Vision-Language Consistency Guided Multi-modal Prompt Learning for Blind AI Generated Image Quality Assessment

Jun Fu, Wei Zhou, Qiuping Jiang, Hantao Liu, Guangtao Zhai

Abstract—Recently, textual prompt tuning has shown inspirational performance in adapting Contrastive Language-Image Pre-training (CLIP) models to natural image quality assessment. However, such uni-modal prompt learning method only tunes the language branch of CLIP models. This is not enough for adapting CLIP models to AI generated image quality assessment (AGIQA) since AGIs visually differ from natural images. In addition, the consistency between AGIs and user input text prompts, which correlates with the perceptual quality of AGIs, is not investigated to guide AGIQA. In this letter, we propose vision-language consistency guided multi-modal prompt learning for blind AG-IQA, dubbed CLIP-AGIQA. Specifically, we introduce learnable textual and visual prompts in language and vision branches of CLIP models, respectively. Moreover, we design a text-to-image alignment quality prediction task, whose learned vision-language consistency knowledge is used to guide the optimization of the above multi-modal prompts. Experimental results on two public AGIQA datasets demonstrate that the proposed method outperforms state-of-the-art quality assessment models. The source code is available at https://github.com/JunFu1995/CLIP-AGIQA.

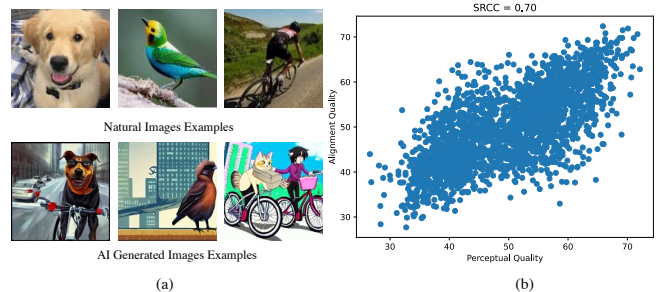*Index Terms*—Multi-modal prompt learning, Vision-language consistency, AGIQA

Fig. 1. (a) Comparison between natural images and AI generated images on the AGIQA-1K dataset [16]; (b) Spearman Rank Correlation Coefficient (SRCC) of the text-to-image alignment quality and perceptual quality on the AIGCIQA-2023 dataset [17].

I. INTRODUCTION

WITH the rapid development of deep generation technology, we have entered the era of artificial intelligence (AI) generated content, where users can obtain images they want by feeding multiple text prompts into deep generative models. However, the quality of AI generated images (AGIs) is highly varied [1]. Therefore, it is necessary to develop an objective image quality assessment (OIQA) metric to automatically filter out unqualified AGIs.

In general, OIQA metrics encompass full-reference (FR) metrics, reduced-reference (RR) metrics, and blind metrics. FR metrics and RR metrics require referencing the original image, whereas blind metrics are reference-free. In real-world scenarios, the original AGI corresponding to user input text

prompts is absent. Therefore, it is essential to develop blind IQA metrics in order to evaluate AGIs effectively.

In the early stage, blind IQA metrics are designed based on handcrafted features, e.g., mean subtracted contrast normalized coefficients [2], [3], [4], visual neuron matrix [5], and edge gradient features [6]. Since manually designing features is a time-consuming and error-prone process, researchers resort to convolutional neural networks [7] or transformers [8], and design more sophisticated IQA models [9], [10], [11]. Recently, Contrastive Language-Image Pre-training (CLIP) models are used to blindly assess the quality of natural images [12], [13], [14], and shows inspirational zero-shot performance and potential to achieve competitive performance through textual prompt tuning [15]. Motivated by the success of CLIP models in natural image quality assessment, we explore using CLIP models to assess the visual quality of AGIs in this letter.

The marriage of CLIP models and AGIQA faces its unique challenges. First, besides textual prompt tuning, it needs to mitigate the domain gap between natural images and AGIs. As shown in Fig. 1. (a), AGIs largely differ from natural images in terms of appearance and style. Second, it needs to explore using vision-language consistency to guide AGIQA. As shown in Fig. 1. (b), the alignment quality of the AGI and the user input text prompt is correlated with the perceived quality of the AGI. The reason for this phenomenon may be that users consider not only image fidelity when evaluating AGI, but also the consistency between the AGI and user input text prompts. Therefore, we believe that vision-language consistency is informative to the quality prediction of AGIs.

