis quality of ICGAN is determined basically the same as that of the class-conditional ICGAN (C-ICGAN) under our evaluation, whereas the FID votes C-ICGAN for the much better one. We thus conduct the other user study to compare two paired generative models. Concretely, we prepare groups of paired images of different generative models and ask 100 individuals to assess which model could produce high-quality images. The same groups are repeated several times by changing the order of images, ensuring the human evaluation is reliable and consistent.

Fig. 8 provides the interface of comparing two paired generative models, users are asked to choose which set of images looks more plausible. Additionally, Fig. 9 shows the pipeline of analyzing the

[1]https://github.com/NVlabs/stylegan3

[2]https://github.com/autonomousvision/stylegan-xl

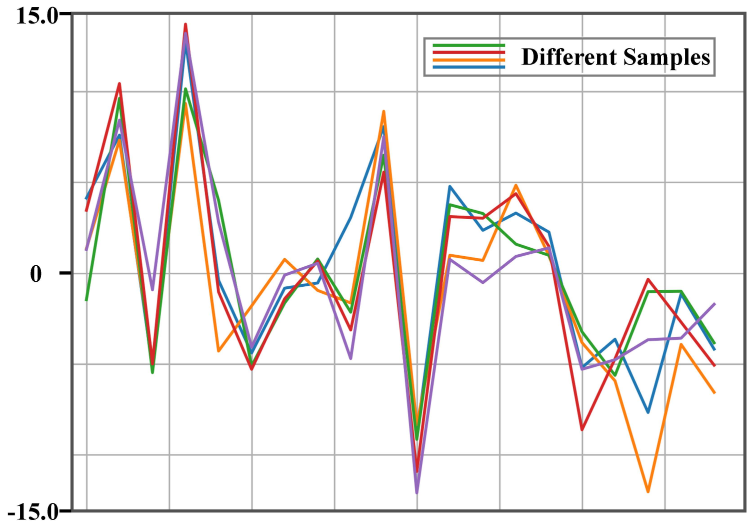


Figure 1: **Visualization of different samples' activations.** The large peaks may dominate the similarity index as their numerical values substantially surpass smaller values.

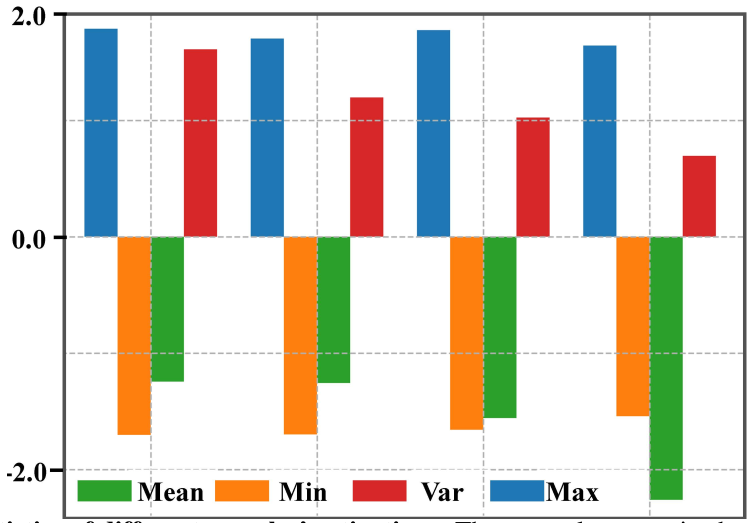


Figure 2: **Statistics of different samples' activations.** There are clear margins between different statistics (*e.g.,* Max and Min) of each sample, suggesting that the activation distribution is very peaky.

paired comparison results. Specifically, the same groups of images are repeated for 4 times in random order and users are shown 16 images from two models to determine the more photorealistic one. In this way, the results of choosing both Projected-GAN and StyleGAN2 two times are identified as indistinguishable for enduring the consistency. Namely, the users choose different rankings between the two sets when the order of images is changed, which does not meet the consistency. Consequently, the final scores for paired comparison are obtained by quantifying the percentage of the human preferences that correlate the consistency.

D More Quantitative and Qualitative Results

In this section, we further provide more results of different semantic levels from various extractors and the curve of different distances evaluated on various data regimes.

