PrefIQA: Human Preference Learning for AI-generated Image Quality Assessment

Hengjian Gao, Kaiwei Zhang, Wei Sun, Chunyi Li, Huiyu Duan, Xiaohong Liu, Xiongkuo Min, Guangtao Zhai*
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
Shanghai, China
Email:{hengjiang,zhangkaiwei,sunguwei,lcysyzxdxc,huiyuduan,xiaohongliu,minxiongkuo,zhaiguangtao}@sjtu.edu.cn

***Abstract*—Despite recent advancements in generative models, the variation in image quality remains a significant concern. To tackle this issue, we propose PrefIQA, an effective human preference learning metric, which can better evaluate the quality of AI-generated images. PrefIQA consists of two units, namely Feature Extraction Unit and Feature Fusion Unit. In Feature Extraction Unit, we introduce a prompt-segmentation module to divide prompts into multiple phrases, enabling a more detailed evaluation of the alignment between images and texts. In Feature Fusion Unit, we introduce a modality-fusion module, which effectively mixes text features and image features to improve the overall performance. In the experiment part, extensive experiments are conducted, demonstrating that PrefIQA surpasses existing text-to-image alignment metrics. We believe that PrefIQA's proposal would facilitate researches on AI-generated image quality assessment, and make a valuable contribution to the field of text-to-image generation.**

***Index Terms*—AI-generated images, text-to-image alignment, image quality assessment**

I. Introduction

Text-to-image generative models have experienced significant advancement in recent years, including GAN-based models like [1]–[3], regressive-based models like [4]–[6], and diffusion-based models like [7]–[9]. However, the quality of images generated by different models often varies widely. Moreover, even when using the same model, the quality of generated images can vary greatly because of different seeds. Therefore, it is important to establish an effective metric for evaluating the quality of generated images.

Since the quality of AI-generated content correlates well with human subjective sensation, subjective human preference can be regarded as the most direct and reliable means of quality assessment of these generated images. Recently several datasets [10]–[13] of human preferences on images generated by text-to-image generative models are proposed. Each example of these datasets includes a prompt, two or more generated images, and labels indicating images rankings. Developing a model to predict human preference becomes essential, as it assists text-to-image generative models in generating images that align better with human preferences. Although some human preference prediction models [10]–[13] have been proposed with above datasets, these models only rely on fine-tuning CLIP [14] or BLIP [15] to build an encoder-based network architecture for learning the alignment similarity between textual and visual modalities. However, they have failed to fully explore the fine-grained and cross-modality features, which may limit the potential of the models to achieve higher levels of task performance.

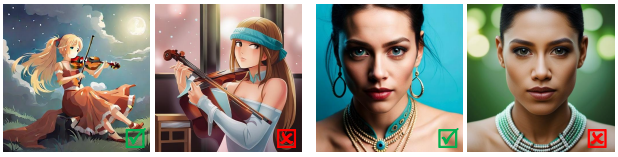

Fig. 1. Examples showing that PrefIQA selects images more consistent with human preference.

In this study, we propose PrefIQA, human Preference Learning for AI-generated Image Quality Assessment. As is shown in Fig. 1, PrefIQA takes a prompt and two generated images as inputs, and then selects the image that exhibits higher quality and better alignment with human preferences. In terms of model architecture, PrefIQA consists of two units: Feature Extraction Unit and Feature Fusion Unit. In Feature Extraction Unit, we use a prompt-segmentation module to divide a prompt into multiple phrases, which are then passed through a transformer encoder [16] to extract features, while the images are passed through ViT [17]. Then in Feature Fusion Unit, the text features and image features are input into a modality-fusion module, where they are mixed to obtain fused text features and image features. Finally, we calculate the cosine similarity of the final text features and image features as the final score predicting human preference.

In the experiment section, we demonstrate that PrefIQA outperforms existing text-to-image alignment metrics, including Aesthetics [18], CLIP [14], HPS [12], HPS v2 [13], ImageReward [10] and PickScore [11]. Ablation studies demonstrate that both prompt-segmentation module and modality-fusion module significantly improve the final prediction results. Consequently, we suggest that PrefIQA be considered as a promising human preference prediction metric, which can better evaluate the quality of AI-generated images.

Our contributions are as follows:

- We propose PrefIQA as an effective human preference predictor, which is suitable for AI-generated image quality assessment. PrefIQA consists of Feature Extraction Unit and Feature Fusion Unit, and utilizes cosine similarity of different modality features as the final score.