To tackle the aforementioned challenges, we propose a

vision-language consistency guided multi-modal prompt learning approach. Specifically, we add learnable prompts to both language and vision branches of CLIP models. In addition, we introduce a text-to-image alignment quality prediction task, whose learned vision-language consistency knowledge is used to guide the optimization of multi-modal prompts. In summary, our contribution encompasses two distinct aspects:

- As far as we know, we are the first one to explore CLIP models for blind AGIQA.
- We study the use of text-to-image alignment information to assist the visual quality prediction of AGIs.

The remainder of this paper is organized as follows. Section II introduces the proposed method in detail. Section III provides experimental results and corresponding analysis. Finally, the paper is concluded in Section IV.

II. METHOD

Our approach, dubbed as CLIP-AGIQA, aims to exploit multi-modal prompt learning to fine-tune CLIP models. Unlike previous methods [18], [19] that optimize multi-modal prompts with only the target task, we introduce an auxiliary task to guide the multi-modal prompt learning. Fig. 2 shows the overall architecture of our proposed framework. As we can see, our approach comprises a perceptual quality prediction task and a text-to-image alignment quality prediction task. Both tasks adopt multi-modal prompt learning to finetune CLIP models. Moreover, there are interactions between the learnable prompts of two tasks. During fine-tuning, the CLIP model is frozen while the rest of the proposed framework is optimized. Below, we first recap the CLIP architecture and then detail the proposed framework.

A. Recap of CLIP Models

Referring to previous prompting methods [15], [20], [18], here we adopt transformer-based CLIP models. In the CLIP model, vision and text encoders are used to generate image and text representations, respectively. The details are introduced below.

For the vision encoder $\mathcal{V}$, the input image I is divided into M fixed-size patches, and each patch is projected into d_v-dimensional latent space. The resulting patch embeddings $G_0 \in \mathbb{R}^{M \times d_v}$ and a learnable class token $c_0 \in \mathbb{R}^{d_v}$ are fed into the transformer block $\mathcal{V}_0$, which later is repeated $k-1$ times. The whole process can be formulated as,

$$[c_i, G_i] = \mathcal{V}_i([c_{i-1}, G_{i-1}]) \quad i = 1, \cdots, k. \quad (1)$$

The image representation $x \in \mathbb{R}^{d_{vl}}$ is obtained by projecting the class token c_k into d_{vl}-dimensional latent space.

For the text encoder $\mathcal{L}$, the input text description is tokenized into words, and each word is projected into d_l-dimensional latent space. The resulting word representations $R_0 = [r_0^1, \cdots, r_0^N] \in \mathbb{R}^{N \times d_l}$ are sequentially processed by k transformer layers, formulated as,

$$[R_i] = \mathcal{L}_i(R_{i-1}) \quad i = 1, \cdots, k. \quad (2)$$

The text representation $z \in \mathbb{R}^{d_{vl}}$ is obtained by projecting the last token r_k^N into the same space as the image representation.

B. Text-to-Image Alignment Quality Prediction

In the AGIQA dataset, the human-annotated text-to-image alignment scores, which reflect the consistency between AGIs and corresponding user input text prompts, are typically available. Since alignment scores are correlated with the perceptual quality of AGIs, we aim to learn the vision-language consistency knowledge to help AGI quality assessment.

A straightforward approach is to add learnable prompts into the vision encoder of the CLIP model and optimize them towards making the similariy between the AGI and user input text prompts close to the alignment score. However, the user input text prompts sometimes only contain several keywords (e.g., the AGIQA-1k dataset [16]), which are not informative. In addition, the user input text prompt is absent in some AGIQA datasets, e.g., the AIGCIQA-2023 dataset [17]. Therefore, we explore a blind setting, where we predict the alignment score without the user input text prompts.

