ANID: How Far Are We? Evaluating the Discrepancies Between AI-synthesized Images and Natural Images through Multimodal Guidance

Renyang Liu[1,2*], **Ziyu Lyu**[1†], **Wei Zhou**[3], **See-Kiong Ng**[2]

[1]School of Cyber Security, Sun Yat-sen University, China
[2]National University of Singapore, Singapore
[3]Research Center of Cyberspace, Yunnan University, China
ryliu@nus.edu.sg, lvzy7@mail.sysu.edu.cn, zwei@ynu.edu.cn, seekiong@nus.edu.sg

Abstract

In the rapidly evolving field of Artificial Intelligence Generated Content (AIGC), one of the key challenges is distinguishing AI-synthesized images from natural images. Despite the remarkable capabilities of advanced AI generative models in producing visually compelling images, significant discrepancies remain when these images are compared to natural ones. To systematically investigate and quantify these discrepancies, we introduce an AI-Natural Image Discrepancy Evaluation benchmark aimed at addressing the critical question: *how far are AI-generated images (AIGIs) from truly realistic images?* We have constructed a large-scale multimodal dataset, the Distinguishing Natural and AI-generated Images (DNAI) dataset, which includes over 440,000 AIGI samples generated by 8 representative models using both unimodal and multimodal prompts, such as Text-to-Image (T2I), Image-to-Image (I2I), and Text *vs.* Image-to-Image (TI2I). Our fine-grained assessment framework provides a comprehensive evaluation of the DNAI dataset across five key dimensions: naive visual feature quality, semantic alignment in multimodal generation, aesthetic appeal, downstream task applicability, and coordinated human validation. Extensive evaluation results highlight significant discrepancies across these dimensions, underscoring the necessity of aligning quantitative metrics with human judgment to achieve a holistic understanding of AI-generated image quality. Code is available at https://github.com/ryliu68/ANID.

Introduction

With the rapid advancement of deep learning techniques, the proliferation of Artificial Intelligence Generated Content (AIGC) has garnered significant attention across various domains, such as e-commerce, gaming, medicine, animation, and autonomous driving (Li et al. 2024; Qian et al. 2024). AI-generated Images (AIGI) have been one of the mainstream forms of AIGC, and a variety of AI image generative models have been proposed to make the generated synthetic images as realistic as natural images, ranging from the earlier generative adversarial networks (GANs) (Tao et al. 2023, 2022), advanced diffusion models (DMs) (Xu et al. 2023b; Wei et al. 2023) to the large multimodal generative model like DALL·E (Ramesh et al. 2022).

Despite the proliferation of AI image generative models, AI-generated images are not qualified for real-world applications, and the discrepancies between generated synthetic images and realistic natural images still exist (Li et al. 2023a). Therefore, many AI-generated image quality assessment and evaluation methods have been proposed (Wang et al. 2023; Xu et al. 2023a; Hu et al. 2023). For example, Wang et al. (Wang et al. 2023) established an AIGC Image Quality Assessment Database called AIGCIQA2023 from 6 text-to-image generative models, and devised a subjective evaluation framework to assess human visual preferences for each AI-generated image from three aspects, e.g., quality, authenticity, and text-to-image correspondence. PKU-I2IQA (Yuan et al. 2023) constructed a human perception-based image-to-image database named PKU-I2IQA and conducted a subject analysis based on both no-reference and full-reference methods. Different from prior studies focusing on establishing the AIGI dataset and conducting a subjective evaluation with human feedback, another group of studies (e.g. Pick-a-Pic (Kirstain et al. 2023), HPS v2 (Wu et al. 2023), TIFA (Hu et al. 2023), and ImageReward (Xu et al. 2023a)) focused on training a unified score model as the automatic evaluation metrics for AIGC image quality assessment (AIGIQA) measures, in contrast with traditional image quality assessment for natural images measures like Inception Score (Salimans et al. 2016) and FID (Heusel et al. 2017).

However, some issues and problems have not yet been fully answered and solved. **First**, prior studies only considered unimodal promopted generation, e.g. Text-to-Image (T2I) or Image-to-Image (I2I), ignoring multimodal prompted content like Text vs.Image-to-Image (TI2I). The data size of the established T2I or I2I dataset is relatively small (around thousands of images), and might be insufficient for comprehensive analysis and evaluation. **Second**, prior studies typically focused on the perceptual quality of AIGC images with traditional image quality assessment metrics, the text-to-image correspondence, and subjective human preference. No studies have fully answered the important questions about what discrepancies exist between

*This work was conducted while Renyang Liu was a Research Intern at the School of Cyber Security, Sun Yat-sen University.

†Corresponding author: Ziyu Lyu (lvzy7@mail.sysu.edu.cn).

generated synthetic images and realistic natural images, and they lacked a systematic and compressive assessment and evaluation to investigate and interpret the discrepancies between AI-generated synthetic images and realistic natural images. The investigation and understanding of the discrepancies are key ways to realize the practical application of AIGC images in real-world scenarios and make breakthroughs in AIGC.

To address the above challenges, we proposed an AI-Natural Image Discrepancy Evaluation Benchmark to investigate and interpret what discrepancies still remained between AI-generated images and natural images, finally answering the important question "How far are AI generative models with respect to the visual forms of images?". Especially, we have two important contributions with reference to the key problems. **Large Multimodal Evaluation Dataset construction:** we have curated a large multimodal evaluation dataset called **DNAI** (Distinguishing Natural and AI-generated) dataset, with about 440,000 AIGI generated from 8 representative generative models based on both unimodal and mutimodal guidance including Text-to-Image (T2I), Image-to-Image (I2I), and Text *vs.* Image-to-Image (TI2I). The data size of our DNAI dataset scales to 100x that of prior datasets. **Fine-grained Evaluation Framework**, we propose a fine-grained evaluation framework to conduct the systematic and comprehensive assessment and evaluation of DNAI, covering 5 diverse and important aspects, including naive visual feature quality, semantic alignment among multimodal generation, aesthetic appeal, downstream applicability, and the coordinated human validation.

Through our DNAI dataset and fine-grained evaluation framework, we conduct extensive benchmark analysis and evaluation and conclude key insights to answer the discrepancy questions:

- **Significant discrepancies in key Areas:** *AI-generated images exhibit substantial discrepancies from natural images in terms of quantitative measures from all aspects, with about 10% to 30%.*
- **Multimodal Alignment:** Different prompted generation might have different semantic alignment scores. *Generated images prompted with texts including both T2I and TI2I demonstrated remarkable semantic alignments.*
- **Downstream Task Applicability:** *Significant differences exist in the usability of AI-generated images vs. natural ones in downstream tasks, highlighting the need for further improvements of generative models to ensure practical applicability in real-world scenarios.*
- **Human Evaluation *vs.* Quantitative Metrics:** *Human evaluation results reveal larger discrepancies compared with quantitative metrics.* It validates the necessity of incorporating human evaluation for coordination.

Related Work

Natural Image Evaluation: Over the past decades, numerous image quality assessment methods have been developed to evaluate traditional natural images (N. et al. 2015). These methods have focused on various visual features and properties, such as perceptual appearance (Zhang et al. 2018), naturalness (Ma et al. 2018), and aesthetics (Esfandarani and Milanfar 2018). For instance, BRISQUE evaluates natural image quality by analyzing spatial domain features, while PIQE assesses perceptual quality through block-based image segmentation, both serving as effective no-reference metrics. Other notable measures include FID and Inception Scores, which assess the quality and diversity of generated images, and SSIM (Wang et al. 2004) and PSNR, which quantify structural similarity and pixel-level accuracy. Beyond quantitative measures, human evaluations have also been leveraged to estimate natural image quality. For example, NIMA (Esfandarani and Milanfar 2018), proposed by Talebi and Milanfar, uses a deep CNN trained on human-rated images to predict aesthetic quality. Similarly, Wong et al. introduced the AVA dataset, which facilitates aesthetic assessment based on human preferences (Murray, Marchesotti, and Perronnin 2012).

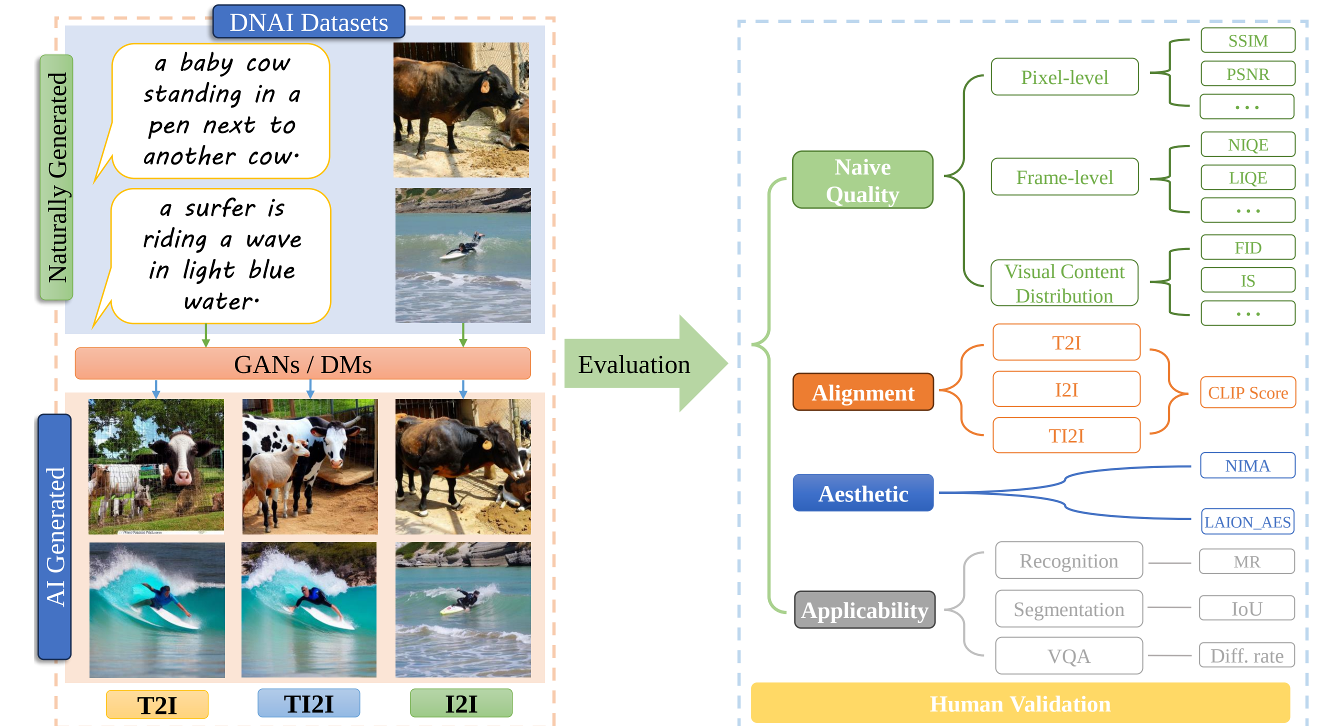
AI-generated Image Evaluation: Recently, AI-generated content (AIGC) has seen significant advancements with the rise of generative models. Several methods and benchmarks have been introduced to evaluate AI-generated images (Zhang et al. 2023; Hu et al. 2023). These methods primarily focus on three aspects: perceptual quality, text-image correspondence, and aesthetics (Xu et al. 2023a; Wang et al. 2023), often training unified models to assess the overall quality of AI-generated images. For example, ImageReward (Xu et al. 2023a) presents a framework for aligning image generation with human preferences, AGIQA-3k (Li et al. 2023a) offers a comprehensive benchmark for assessing AI-generated image quality, and QBench (Wu et al. 2024) evaluates content quality across multiple dimensions.