*Corresponding author

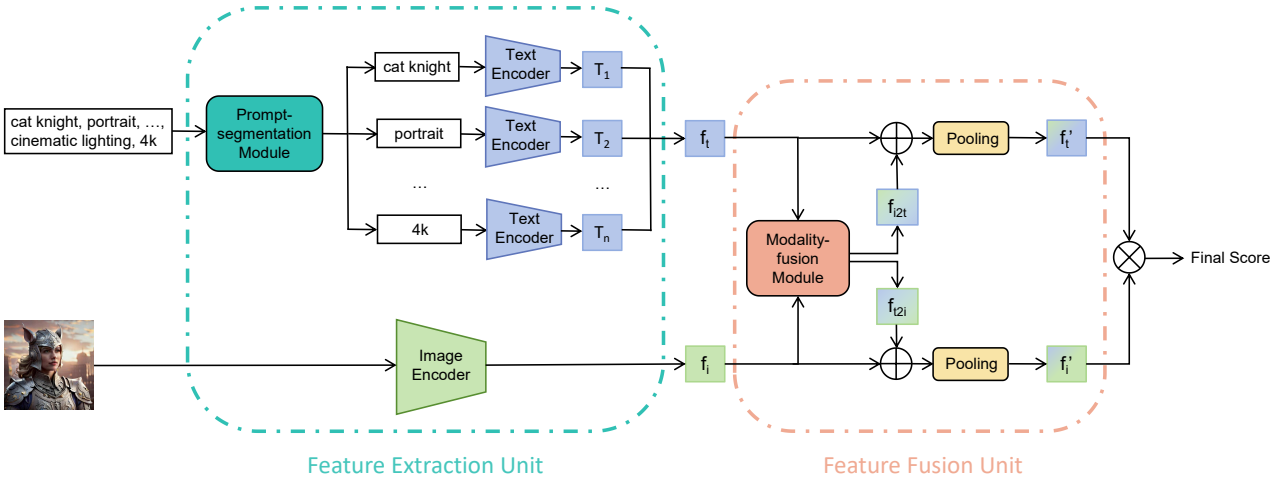

Fig. 2. The architecture of PrefIQA. In Feature Extraction Unit, a prompt is segmented into multiple phrases using a prompt-segmentation module. These phrases are then fed into a text encoder to extract features, while the image is processed by an image encoder. Subsequently, in Feature Fusion Unit, the text features and image features are combined in a modality-fusion module to generate fused text features and image features. Finally, the cosine similarity between the two final features is calculated as the final prediction score.

- In Feature Extraction Unit, we introduce a prompt-segmentation module to segment a prompt into multiple phrases. This detailed segmentation enables the model to better understand the input prompt and evaluate the consistency between images and texts in more detail.
- In Feature Fusion Unit, we propose a modality-fusion module to fuse text features and image features. The fusion operation helps the encoders learn more cross-modality feature representations, leading to improved performance in the final prediction.
- We conduct extensive experiments to validate our approach, demonstrating that PrefIQA outperforms existing text-to-image alignment metrics, and highlighting the importance of the two key components, prompt-segmentation module and modality-fusion module.

II. Related Work

A. AI-generated Image Quality Datasets

With the rapid development of text-to-image generative models, several AI-generated image quality datasets have emerged in recent years. The earliest one, DiffusionDB [19], consists of over 1.8 million Text-Image pairs generated by the StableDiffusion model. Despite lacking subjective scoring, its extensive collection of images and prompts has provided a foundation for subsequent subjective databases. AGIQA-1K [20], AGIQA-3k [21] and AIGCIQA2023 [22] are the subjective databases for assessing AI-Generated image quality, utilizing fine-grained scoring through MOS. However, the limited scale makes it difficult to generalize in a wide range of AI-generated images.

ImageReward [10], Pick-a-Pic [11], HPD v1 [12], HPD v2 [13] have expanded the scale of images and prompts. HPD v2 is the largest database including 430k images from a wide range of sources, but it has not been made available as an open-source dataset. Pick-a-Pic contains 500k images and HPD v1 contains 98k images, which crawl results generated by Stable-Diffusion or directly utilize Text-Image pairs from DiffusionDB. ImageReward, containing 137k images, considers an auto-regression based model when generating images, as well as diffusion-based models. Due to the smaller scale of HPD v1 and the unavailability of HPD v2, we choose to use Pick-a-Pic and ImageReward for our training and testing.

