

Supplementary Materials: AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception

Anonymous Authors

AesExpert: Towards Multi-modality Foundation Model for Image Aesthetics Perception

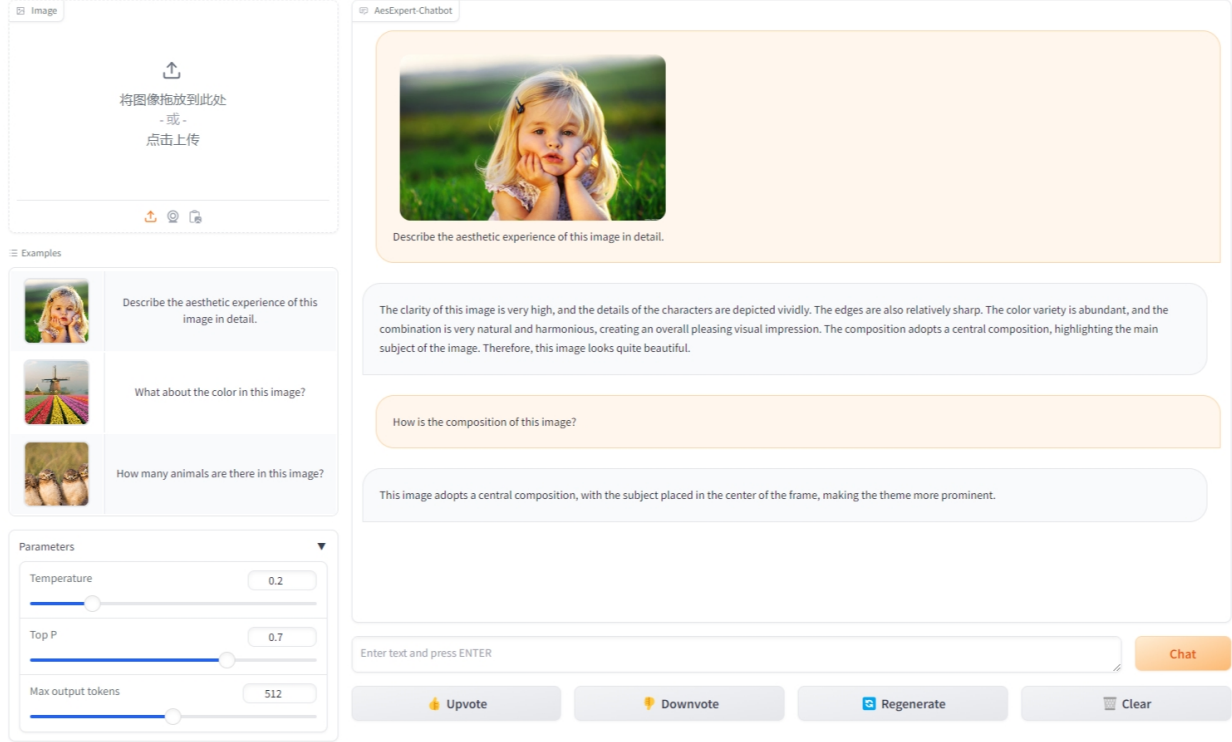


Figure 1: The screenshot of the designed AesExpert-Chatbot demo, which will be publicly available.

1 BUILDING THE AESEXPERT-CHATBOT

To reduce the difficulty of deploying multi-modal large language models (more than 12GB graphic memory), we further design a user-interactive Chatbot for the proposed multi-modality Aesthetic Expert models (named **AesExpert-Chatbot**), as shown in Figure 1. AesExpert-Chatbot can be used directly on devices with a browser without additional graphic memory (not limited to Windows, Ubuntu, macOS, or Android), facilitating human-machine interaction in aesthetics tasks. Technical details are as follows.

System Configuration: The AesExpert-Chatbot model is performed based on Gradio [3], which is an open-source Python library. Considering the inference speed, we deployed the well-trained AesExpert model on 2 NVIDIA RTX 4090 24G GPUs with a good network environment to ensure the stability of operation.

User Interface Design: The principles of the user interface include simplicity, usability, and comprehensiveness. First, it has a simple layout with text boxes for user input and chatbot response.

For the input image, the interface provides three different selection methods, including dragging an image into the specified area, selecting an image from the specified file, and real-time acquisition by the camera. In addition, we also provide three examples of different question types to allow users to quickly master how to use it. Finally, we optimize the interface for response speed and also ensure compatibility across different devices and screen sizes.

Deployment and Evaluation: We leverage the function of network sharing in Gradio for online deployment to load the user interface on a publicly accessible URL. Then, we conduct comprehensive testing for AesExpert-Chatbot and optimize it based on collected user feedback. Finally, we provide the stable AesExpert-Chatbot, as shown in Figure 1. We hope this work can provide high-quality resources to the community to promote the further development of multi-modal large language models in image aesthetics perception.