Similarities between various representation spaces. Recall that we filtered out extractors that define similar representation spaces to avoid redundancy in the main paper. The correlation between representations of high dimension in different feature extractors is calculated following [7]. In particular, a considerable number of images (*i.e.,* $10K$ images from ImageNet) are fed into these

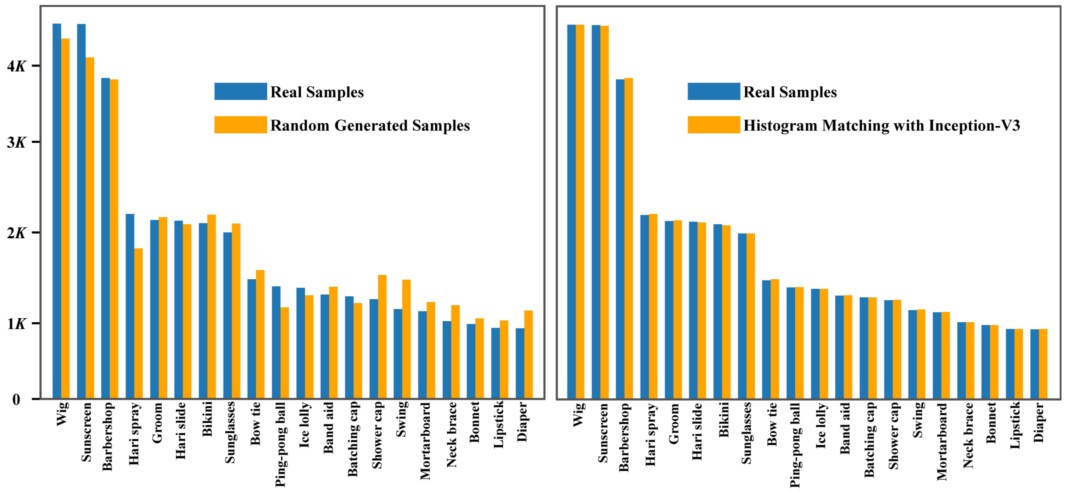


Figure 3: **The class distribution of randomly generated images (*left*)** and **histogram matched images (*right*),** predicted by the fully-supervised Inception-V3 [13].

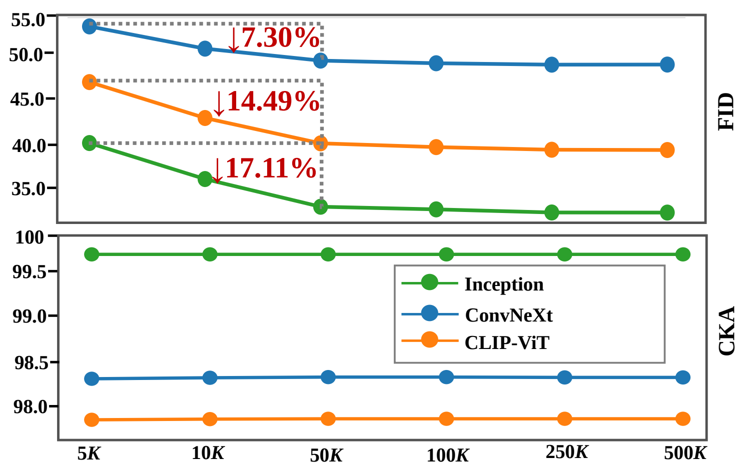


Figure 4: **Fréchet Distance (FD) and Centered Kernel Alignment (CKA) scores evaluated under various data regimes on ImageNet dataset.** FID scores are scaled for better visualization. ↓ denotes the results fluctuate downward. The percentages represent the magnitude of the numerical variation.

extractors for computing their correspondence. Fig. 10 shows the similarity of their representations. Obviously, the representations of CLIP-ResNet and MoCo-ResNet have higher similarity with other extractors. Considering these two extractors are both CNN-based and they capture similar semantics with other CNN-based extractors, we remove the CLIP-ResNet and MoCo-ResNet to avoid redundancy. Accordingly, we obtain a set of feature extractors that 1) capture rich semantics in a complementary way, 2) are robust toward the histogram matching attack, and 3) define meaningful and distinctive representation spaces for synthesis comparison. The following table presents these feature extractors. These extractors, including both CNN-based and ViT-based architectures,

CNN-based	ConvNeXt [9], SWAV [2], RepVGG [5]
ViT-based	CLIP-ViT [12], MoCo-ViT [3], ViT [6]

have demonstrated strong performance in pre-defined and downstream tasks, facilitating more comprehensive and reliable evaluation. Notably, the inclusion of self-supervised extractors SWAV, CLIP-V, and MoCo-V aligns with previous findings [10, 8, 1]. This selection of feature extractors provides a diverse and complementary set of representations, enabling a more comprehensive and reliable evaluation of synthesis quality in generative models.