B. AI-generated Image Quality Metrics

Inception Score (IS) [23] and Fréchet Inception Distance (FID) [24] are initially proposed and widely used in the evaluation of text-to-image generative models. However, they are not suitable for evaluating the no reference quality assessment, and lack a strong correlation with human preferences for images produced by recent text-to-image models [11]–[13].

To address this limitation, recent studies [10]–[13] propose a different approach. They suggest fine-tuning visual-language models, like CLIP [14] and BLIP [15], based on human preference for images generated from the same prompt. However, these metrics almost rely on training with large amounts of data to improve accuracy, without extensive analysis of the model structure. Therefore, taking into account the limitations, we consider designing a novel model architecture, namely PrefIQA.

III. Approach

The framework of PrefIQA is shown in Fig. 2, consisting of two units: Feature Extraction Unit and Feature Fusion Unit.

A. Feature Extraction Unit

1) Text Feature Extraction: When reading a sentence, it is common for individuals to first divide it into shorter phrases using punctuation marks, and then comprehend each of these phrases individually. Additionally, our extensive observation on prompts in DiffusionDB [19] shows that most prompts separate different key points using commas or periods, while

the usage of long sentences without any punctuation marks is less frequent. Therefore, we propose a prompt-segmentation module to divide the prompt based on commas and periods.

Prompt-segmentation Module: Let T_0 be a given prompt, and we split it into phrases $p_1, p_2, \cdots, p_n$, where n is fixed. We fix the value of n to simplify code writing and facilitate model learning. Specifically, we first split T_0 into $t_1, t_2, \cdots, t_k$ based on commas and periods. The value of k may vary depending on the content of T_0. Then we compare the value of k and n: if $k < n$, we fill T_0 to make the total number of phrases equal to n; if $k > n$, we combine the last $k - n + 1$ phrases $t_n, t_{n+1}, ..., t_k$ into a single phrase p_n. The reason for merging the last $k-n+1$ phrases is that the last few phrases of a prompt generally provide supplementary information about the style, resolution, and other aspects of the images that are comparatively less crucial than the former phrases. The equations are as follows:

$$(t_1, t_2, \cdots, t_k) = \text{Split}(T_0) \tag{1}$$

$$(p_1, p_2, \cdots, p_n) = \begin{cases} (t_1, \cdots, t_k, T_0, \cdots, T_0), & k < n \\ (t_1, \cdots, t_{n-1}, \{t_n, \cdots, t_k\}), & k > n \\ (t_1, \cdots, t_n), & k = n \end{cases} \tag{2}$$

Hence, we successfully segment a prompt into multiple phrases. This prompt-segmentation module poses two benefits: Firstly, it assists the model in learning semantic information of prompts more accurately by providing finer-grained representations. Secondly, it enables a more detailed assessment of the consistency between the image and text, which positively impacts the final preference prediction.

Next, we use the transformer encoder to extract features for each phrase $(p_1, p_2, ..., p_n)$ and concatenate them to obtain the text feature f_t:

$$f_t = \text{Concat}(\Phi_{text}(p_1), \cdots, \Phi_{text}(p_n)) \tag{3}$$

where Φ_{text} refers to the text encoder, and Concat$(\cdot)$ denotes a function concentrating multiple arrays into one.

2) Image Feature Extraction: We directly use ViT-H [17] as the image encoder to extract the image feature f_i from the given image I_0:

$$f_i = \Phi_{img}(I_0) \tag{4}$$

where Φ_{img} refers to the image encoder.

B. Feature Fusion Unit

CLIP model uses two independent encoders to extract features from images and texts, which are then directly used to compute similarity. To a certain extent, the lack of information integration between modalities could lead to significant differences in features extracted by the two encoders, making it difficult for the model to fully learn effective information. Therefore, we introduce a modality-fusion module after the encoders to better combine information from texts and images and improve model learning efficiency, as shown in Fig. 3.

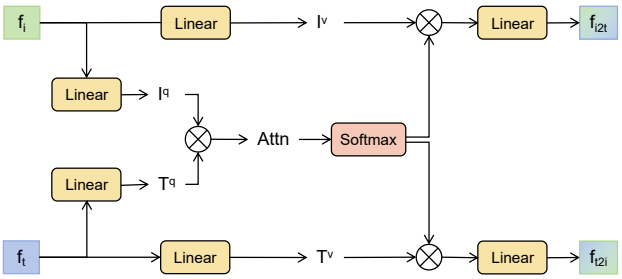

Fig. 3. The architecture of modality-fusion module.

