VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation

Anonymous ACL submission

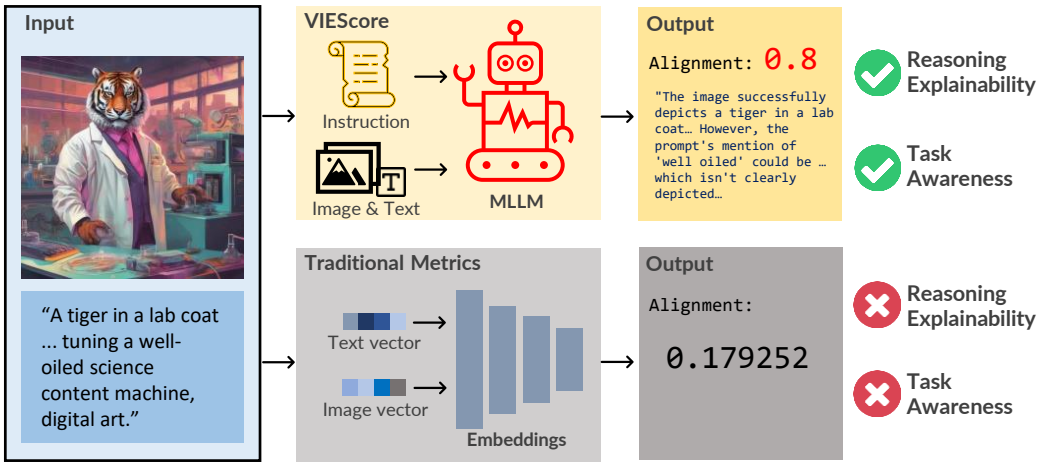

Figure 1: Which method (VIESCORE or traditional metrics) is "closer" to the human perspectives? Metrics in the future would provide not just the score but also the rationale, enabling the understanding of each judgment.

Abstract

In the rapidly advancing field of conditional image generation research, challenges such as limited explainability lie in effectively evaluating the performance and capabilities of various models. This paper introduces VIESCORE, a Visual Instruction-guided Explainable metric for evaluating any conditional image generation tasks. VIESCORE leverages general knowledge from Multimodal Large Language Models (MLLMs) as the backbone and does not require training or fine-tuning. We evaluate VIESCORE on seven prominent tasks in conditional image tasks and found: (1) VIESCORE (GPT4-v) achieves a high Spearman correlation of 0.3 with human evaluations, while the human-to-human correlation is 0.45. (2) VIESCORE (with open-source MLLM) is significantly weaker than GPT-4v in evaluating synthetic images. (3) VIESCORE achieves a correlation on par with human ratings in the generation tasks but struggles in editing tasks. With these results, we believe VIESCORE shows its great potential to replace human judges in evaluating image synthesis tasks.

1 Introduction

Diffusion models have become a focal point in AI research for image synthesis. Over the past year, several new models (Kumari et al., 2023; Ruiz et al., 2023; Li et al., 2023c; Zhang and Agrawala, 2023) have been introduced to enhance control over image generation. However, comprehensively evaluating AI-synthesized images remains a challenging and unresolved issue. While metrics like LPIPS (Zhang et al., 2018), CLIP-Score (Hessel et al., 2021), and DreamSim (Fu et al., 2023b) were proposed, they have certain limitations: (1) these metrics are agnostic the end task, which can fail to measure the desired aspects of the generated images, (2) the score is opaque with limited explainability. These limitations heavily restrict their effectiveness in assessing conditional image generation. Some research work (Denton et al., 2015; Isola et al., 2017; Meng et al., 2021; Chen et al., 2023; Sheynin et al., 2023) relied on human-driven evaluation methods. While humans excel at understanding and interpreting visual content, such methods in the context are facing challenges such as scalability limits and preference subjectivity issues. This reliance on