Specifically, we use a prompt pairing strategy to estimate the alignment score of the AGI. Let us denote t_1^{align} and t_2^{align} as a pair of antonym prompts, i.e., "Aligned photo." and "Misaligned photo.". We first compute the cosine similarity between manually designed antonym prompts and the AGI as follows,

$$s_i^{align} = \frac{\mathcal{V}(I) \odot \mathcal{L}(t_i^{align})}{\|\mathcal{V}(I)\| \cdot \|\mathcal{L}(t_i^{align})\|}, i \in \{1, 2\}, \quad (3)$$

where $\|\cdot\|$ denotes l_2 norm and $\odot$ represents the vector dot-product. Then, we estimate the alignment score as follows,

$$q_{align} = \frac{e^{s_1^{align}}}{e^{s_1^{align}} + e^{s_2^{align}}}. \quad (4)$$

Since hand-crafted antonym prompts are often not optimal, we introduce b learnable prompts $P_{i-1}^{align} \in \mathbb{R}^{b \times d_l}$ into each transformer layer of the text encoder, formulated as,

$$[\_\_, R_i] = \mathcal{L}_i([P_{i-1}^{align}, R_{i-1}]) \quad i = 1, \cdots, k. \quad (5)$$

In addition, since CLIP models, pretrained on natural images, are limited to capture distinguishable image representations for AGIs, we also introduce b learnable prompts $Q_{i-1}^{align} \in \mathbb{R}^{b \times d_l}$ into each transformer layer of the vision encoder, formulated as,

$$[c_i, G_i, \_\_] = \mathcal{V}_i([c_{i-1}, G_{i-1}, Q_{i-1}^{align}]) \quad i = 1, \cdots, k. \quad (6)$$

C. Perceptual Quality Prediction

Like the text-to-image alignment quality prediction task, we also use the prompt pairing strategy to estimate the perceptual quality of the AGI. Let us denote $t_1^{percept}$ and $t_2^{percept}$ as "Good photo." and "Bad photo.", respectively. The predicted perceptual quality is computed as follows,

$$q_{percept} = \frac{e^{s_1^{percept}}}{e^{s_1^{percept}} + e^{s_2^{percept}}}. \quad (7)$$

In addition, we also adopt multi-modal prompt learning to fine-tune the CLIP model. Specifically, the formulation of the text encoder is defined as,

$$[\_\_, R_i] = \mathcal{L}_i([P_{i-1}^{percept}, R_{i-1}]) \quad i = 1, \cdots, k, \quad (8)$$
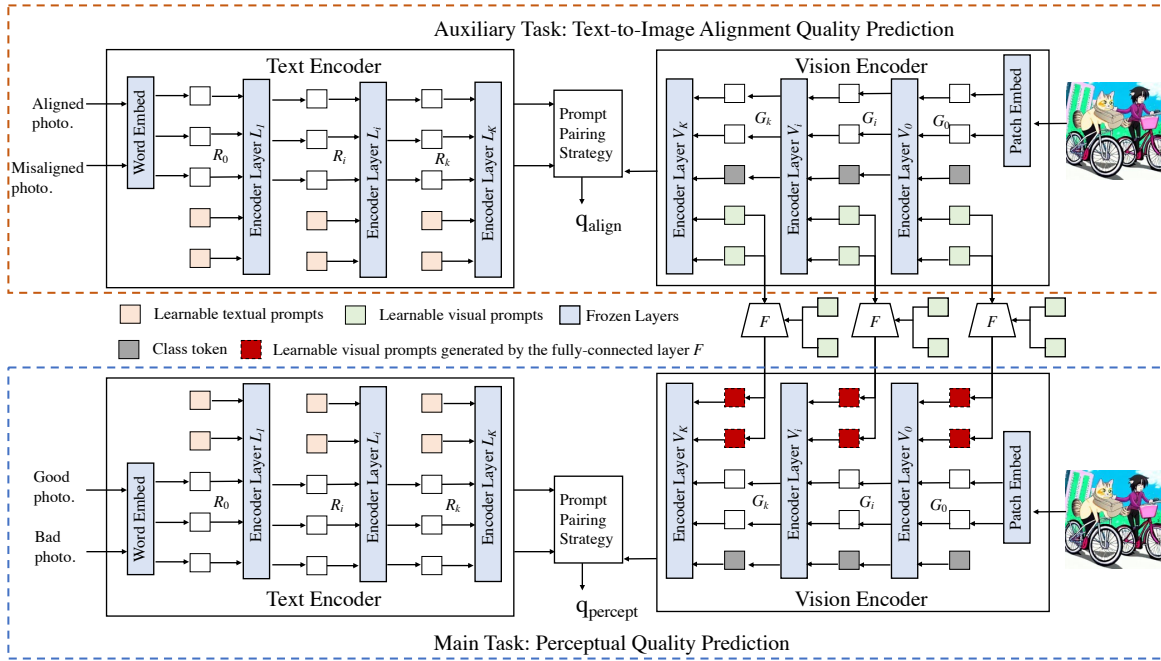
Fig. 2. Framework of our proposed method. It includes an auxiliary task and a main task. Both tasks are based on CLIP models and involve multi-modal prompt learning. Moreover, learnable visual prompts of the main task are conditioned on those of the auxiliary task.