Despite advancements, little research has delved into the fine-grained differences between AI-generated and natural images, and existing datasets are often too small for comprehensive evaluation. To address these gaps, we propose an AI-Natural Image Difference Evaluation Benchmark, featuring a large Distinguishing AI-Natural Image Dataset and a systematic, fine-grained evaluation framework to thoroughly explore these differences.

AI-Natural Image Discrepancy Evaluation Benchmark

To tackle the challenges of current AI-generated image evaluation, we construct an AI-Natural Image Discrepancy evaluation benchmark (**ANID**) to evaluate the potential discrepancies between AI-generated images and natural images. Figure 1 shows the overview of our ANID benchmark.

Our ANID benchmark contributes to two core components: (1) A Distinguishing AI-Natural Image (DANI) Dataset is constructed, in which we collect AI-generated images from diverse generative models for real natural images in MS COCO dataset (Lin et al. 2014) based on three types of guidance prompts e.g. text-only, image-only and text-image prompts; (2) We devise a systematic and comprehensive evaluation framework to measure and evaluate the potential differences between AI-generated images and real



(a) Dataset (b) Framework

Figure 1: **Overview of AI-Natural Image Discrepancy Evaluation Benchmark (ANID):** (a) A large Distinguishing AI-Natural Image (DANI) dataset is curated, guided by three types of prompts. (b) A fine-grained evaluation framework is proposed to measure and evaluate the potential discrepancies between AI-generated images and natural images from five aspects covering naive image quality, semantic alignment, aesthetic appeal, downstream applicability, and coordinated human assessment.

natural images from five aspects including naive image quality, semantic alignment, aesthetic appeal, downstream applicability, and the coordinated human assessment. In the following parts, we introduce the DANI Dataset and the evaluation framework in detail.

Distinguishing AI-Natural Image Dataset

We construct the Distinguished Natural and AI-generated Image Dataset based on the classical natural image dataset MS COCO dataset. We select 5,000 images covering diverse captions from the MS COCO dataset and collect a total of 25,000 text-image pairs by pairing each image with five different pieces of text descriptions. The 25,000 text-image pairs are seen as the **referenced natural image set**.

For each text-image sample in the referenced natural image set, we collect AI-generated images from 8 representative generative models, guided by three types of prompts, including text-only prompts, image-only prompts, and text-image prompts. The 8 representative generative models are from diverse generative models ranging from the earlier generative adversarial networks (GANs) (e.g. GALIP (Tao et al. 2023), DF-GAN (Tao et al. 2022)) to the recent benchmarking diffusion models (DMs) like Stable Diffusion v1.4, v1.5, v2.1, XL (Rombach et al. 2022), Versatile-Diffusion (Xu et al. 2023b) and OpenAI DALL·E 2 (Ramesh et al. 2022). The referenced natural images and the generated images comprise our distinguishing AI-Natural image dataset. More details about data construction and collection are presented in Section Appendix A.

Table 1 shows the statistics and properties of our DANI Dataset compared with the existing evaluation dataset. Our DANI Dataset has a total of 440,000 AI-generated images from 8 representative generated models. For details, please refer to Appendix Sec. A. Compared with state-of-the-art evaluation datasets, our DANI Dataset has convincing contributions:

- DANI Dataset is the most extensive dataset for AI-generated image evaluation, and the **data size scales to 100x**.
- DANI Dataset is guided by both **unimodal prompts and multi-modal prompts (text-image)**. Prior datasets barely considered text prompts.
- Diverse generative models are exploited, ranging from the earlier GAN models, the recent popular diffusion models, and the influential large commercial generative model DALL·E 2 issued by OpenAI.

Our large and comprehensive dataset works as the foundation of the evaluation benchmark and enables thorough evaluation and analysis of the differences between AI-generated images and realistic natural images.

Dataset	Model	Text	Image	Text *vs.* Image	AIGI
AGIQA-1K	2	1,080	-	-	1,080
AGIQA-3K	6	300	-	-	2,982
AIGCIQA2023	6	100	-	-	2,400
PKU-I2IQA	2	200	200	-	1,600
DNAI (Ours)	**8**	**25,000**	**5,000**	**25,000**	**440,000**

Table 1: Comparison of DANI dataset and Counterparts.

Fine-Grained Evaluation Framework

On top of the DANI dataset, we construct a fine-grained evaluation framework to systematically measure and interpret the differences between AI-generated images and realistic natural images from five different aspects, including naive image quality, semantic alignment, aesthetic appeal, downstream applicability, and coordinated human assessment.

1. Naive Image Quality: We leverage traditional image quality assessment methods for natural images to measure the AI-Natural image differences. In order to perform fine-grained analysis and exploration, we estimate the AI-Natural image differences from three levels by considering the inherent properties of image-style content, ranging from the **low-level visual features, frame-level visual features to the holistic content distribution**.

- First, we use the **pixel-level similarity measures** like SSIM, LPIPS, and DISTS to compute the **visual similarities** between the paired AI-generated image and natural image and use the inverse similarity values to quantify the AI-Natural image differences from the low-level visual view.
- Second, we use **frame-level visual features** to qualify the perceptional quality of images and compare the global-level visual quality differences between AI-generated images and natural images from the high-level visual view. The exploited frame-level visual features include PIQE (N. et al. 2015), IL-NIQE (Zhang, Zhang, and Bovik 2015), MUSIQ (Ke et al. 2021), DBCNN (Zhang et al. 2020), LIQE (Ma et al. 2018), Inception Score (Salimans et al. 2016), CLIPIQA (Radford et al. 2021a), TReS (Gu et al. 2015), HyperIQA (Su et al. 2020), UNIQUE (Zhang et al. 2021), BRISQUE , NIQE (Mittal, Soundararajan, and Bovik 2013), NRQM (Ma et al. 2018), etc. These metrics provide insights into global-level visual quality differences.
- Third, we investigate the **visual content distribution** to capture the holistic differences between AI-generated images and natural images by computing the FID and Inception Score of the whole AI-generated image dataset and the referenced natural image set.

Based on the different levels of image quality measures, our framework can evaluate the fine-grained AI-Natural image differences, covering both low-level and high-level. More details about the utilized traditional image quality assessment measures are described in Appendix Sec. C.

2. Semantic Alignment: Generative models require the referenced prompts as semantic guidance to generate images. Therefore, semantic alignment can be an important quality indicator of the AIGIs. We use the widely used CLIP model (Radford et al. 2021b) to measure semantic alignment and compare the different CLIP Scores from natural text, image, and the paired AI-generated image-text prompt. The studied AI-generated images were originally generated from the 8 Representative generative models based on three types of prompts as described in Appendix Sec. B; the semantic alignment evaluation can provide a fine-grained analysis of the potential differences in semantic alignment with respect to both unimodal and multimodal guidance.

3. Aesthetic Appeal: Aesthetic Appeal evaluation is to estimate the visual appeal and artistic quality of images, reflecting the visual attractiveness and artistic quality of images. We utilize the classical aesthetic measures NIMA (Esfandarani and Milanfar 2018) and LAION-AES (Schuhmann et al. 2022) as the quantitative metrics and compare the aesthetic scores for the paired AI-generated and natural images.

4. Downstream Applicability: This aspect is devised to investigate the practical utility of AI-generated images in downstream tasks and evaluate whether AI-generated images can have different practical utilities with realistic natural images in downstream application tasks. We mainly focus on two classical downstream tasks, e.g., image recognition and object segmentation. For the image recognition task, we evaluate the Classification Mismatch Rate (MR) rates between AI-generated images and natural images when using a pre-trained image recognizer like ResNet-152 (He et al. 2016) model. For the object segmentation task, we use the Intersection over Union (IoU) metric (Everingham et al. 2010) as quantitative measures and investigate the segmentation results for AI-generated images with the segmentation model U2NET (Qin et al. 2020).

5. Human Assessment: We involve human assessments to coordinate the above evaluation aspects. We develop the human assessment interface to demonstrate AI-generated images with the referenced natural images and text descriptions and collect human ratings (on a scale from 1 to 5) alongside the following three aspects: naive image quality, semantic alignment, and Aesthetic Appeal as human assessments. Human participants are provided with example images before the evaluation to understand high and low scores without being given specific numerical values, ensuring unbiased and informed ratings. Detailed information about human assessment procedures is provided in Appendix Sec. D.

