Supplemental Material for AI-generated Image Quality Assessment in Visual Communication

Yu Tian[1], Yixuan Li[1], Baoliang Chen[2], Hanwei Zhu[1], Shiqi Wang[1], Sam Kwong[3]
[1]City University of Hong Kong, Hong Kong SAR, China
[2]South China Normal University, Guangzhou, China
[3]Lingnan University, Hong Kong SAR, China
{ytian73-c, hanwei.zhu}@my.cityu.edu.hk, yixuanli423@gmail.com, blchen@scnu.edu.cn, shiqwang@cityu.edu.hk, samkwong@ln.edu.hk

More Details of AIGI-VC Database

Prompt Generation

Our prompts consist of three components: subject, attributes, and setting. In the process of generating prompts, we first determine the subject, which refers to the product in a product advertisement (ad) or the topic in a public service announcement (PSA) that we want to generate. After that, we use GPT-4v (OpenAI 2023) using the following the query:

> *I want you to act as a graphic designer for advertising. I will provide a product/topic. Your task is to design 10 diverse, detailed and creative ads designed for the product/topic. Please describe your design contains the following elements: 1) What is the actual object presented (Subject) 2) What are the attributes of the subject (Attributes) 3) What is the setting in which the product/topic could be presented well (Setting) 4) What is the dominant emotion that this ad aims to evoke in the audience? The answer must be selected from the following emotions: amusement, awe, contentment, excitement, anger, disgust, fear and sadness. Convert the subject, attributes and setting to a description (within 35 words) for text-to-image generation. The output should strictly following: Description: {} Emotion:{}.*

We use BLEUScore (Papineni et al. 2002) to calculate the similarity of textual information between two prompts and remove one of the prompts if the similarity score is greater than 0.9. This process is repeated until the number of prompts reaches 500.

Interface for Subjective Experiments

The interface for the subjective experiments is built using Flask 2.0.2 and set up locally. Complying with the ITU-R BT.500 (Series 2012) standard, we set up the experimental environment as a normal indoor home setting, with normal lighting levels. Before starting the experiments, the participants receive a brief introduction explaining the purpose of the study and the evaluation process. A screenshot of the interface is shown in Fig. 1. Participants are presented with two images generated from the same prompt, displayed side by side. They are then required to choose the preferred image based on the two questions provided below.

Figure 1: The interface for subjective experiments. Participants are required to choose the preferred image from a pair based on two questions.

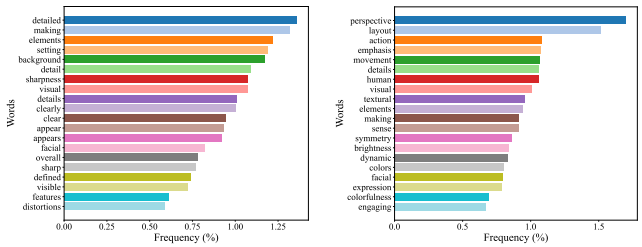
Statistical Analyses for Fine-grained Descriptions

We present the frequently occurring words in the proposed AIGI-VC dataset, highlighting terms significant for information clarity and emotional interaction. As shown in Fig. 2, we can observe that 1) some high-frequency words, such as "visual", "details", "elements", and "facial", are common across both two evaluation dimensions, indicating the critical role of these fundamental visual elements in assessing AI-generated image quality; 2) for information clarity, the high-frequency words like "detailed", "appear", "setting", and "background" illustrate a strong emphasis on the precise depiction and clarity of image content; 3) for emotional interaction, the high-frequency words, such as "perspective", "layout", "emphasis" and "action", demonstrate the organization and presentation of visual elements are essential for effectively conveying emotions.

More Details of Evaluation

Prompt Settings for Preference Prediction

When evaluating the performance of LMMs in preference prediction, we use the predefined queries of information clarity and emotional interaction, respectively. Let `<Image>` denote the image tokens. `<Text>` and



(a) Information Clarity (b) Emotional Interaction

Figure 2: Statistics of top-20 frequently-used words in the AIGI-VC database.

Model	Dimension	*B.*↑	*O.*↑	*R.-L*↑
LLaVA-v1.5-13B	IC	0.0294	0.0361	**0.2479**
	EI	0.0075	0.0093	0.1995
	Overall	0.0185	0.0227	**0.2237**
BakLLava	IC	0.0187	0.0227	0.2364
	EI	0.0149	0.0183	0.2029
	Overall	0.0168	0.0205	0.2197
mPLUG-Owl2	IC	0.0215	0.0265	0.2352
	EI	0.0094	0.0115	0.1981
	Overall	0.0154	0.0190	0.2167
InternLM-XC.2-vl	IC	0.0142	0.0175	0.2125
	EI	0.0115	0.0139	0.1972
	Overall	0.0129	0.0157	0.2049
IDEFICS-Instruct	IC	0.0142	0.0177	0.2223
	EI	0.0055	0.0069	0.1478
	Overall	0.0098	0.0123	0.1851
Qwen-VL-Chat	IC	0.0176	0.0214	0.2245
	EI	0.0054	0.0066	0.1657
	Overall	0.0115	0.0140	0.1951
GPT-4o	IC	**0.0562**	**0.0623**	0.2228
	EI	**0.0471**	**0.0514**	**0.2047**
	Overall	**0.0516**	**0.0568**	0.2138

Table 1: Comparison of the **preference interpretation** abilities of LMMs using automatic metrics. IC: Information clarity. EI: Emotional interaction. The best two results are highlighted in bold and underlined, respectively.