human judgment highlights the need for more uniform evaluation methods in the field. To solve the mentioned issues, we formulate the problem definition with our desired properties, as presented in equation 1. The function f takes an instruction I, a synthesized image O, and C^* which is a set of conditions (e.g. style, subject, background, canny-edge, etc). The score function should produce the intermediate rationale in the form of natural language before generating the final score according to the prompt instruction I:

$$f_{\text{VIE}}(I, O, C^*) = (\text{rationale}, \text{score}) \quad (1)$$

The function f can be any Multimodal Large Language Model (MLLM) such as GPT-4 (OpenAI, 2023) and LLaVA (Liu et al., 2023a), which can take input images to generate human-like text responses. Unlike the automatic metrics, MLLM can receive human instructions and produce rationale. With such motivation, we introduce VIESCORE (Visual Instruction-guided Explainable Score), a framework to assess synthetic images in different conditional image generation tasks. VIESCORE has multiple advantages compared to auto-metrics and human evaluation. It includes:

Task Awareness. Existing metrics were often designed to measure a certain aspect of generated images. For example, LPIPS measures the perceptual similarity of a pair of images, while CLIP-Score measures the text alignment of one single image. As a consequence, these metrics cannot be adapted to evaluate other tasks. VIESCORE acts as a silver bullet to tackle all conditional image generation evaluation processes due to its instruction-guiding property. It can be carefully adjusted with different instruction requirements.

Explainability. The existing metrics normally output a single float-point score, which cannot offer detailed insights into the 'rationale' behind its evaluations. Such a score makes it difficult to interpret the decisions from the metric output. Instead, VIESCORE can offer the rationale in the form of natural languages to help humans understand the reasoning process. As depicted in Figure 1, the rationale can significantly improve the trustworthiness of VIESCORE.

While the ultimate goal is to derive an MLLM that can rate images like humans, in this paper we also explore how well MLLMs can assess synthetic images compared to human evaluation and present insights and challenges on state-of-the-art MLLMs towards human evaluators, as shown in Figure 2.

2 Related Works

2.1 Conditional Image Synthesis

With recent advancements in Image Synthesis research (Goodfellow et al., 2016; Ho et al., 2020; Dhariwal and Nichol, 2021), researchers proposed different methods and contributed a large amount of controllable image synthesis models with conditional inputs. Prevalent tasks include Text-To-Image generation (Saharia et al., 2022; Rombach et al., 2022; stability.ai, 2023) (known as text-guided image generation), Inpainting (Avrahami et al., 2022; Lugmayr et al., 2022) (known as mask-guided image editing) and Text-guided image editing (Brooks et al., 2023; Couairon et al., 2022; Wu and la Torre, 2023).

More recent works proposed new tasks such as Subject-driven image generation and editing (Gal et al., 2022; Ruiz et al., 2023; Li et al., 2023c) to inject one specific subject into a synthesized image, while Multi-concept image composition (Kumari et al., 2023; Liu et al., 2023b) allows multiple specific subjects into the synthesized image. Control-guided image generation (Zhang and Agrawala, 2023; Qin et al., 2023) allows additional conditions alongside the text prompt to guide the image synthesis. Our work uses MLLM to access all the discussed tasks on synthetic image evaluation.

2.2 Synthetic Images Evaluation

Various metrics are proposed to evaluate the quality of AI-generated images. Traditional measures like the Inception Score (IS) (Salimans et al., 2016) and the Frechet Inception Distance (FID) (Heusel et al., 2017) are commonly employed to measure image fidelity. On the other hand, to measure the alignment between the generated image and the text prompt, several metrics (Kim et al., 2022; Kynkäänniemi et al., 2019; Park et al., 2021; Sajjadi et al., 2018) have been introduced. The CLIP score (Hessel et al., 2021) and BLIP score (Li et al., 2022) are the most commonly used. Recently, approaches such as (Cho et al., 2023) and (Lu et al., 2023c) aim to provide a fine-grained evaluation framework, while the HEIM-benchmark (Lee et al., 2023) assesses text-to-image models across multiple aspects, such as toxicity and safety. Other methods, such as projective-geometry (Sarkar et al., 2023), evaluate images' physical and geometric realism. However, these metrics are primarily focused on text-to-image generation and remain narrow in scope. General image generation tasks like