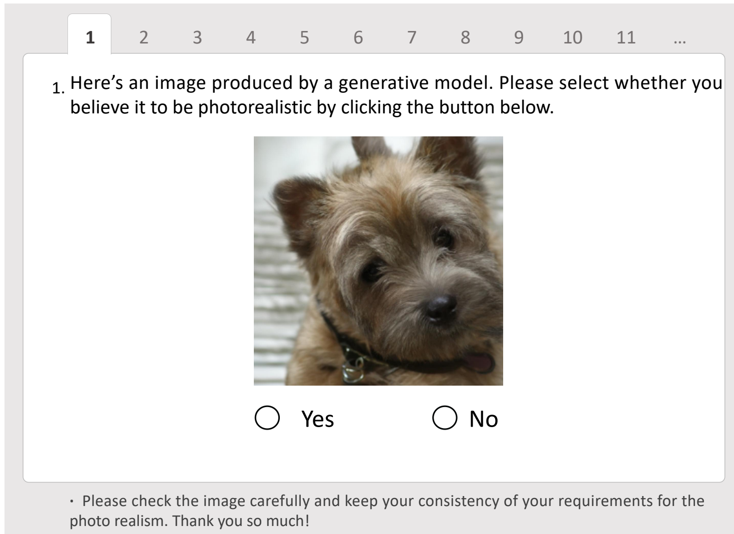
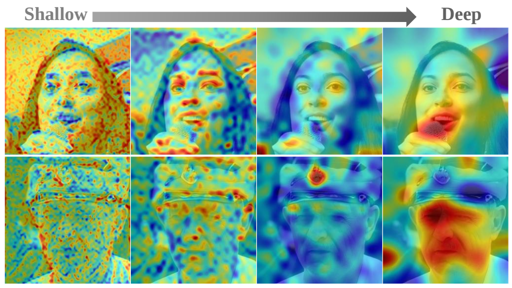


Figure 5: **User interface for benchmarking the synthesis quality.**

Figure 6: **Human judgment results of various generative models on FFHQ.** $2K$ images randomly generated by different models are selected for comparison.

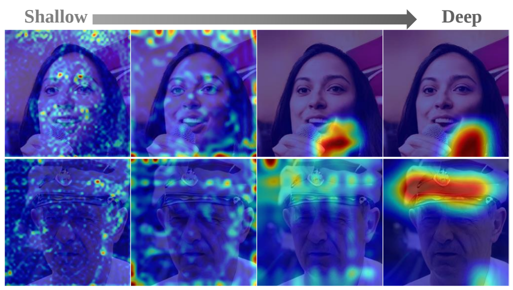
More results of hierarchical levels from various extractors. Tab. 4, Tab. 5, Tab. 6, Tab. 7, and Tab. 8 respectively present the heatmaps and quantitative results of various semantic levels. We could tell that despite the Fréchet Distance (FD) scores consistently reflect synthesis quality, their numerical values fluctuate dramatically. On the contrary, CKA provides normalized distances *w.r.t* the numerical scale across various levels. Also, the heatmaps from various semantic levels reveal that hierarchical features encode different semantics. Such observation provides interesting insights that feature hierarchy should be also considered for synthesis comparison. Notably, benefiting from the bounded quantitative results, CKA demonstrates great potentials for comparison across hierarchical layers.

Table 4: **Heatmaps from various semantic levels on FFHQ dataset (*left*)** and **quantitative results of Fréchet Distance (FD $\downarrow$) and Centered Kernel Alignment (CKA $\uparrow$) on ImageNet dataset (*right*)** . ConvNext [9] serves as the feature extractor for hierarchical evaluation here.