where $P_{i-1}^{percept} \in \mathbb{R}^{b \times d_l}$ is learnable textual prompts. The vision encoder can expressed as,

$$[c_i, G_i, \_\_] = \mathcal{V}_i([c_{i-1}, G_{i-1}, \tilde{Q}_{i-1}^{percept}]) \quad i = 1, \cdots, k,$$
$$\tilde{Q}_{i-1}^{percept} = \mathcal{F}_{i-1}([Q_{i-1}^{align}, Q_{i-1}^{percept}]) \quad i = 1, \cdots, k, \quad (9)$$

where $\mathcal{F}_{i-1}$ denotes the fully-connected layer, $Q_{i-1}^{percept}$ represents learnable visual prompts in the perceptual quality prediction task. As shown in Equation 9, we explicitly condition $Q_{i-1}^{percept}$ on the learnable visual prompts Q_{i-1}^{align} in the text-to-image alignment score prediction task. The motivation behind this is that Q_{i-1}^{align} contains the vision-language consistency knowledge which is informative to the perceptual quality prediction. Notably, we empirically find that adding such conditions to textual learnable prompts brings limited gains.

D. Loss Function

The loss function for the alignment score prediction is defined as,

$$L_{align} = \frac{1}{N} \sum_{i=1}^{N} \|q_{align}^i - g_{align}^i\|_2^2, \quad (10)$$

where N is the batch size and g_{align}^i is the ground-truth alignment score of i-th AGI. The loss function for the perceptual quality prediction is defined as,

$$L_{percept} = \frac{1}{N} \sum_{i=1}^{N} \|q_{percept}^i - g_{percept}^i\|_2^2, \quad (11)$$

where $g_{percept}^i$ is the ground-truth perceptual quality of i-th AGI. The final loss function is defined as,

$$L = L_{percept} + \lambda L_{align}, \quad (12)$$

where λ is a hyperparameter.

III. EXPERIMENTS

A. Database and Evaluation Criteria

We conduct extensive experiments on two public AGIQA datasets, i.e., AGIQA-3K [21] and AIGCIQA-2023 [17]. The AGIQA-3K database contains 2982 AGIs which are generated by Glide [22], Stable Diffusion [23], Stable Diffusion XL [24], Midjourney [25], AttnGAN[26], and DALLE2 [27]. The AIGCIQA-2023 dataset uses Glide [22], Lafite [28], DALLE2 [27], Stable Diffusion [23], Unidiffuser [29], and Controlnet [30] to generate 2400 AGIs. In both datasets, each AGI is accompanied with a perceptual quality and alignment score, which are annotated by subjects. Notably, the user input text prompts are not available in the AIGCIQA-2023 dataset.

We use Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall's Rank Correlation Coefficient (KRCC) to compare IQA metrics. Good IQA methods generally achieve high scores in all three evaluation metrics. Since the AIGCIQA dataset is limited in scale, we evaluate each IQA model 10 times, and report the average performance.