Table 1: Hyper-parameters for AesExpert models.

Model	LLaVA-1.5	mPLUG-Owl2
input size	336×336	448×448
batch size	128	256
gradient ACC	1	16
lr max	2e-5	2e-5
lr schedule	cosine decay	cosine decay
warmup epochs	0.03	0.03
weight decay	0	0
numerical precision	bfloat16	bfloat16
epoch	1	1
optimizer	AdamW	AdamW
optimizer sharding	True	True
activation checkpointing	True	True

2 PROMPTS FOR BUILDING AESMMIT WITH GPT-4

While the corpus-rich aesthetic critiques in the **AesFeedback** dataset can provide rich knowledge for aesthetic instruction tuning of MLLMs, we further leverage GPT to transform human feedback into instruction-following formats [1]. GPT prompts for instruction generation are as follows:

#System: You are a helpful assistant that can generate a question-and-answer pair based on the given description of image aesthetics perception. The purpose of these questions is to ask what can be found in the images, so these questions should not contain text information to aid judgment. The options should be concise and only contain the core information with a minimum of words. You should also generate several false answers for each question under the key of false candidates, which are also reasonable given the question contradicts the description. Organize the output a list in JSON format and when you respond, please only output the JSON, no other words are needed:

Yes-or-No-style questions: *#User: These questions are with the beginning of Does. The given description: [\$DESC].*

Why-style questions: *#User: These questions are with the beginning of Why. The given description: [\$DESC].*

What-style questions: *#User: These questions are with the beginning of What. The given description: [\$DESC].*

How-style questions: *#User: These questions are with the beginning of How. The given description: [\$DESC].*

3 HYPER-PARAMETERS DURING FINE-TUNING

In this work, we directly use the proposed AesMMIT dataset to perform supervised instruction fine-tuning on the models pre-trained on general-purpose visual tasks [2, 5]. For LLaVA-v1.5 (7B/13B) [4], the process of instruction tuning is conducted on 8 NVIDIA Tesla A100 80G GPUs (requiring about 6 hours for 7B, 10 hours for 13B). For mPLUG-Owl2 [6], the process of instruction tuning is also conducted on 8 NVIDIA Tesla A100 80G GPUs (requiring about 6.5 hours). All hyper-parameters are recorded in Tabel 1.

4 QUALITATIVE ANALYSIS

To visually verify the aesthetic perception performance of the proposed AesExpert model, we invited some aesthetic researchers to conduct tests in actual application scenarios. These conversations exhibit its excellent abilities on various aesthetic tasks, including aesthetic evaluation and aesthetic interpretation (in Figure 2), composition perception (in Figure 3), emotion understanding (in Figure 4) and enhancement suggestion (in Figure 5). We will make the AesExpert demo publicly available, and we hope to collect more user suggestions to improve the aesthetic perception capabilities of multi-modality large language models.

REFERENCES

- [1] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793* (2023).
- [2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500* (2023).
- [3] Gradio. 2024. *Build Share Delightful Machine Learning Apps*. <https://www.gradio.app>
- [4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. *arXiv preprint arXiv:2310.03744* (2023).
- [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023).
- [6] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *arXiv preprint arXiv:2311.04257* (2023).