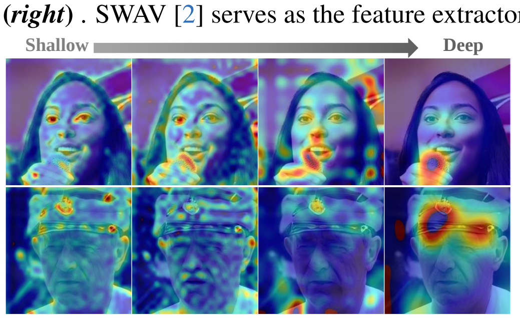
Model	BigGAN		BigGAN-deep		StyleGAN-XL	
Layer	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$
Layer$_1$	2.64	96.08	2.56	96.35	0.58	98.24
Layer$_2$	40.20	-	32.32	-	11.84	-
Layer$_3$	687.40	58.76	364.95	60.25	264.87	62.53
Layer$_4$	140.04	68.86	102.26	69.27	19.22	70.52
Overall	N/A	74.57	N/A	75.29	N/A	77.10

Table 5: **Heatmaps from various semantic levels on FFHQ dataset (*left*)** and **quantitative results of Fréchet Distance (FD $\downarrow$) and Centered Kernel Alignment (CKA $\uparrow$) on ImageNet dataset (*right*)** . RepVGG [5] serves as the feature extractor for hierarchical evaluation here.



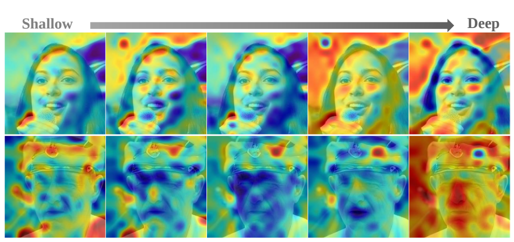
Model	BigGAN		BigGAN-deep		StyleGAN-XL	
Layer	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$
Layer$_1$	0.35	96.92	0.32	97.51	0.04	98.79
Layer$_2$	0.35	96.19	0.33	96.32	0.03	98.32
Layer$_3$	0.23	90.05	0.18	91.15	0.04	93.51
Layer$_4$	67.53	74.40	58.85	76.68	15.93	80.28
Overall	N/A	89.39	N/A	90.42	N/A	92.73

Table 6: **Heatmaps from various semantic levels on FFHQ dataset (*left*)** and **quantitative results of Fréchet Distance (FD $\downarrow$) and Centered Kernel Alignment (CKA $\uparrow$) on ImageNet dataset (*right*)** . SWAV [2] serves as the feature extractor for hierarchical evaluation here.



Model	BigGAN		BigGAN-deep		StyleGAN-XL	
Layer	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$
Layer$_1$	0.67	99.90	0.46	99.91	0.07	99.99
Layer$_2$	0.87	97.89	0.60	98.87	0.31	99.51
Layer$_3$	16.15	95.60	12.02	96.21	1.90	98.15
Layer$_4$	11.18	86.10	8.69	87.71	1.85	92.54
Overall	N/A	94.87	N/A	95.68	N/A	97.55

Table 7: **Heatmaps from various semantic levels on FFHQ dataset (*left*)** and **quantitative results of Fréchet Distance (FD $\downarrow$) and Centered Kernel Alignment (CKA $\uparrow$) on ImageNet dataset (*right*)** . ViT [6] serves as the feature extractor for hierarchical evaluation here.



Model	BigGAN		BigGAN-deep		StyleGAN-XL	
Layer	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$
Layer$_1$	0.20	99.62	0.19	99.67	0.01	99.97
Layer$_2$	1.31	97.75	1.19	97.92	0.18	99.76
Layer$_3$	6.93	97.53	6.06	97.63	1.22	99.67
Layer$_4$	29.95	96.49	23.98	97.20	8.51	98.72
Overall	N/A	97.85	N/A	98.11	N/A	99.53