B. Implement Details

We use a ViT-B/32 based CLIP model, where the length of the learnable multi-modal prompt is set to 8. The hyperparameter λ is empirically set to 0.1. The dataset is partitioned into training and testing sets at an 8:2 ratio, ensuring that images with the same user prompts are grouped together. During the training process, 64 patches with a size of 224×224 are fed into the CLIP model at each iteration. We employ Adam algorithm [31] to optimize the learnable parameters of the model. The learning rate and training epoch are configured as

TABLE I
PERFORMANCE COMPARISONS OF OBJECTIVE QUALITY METRICS ON
AGIQA-3K AND AIGCIQA-2023 DATABASES.

Type	Method	AGIQA-3k Database			AIGCIQA-2023 Database		
		SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
Handcrafted feature based	BRISQUE [2]	0.4932	0.5399	0.3348	0.6309	0.5977	0.4348
	NIQE [3]	0.5151	0.5241	0.3499	0.4870	0.4576	0.3270
	ILNIQE [4]	0.5935	0.6240	0.4183	0.5576	0.4933	0.3762
Deep learning-based	CNNIQA [9]	0.7437	0.8332	0.5516	0.6974	0.7011	0.4873
	ResNet50 [7]	0.8445	0.9033	0.6631	0.8113	0.8416	0.5956
	HyperIQA [10]	0.8433	0.9013	0.6612	0.8174	0.8459	0.6032
CLIP based	CLIPIQA [12]	0.6846	0.6987	0.4915	0.4171	0.3970	0.2823
	CLIPIQA$^+$ [12]	0.8428	0.8879	0.6556	0.8072	0.8280	0.5905
	CLIP-AGIQA	**0.8747**	**0.9190**	**0.6976**	**0.8324**	**0.8604**	**0.6220**

1e-4 and 50, respectively. During the testing, we calculate the quality score of the input AGI using a patch-based evaluation fashion [32], [33], [34]. We implement our method based on PyTorch [35], and run all experiments on a NVIDIA RTX 4090 GPU platform with an Intel Core i9-13900KF CPU.

C. Performance Comparisons

To validate the efficacy of the proposed approach, we conduct a comparative analysis with three hand-crafted feature based methods [2], [3], [4], three deep-learning based approaches [7], [9], [10], and two CLIP based metrics [12]. The results are reported in Table I. Based on the data provided in Table I, we can draw the following conclusions. First, handcrafted feature based methods achieve poor performance on both AGIQA datasets. This is because AGIs largely differ from natural images for which handcrafted features are designed. Second, the deep learning-based methods achieve relatively higher accuracy. This verifies the superiority of learned features over handcrafted ones. Third, CLIPIQA shows impressive zero-shot performance, and CLIPIQA$^+$ further improves the performance through textual prompt learning. This shows the promising potential of exploring CLIP models for AGIQA. Lastly, the proposed method called CLIP-AGIQA shows a clear advantage over CLIPIQA$^+$ on both datasets. This confirms the effectiveness of the proposed method.

TABLE II
ABLATION STUDY ON EACH COMPONENT OF THE PROPOSED METHOD.
THE TRAINING TIME AND TESTING TIME ARE CALCULATED ON IMAGES
OF SPATIAL SIZE 224×224.

Components / Method	CLIPIQA	A_1	A_2	CLIP-AGIQA
Handcrafted Text Prompts	✓			
Textual Prompt Learning		✓	✓	✓
Visual Prompt Learning			✓	✓
Vision-Language Consistency				✓
SRCC	0.6846	0.8473	0.8696	**0.8747**
PLCC	0.6987	0.8929	0.9186	**0.9190**
KRCC	0.4915	0.6611	0.6919	**0.6976**
Training time per epoch (s)	**0**	5.560	6.823	8.062
Testing time per epoch (s)	**5.531**	5.567	5.617	5.772

D. Ablation Study

We first evaluate the efficacy of each component in the proposed method. The findings are presented in Table II, from which we can infer the following conclusions. First, the variant method A_1, which only uses textual prompt learning, achieves better performance than CLIPIQA which uses handcrafted text prompts. This confirms the necessity of using textual prompt learning. Second, the variant method A_2, which uses

textual and visual prompt learning, is superior to A_1. This confirms the advantage of multi-modal prompt learning over textual prompt learning. Third, the proposed method slightly outperforms A_2. This shows that the vision-language consistency knowledge is informative to AGIQA. Fourth, while the proposed method has much higher training cost than CLIPIQA which does not require training, its inference cost is comparable to CLIPIQA's. This is mainly because the auxiliary task can be discarded in the testing phase.