subject-driven image generation and image editing (Ruiz et al., 2023; Li et al., 2023c) still lack effective automatic metrics. One traditional, yet effective method to evaluate AI-generated image performance is to have human annotators assess visual quality. Recent works like ImagenHub (Ku et al., 2023), and HEIM (Lee et al., 2023) attempt to standardize human evaluation across various image generation tasks, though scalability remains a challenge. Our research aims to identify the challenges in mimicking human perception in synthetic image evaluation and address these gaps by developing auto-metrics that align with human judgment across common image evaluation tasks.

2.3 Large Language Models as Evaluators

Large language models (LLMs) are often used to evaluate the quality of model-generated outputs. Recent works used LLMs as an evaluator demonstrating their great ability in text generation evaluation (Zheng et al., 2023; Dubois et al., 2023). This ability for evaluation naturally emerges (Fu et al., 2023a) and stems from LLM's great reasoning ability and instruction-following ability. Recent works also tried to devise a smaller but explicitly fine-tuned LLM(Touvron et al., 2023) that achieves similar evaluation results on natural language generation (Xu et al., 2023; Jiang et al., 2023; Li et al., 2023b). Besides text evaluation, LLMs with visual features have been used as evaluators on images (Lu et al., 2023d; Huang et al., 2023). GPT-4v, regarded as the state-of-the-art LLM with visual features, also reported a decent ability on image evaluation, especially in text-image alignment (Zhang et al., 2023b). However, the GPT-4v is not perfect for image evaluation. A comprehensive study on GPT-4v's vision ability reported that GPT-4v makes mistakes on image evaluation tasks (Yang et al., 2023). For example, it failed to provide proper reasonings for spotting the difference between two similar images.

3 Preliminary

3.1 Evaluation Benchmark

ImagenHub (Ku et al., 2023) is a standardized benchmark for evaluating conditional image generation models with human raters. The framework covered mainstream tasks, including image generation, editing, and several conditioned tasks. In this section, we visit how humans assess images in the ImagenHub framework. Images are rated in two aspects: (1) Semantic Consistency (SC) assesses how well the generated image aligns with the given conditions, such as prompts and subject tokens, ensuring coherence and relevance to the specified criteria according to the task. (2) Perceptual Quality (PQ) evaluates the extent to which the generated image appears visually authentic and conveys a sense of naturalness.

ImagenHub curated a human evaluation dataset for each task, in which the dataset contains around 100 to 200 conditional inputs for generating synthesized images. Then each image was rated by three human raters according to the guidelines of the defined task, and a final score in the range [0.0, 1.0] was reported for the average score in semantic consistency (SC) and perceptual quality (PQ) respectively, with another overall score (O) derived from the geometric mean of semantic consistency and perceptual quality at the instance level. ImagenHub covered 30 image synthesis models and reported 0.4 Krippendorff's alpha on the inter-worker agreement of their human rating.

3.2 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) typically denote LLMs with integrated visual capabilities (Yin et al., 2023). This visual proficiency opens up the potential to perform image analysis and evaluation. However, for a comprehensive assessment of synthetic images, multiple images may be examined in one pass due to complex conditions. The prompt will also be extensive to comprehensively describe the rating process. Therefore, the MLLM candidate should possess specific capabilities: (1) The model must efficiently process and interpret multiple images simultaneously. (2) The model needs to comprehend and respond to lengthy text prompts while matching all requirements.

Recent popular open-source MLLMs, including LLaVA (Liu et al., 2023a), InstructBLIP (Dai et al., 2023), Fuyu (Bavishi et al., 2023), and CogVLM (Wang et al., 2023), can only accept a single image as input along with text instruction. To feed multiple images, a workaround is to merge and concatenate multiple images horizontally and feed as one image. More recent MLLMs such as Open-Flamingo (Awadalla et al., 2023), Kosmos-2 (Peng et al., 2023), and QwenVL (Bai et al., 2023) can accept multiple images in an interleaved image-text format. For closed-source MLLM, according to OpenAI's product description, GPT-4v (OpenAI, 2023) can only take up to 10 images.