Through the integration of diverse and comprehensive evaluation dimensions, our fine-grained evaluation framework offers a systematic assessment solution to investigate and interpret the differences that still remain between AI-generated images and natural images.

Evaluation

With respect to the 5 different aspects, we conducted benchmark experiments and performed experimental evaluation

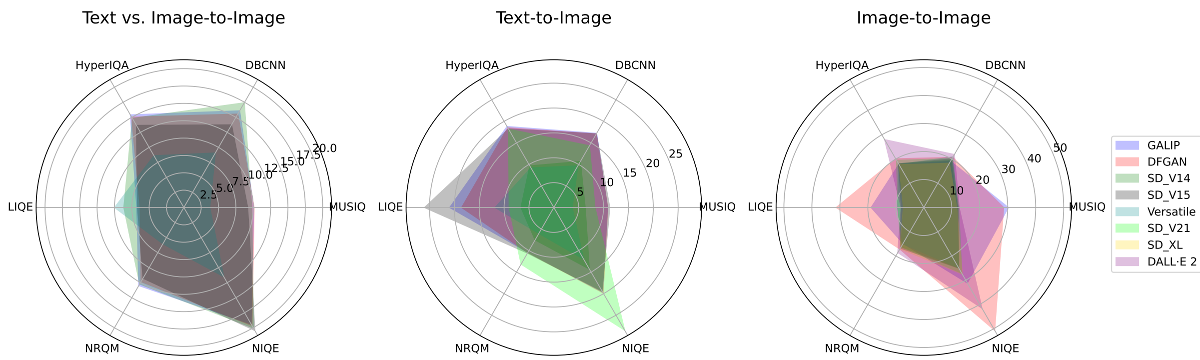


Figure 2: We evaluate eight image-generating models from 6 aspects. The numerical values in the radar chart represent the mean difference rate of each model.

alongside the following research questions.

RQ1: What are the fine-grained discrepancies between AIGIs and natural images?

- **RQ1.a:** What discrepancies between I2I-guided AIGIs and their natural counterparts can be interpreted via the pixel-level quality measures?
- **RQ1.b:** What discrepancies between multimodal-guided AIGIs and their natural counterparts can be revealed via the frame-level metrics?
- **RQ1.c:** How do the structural visual content distributions differ between AIGIs and natural images?

RQ2: How significant are the discrepancies in semantic alignment between AIGIs and natural images in different types of guided prompts?

RQ3: What are the discrepancies in aesthetic appeal between AIGIs and natural images?

RQ4: How do AIGIs differ from natural images in downstream task applicability?

RQ5: Are human assessment results consistent with quantitative measures? What are the discrepancies revealed from human evaluation?

Experimental Setting

We use the described quantitative evaluation metrics, and details of all metrics are presented in Appendix Sec. B.

Full-reference metrics: For full-reference metrics like pixel-level image quality, we directly report the calculated value cause it already can present the discrepancy.

No-reference metrics We calculate *Difference Rate (DR)* for each quantitative metric in non-reference scenarios, by subtracting the value of each AI-generated image from its corresponding natural reference image and then averaging these difference rates, defined as follows:

$$\textit{Difference Rate} = \frac{1}{n}\sum_{i=1}^{n}\frac{|m(\mathbf{X}_i) - m(\mathbf{N}_i)|}{m(\mathbf{N}_i)} \qquad (1)$$

where n is the number of images, m represents the metric, $\mathbf{X}$ denotes the generated images, and $\mathbf{N}$ refers to the referenced natural images.

	SD_V14	SD_V15	Versatile	SD_V21	SD_XL	DALL·E 2
SSIM ↑	0.35	0.35	0.46	0.44	0.70	0.38
PSNR ↑	13.81	13.80	16.52	16.00	23.25	12.42
VIF ↑	0.02	0.02	0.05	0.04	0.17	0.03
VSI ↑	0.84	0.84	0.89	0.87	0.96	0.83
FSIM ↑	0.61	0.61	0.71	0.67	0.86	0.61
LPIPS ↓	0.60	0.60	0.39	0.52	0.19	0.61
DISTS ↓	0.27	0.27	0.17	0.23	0.10	0.24
MAD ↓	211.61	211.73	196.86	202.29	143.66	214.86

Table 2: Full-reference Evaluation Results (I2I).

RQ1: Naive Quality Results

To evaluate the image quality of AI-generated images versus natural images, we conducted a comprehensive analysis using a variety of metrics that assess visual feature quality, naturalness, and similarity of images. ***Our findings reveal substantial discrepancies between AIGIs and natural images across these dimensions***.

RQ1.a Pixel-level We evaluated the visual similarity between AI-generated and natural images using structural metrics like SSIM, LPIPS, DISTS, and PSNR. The results, as shown in Table 2, indicate that ***AI-generated images exhibit significantly lower similarity to natural counterparts***. AI-generated images show a 20% to 50% reduction in SSIM, highlighting a major loss in structural fidelity. LPIPS and DISTS further reveal notable perceptual dissimilarities, with LPIPS scores ranging from 0.19 to 0.60. Additionally, PSNR values, between 12.42 and 23.25, suggest that AI-generated images have higher noise levels, reducing their overall visual similarity.

RQ1.b Frame-level We assessed the frame-level quality of AI-generated and natural images using a suite of metrics, including MUSIQ, DBCNN, HyperIQA, LIQE, BRISQUE, NRQM and NIQE. Figure 2 uses radar figures to visualize the results from the six metrics, and indicates that ***there are significant quality differences between AI-generated and natural images***. These deviations typically range from

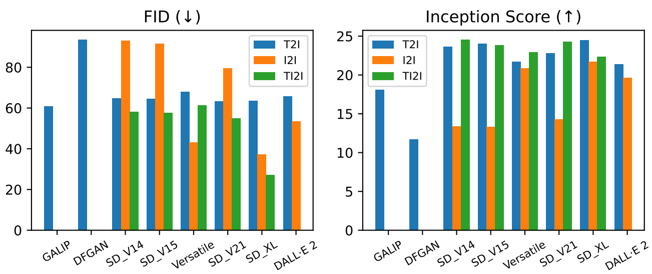


Figure 3: The FID and Inception Score of AI-generated images from different source models and guidance.

10% to 20%, irrespective of the guidance method used, with some instances showing quality differences as high as 30%. This underscores the considerable gaps between AIGIs and natural images in terms of global image quality. Moreover, the results of NRQM and NIQE reveal that ***AI-generated images significantly differ from natural images in terms of naturalness***. Specifically, AI-generated images show notable deviations from natural images in visual realism and adherence to natural scene statistics. The NRQM difference rates range from 5% to 20%, while the NIQE difference rates range from 10% to 50%, highlighting the challenges AI-generated images face in replicating the inherent naturalness of real-world scenes. More detailed results with more metrics are presented in Appendix Sec. C.

RQ1.c Visual Content Distribution We present the FID and Inception Score results, presented in Figure 3, suggesting that AI-generated images often exhibit relatively higher FID scores and lower Inception Scores. This finding indicates that current generative models, particularly GAN-based models and earlier versions of Stable Diffusion, struggle to consistently produce high-quality images. Notably, these models tend to generate better images when a reference image is provided (TI2I), as it has more detailed guidance.

In conclusion, our evaluation of image visual feature quality, naturalness, and similarity reveals significant disparities between AI-generated and natural images. These differences highlight the current limitations of AIGC technologies in matching the visual quality and realism of natural images.

RQ2: Alignment Results

In our evaluation of alignment results, we assessed how well AI-generated images correspond to their respective reference prompts across different guidance types: Text-to-Image (T2I), Image-to-Image (I2I), and Text-and-Image-to-Image (TI2I). And we find that ***AI-generated images often struggle to maintain high alignment with their prompts, especially the image-only reference provided***. The results, shown in Table 3, indicate that AI-generated images typically achieve lower CLIP Score compared to natural images, with alignment discrepancies ranging from 4.02% to 36.69%. We also find that ***even if the AIGIs can get a higher CLIP score, especially with text and image-guided, they are looking stranger to humans*** as Table 3 illustrated. The CLIP score, which measures the alignment between text descriptions and generated images by embedding both into a shared space and calculating cosine similarity, revealed a noticeable gap in how accurately AIGC images reflect the input prompts. This suggests that current AIGC models often struggle to maintain high fidelity to the provided references, leading to variations that may impact the intended alignment. Our findings underscore the challenges that these models face in consistently generating content to accurately reflect the specified input conditions, highlighting both their strengths and limitations in real-world applications.

	GALIP	DFGAN	SD_V14	SD_V15	Versatile	SD_V21	SD_XL	DALL·E 2
T2I	10.72	18.51	9.78	9.69	9.71	9.93	10.20	9.46
I2I	-	-	36.46	36.69	4.92	26.96	3.98	5.41
TI2I	-	-	9.21	9.20	8.03	8.48	4.02	-

Table 3: The difference rate of CLIP score (%).

RQ3: Aesthetic Appeal We assessed the aesthetic appeal of AI-generated and natural images using NIMA and LAION-AES metrics and found that ***AI-generated images generally fall short of natural images in terms of aesthetic quality***. Figure 4 presents the results, which reveal that AI-generated images receive lower scores for aesthetic appeal, with discrepancies ranging from 3.34% to 23.85%. These findings indicate that while AI models can produce visually appealing images, they often lack the nuanced artistic quality and emotional impact that are characteristic of natural images.

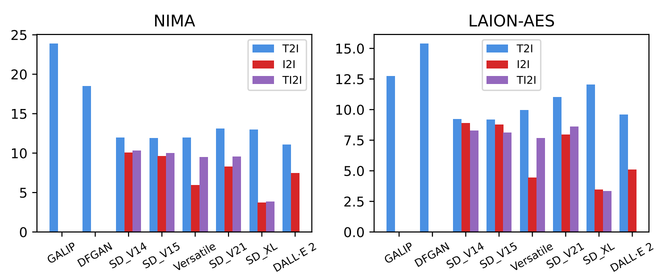


Figure 4: The aesthetic difference rate between AI-generated and naturally generated images.