Modality-fusion Module: Given text features f_t and image features f_i, we obtain fused features f_{i2t} and f_{t2i} through the following equations:

$$I^q = f_i W_i^q, I^v = f_i W_i^v, T^q = f_t W_t^q, T^v = f_t W_t^v \tag{5}$$

$$Attn = I^q (T^q)^\mathsf{T} / \sqrt{d}, \tag{6}$$

$$f_{t2i} = \text{Softmax}(Attn) T^v W_i^{out} \tag{7}$$

$$f_{i2t} = \text{Softmax}(Attn^\mathsf{T}) I^v W_t^{out} \tag{8}$$

where $W_t^\theta, W_i^\theta (\theta \in \{q, v, out\})$ are the projection matrices which function similarly to the linear layers in Multi-Head Attention [16].

With the modality-fusion module, we can make the learned image features language-aware, and the learned text features vision-aware.

Then we add the original features f_i, f_t and fused features f_{i2t}, f_{t2i} together, and then use a pooling module, the same as CLIP, to obtain the final features f_t' and f_i':

$$f_t' = \text{pooling}(f_t + f_{i2t}), f_i' = \text{pooling}(f_i + f_{t2i}) \tag{9}$$

Finally, we calculate the cosine similarity of f_t' and f_i' as the final score:

$$s = \cos(f_t', f_i') \tag{10}$$

C. Loss Function

Now given a prompt T_0, two images I_1, I_2, we can successfully obtain the final score s_1, s_2. Let p_1, p_2 be the preference label indicating human preference over the two images, the objective is to optimize the model's parameters by minimizing the KL-divergence between p_1, p_2 and the softmaxed scores of s_1, s_2:

$$L_{pref} = \sum_{i=1}^{2} p_i (log p_i - log \hat{p_i}), s.t. \hat{p_i} = \frac{e^{s_i}}{e^{s_1} + e^{s_2}} \tag{11}$$

IV. Experiment

A. Experiment Settings

We evaluate PrefIQA by assessing its performance in predicting human preferences. To demonstrate the effectiveness of our model, we compare it with well-known metrics like Aesthetic Score [18], CLIP-H [14], and the recently proposed metrics like HPS [12], HPS v2 [13], ImageReward [10] and PickScore [11]. The prediction accuracy of human experts is also included. These comparisons are conducted on Pick-a-Pic test set and ImageReward test set.

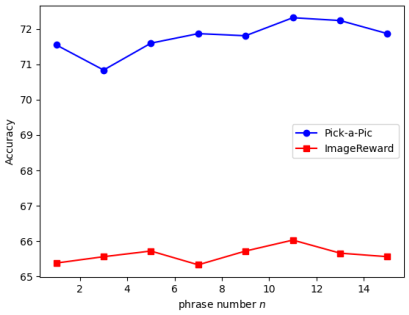

Fig. 4. The trend of accuracy variation with phrase number n.

TABLE I
PREFERENCE PREDICTION ACCURACY ON PICK-A-PIC TEST SET AND IMAGEREWARD TEST SET.

Model	Pick-a-Pic	ImageReward
Human Experts	68.0	65.3
Aesthetics [18]	56.8	57.4
CLIP-H [14]	60.8	54.8
HPS [12]	66.7	61.2
ImagReward [10]	61.1	65.1
PickScore [11]	70.2	62.9
HPS v2 [13]	69.8	65.7
Ours	**72.3**	**66.0**

1) Training Detail: We load the pretrained checkpoints of CLIP-H(ViT-H for image encoder, 24-layers transformer encoder for text encoder) and fix 50% of transformer layers. The model is trained for 4000 steps with a learning rate of 3e-6 and a batch size of 32. The model follows a linearly decaying learning rate, with a warm up period of 500 steps. Weight decay is set as 0.3.

2) The selection of phrase number: In the prompt-segmentation module, we divide a prompt into n phrases. To study the impact of different values of n on the results, we conducted experiments on Pick-a-Pic and ImageReward test sets by changing n from 1 to 15 with an interval of 2. The results is shown in Fig. 4, which approximately aligns with our expectations. When n is small, the prompt segmentation is not detailed enough, while when n is large, the prompts are too finely divided, which may have a negative impact on the model's performance on the contrary. Following the results in Fig. 4, we choose $n = 11$ to achieve a better performance.