3.3 Existing Auto-metrics

Here we list some prominent automatic metrics:

Image-Text Alignment. CLIP-Score (Hessel et al., 2021) computes the average cosine similarities between prompt and generated image CLIP embeddings. One disadvantage of CLIP-Score is that the score is biased towards the training distribution (Kim et al., 2023). Moreover, in practical evaluation, the average CLIP-Score result of a decent method will always fall in the range [0.25, 0.35] even though a single CLIP-Score is within [0, 1]. Such a narrow range may not offer enough differentiation to know which model is better. Moreover, image-text alignment is not the only considered aspect of semantic consistency. For example, it cannot examine the degree of overediting in text or mask-guided image editing tasks.

Perceptual Distance. LPIPS (Zhang et al., 2018) measures the resemblance between two images in a manner that aligns with human perception. With its sensitivity to distortions, it is an often used metric in image synthesis research such as image editing tasks and control-guided Image Generation task (Meng et al., 2021; Qin et al., 2023), to measure between the input (or ground truth) and the generated image. However, in the image editing context, the image's naturalness (e.g. shadow, lighting, sense of distance) is often required in the human perspective of perceptual quality, which is missed in the LPIPS metric. It is also difficult to access a model's performance by distortion level, as in the current state of research the models often can process high-quality editing without artifacts.

Subject Fidelity. CLIP-I computes the average pairwise cosine similarities between CLIP embeddings of generated and real images, first proposed in Textual-Inversion (Gal et al., 2022). However, CLIP-I cannot distinguish between different subjects that may have highly similar text descriptions, and it is less sensitive to shape consistency as it compares the semantic similarity between images. DINO metric was proposed in DreamBooth (Ruiz et al., 2023). The metric is computed by the mean cosine similarities calculated pairwise between the DINO embeddings of ViT-S/16 (Caron et al., 2021) for both synthesized and authentic images. In contrast to CLIP-I, the DINO metric is sensitive to differences between subjects of the same class due to the self-supervised training objective of DINO. These two became popular metrics reported in research on subject-driven image generation and editing tasks (Li et al., 2023c; Lu et al., 2023a).

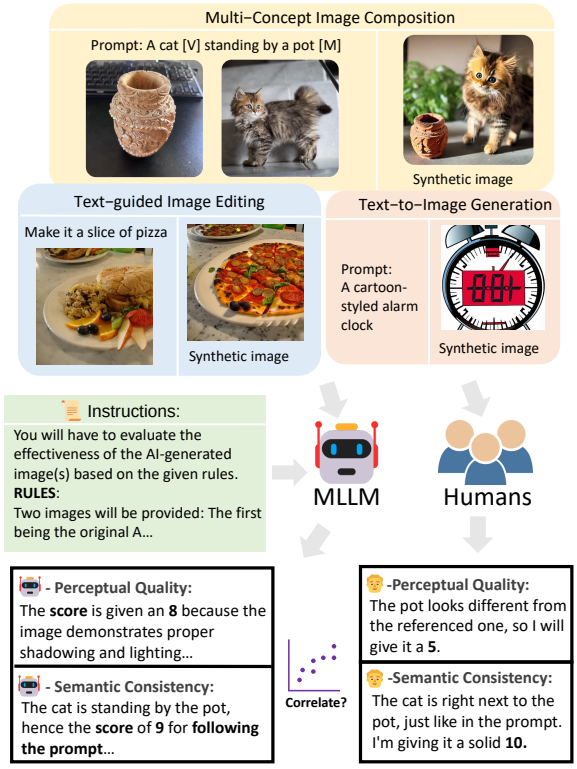

Figure 2: We study the correlation between MLLMs and human perspectives on rating images.