TABLE III
THE IMPACT OF VISION-LANGUAGE CONSISTENCY ON THE
PERFORMANCE.

Method / Metrics	SRCC	PLCC	KRCC
B_1	0.8741	0.9182	0.6965
CLIP-AGIQA	**0.8747**	**0.9190**	**0.6976**

Subsequently, we conduct an investigation into the vision-language consistency. The results are shown in Table III. B_1 adopts the same framework as the proposed method, while learning the vision-language consistency knowledge with user input text prompts. More specifically, in the text-to-image quality prediction task, it feeds user input text descriptions into the text encoder without learnable textual prompts. According to Table III, we find that the proposed method exhibits a slight superiority over B_1. The possible reason for this phenomenon is as follows. For the text-to-image alignment quality prediction, the blind setting (i.e., without user input text prompts) is usually more challenging than the non-blind setting (i.e., with user input text prompts), which helps us learn non-trivial vision-language consistency knowledge.

TABLE IV
COMPARISON WITH COMPETITIVE PROMPT LEARNING METHODS.

Metrics / Method	CoCoOP [20]	MaPLe [18]	CLIP-AGIQA
SRCC	0.8582	0.8713	**0.8747**
PLCC	0.9079	0.9176	**0.9190**
KRCC	0.6759	0.6939	**0.6976**

Finally, we compare the proposed method with two competitive prompt learning methods, i.e., CoCoOP [20] and MaPLe [18]. For fair comparison, these two methods share the same training and testing settings as our proposed method. The results are presented in Table IV. As shown, the proposed method exceeds CoCoOP by a clear margin. Moreover, the proposed method slightly outperforms the multi-modal prompt learning approach known as MaPLe. This can be attributed to the learned vision-language consistency knowledge.

IV. CONCLUSION

In this letter, we propose vision-language consistency guided multi-modal prompt learning to adapt CLIP models to blindly assess the visual quality of AI generated images. Experiments evince that our approach achieves more accurate predictions than existing IQA metrics, and each technical component in our method plays a crucial role. However, since the auxiliary task is designed as text-to-image alignment quality prediction, our method cannot be applied to the scenario where alignment quality scores are unavailable. Therefore, we will explore better auxiliary tasks in the future.

REFERENCES

[1] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[2] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[3] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[4] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[5] H.-W. Chang, X.-D. Bi, and C. Kai, "Blind image quality assessment by visual neuron matrix," *IEEE Signal Processing Letters*, vol. 28, pp. 1803–1807, 2021.

[6] C. Feichtenhofer, H. Fassold, and P. Schallauer, "A perceptual image sharpness metric based on local edge gradient analysis," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 379–382, 2013.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[9] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[10] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[11] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 433–442.

[12] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

[13] T. Miyata, "Interpretable image quality assessment via clip with multiple antonym-prompt pairs," *arXiv preprint arXiv:2308.13094*, 2023.

[14] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 071–14 081.

[15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[16] 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.

[17] 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:

[22] 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.

from the perspectives of quality, authenticity and correspondence," *arXiv preprint arXiv:2307.00211*, 2023.

[18] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.

[19] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia*, 2023.

[20] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[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.

[23] 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.

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

[25] D. Holz, "Midjourney," *https://www.midjourney.com/*, 2023.

[26] 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.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[28] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 907–17 917.

[29] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," *arXiv preprint arXiv:2303.06555*, 2023.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] J. Fu, "Scale guided hypernetwork for blind super-resolution image quality assessment," *arXiv preprint arXiv:2306.02398*, 2023.

[33] W. Zhou, Q. Jiang, Y. Wang, Z. Chen, and W. Li, "Blind quality assessment for image superresolution using deep two-stream convolutional networks," *Information Sciences*, vol. 528, pp. 205–218, 2020.

[34] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3946–3958, 2019.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.