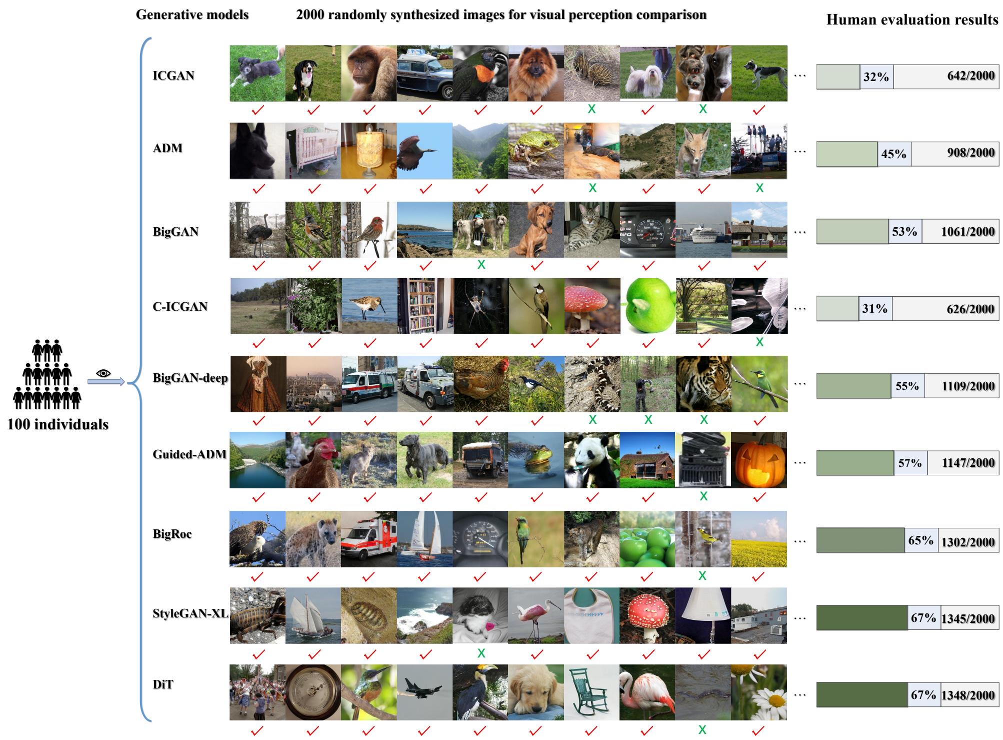


Figure 7: **Human judgment results of various generative models on ImageNet.** $2K$ images randomly generated by different models are selected for comparison.

Table 8: **Heatmaps from various semantic levels on FFHQ dataset (*left*)** and **quantitative results of Fréchet Distance (FD ↓) and Centered Kernel Alignment (CKA ↑) on ImageNet dataset (*right*)** . MoCo-ViT [3] serves as the feature extractor for hierarchical evaluation here.

Model	BigGAN		BigGAN-deep		StyleGAN-XL	
Layer	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$	$\text{FD}_{\downarrow}$	$\text{CKA}_{\uparrow}$
Layer_1	0.10	98.62	0.05	99.04	0.04	99.97
Layer_2	1.01	97.15	0.68	97.30	0.43	99.64
Layer_3	9.18	96.07	9.01	96.77	4.30	99.11
Layer_4	3.35	97.25	3.22	97.82	1.85	99.00
Overall	N/A	97.27	N/A	97.73	N/A	99.43

References

[1] E. Betzalel, C. Penso, A. Navon, and E. Fetaya. A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*, 2022.

[2] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inform. Process. Syst.*, pages 9912–9924, 2020.

[3] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, pages 9640–9649, 2021.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009.

[5] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun. Repvgg: Making vgg-style convnets great again. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13733–13742, 2021.

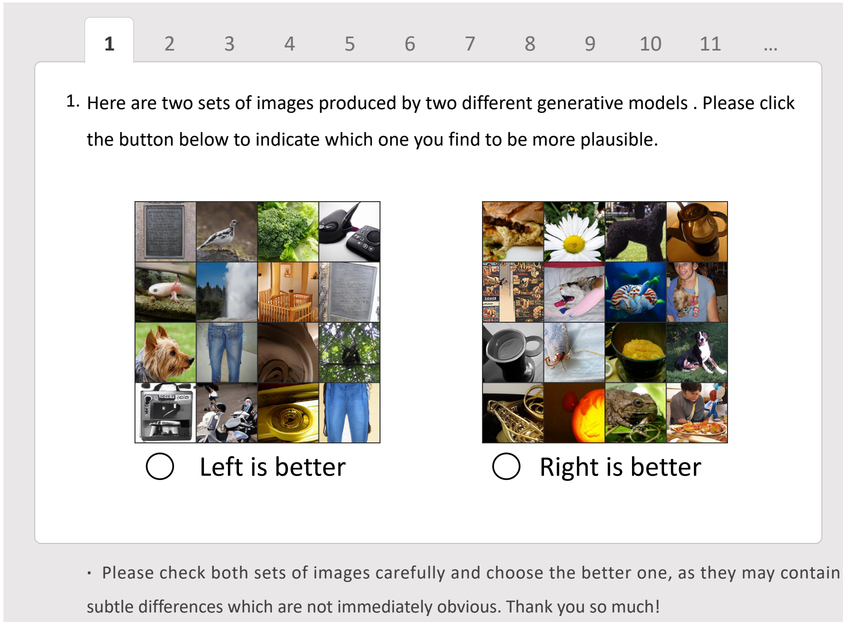


Figure 8: **User interface for comparing the synthesis quality of two paired generative models.** People are asked to determine which set of images look more photorealistic.