4 Method

During the experiment, we select 29 models evaluated in ImagenHub (Ku et al., 2023) to compare the correlations with human ratings. See Appendix B for the details.

Rating instructions. In ImagenHub, each image in one rating aspect is rated by picking an option from List[0, 0.5, 1] by three human raters. While such simple rating instruction is human-friendly and offers enough granularity, the simplicity of the scale can lead to less accurate representations of opinions, as given the broad spectrum covered by the rating aspects of semantic consistency (SC) and perceptual quality (PQ). We propose a more rigorous rating instruction toward comprehensive evaluation for each type of task. We split the rating of semantic consistency (SC) and perceptual quality (PQ) into multiple sub-scores, which SC contains multiple scores according to the tasks.

For example, in the multi-concept image composition task as shown in Figure 3, two images (known as concepts) and a text prompt are provided as input, and the desired synthesized image will contain the two concept objects in the image

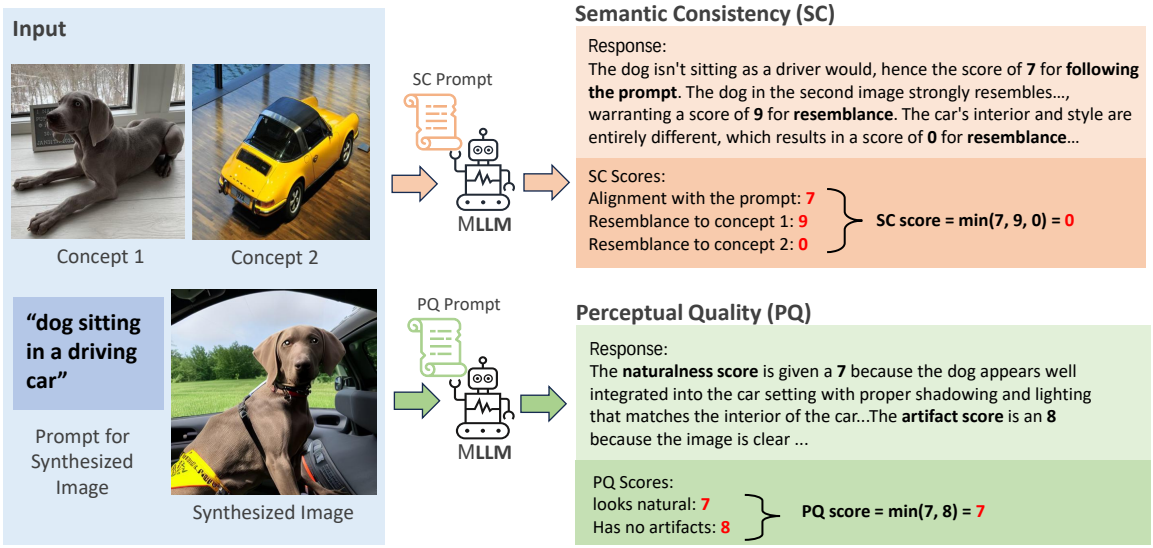

Figure 3: Process of MLLM evaluation on one synthetic image. All input conditions, synthesized images, and rating instructions are fed together to the MLLM in one pass. Multi-concept image composition task is used here as an example. The final overall score of the image is derived with equation 2.

in actions according to the text prompt. Thus SC will be split into 3 sub-scores: (1) Is the image aligning with the prompt? (2) Does the object in the image resemble the first concept? (3) Does the object in the image resemble the second concept? For PQ, the naturalness level and distortion level will be accessed separately, resulting in 2 sub-scores: (i) Does the image give an unnatural feeling such as a wrong sense of distance, wrong shadow, or wrong lighting? (ii) Does the image contain a portion of distortion, watermark, scratches, etc.? Our proposed rating system enhances the evaluation of tasks by dividing SC and PQ into distinct sub-scores. The details of prompt templates are available in Appendix A.

$$O = [\min(\alpha_1, ..., \alpha_i) \min(\beta_1, ..., \beta_i)]^{\frac{1}{2}} \quad (2)$$