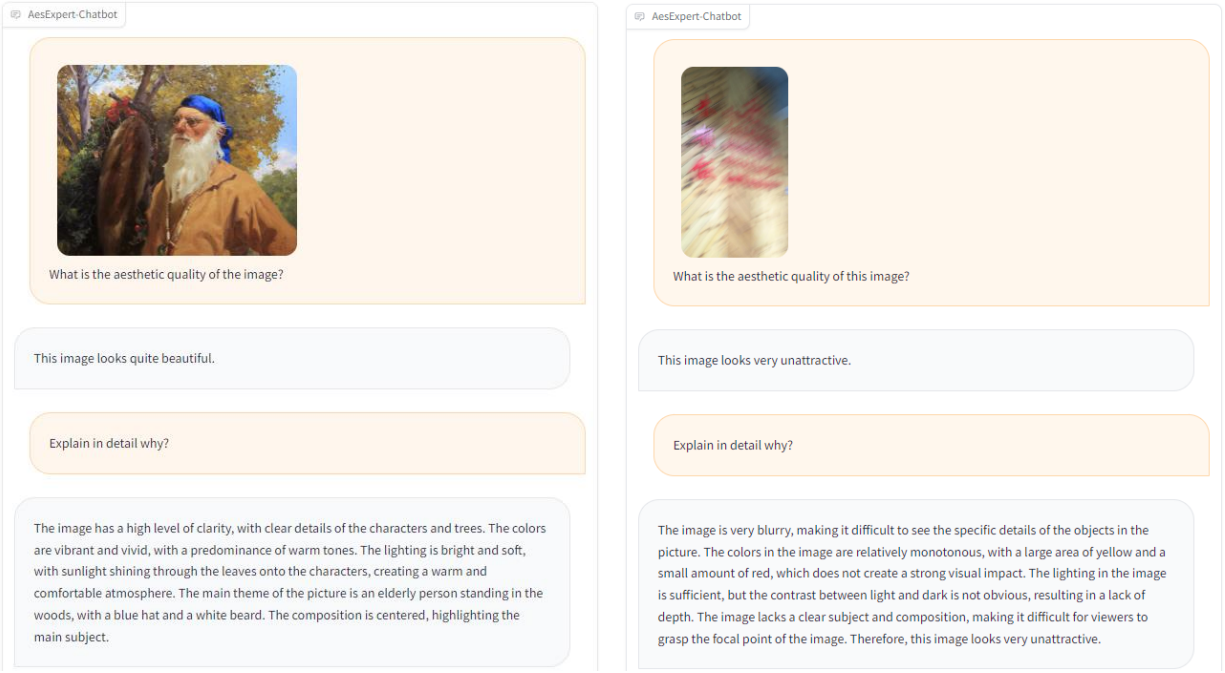


Figure 2: Aesthetic evaluation and aesthetic interpretation.

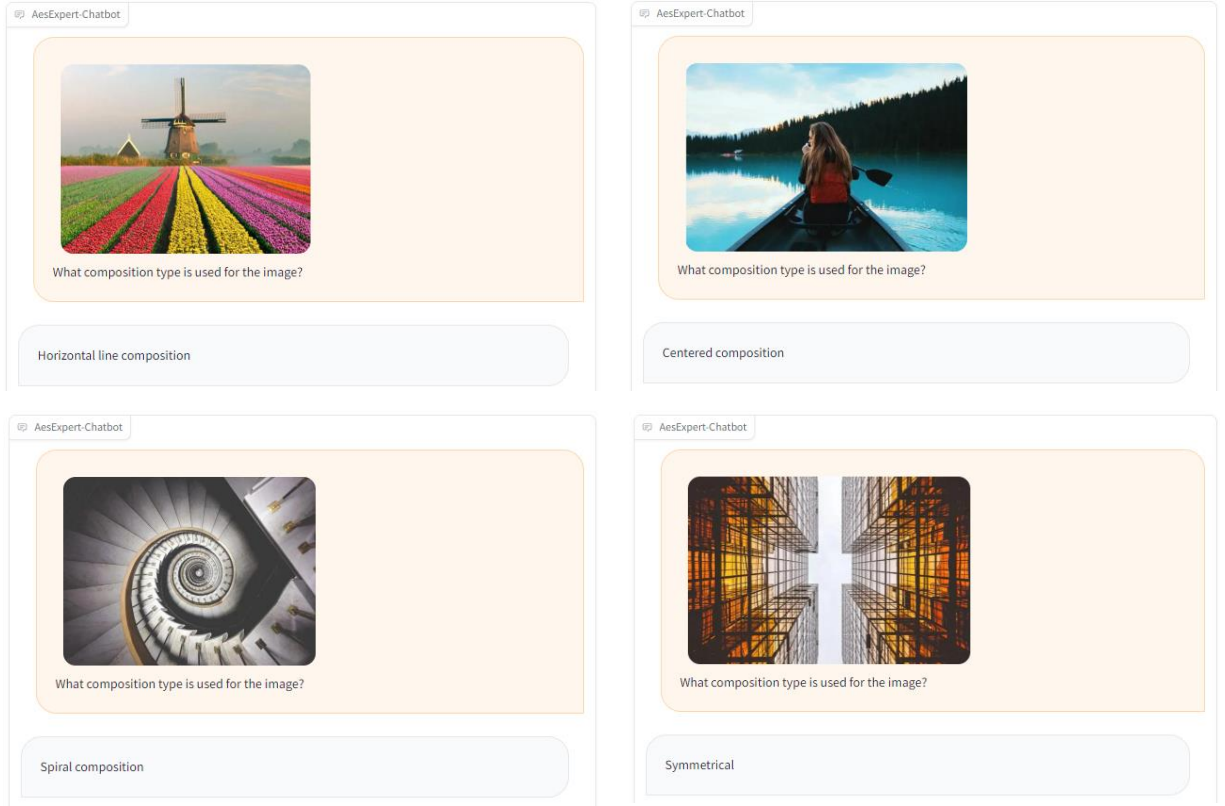


Figure 3: Composition perception.

