

Supplementary Materials for T2I-Scorer: Quantitative Evaluation on Text-to-Image Generation via Fine-Tuned Large Multi-Modal Models

Anonymous Authors

1 RELEVANCE TO ACMMM CONFERENCE

This pioneering study leverages the capabilities of Large Multi-modal Models (LMMs) to rigorously assess the generation of multimedia content, specifically through text-to-image processes. This novel approach is particularly significant because both the evaluation mechanism and the subject of the evaluation are intrinsically multimodal. The exploration of LMMs in this context is not only timely but is poised to set a new benchmark in the field of multimedia generation. The findings from this study are expected to contribute valuable insights into the advancements of AI-driven multimedia tools, making a significant impact on both academic research and practical applications in various industries.

2 PROMPTS FOR GPT-4V TO ANNOTATE

2.1 On Single Images.

Analyze whether the generated image has good visual quality. Do this in the following steps. First, generate a JSON file that asks questions about whether the objects (main object, background, etc) in the image meets criteria related to a high-quality image, and answer with "Yes" or "No", with the format: ["question": str, "answer": str ("Yes" or "No"), ...]. Then, generate a JSON file that asks and rates the multi-dimensional quality (e.g. overall, clarity, lighting, structure,) of the image with the format: ["question": str, "answer": str ("Good", "Fair" or "Poor"), ...] Please only generate the two lists in your answer.

2.2 On Image Pairs.

Analyze which generated image has better visual quality. Do this in the following steps. First, generate a JSON file that asks questions COMPARING the two images on the criteria related to quality, and answer with only "Yes" or "No", with the format: ["question": str, "answer": str ("Yes" or "No"), ...]. Then, generate a JSON file that asks and compares the multi-dimensional quality (e.g. overall, clarity, lighting, structure,) of the image with the format: ["question": str, "answer": str ("First image", "Second image" or "Tie"), ...] Please only generate the two lists in your answer.

3 TRAINING HYPER-PARAMETERS

The T2I-Scorer consists of 8.2B parameters, which are fully updated during the training process. The hyper-parameters for the two training stages are listed as in Tab. 1 and Tab. 2.

Hyper-parameter	T2I-Scorer-IT (Training Stage 1)
ViT init.	CLIP-ViT-Large-psz14
LLM init.	LLaMA-2-7B
LMM init.	mPLUG-Owl2
image resolution	448 × 448
batch size	192
lr max	2e-5
lr schedule	cosine decay
lr warmup ratio	0.03
weight decay	0
gradient acc.	2
numerical precision	bfloat16
epoch	1
warm-up epochs	0.03
optimizer	AdamW
optimizer sharding	✓
activation checkpointing	✓
model parallelism	2
pipeline parallelism	1

Table 1: Hyper-parameters for training T2I-Scorer-IT.

Hyper-parameter	T2I-Scorer (Training Stage 2)
ViT init.	CLIP-ViT-Large-psz14
LLM init.	LLaMA-2-7B
LMM init.	T2I-Scorer-IT
image resolution	448 × 448
batch size	192
lr max	2e-5
lr schedule	cosine decay
lr warmup ratio	0.03
weight decay	0
gradient acc.	2
numerical precision	bfloat16
epoch	1
warm-up epochs	0.03
optimizer	AdamW
optimizer sharding	✓
activation checkpointing	✓
model parallelism	2
pipeline parallelism	1

Table 2: Hyper-parameters for training T2I-Scorer.

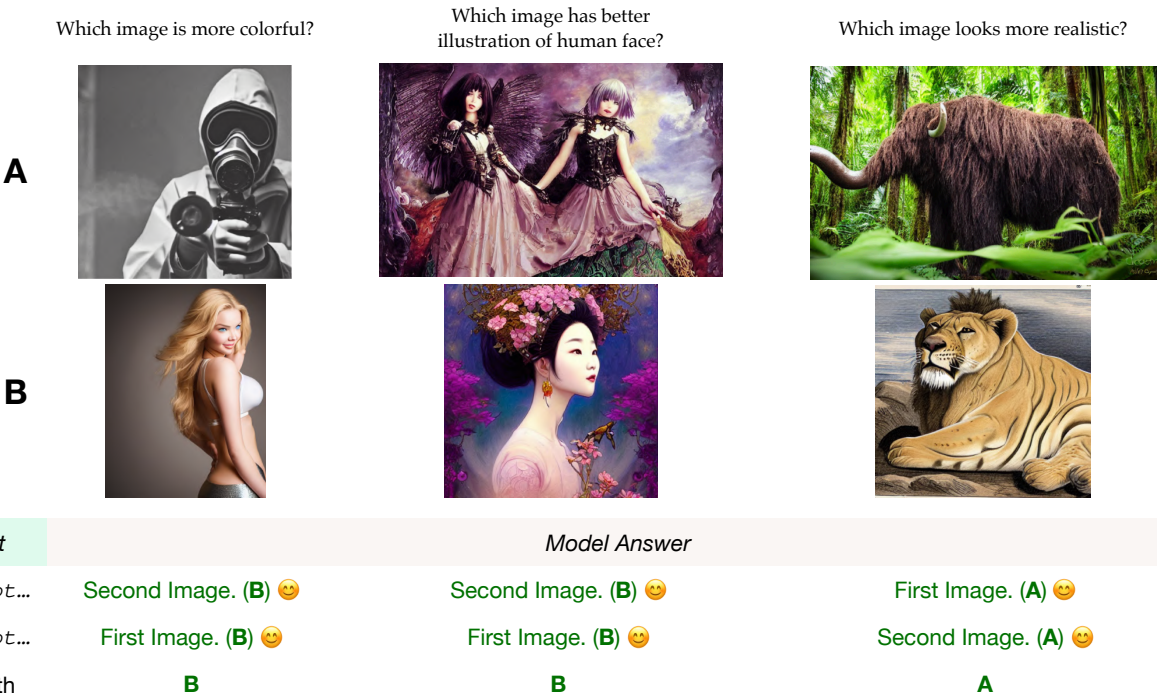


Figure 1: Qualitative study on the pairwise comparison ability of T2I-Scorer-IT.

4 T2I-SCORER-IT ON IMAGE PAIRS

While the first training stage has involved instruction tuning image pairs, our main motivation is to better enhance the general quality perception ability on T2I-generated images, which is discussed in Tab. 5 and Tab. 6 of our main paper. In Fig. 1, we further prove that the proposed T2I-Scorer-IT has learnt to effectively compare the quality-related aspects between two generated images. We

hope to explore how to integrate the pairwise evaluation into real applications for T2I evaluations in the future.

5 ETHICAL STATEMENT ON HUMAN EXAMINATION OF T2I-IPT

During the human examination process of T2I-IPT, we set a REPORT button for human experts to report any inappropriate or violent content. We have not received such reports during examination.