Our overall score is derived as shown in equation 2. We assume each sub-score weights the same and used $\min$ operation to emphasize the importance of meeting all criteria without exception. α_i is a sub-score in SC and β_i is a sub-score in PQ. The final rating scores of SC and PQ provided by MLLMs are on a scale of 0 to 10. The design rationale is that in ImagenHub's human rating method, the possible results when the answers of three human raters, each picking an option from List[0, 0.5, 1], are added together and then divided by 3, will fall into one of the options: List[0.0, 0.17, 0.33, 0.5, 0.67, 0.83, 1.0]. Thus we simply use a scale of 0 to 10 and normalized in the range [0.0, 1.0] when comparing with human ratings. Input conditions and synthetic image are fed into the MLLM together during the rating process of SC, while in the PQ rating process, only the synthetic image is fed into the MLLM. This is to avoid the model getting confused by the input conditions in the PQ rating process, as to be discussed in section 5.1.

Backbone	M-H$^{SC}_{corr}$	M-H$^{PQ}_{corr}$	M-H$^{O}_{corr}$
Across All 7 Tasks			
Human Raters	0.4700	0.4124	0.4558
VIESCORE			
GPT-4v$_{0shot}$	0.3655	0.3092	0.3266
GPT-4v$_{1shot}$	0.2689	0.2338	0.2604
LLaVA$_{0shot}$	0.1046	0.0319	0.0925
LLaVA$_{1shot}$	0.1012	0.0138	0.0695
Qwen-VL$_{0shot}$	0.0679	0.0165	0.0920
BLIP2$_{0shot}$	0.0504	-0.0108	0.0622
InstructBLIP$_{0shot}$	0.0246	0.0095	0.0005
Fuyu$_{0shot}$	-0.0110	-0.0172	0.0154
CogVLM$_{0shot}$	-0.0228	0.0514	-0.0050
OpenFlamingo$_{0shot}$	-0.0037	-0.0102	-0.0122

Table 1: Correlations across all tasks with different backbone models. We highlight the highest correlation numbers in green. See Appendix C for details.

5 Experimental Results

5.1 Correlation Study

For all presented correlations, we applied Fisher Z-transformation to estimate the average Spearman correlation $\in [-1, 1]$ across models and tasks.

Method	M-H$^{SC}_{corr}$	M-H$^{PQ}_{corr}$	M-H$^{O}_{corr}$
Text-guided Image Generation (5 models)			
Human Raters	0.5044	0.3640	0.4652
CLIP-Score	-0.0817	-0.0114	-0.0881
VIESCORE			
GPT-4v$_{0shot}$	0.4885	0.2379	0.4614
GPT-4v$_{1shot}$	0.4531	0.1770	0.3801
LLaVA$_{0shot}$	0.1809	0.0306	0.1410
LLaVA$_{1shot}$	0.1789	-0.0020	0.1309
Mask-guided Image Editing (4 models)			
Human Raters	0.5390	0.5030	0.4981
LPIPS	-0.1012	0.0646	-0.0694
VIESCORE			
GPT-4v$_{0shot}$	0.4508	0.2859	0.4069
GPT-4v$_{1shot}$	0.4088	0.2352	0.3810
LLaVA$_{0shot}$	0.1180	-0.0531	0.0675
LLaVA$_{1shot}$	0.1263	-0.0145	0.1040
Text-guided Image Editing (8 models)			
Human Raters	0.4230	0.5052	0.4184
LPIPS	0.0956	0.2504	0.1142
VIESCORE			
GPT-4v$_{0shot}$	0.2610	0.4274	0.2456
GPT-4v$_{1shot}$	0.2428	0.3402	0.2279
LLaVA$_{0shot}$	0.0448	0.0583	0.0273
LLaVA$_{1shot}$	0.0185	-0.0107	0.0258

Table 2: Correlations comparison of available methods. We highlight the best method and the correlation numbers closest to human raters. Continue in Table 3.