RQ4: Applicability Results

When evaluating the applicability of AI-generated images, we focused on two crucial downstream tasks: image recognition and semantic segmentation. Our findings indicate ***significant discrepancies between the performance of AIGC images and natural images in these tasks.***

Image Recognition: For image recognition, we measured the classification mismatch rate, which quantifies the differences in predicted labels between AIGC images and their natural counterparts. The results in Table 4 showed a substantial gap, with the mismatch rate ranging from 29.89% to 94.44%. This indicates that AI-generated images often fail to achieve the same level of accuracy as natural images, highlighting the limitations of current AIGC models in producing reliable content for image recognition tasks.

	GALIP	DFGAN	SD_V14	SD_V15	Versatile	SD_V21	SD_XL	DALL·E 2
T2I	72.02	87.43	94.44	94.29	43.72	84.42	30.41	69.74
I2I	-	-	68.62	68.33	70.73	68.01	67.73	53.42
TI2I	-	-	62.84	62.35	62.28	58.56	29.89	-

Table 4: The results of Classification Mismatch Rate (MR).

	GALIP	DFGAN	SD_V14	SD_V15	Versatile	SD_V21	SD_XL	DALL·E 2
T2I	0.23	0.23	0.25	0.25	0.26	0.25	0.26	0.26
I2I	-	-	0.38	0.38	0.69	0.50	0.82	0.42
TI2I	-	-	0.43	0.44	0.30	0.53	0.82	-

Table 5: The results of Segmentation IoU.

The high mismatch rate underscores the need for further advancements in generative models to improve their applicability in real-world recognition tasks.

Overall, our results indicate that AI-generated images currently exhibit significant limitations in applicability for both image recognition and semantic segmentation tasks. The high classification mismatch rates and variable mean IoU values highlight the gap between AIGC and natural images in practical applications. These findings emphasize the importance of ongoing research and development to enhance the reliability and accuracy of AIGC models, ensuring they can be effectively used in real-world scenarios.

Semantic Segmentation: In semantic segmentation, we evaluated the performance using the mean Intersection over Union (IoU) metric and show the results in Table 5. The mean IoU values for AIGC images varied between 0.23 and 0.82, indicating a wide range of segmentation accuracy. While some generated images approached the performance of natural images, a significant portion of AIGIs fell short, demonstrating poor segmentation results. This variability suggests that while AIGC models can sometimes produce high-quality segmentation, they often struggle to consistently match the accuracy of natural images.

Visual Question Answering: To further evaluate the performance of AI-generated images and natural images in downstream tasks, we conducted tests on the VQA model BLIP-2 (Li et al. 2023b). The evaluation involved three different types of questions: (1) identifying objects in the image (Object), (2) determining whether the image aligns with a given text by answering "yes" or "no" (Align), and (3) predicting the similarity probability between the image and the text (Prob.). Table 6 summarizes the discrepancies (%) between natural and AI-generated images across various models. For Object questions, the highest discrepancies occurred in the I2I setting, with SD_V14 and SD_V15 reaching 77.27% and 75.51%, respectively, while T2I discrepancies ranged from 36.08% (DALL·E 2) to 68.96% (DFGAN). Align questions showed lower discrepancies, with T2I values ranging from 24.72% (Versatile) to 34.65% (DFGAN) and I2I values from 15.90% (Versatile) to 29.24% (SD_V15). For similarity probability (Prob.), T2I discrepancies varied between 36.61% (DALL·E) and 62.01% (DFGAN), while I2I results ranged from 31.03% (DALL·E) to 62.80% (SD_V14), highlighting model-dependent variations.

		GALIP	DFGAN	SD_V14	SD_V15	Versatile	SD_V21	SD_XL	DALL·E 2
Object	T2I	47.84	68.96	39.70	38.97	43.46	38.50	38.42	36.08
	I2I	-	-	77.27	75.51	28.78	71.59	28.37	26.03
	TI2I	-	-	41.18	38.06	37.64	39.27	26.69	-
Align	T2I	29.81	34.65	27.52	28.03	24.72	31.99	29.78	29.79
	I2I	-	-	28.48	29.24	15.90	27.07	16.33	25.50
	TI2I	-	-	27.52	25.79	19.26	27.31	15.85	-
Prob.	T2I	54.61	62.01	54.74	52.69	49.67	54.84	58.68	36.61
	I2I	-	-	62.80	58.39	36.16	56.14	42.88	31.03
	TI2I	-	-	52.75	48.71	43.69	46.30	43.24	-

Table 6: The results of VQA task *w.r.t* Object, Alignment, and Similarity probability.

	SD_V21			SD_XL			Versatile			DALL·E 2	
	T2I	I2I	TI2I	T2I	I2I	TI2I	T2I	I2I	TI2I	T2I	I2I
Quality	3.54	3.68	2.88	3.44	2.40	4.39	4.11	3.49	3.31	4.37	3.32
Alignment	3.45	3.43	2.48	3.16	2.44	4.47	4.22	3.59	3.40	4.50	2.96
Aesthetic	3.38	3.55	2.70	3.21	2.35	4.29	3.96	3.32	3.19	4.28	3.23

Table 7: Human evaluation results.

These results indicate that AI-generated images, especially those produced under certain generative settings, face significant challenges in downstream VQA tasks compared to natural images. The findings also reveal performance variations across generative models, underscoring the need for improved alignment and semantic understanding in AI-generated images.

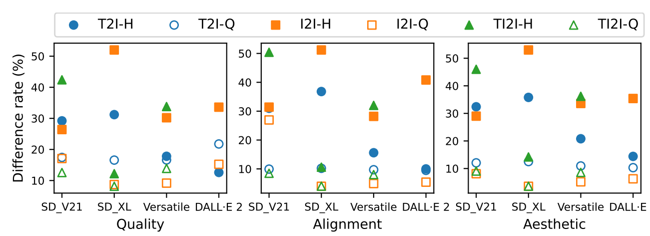


Figure 5: The difference rate provided by human & quantitative statistics, where 'H' represents human and 'Q' represents quantitative.

Human Validation

To rigorously assess the differences between AI-Natural images and to evaluate the effectiveness of existing metrics in measuring AIGC image quality, we conducted human evaluations focusing on three key aspects: Image Quality, Alignment, and Aesthetic Appeal. We excluded results for Applicability since this dimension inherently involves human-generated ground truth labels. Participants rated images on a scale from 1 (poor) to 5 (excellent). Across all three dimensions, AI-generated images generally received lower scores, often between 2 and 3, compared to natural images.

Image Quality: Participants assessed the overall quality of the images, considering factors such as clarity, detail, and the presence of artifacts. The results revealed a noticeable gap between AIGC images and their natural counterparts, with the generated images often lacking the same level of detail and clarity.

Alignment: This aspect measured how well the generated images matched the given prompts, revealing a notable discrepancy. Generated images frequently deviated from the descriptions, indicating issues with model adherence to specified conditions.

Aesthetic: The aesthetic appeal of the images was rated based on visual attractiveness, composition, and overall artistic quality. While some generated images displayed impressive artistic qualities, the overall consistency and aesthetic appeal were generally considered inferior to that of natural images.

Overall, our human evaluation results are shown in Table 7 and Figure 5 reveal ***significant discrepancies between AI-generated images and natural images across all three assessed aspects: quality, alignment, and aesthetic appeal***. AIGC images consistently received lower scores, emphasizing the need for further advancements to narrow these gaps. These findings underscore the importance of continued research and development in AI image generation to bring AIGC images closer to the standards of natural images.

Moreover, the results from human validation indicate that ***traditional quantitative metrics often fail to align with human preferences, suggesting that they are not fully compatible with assessing AIGC images***. This highlights the necessity for developing new measures tailored specifically to the unique characteristics of AI-generated content.

More Evaluations

To analyze the differences between AI-generated and natural images across various categories, we evaluated the performance of generated data from eight categories (*'person', 'animal', 'indoor', 'outdoor', 'vehicle', 'food', 'sports', 'accessory'*) across three key aspects: (1) **Naive Quality**, including pixel-level similarity (SSIM), frame-level quality (CLIPIQA), and content distribution (FID); (2) **Alignment**, measured by CLIP Score; and (3) **Aesthetic Quality**, assessed by LAION-AES. Partial results are presented in Figures 6 and 7, while the full experimental findings are provided in Appendix Sec. E (Figures 13 and 14).

From Figure 6, we observe substantial variability in the differential rates for categories such as "food" and "sports". Compared to "sports" images, the "food" demonstrate significantly higher discrepancies in image quality (CLIPIQA), revealing AI models struggle to generate natural food images. In Figure 7, the FID and SSIM metrics highlight distinct trends across categories. The "accessory" category exhibits the highest FID values, indicating poorer content distribution, while the "animal" category consistently achieves higher SSIM scores, reflecting better pixel-level similarity. These results suggest that certain categories, such as "accessory" and "indoor", remain particularly challenging for AIGC models to reproduce accurately, especially in maintaining semantic consistency and achieving balanced content generation.

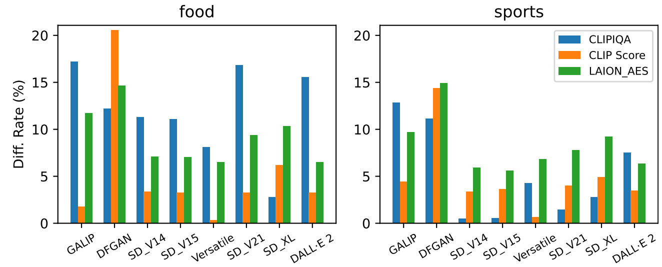


Figure 6: Differential rates of AI-generated images in the food and sports categories across naive image quality, alignment, and aesthetic metrics.