B. Experiment Results

The experiment results are shown in Tab I. As demonstrated, PrefIQA significantly outperforms other metrics on both benchmarks. Specifically, PrefIQA outperforms PickScore by 2.1% on Pick-a-Pic dataset, and outperforms HPS v2 by 0.3% on ImageReward dataset.

TABLE II
ABLATION STUDIES ON PICK-A-PIC AND IMAGEREWARD

Model	Pick-a-Pic	ImageReward
Ours w/o Prompt-segmentation Module & Modality-fusion Module	70.2	64.8
Ours w/o Prompt-segmentation Module	71.7	65.4
Ours w/o Modality-fusion Module	71.9	65.5
Ours	72.3	66.0

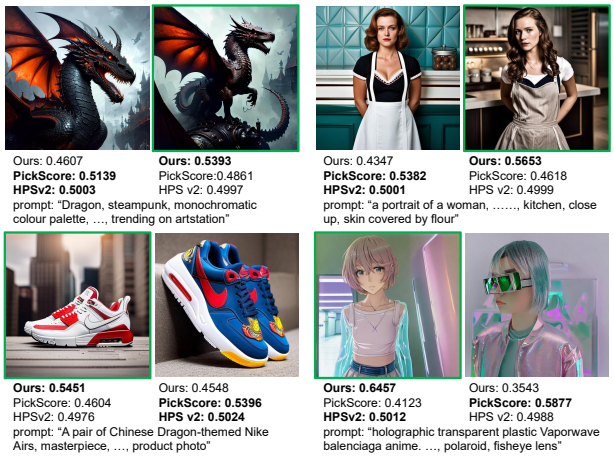

Fig. 5. Examples of preference scores obtained by PrefIQA, PickScore and HPS v2. All scores are softmaxed for better visualization. The ground truth images are indicated by a green border.

It is worth noting that human preferences are not constant: different individuals may have different preferences. To some extent, the difference restricts the improvement of models' prediction accuracy. By comparing the performance of our model with human experts, we find that the model's prediction accuracy surpasses that of the human experts(by 4.3% on Pick-a-Pic, by 0.7% on ImageReward), demonstrating its strong generalization capabilities.

For a more intuitive comparison, we also compare the scorings of different preference prediction models on specific images, as is shown in Fig. 5. It can be seen that PrefIQA tends to choose images that are more aesthetic pleasing, rather than solely focusing on the text-to-image alignment.

C. Ablation Study

In this section, we aim to validate the effectiveness of our proposed model by conducting an ablation experiment on Pick-a-Pic and ImageReward datasets. We systematically remove the prompt-segmentation or modality-fusion module and measure the prediction performance. As is shown in Tab II, it is evident that removing either prompt-segmentation module or modality-fusion module would lead to a decrease in model prediction accuracy, indicating that both modules play a significant role in improving the overall performance.

V. CONCLUSION

In this work, we propose a human preference learning metric PrefIQA for AI-generated image quality assessment . PrefIQA achieves improved prediction accuracy by utilizing the prompt-segmentation module and the modality-fusion module. The prompt-segmentation module allows for a more fine-grained evaluation of text-to-image alignment, while the modality-fusion module enables the learned image features language-aware and the learned text features vision-aware. Experimental results demonstrate that PrefIQA aligns better with human judgements than any other existing metric, bringing a promising future to the field of AI-generated image quality assessment.

REFERENCES

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[2] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[3] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[4] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[6] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, p. 5, 2022.

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

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] R. Rombach, A. Blattmann, and B. Ommer, "Text-guided synthesis of artistic images with retrieval-augmented diffusion models," *arXiv preprint arXiv:2207.13038*, 2022.

[10] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *arXiv preprint arXiv:2304.05977*, 2023.

[11] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *arXiv preprint arXiv:2305.01569*, 2023.

[12] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Better aligning text-to-image models with human preference," *arXiv preprint arXiv:2303.14420*, 2023.

[13] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.

[14] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] 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," *arXiv preprint arXiv:2010.11929*, 2020.

[18] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[19] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022.

[20] Z. Zhang, C. Li, W. Sun, X. Liu, X. Min, and G. Zhai, "A perceptual quality assessment exploration for aigc images," *arXiv preprint arXiv:2303.12618*, 2023.

[21] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *arXiv preprint arXiv:2306.04717*, 2023.

[22] J. Wang, H. Duan, J. Liu, S. Chen, X. Min, and G. Zhai, "Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence," *arXiv preprint arXiv:2307.00211*, 2023.

[23] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

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