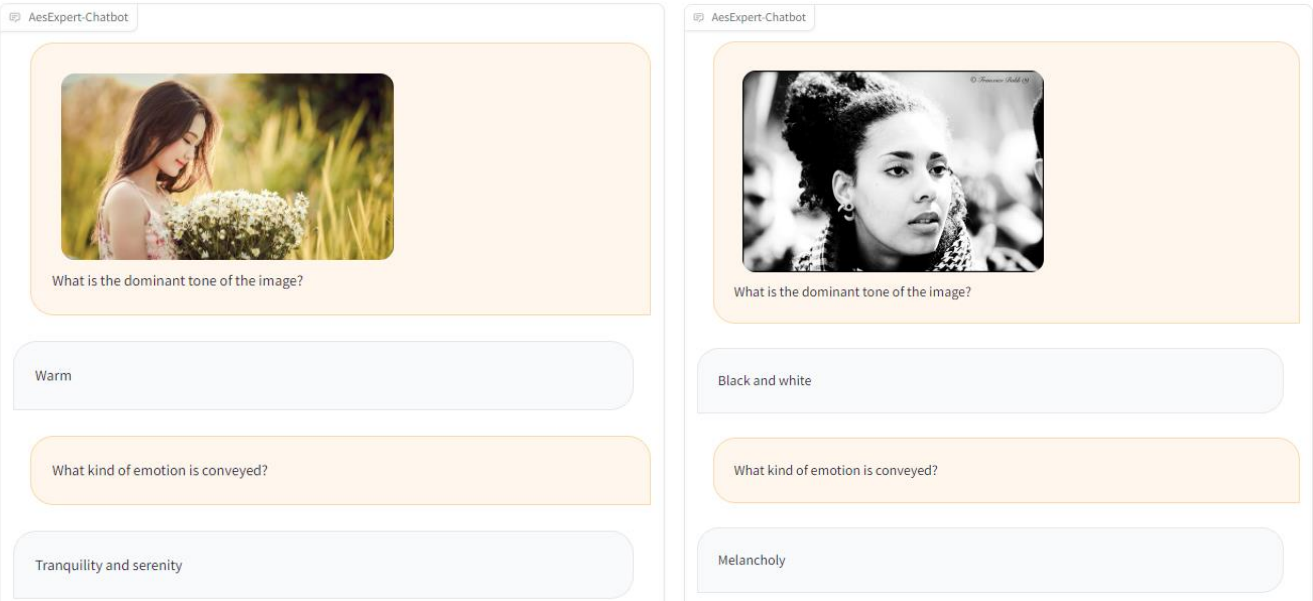


Figure 4: Emotion understanding.

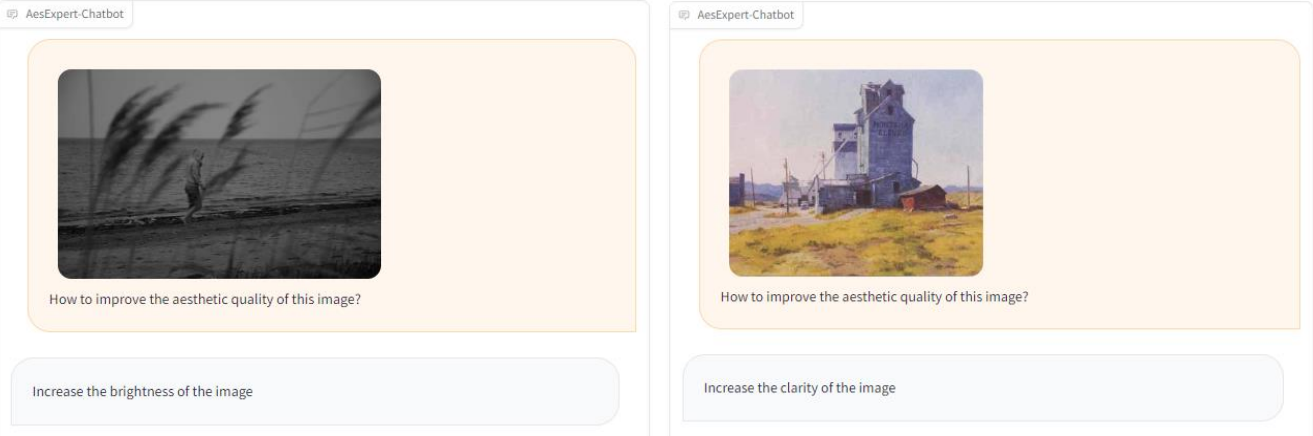


Figure 5: Enhancement suggestion.