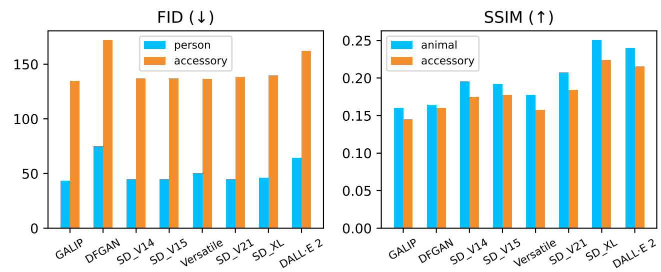


Figure 7: FID and SSIM metrics for AI-generated images across different source models and categories. FID evaluates distribution-level image quality, while SSIM measures pixel-level similarity to assess overall image quality.

Overall, these findings demonstrate that AI-generated images still fall short of matching natural images in terms of naturalness and image quality. Additionally, the pronounced performance discrepancies across categories underscore an imbalance in the training data used for AIGC models. Addressing these discrepancies will require more balanced datasets and targeted improvements in model architecture and training strategies.

Conclusion and Discussion

In this paper, we develop a systematic and comprehensive framework to evaluate the discrepancies between AI-Natural images using the formulated DNAI dataset. Our findings reveal significant disparities in naive image quality, semantic alignment, aesthetics, and downstream task applicability. While AIGC models have achieved impressive innovations, AIGIs still lag behind natural images in these critical areas. Besides, human evaluations also reveal lower scores for AIGC in terms of visual quality, semantic alignment, and aesthetics, emphasizing the gaps between AIGIs and natural images. Finally, the poor utility of AIGIs in downstream tasks indicates practical application challenges for AIGC.

This study also has limitations, including reliance on specific quantitative metrics and models available at the time. Future studies should incorporate more generative models, larger datasets, and sophisticated metrics to capture further discrepancies. Addressing these limitations and improving models will help reduce the perceptual gaps between AIGC images and natural images, facilitating better integration into real-world applications.

References

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5): 2567–2581.

Esfandarani, H. T.; and Milanfar, P. 2018. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.*, 27(8): 3998–4011.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 88(2): 303–338.

Gu, K.; Zhai, G.; Yang, X.; and Zhang, W. 2015. Using Free Energy Principle For Blind Image Quality Assessment. *IEEE Trans. Multim.*, 17(1): 50–63.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 6626–6637.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *ICCV*, 20349–20360.

Kang, L.; Ye, P.; Li, Y.; and Doermann, D. S. 2014. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *CVPR*, 1733–1740.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*, 5128–5137.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *NeurIPS*.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19(1): 011006.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023a. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *CoRR*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, volume 202, 19730–19742.

Li, M.; Cai, T.; Cao, J.; Zhang, Q.; Cai, H.; Bai, J.; Jia, Y.; Li, K.; and Han, S. 2024. Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *CVPR*, 7183–7193.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, 740–755.

Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; and Zuo, W. 2018. End-to-End Blind Image Quality Assessment Using Deep Neural Networks. *IEEE Trans. Image Process.*, 27(3): 1202–1213.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.*, 20(3): 209–212.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2408–2415.

N., V.; D., P.; Bh., M. C.; Channappayya, S. S.; and Medasani, S. S. 2015. Blind image quality evaluation using perception based features. In *NCC*, 1–6.

Qian, Y.; Cai, Q.; Pan, Y.; Li, Y.; Yao, T.; Sun, Q.; and Mei, T. 2024. Boosting Diffusion Models with Moving Average Sampling in Frequency Domain. In *CVPR*, 8911–8920.

Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaïane, O. R.; and Jägersand, M. 2020. U^2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.*, 106: 107404.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021b. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685.

Salimans, T.; Goodfellow, I. J.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *NeurIPS*, 2226–2234.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Sheikh, H. R.; and Bovik, A. C. 2006. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2): 430–444.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *CVPR*, 3664–3673.

Tao, M.; Bao, B.; Tang, H.; and Xu, C. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *CVPR*, 14214–14223.

Tao, M.; Tang, H.; Wu, F.; Jing, X.; Bao, B.; and Xu, C. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *CVPR*, 16494–16504.

Wang, J.; Duan, H.; Liu, J.; Chen, S.; Min, X.; and Zhai, G. 2023. AIGCIQA2023: A Large-Scale Image Quality Assessment Database for AI Generated Images: From the Perspectives of Quality, Authenticity and Correspondence. In *CAAI*, volume 14474, 46–57.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *ICCV*, 15897–15907.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; and Lin, W. 2024. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *ICLR*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *CoRR*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023a. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *NeurIPS*.

Xu, X.; Wang, Z.; Zhang, E. J.; Wang, K.; and Shi, H. 2023b. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In *ICCV*, 7720–7731.

Yuan, J.; Cao, X.; Li, C.; Yang, F.; Lin, J.; and Cao, X. 2023. PKU-I2IQA: An Image-to-Image Quality Assessment Database for AI Generated Images. *CoRR*, abs/2311.15556.

Zhang, L.; Shen, Y.; and Li, H. 2014. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Trans. Image Process.*, 23(10): 4270–4281.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.*, 24(8): 2579–2591.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.*, 20(8): 2378–2386.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 586–595.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Trans. Circuits Syst. Video Technol.*, 30(1): 36–47.

Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-Aware Blind Image Quality Assessment in the Laboratory and Wild. *IEEE Trans. Image Process.*, 30: 3474–3486.

Zhang, Z.; Li, C.; Sun, W.; Liu, X.; Min, X.; and Zhai, G. 2023. A Perceptual Quality Assessment Exploration for AIGC Images. In *ICMEW*, 440–445.

Appendix

Appendix Overview

The appendix provides supplementary details and additional experimental results that were not included in the main paper due to space limitations. It is organized as follows:

- **Section A:** In-depth Description of the DNAI Dataset.
- **Section B:** Detailed Overview of the Evaluated Generative Models.
- **Section C:** Comprehensive Summary of Evaluation Metrics.
- **Section D:** Extended Information on Human Perceptual Evaluations.
- **Section E:** Discussion on the Safety Mechanisms in Existing Diffusion Models.

A Details of the DNAI Dataset

Overview

The DNAI dataset is a carefully curated collection of both natural and AI-generated images, designed to support comprehensive evaluations of image generation models. The natural images are sourced from the COCO validation set. AI-generated images were produced using eight different models. These images were generated under three distinct guidance modes: Text-to-Image (T2I), Image-to-Image (I2I), and Text *vs.* Image-to-Image (TI2I). Representative samples from the DNAI dataset are shown in Figure 8. For a detailed breakdown of the DNAI dataset composition, please refer to Table 8.

Natural Images

The DNAI dataset features a carefully curated selection of natural images sourced from the COCO validation set. This subset comprises 5,000 images, each paired with at least five unique captions, totaling 25,014 captions. These rich and varied textual descriptions enhance the dataset's utility across a wide range of image-to-text and text-to-image tasks, making it an invaluable resource for evaluating and training image generation and understanding models. In this study, we use these images and their accompanying (five) captions to guide the generative models in creating new images.

Generated Images

The DNAI dataset includes AI-generated images created using several advanced generative models across three distinct guidance modes: Text-to-Image (T2I), Image-to-Image (I2I), and Text-and-Image-to-Image (TI2I). These modes collectively contributed a significant volume of images to the dataset. The models utilized in this process include DF-GAN, GALIP, various versions of Stable Diffusion (v1.4, v1.5, v2.1, XL), Versatile Diffusion, and DALL·E 2. Each model, except DALL·E 2, generated 25,000 images per guidance mode. DALL·E 2 produced 5,000 images per mode, resulting in a total of 440,000 generated images. It is important to note that DF-GAN and GALIP were used only for T2I, while DALL·E 2 was employed for both T2I and I2I modes. For detailed information, please refer to Table 8.

Image Resolution

The resolution of the generated images varies depending on the model used. Images generated by GALIP are 224x224 pixels, while those produced by DF-GAN are 256x256 pixels. The other models, including various versions of Stable Diffusion and DALL·E 2, generate images at a resolution of 512x512 pixels.

Dataset Composition

The DNAI dataset offers a robust foundation for evaluating the performance and capabilities of various AIGC models under different conditions and guidance modes. By encompassing a broad spectrum of images and resolutions, the dataset ensures comprehensive coverage of the potential use cases and challenges associated with AI-generated content. The combination of natural and generated images, alongside diverse textual descriptions, facilitates a thorough assessment of image generation models, supporting the development of more accurate and reliable AI systems.

B Evaluated Generative Models

GALIP (Generative Adversarial Latent Image Processing): (Tao et al. 2023) is a generative model that produces high-quality images by leveraging adversarial techniques. It utilizes latent image processing to enhance both the quality and fidelity of generated content. GALIP's architecture involves a generator and discriminator, where the generator creates images from latent vectors, and the discriminator evaluates their authenticity. Through iterative training, GALIP achieves highly detailed and realistic image synthesis.

DFGAN (Deep Fusion Generative Adversarial Network): (Tao et al. 2022) focuses on generating images from text descriptions using deep fusion techniques. The model integrates textual information into the image generation process through multiple layers of fusion, ensuring alignment with input descriptions. DFGAN operates in two stages: the first generates a coarse image based on the text, and the second refines the image, enhancing detail and coherence.

Stable Diffusion v1.4 (SD_V14): (Rombach et al. 2022) is part of the Stable Diffusion family, designed for image synthesis through iterative diffusion processes. Starting with pure noise, the model refines it into coherent images through a series of steps, guided by a learned model that predicts noise distribution. SD_V14 is recognized for producing high-resolution images with fine textures and complex structures.

Stable Diffusion v1.5 (SD_V15): (Rombach et al. 2022) builds on SD_V14, introducing enhancements in architecture and training techniques to improve image quality, consistency, and the handling of complex scenes. While continuing to use the diffusion process, SD_V15 incorporates better optimization strategies and larger datasets, yielding superior results.