Method	M-H$^{SC}_{corr}$	M-H$^{PQ}_{corr}$	M-H$^{O}_{corr}$
Subject-driven Image Generation (4 models)			
Human Raters	0.4780	0.3565	0.4653
DINO	0.4160	0.1206	0.4246
CLIP-I	0.2961	0.1694	0.3058
VIESCORE			
GPT-4v$_{0shot}$	0.3979	0.1903	0.3738
GPT-4v$_{1shot}$	0.2757	0.2261	0.2753
LLaVA$_{0shot}$	0.0326	-0.0303	0.1219
LLaVA$_{1shot}$	0.1334	0.0858	0.1248
Subject-driven Image Editing (3 models)			
Human Raters	0.4887	0.2986	0.4747
DINO	0.3022	-0.0381	0.3005
CLIP-I	0.2834	0.1248	0.2813
VIESCORE			
GPT-4v$_{0shot}$	0.3274	0.2960	0.1507
GPT-4v$_{1shot}$	-0.0255	0.1572	-0.0139
LLaVA$_{0shot}$	0.0360	-0.0073	0.0168
LLaVA$_{1shot}$	0.0587	-0.0249	0.0309
Multi-concept Image Composition (3 models)			
Human Raters	0.5927	0.5145	0.5919
DINO	0.0979	-0.1643	0.0958
CLIP-I	0.1512	-0.0963	0.1498
VIESCORE			
GPT-4v$_{0shot}$	0.3209	0.3025	0.3346
GPT-4v$_{1shot}$	0.1859	0.1185	0.1918
LLaVA$_{0shot}$	0.1022	0.1194	0.1070
LLaVA$_{1shot}$	0.0828	0.0379	0.0293
Control-guided Image Generation (2 models)			
Human Raters	0.5443	0.5279	0.5307
LPIPS	0.3699	0.4204	0.4133
VIESCORE			
GPT-4v$_{0shot}$	0.4360	0.4975	0.3999
GPT-4v$_{1shot}$	0.3892	0.4132	0.4237
LLaVA$_{0shot}$	0.2207	0.1060	0.1679
LLaVA$_{1shot}$	0.1121	0.0247	0.0416

Table 3: Continued from Table 2.

Metric-to-Human (M-H) correlations. In Table 2 and 3, we first verified the reliability of ImagenHub human ratings by computing the Human-to-Human (H-H) correlation, as the correlation goes around 0.5, expected to be the highest value compared to MLLMs. Then we benchmark the MLLMs according to our designed method to compute the Metric-to-Human (M-H) correlation. We noticed only GPT4v and LLaVA were able to follow our instructions clearly while other MLLMs were not able to produce any meaningful results according to our setup. For example, BLIP-2, while able to output the correct format, the scores provided are constant zeros. Qwen-VL and InstructBLIP could only produce a portion of responses for semantic consistency but failed to generate any results for perceptual quality evaluation. From overall performance, we found that GPT-4v reports a significantly higher correlation than LLaVA, while LLaVA's correlation is much less than human raters. It seems that LLaVA is less effective in these specific tasks compared to GPT-4v. It is worth mentioning that GPT4v shows satisfactory performance on nearly all tasks with a difference of less than 0.2 towards human correlations, even on par with humans in text-guide image generation tasks. Both GPT-4v and LLaVA demonstrated the weakest performance in the text-guided and subject-driven image editing tasks. This suggests that GPT-4v is a capable model in some tasks although it is still not on par with human performance.