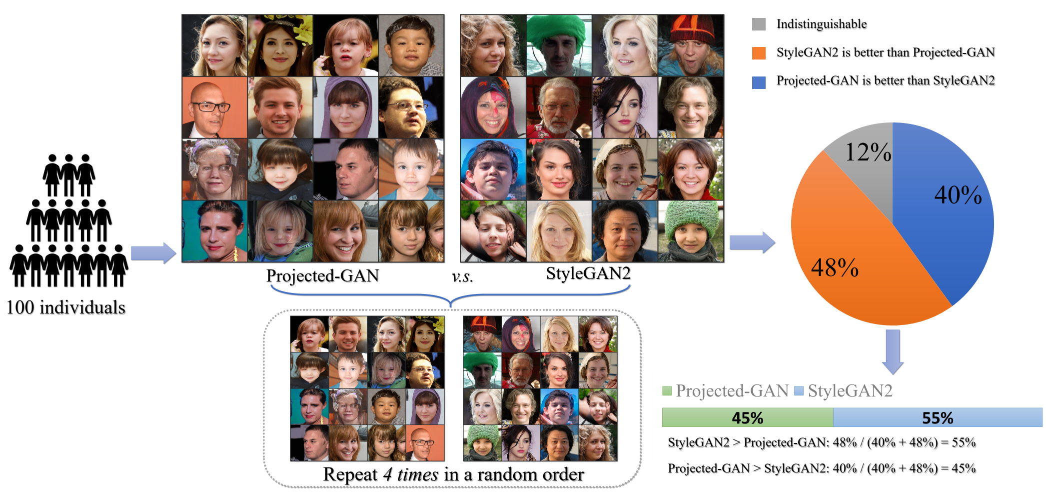


Figure 9: **The pipeline of analyzing the paired comparison results.**

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020.

[7] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *Int. Conf. Mach. Learn.*, pages 3519–3529, 2019.

[8] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen. The role of imagenet classes in fr\'echet inception distance. In *arXiv preprint arXiv:2203.06026*, 2022.

[9] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11976–11986, 2022.

[10] S. Morozov, A. Voynov, and A. Babenko. On self-supervised image representations for gan evaluation. In *Int. Conf. Learn. Represent.*, 2021.

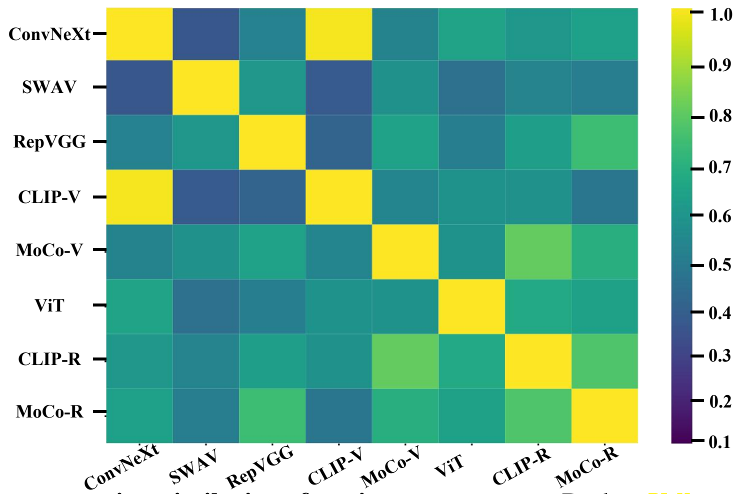


Figure 10: **Representation similarity of various extractors.** Darker Yellow denotes higher similarity.

[11] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11410–11420, 2022.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2818–2826, 2016.

[14] T. A. Tero Karras, Samuli Laine. Flickr-faces-hq dataset (ffhq). URL https://github.com/NVlabs/ffhq-dataset.

[15] P. Wang, Y. Li, and N. Vasconcelos. Rethinking and improving the robustness of image style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 124–133, 2021.

[16] C. Yang, Y. Shen, Y. Xu, and B. Zhou. Data-efficient instance generation from instance discrimination. *Adv. Neural Inform. Process. Syst.*, pages 9378–9390, 2021.

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