Versatile Diffusion: (Xu et al. 2023b) is a diffusion-based model designed for a variety of image generation tasks, including style transfer, image inpainting, and super-resolution. Its versatility lies in its ability to adapt the diffu-

Figure 8: Examples of AI-generated content (AIGC) created using five state-of-the-art generative models: SD_v1.4, SD_v1.5, Versatile Diffusion, SD_v2.1, and SD_XL. Each case is illustrated with three rows corresponding to Text-to-Image (T2I), Image-to-Image (I2I), and Text-and-Image-to-Image (TI2I) generation modes. For the complete set of AIGC images, please refer to our DNAI dataset, which will be available publicly soon.

Types	DFGAN	GALIP	SD_V1.4	SD_V1.5	Versartile	SD_V2.1	SD_XL	DALL·E 2
T2I	25,000	25,000	25,000	25,000	25,000	25,000	25,000	5,000
I2I	-	-	25,000	25,000	25,000	25,000	25,000	5,000
TI2I	-	-	25,000	25,000	25,000	25,000	25,000	-

Table 8: The details of the DNAI dataset.

sion process to the specific requirements of each task, making it suitable for diverse applications.

Stable Diffusion v2.1 (SD_V21): (Rombach et al. 2022) is an advanced iteration of the Stable Diffusion series, offering significant improvements in image fidelity and generation speed. With optimizations in the diffusion process, better noise handling, and enhanced training algorithms, SD_V21 produces more realistic and high-quality images, making it highly effective for a range of creative and practical uses.

Stable Diffusion XL (SD_XL): (Rombach et al. 2022) represents the most advanced model in the Stable Diffusion series, capable of generating extremely high-resolution images with intricate details. SD_XL uses an extended number of diffusion steps and a larger network architecture to manage increased complexity, making it ideal for applications requiring ultra-high resolution and precision, such as detailed artworks and large-scale prints.

DALL·E 2: (Ramesh et al. 2022) is a generative model developed by OpenAI that excels in creating images from textual descriptions. As an advancement over the original DALL·E, it offers improved capabilities in generating high-quality, diverse, and coherent images. DALL·E 2 combines CLIP (Contrastive Language-Image Pre-training) with a transformer-based generative model to translate complex and abstract textual inputs into visual content, making it renowned for its creative and realistic outputs.

These generative models represent significant advancements in AI-driven image synthesis, each with unique strengths tailored to specific applications. From text-to-image generation and high-resolution synthesis to versatile image processing, these models push the boundaries of AI-generated content.

C Evaluation Metrics

Overview

In this study, we selected a comprehensive set of metrics to evaluate the image quality, alignment, and aesthetics of both natural and AI-generated images. These metrics, summarized in Table 9, are carefully chosen to highlight the differences across multiple dimensions, providing a robust framework for assessing the quality and realism of AIGC in comparison to natural images.

To evaluate image quality, we divided the assessment into three sub-aspects: pixel-level similarity, frame-level visual metrics, and visual content distribution.

- **Pixel-level:** Metrics such as PSNR, LPIPS, DISTS, VIF, VSI, FSIM, and MAD assess structural, perceptual, and textural fidelity. These metrics measure how closely the generated images match natural ones in terms of structural integrity and perceptual similarity.
- **Frame-level:** Metrics including PIQE, IL-NIQE, MUSIQ, DBCNN, CNNIQA, CLIPIQA, BRISQUE, TReS, HyperIQA, LIQE, and UNIQUE provide no-reference assessments of image quality. They evaluate attributes such as distortion levels, natural scene statistics, and learned features to predict the overall quality of the images. Additionally, naturalness metrics like NIQE and NRQM help gauge how closely AI-generated images resemble natural scenes based on statistical properties.
- **Visual Content Distribution:** Metrics such as FID and Inception Score examine overall content distribution differences between AI-generated and natural images across the entire dataset. These metrics capture broad differences in content quality.

For alignment, we utilize the CLIP Score (Radford et al. 2021b) to measure the correspondence between text descriptions and the generated images. This metric evaluates how well the generated content aligns with the given prompts by embedding both text and image into a shared space and calculating their similarity.

In terms of aesthetic evaluation, we employ NIMA and LAION-AES metrics to assess the visual appeal and artistic quality of the images. These metrics predict human perception of image aesthetics based on deep learning models and various visual features.

By integrating these diverse metrics, our framework offers a detailed and multi-faceted evaluation of AI-generated images, capturing the nuances and differences between natural and synthetic content. The comprehensive nature of these metrics ensures a robust and reliable assessment, guiding improvements in AIGC technologies.

The chosen metrics, as detailed in Table 9, collectively provide a comprehensive toolkit for evaluating image quality, alignment, and aesthetics, each focusing on different aspects of visual perception and fidelity.

Naive Image Quality

We evaluated the naive image quality of AI-generated images versus natural images using a range of metrics. These metrics assess both pixel-level and frame-level qualities, as well as visual content distribution, to capture the holistic differences between AI-generated and natural images.

Pixel-Level Quality Metrics:

- **SSIM** (Wang et al. 2004) (Structural Similarity Index) compares the structural information between a reference

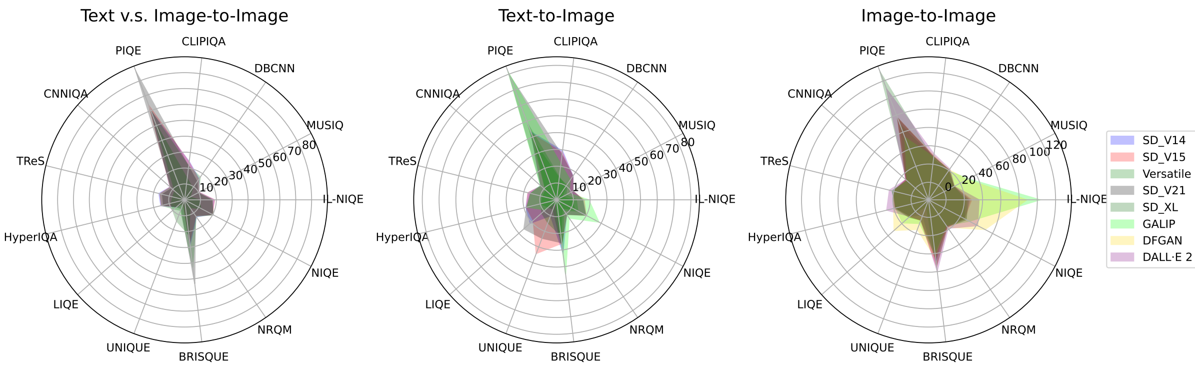


Figure 9: We evaluate eight image-generating models from 13 aspects. The numerical values in the radar chart represent the mean difference rate of each model.

and a test image, considering luminance, contrast, and structure for a comprehensive measure of similarity.

- **PSNR** (Peak Signal-to-Noise Ratio) measures the ratio between the maximum possible power of an image and the power of corrupting noise, with higher values indicating better image quality.
- **VIF** (Sheikh and Bovik 2006) (Visual Information Fidelity) quantifies the amount of visual information preserved in the test image relative to the reference, based on natural scene statistics and the human visual system.
- **VSI** (Zhang, Shen, and Li 2014) (Visual Saliency-based Index) evaluates perceptual similarity by focusing on how salient (prominent) features in the images compare.
- **FSIM** (Zhang et al. 2011) (Feature Similarity Index) measures similarity using phase congruency and gradient magnitude, emphasizing feature similarity, which is crucial for human perception.
- **LPIPS** (Zhang et al. 2018) (Learned Perceptual Image Patch Similarity) uses deep network features to assess perceptual similarity, capturing human visual perception more effectively by comparing feature activations in a deep neural network.
- **DISTS** (Ding et al. 2022) (Deep Image Structure and Texture Similarity) combines structural and textural similarity metrics using deep learning features to evaluate overall image similarity.
- **MAD** (Larson and Chandler 2010) (Mean Absolute Deviation) provides a straightforward assessment of pixel-level differences, with lower values indicating greater similarity between images.

Frame-Level Quality Metrics:

- **PIQE** (N. et al. 2015) (Perception-based Image Quality Evaluator) provides a no-reference quality assessment by analyzing image blocks and estimating distortion levels.
- **IL-NIQE** (Zhang, Zhang, and Bovik 2015) (Integrated Local Natural Image Quality Evaluator) evaluates image quality based on natural scene statistics and local features, offering a no-reference assessment.
- **MUSIQ** (Ke et al. 2021) (Multi-Scale Image Quality) assesses image quality at multiple scales, considering both local and global features for a comprehensive no-reference evaluation.
- **DBCNN** (Zhang et al. 2020) (Deep Bilinear Convolutional Neural Network) predicts image quality using deep learning, based on bilinear pooling of convolutional features.
- **LIQE** (Ma et al. 2018) (Learning-based Image Quality Evaluator) combines traditional quality metrics with learned features for robust no-reference quality assessment.
- **CNNIQA** (Kang et al. 2014) (Convolutional Neural Network Image Quality Assessment) leverages deep convolutional networks for no-reference image quality assessment using learned features.
- **CLIPIQA** (Radford et al. 2021a) (CLIP Image Quality Assessment) uses CLIP model embeddings to assess image quality by comparing the alignment between image and text descriptions.
- **TReS** (Radford et al. 2021a) (Textural and Edge-based Similarity) evaluates image quality by focusing on textural and edge-based features, providing a no-reference assessment.
- **HyperIQA** (Su et al. 2020) (Hyper Network-based Image Quality Assessment) uses a hypernetwork to predict image quality, leveraging deep learning for no-reference quality assessment.
- **UNIQUE** (Zhang et al. 2021) (Universal Quality Index with Deep Features) evaluates unique aspects of image quality using deep learning features, offering a no-reference assessment.
- **BRISQUE** (Mittal, Moorthy, and Bovik 2012) (Blind/Referenceless Image Spatial Quality Evaluator) assesses image quality based on natural scene statistics, providing a no-reference measure.
- **NIQE** (Mittal, Soundararajan, and Bovik 2013) (Natural Image Quality Evaluator) evaluates how natural an image appears by comparing its features with those derived from a dataset of natural images, without needing a reference image.