`<Emotion>` represent the textual information and emotion category of the image. For information clarity, the query is described as follows:

> *This is the first image: `<Image>`. This is the second image: `<Image>`. Information clarity measures whether all elements in a given prompt are visually presented in the image, and how clearly the image presents these details of each of these elements. The given prompt is: `<Text>`, which image has higher information clarity? Please respond with either "the first" or "the second".*

For emotional interaction, the query is described as follows:

> *This is the first image: `<Image>`. This is the second image: `<Image>`. Which image evokes more `<Emotion>`? Please respond with either "the first" or "the second".*

Model	Dimension	*B.*↑	*O.*↑	*R.-L*↑
LLaVA-v1.5-13B	IC	0.0144	0.0375	0.2185
	EI	**0.0121**	**0.0146**	**0.2142**
	Overall	0.0132	0.0260	0.2163
BakLLava	IC	0.0067	0.0083	0.1934
	EI	0.0054	0.0066	0.1817
	Overall	0.0060	0.0075	0.1876
mPLUG-Owl2	IC	**0.0398**	**0.0486**	0.2446
	EI	0.0057	0.0071	0.1816
	Overall	0.0227	0.0279	0.2131
InternLM-XC.2-vl	IC	0.0372	0.0454	**0.2614**
	EI	0.0089	0.0110	0.2064
	Overall	**0.0230**	**0.0282**	**0.2339**
IDEFICS-Instruct	IC	0.0270	0.0337	0.2307
	EI	0.0028	0.0035	0.1393
	Overall	0.0149	0.0186	0.1850
Qwen-VL-Chat	IC	0.0149	0.0183	0.2079
	EI	0.0032	0.0040	0.1657
	Overall	0.0090	0.0112	0.1868
GPT-4o	IC	0.0317	0.0385	0.2431
	EI	0.0072	0.0088	0.1961
	Overall	0.0194	0.0237	0.2196

Table 2: Comparison of the **preference reasoning** abilities of LMMs using automatic metrics. IC: Information clarity. EI: Emotional interaction. The best two results are highlighted in bold and underlined, respectively.

Details on Competing LMMs

In experimental section, we employ seven LMMs to evaluate their performance in assessing the communicability of AIGIs in visual communication. Typically, LMMs consist of three core components: a modality encoder, a language model, and a modality interface for cross-modal interactions. The detailed structures for these LMMs are provided in Table 3.

Additional Experiment

Performance on Interpretation and Reasoning

In the main paper, we use the GPT-assisted evaluation method to evaluate the performance of LMMs in interpretation and reasoning. GPT assesses text similarity based on semantic understanding and contextual awareness. Additionally, we employ three automatic metrics, including BLEUScore (*B.*) (Papineni et al. 2002), ORANGE (*O.*) (Lin and Och 2004), and ROUGE-L (*R.-L*) (Lin 2004). BLEUScore and ORANGE measure the performance of LMMs by comparing n-gram matches between the LMM responses and reference texts. ROUGE-L compares the LMM responses against the references via the longest common subsequence. The higher values of BLEUScore, ORANGE, and ROUGE-L indicate the higher similarity between the LMM responses and the golden descriptions. The results are shown in Tables 1&2. From the results in Table 1, we can see that GPT-4o achieves the most top performance in preference interpretation. More specifically, GPT-4o achieves the best performance 7 times, followed by LLaVa-v1.5-13B (2 times). From the results in Table 2, InternLM-XC.2-v

Model	Visual Encoder	Visual-Language Alignment	Language Model
LLaVA-v1.5-13B	CLIP-ViT-Large/14	MLP	Vicuna-v1.5-13B
BakLLava	CLIP-ViT-Large/14	MLP	Mistral-7B
mPLUG-Owl2	CLIP-ViT-Large/14	MAM	LLaMA2-7B
InternLM-XC.2-vl	CLIP-ViT-Large/14	Partial LoRA	InternLM2-7B
IDEFICS-Instruct	CLIP-ViT-Large/14	Cross-Attention	LLaMA-7B
Qwen-VL-Chat	Openclip-ViT-bigG	Cross-Attention	Qwen-7B

Table 3: Overview of the competing LMMs. MLP and MAM are the multilayer perception and the modality-adaptive module, respectively.

achieves the best overall quality in preference reasoning across all automatic metrics.

References

Lin, C.-Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, 74–81.

Lin, C.-Y.; and Och, F. J. 2004. ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *COLING: Proc. Int. Conf. Computational Linguistics*, 501–507.

OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proc. 40th Annual Meeting of The Association for Computational Linguistics*, 311–318.

Series, B. T. 2012. Methodology for The Subjective Assessment of The Quality of Television Pictures. *Recommendation ITU-R BT*, 500(13).