

Align-IQA: Aligning Image Quality Assessment Models with Diverse Human Preferences via Customizable Guidance

Anonymous Author(s)

1 APPENDIX

1.1 The effectiveness of each component of Align-IQA

As reflected in Tab. 1, we conduct individual experiments on the AI-generated content (AIGC) dataset to analyze the effectiveness of each component of our proposed Align-IQA. The results demonstrate that our CGI module effectively incorporates quality-aware prior knowledge into the general-purpose pre-trained model, guiding it to prioritize extracting quality-aware features over semantic-aware features. By injecting specializable quality-aware prior knowledge into general-purpose pre-trained models, our Align-IQA allows for different adjustments of features to be consistent with diverse human preferences for various types of visual content. Meanwhile, our MSFA module effectively enhances quality-aware features from a human perception perspective, underscoring the importance of integrating multi-scale information for precise quality assessment. By simulating the multi-scale mechanism in the human visual system, our Align-IQA achieves further improvements in accuracy. Notably, our Align-IQA performs best when combining the CGI and MSFA modules. This indicates that our approach accurately replicates the functionality of the human eye, considering numerous intricate factors that impact visual quality perception.

1.2 The effectiveness of different backbone networks

We conduct comparative experiments on the AGIQ-3K dataset (AGIQ-3K) to evaluate the performance of three different types of backbone networks. These include ViT-Tiny/16, ViT-Small/16, and ViT-Base/16[1]. The input image size for all backbone networks is standardized at 224×224 , with the image patch shape defined as 16×16 . Our results, which are provided in Tab. 2, reveal that utilizing ViT-Base/16 yield the best performance among the tested backbone networks. This demonstrates that employing a deeper and wider backbone network allows for capturing richer quality-aware representations, ultimately leading to improved overall performance in our experiments.

1.3 The effectiveness of different injection strategies

To evaluate the effectiveness of our CGI module on the AGIQ-3K dataset, two additional injection strategies [2] (Fig. 1 (a) & (b)) are selected to introduce guidance tokens into the general-purpose pre-trained model. One strategy involves the summation of the input guide and image embedding. The other strategy entails concatenating the input guide and image embedding. It is important to note that for AI-generated images, the input guide is replaced with the embedding of the visual-semantic relation information obtained from the vision-language model (e.g., [4]). The performance comparison between these two injection strategies and our

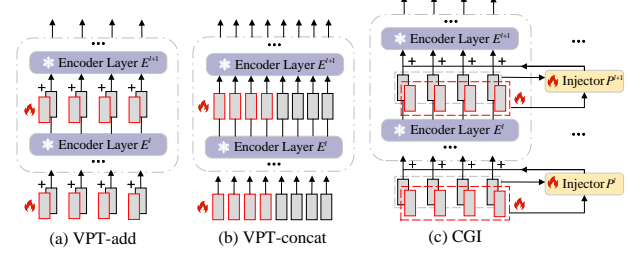


Figure 1: Variants of vanilla injection-structure and our CGI model.

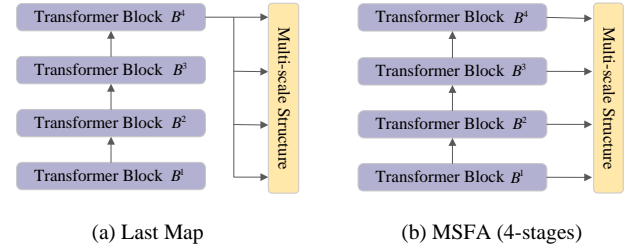


Figure 2: Building a feature pyramid on the vanilla ViT. (a) Using solely the final feature map of the vanilla ViT. (b) Our MSFA module: the vanilla ViT is artificially segmented into multiple stages.

Table 1: Comparison of each component of our Align-IQA on the AGIQ-3K dataset.

CGI	MSFA	AGIQ-3K	
		SRCC	PLCC
×	×	0.850	0.915
✓	×	0.866	0.918
×	✓	0.871	0.922
✓	✓	0.874	0.924

Table 2: Comparison of different backbone networks employed by our Align-IQA on the AGIQ-3K dataset.

	AGIQ-3K	
	SRCC	PLCC
ViT-Tiny/16	0.836	0.901
ViT-Small/16	0.860	0.914
ViT-Base/16	0.874	0.924

CGI module is presented in Tab. 3. Our analysis reveals that our

Table 3: Comparison of different injection strategies employed by our Align-IQA on the AGIQA-3K dataset.

	AGIQA-3K	
	SRCC	PLCC
VPT-add	0.857	0.915
VPT-concat	0.863	0.920
CGI	0.874	0.924

Table 4: Comparison of different strategies for building a feature pyramid on the AGIQA-3K dataset.

Strategy	Last Map		MSFA(4-stages)	
	SRCC	PLCC	SRCC	PLCC
AGIQA-3K	0.859	0.919	0.874	0.924

CGI module outperforms these alternative injection strategies. This shows that our approach effectively leverages guidance tokens to improve the capabilities of general-purpose pre-trained models in acquiring quality-aware features aligned with human preference for AI-generated images.

1.4 The effectiveness of different strategies for extracting multi-scale features

To test the effectiveness of our MSFA model in building a feature pyramid using the multi-level output of ViT, an additional strategy [3] (refer to Fig. 2 (a)) is selected for comparison. In this strategy, only the output from the last layer of ViT is utilized to build the feature pyramid. The SRCC and PLCC results are summarized in Tab. 4. It is observed that enhanced performance is achieved through the use of multi-level output for the extraction of multi-scale features. This suggests that our Align-IQA is capable of more robustly assessing image quality by comprehensively considering the diverse and informative multi-level features provided by ViT.

REFERENCES

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [2] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.
- [3] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*. Springer, 280–296.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.