- **NRQM** (Ma et al. 2018) (Naturalness Image Quality Metric) assesses how closely an image aligns with the statistical properties of natural scenes based on natural scene statistics and perceptual models.

Visual Content Distribution Metrics:

- **FID** (Heusel et al. 2017) (Fréchet Inception Distance) measures the similarity between two sets of images by comparing the means and covariances of their feature representations, providing a comprehensive assessment of both quality and diversity.
- **Inception Score** (Salimans et al. 2016) evaluates the quality of generated images by considering the confidence of the classifier and the diversity of the generated samples, with higher scores indicating better quality and diversity.

We present the additional results of the Naive Image Quality assessment in Figure 9.

Alignment

To assess alignment, we employ the **CLIP Score** (Contrastive Language-Image Pre-training Score), which measures the correspondence between text descriptions and generated images. This metric evaluates how well the visual content aligns with the textual prompts by embedding both into a shared space and calculating their cosine similarity.

Aesthetic

- **NIMA** (Esfandarani and Milanfar 2018) (Neural Image Assessment) is a deep learning-based model that predicts the aesthetic quality of images. It generates a score that reflects human perception of visual appeal.
- **LAION-AES** (Schuhmann et al. 2022) (LAION Aesthetic Score) evaluates images' aesthetic quality using various visual features, providing a numerical value to indicate their aesthetic appeal.

Together, these metrics offer a robust toolkit for evaluating images, focusing on alignment with prompts and aesthetic quality, thereby covering key aspects of visual perception and fidelity.

D Human Perceptual Evaluation

Interface for Human Evaluation

The human evaluation of our DNAI dataset was conducted using a custom-designed interface (Figure 11). This interface allowed participants to evaluate AI-generated images in comparison to reference content across three key aspects: Quality, Alignment, and Aesthetics. The reference content included prompts used to generate the images, categorized into three types: 1) text, 2) image, and 3) text and image. The AI-generated images evaluated were produced by four different models: SD_V2.1, SD_XL, Versatile Diffusion, and DALL·E 2.

Aspects	Sub_aspects	Netrucs	Reference
Naive Quality	Pixel-level	SSIM	TRUE
		PSNR	TRUE
		LPIPS	TRUE
		DISTS	TRUE
		VIF	TRUE
		VSI	TRUE
		FSIM	TRUE
		MAD	TRUE
	Frame-level	PIQE	FALSE
		IL-NIQE	FALSE
		MUSIQ	FALSE
		DBCNN	FALSE
		CNNIQA	FALSE
		CLIPIQA	FALSE
		BRISQUE	FALSE
		TReS	FALSE
		HyperIQA	FALSE
		LIQE	FALSE
		UNIQUE	FALSE
		NIQE	FALSE
		NRQM	FALSE
	Content Distribution	FID	TRUE
		Inception Score	TRUE
Alignment		CLIP Score	FALSE
Aesthtic		NIMA	FALSE
		LAION-AES	FALSE

Table 9: The detail of evaluation metrics.

Evaluation Process

A total of 58 volunteers participated in the evaluation process. Each volunteer assessed the images based on the provided reference content. To ensure the validity of the data collected, we included duplicate images in the evaluation system. By comparing the scores given to these identical images, we could determine the consistency of each participant's evaluations. Only data from participants who showed consistent scoring were considered valid, resulting in 47 effective samples.

Participant Demographics

The volunteers who participated in the evaluation ranged in age from 19 to 46 years old, with a gender ratio of 6:4 (male to female). All participants held at least an undergraduate degree, ensuring a well-educated sample group. Alongside scoring data, we collected demographic details such as age and gender. Upon completing their evaluations, participants were also invited to provide feedback on the evaluation system and their experience with assessing the differences between natural and AI-generated images.

Evaluation Metrics

Participants were instructed to evaluate each image based on three aspects:

1. **Quality:** Overall visual fidelity and clarity of the image.

2. **Alignment:** Accuracy in matching the image to the provided prompt.
3. **Aesthetic:** Visual appeal and artistic quality of the image.

To facilitate the scoring process, we provided reference examples within the evaluation system (Figure 12). These examples illustrated what constituted good or poor performance in each aspect without giving specific scores, thereby preventing any potential bias in participant evaluations.

Summary

The comprehensive human evaluation process ensured that the collected data was reliable and reflected genuine human perception. The demographic diversity and the rigorous validation of participant responses provided a robust foundation for assessing the performance of various AIGC models in generating high-quality, aligned, and aesthetically pleasing images. This approach not only highlighted the strengths and weaknesses of the evaluated models but also offered valuable insights into the perceptual differences between AI-generated and natural images.

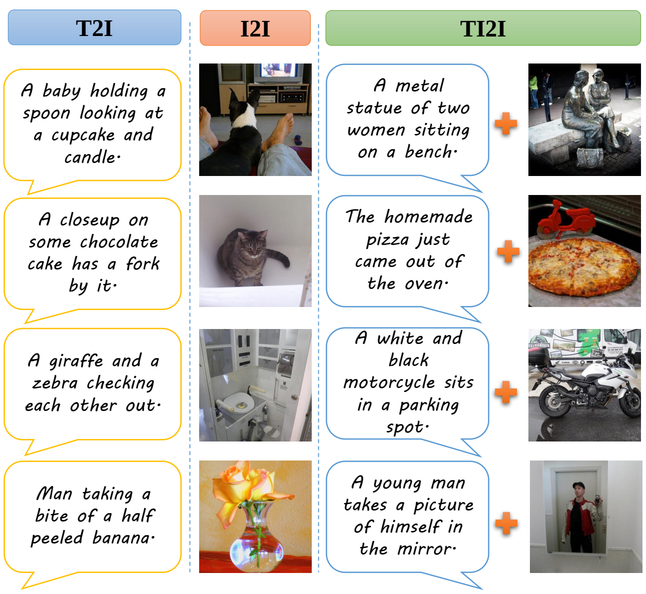


Figure 10: Examples for unsafe references.

E More Evaluations

This section provides a detailed evaluation of AI-generated images across different categories, focusing on naive image quality, alignment, aesthetics, pixel-level similarity, and content distribution. These analyses aim to uncover discrepancies between AI-generated and natural images and identify category-specific challenges.

From Figure 13, we observe notable variability in the differential rates across categories such as "indoor" and "sports" among different generative models. Images in the "indoor" category exhibit significantly higher discrepancies in alignment (CLIP Score) and aesthetics (LAION-AES)

	T2I	I2I	TI2I	T2I(%)	I2I(%)	TI2I(%)
SD v1.4	104	263	141	4.16	10.51	5.64
SD v1.5	133	270	145	5.32	10.79	5.80
DALL·E 2	20	-	-	4	-	-

Table 10: Unsafe Prompt.

compared to "sports", indicating that AI-generated sports images are more aligned with natural images. In contrast, generating natural-looking indoor images remains a challenge due to their complex structures and diverse visual features.

Figure 14 presents trends in SSIM and FID metrics across categories and models, where SSIM evaluates pixel-level similarity, and FID measures distribution-level content alignment. The "accessory" category consistently shows higher FID values, reflecting poorer content distribution, while the "animal" category achieves higher SSIM scores, demonstrating better pixel-level similarity. These findings suggest that "accessory" and "indoor" images pose greater challenges for generative models in maintaining semantic consistency and achieving balanced content generation compared to other categories.

Overall, these results indicate that AI-generated images still lag behind natural images in achieving comparable levels of naturalness and image quality. Moreover, the substantial performance discrepancies across categories highlight an imbalance in the training data used for generative models. This imbalance likely limits their ability to generalize effectively across diverse content types, underscoring the need for more balanced datasets and targeted advancements in model design.

F Discussion

Safety Mechanism Activation in SD 1.4, 1.5 DALL·E 2

In the rapidly evolving field of AI-generated content (AIGC), the balance between model robustness and safety mechanisms remains a critical area of study. A noteworthy issue has emerged with popular image synthesis models, such as Stable Diffusion v1.4, v1.5, and DALL·E 2, which occasionally fail to generate images in response to seemingly benign prompts. Some examples of these prompts are shown in Figure 10. This unexpected behavior suggests an oversensitivity in the models' safety filters, which are intended to prevent the creation of inappropriate or sensitive content.

This case study delves into instances where typical prompts, which ostensibly do not contain objectionable content, nonetheless trigger these safety mechanisms. Through a series of experiments, we systematically presented a variety of benign prompts to both model versions and documented the conditions under which it is hard to output the AIGC images (Table 10). Our findings suggest that certain benign text, benign image, or combinations of text and image are misinterpreted by the models' safety algorithms as

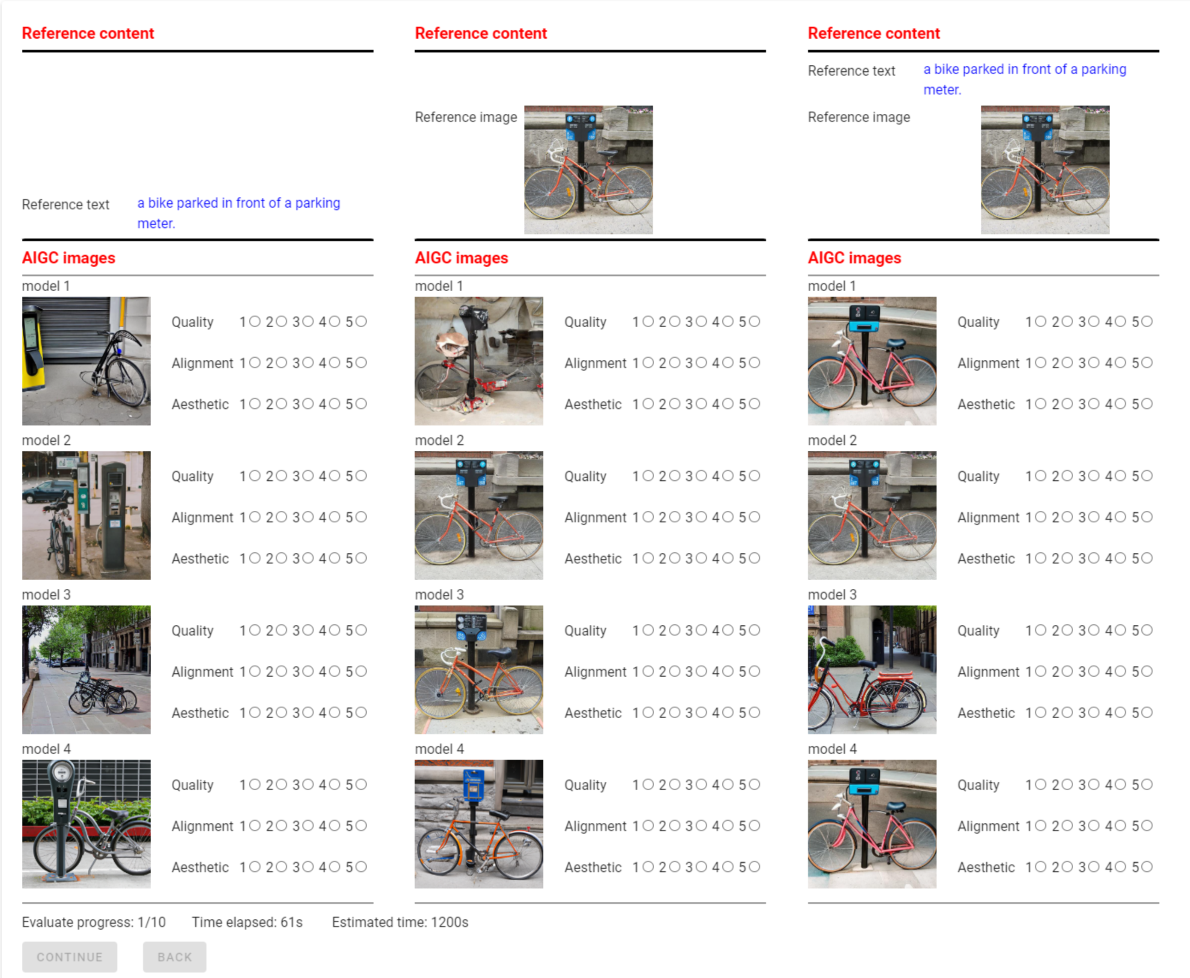


Figure 11: Interface for Human Evaluation. Reference content: the prompt guidance used to generate images, including three types: 1) text, 2) image, and 3) text and image. AIGC images: AI-generated images that come from four different models. Participants can rely on the reference content to evaluate each image on three aspects, i.e., Quality, Alignment, and Aesthetic.

being potentially harmful or sensitive.

The implications of these findings are twofold. First, they highlight a critical need for refining the sensitivity of safety algorithms to reduce false positives, thereby ensuring that the AIGC models do not unduly limit creative expression or practical application. Second, they serve as a stark reminder of the challenges inherent in balancing content safety with the functional robustness of generative models. This case study aims to contribute to the ongoing discussion on optimizing safety mechanisms in AIGC models, striving to enhance their utility and accessibility without compromising essential safeguards.

Figure 12: Before the evaluation process, the participant will be given some examples to learn which image should be good or bad in a specific aspect.

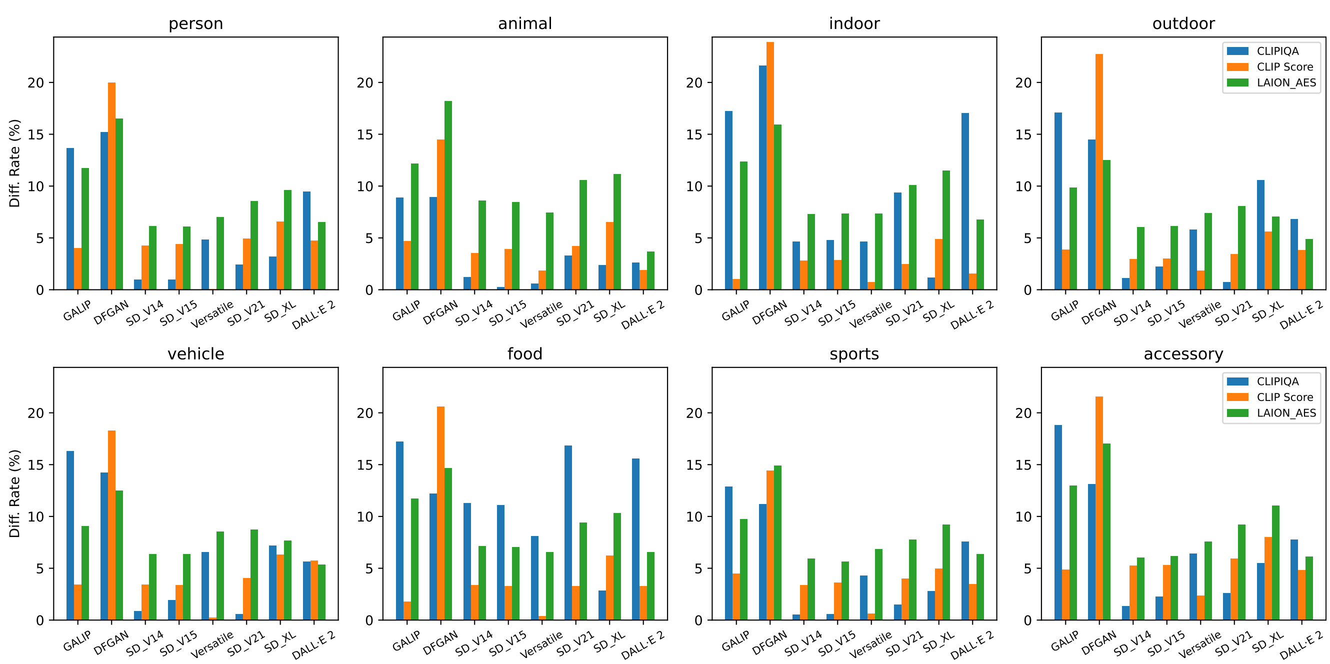


Figure 13: The differential rate of data from different categories models across naive image quality, alignment, and aesthetic.

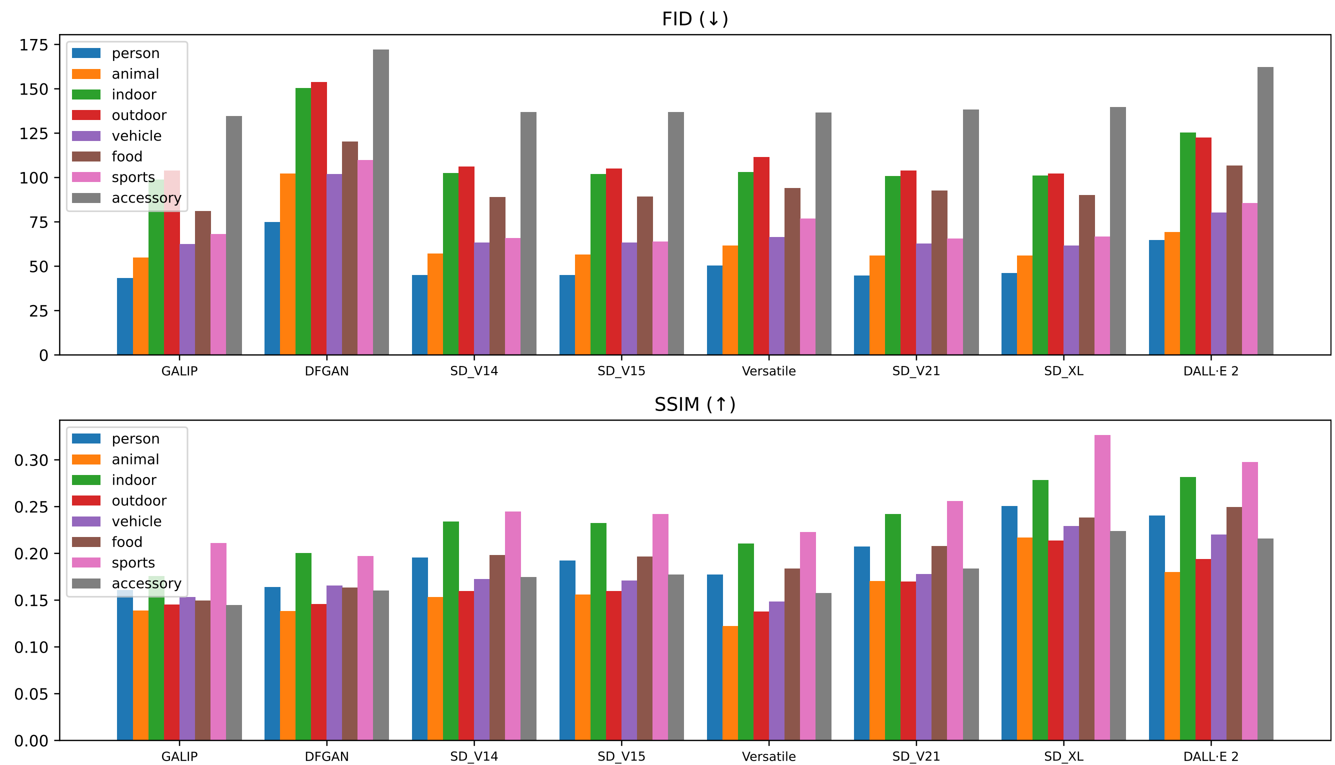


Figure 14: The FID and SSIM of AI-generated images from different source models and categories. The FID and SSIM are used to evaluate image quality at the pixel level and content distribution level.