Extra visuals resulted in a decline in performance. Numerous studies (Brown et al., 2020; Parnami and Lee, 2022; Liu et al., 2021) have

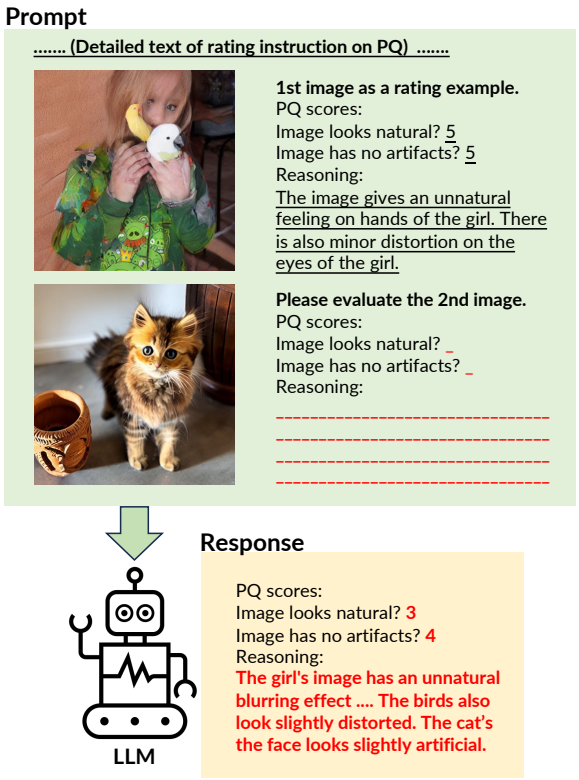

Figure 4: MLLM making mistakes on rationale when prompted with extra images as examples.

highlighted that In-Context Learning (ICL) allows LLMs to tackle novel tasks effectively without requiring the traditional fine-tuning process. We applied In-Context Learning in our prompting method with the expectation of increasing the correlation scores, but we observed the opposite. In Table 2 and 3, there is an observable general trend of diminishing correlation scores. The overall correlation score in subject-driven image generation and editing, and the multi-concept image composition task dropped significantly. Only the mask-guided image editing task and control-guided image generation task reported a subtle increase in correlation score.

Looking into the rationale, we found that the MLLMs tend to get confused by the example images, as illustrated in Figure 4. Such behavior is observed in both GPT4v and LLaVA rationale. Another recent work (Lu et al., 2023b) also reported a similar issue where the model attempted to consider the example when answering the visual question. This explains the deterioration of the correlation on both GPT4v and LLaVA when the ICL prompting technique is used. This also implied the low correlation scores on image editing tasks were due to the limited capability of state-of-the-art MLLMs for multiple image understanding.

Ablation study on PQ rating setting. As providing multiple images could potentially decrease the performance, we attempt to minimize the workload of MLLM by only providing the synthetic image in the PQ rating process instead of including the input conditions. We report the correlation score in the two different settings with GPT-4v in Table 4 to examine the impact. We spotted a significant improvement in correlation after taking away the inputs in the PQ rating process.

PQ Prompting Method	TIE M-H$_{\text{corr}}^{PQ}$	MCIC M-H$_{\text{corr}}^{PQ}$
Human (with inputs)	0.5052	0.5145
without inputs	0.4274	0.3025
with inputs	0.2256	0.0731

Table 4: Correlations of GPT4v when including inputs in the PQ prompt in TIE (Text-guided Image Editing) and MCIC (Multi-concept Image Composition) task. See 6 for a detailed comparison in the Appendix.

Ranking image models. Besides rating score correlations, we also compared the model ranking from the ImagenHub human evaluation leaderboard and the model ranking suggested by the MLLMs, shown in Table 5. We computed Spearman's footrule $d_{SF}(r, r_*) \in [0, +\infty)$ and Spearman's rho $\rho_S(r, r_*) \in [-1, 1]$ to examine the ranking correlation. Both GPT4v and LLaVA can align to ImagenHub rankings on the multi-concept image composition task and control-guided image generation task, and with only one model difference in the subject-driven image editing task. While the results vary significantly across other tasks, GPT4v generally maintains a stronger alignment with the ImagenHub rankings compared to LLaVA.

5.2 Insights and Challenges on VIESCORE

MLLMs are weak at capturing image nuances in edited images. From Table 2, we noticed the correlation scores on editing tasks are generally lower than generation tasks. Upon investigation, it was found that MLLMs often fail to detect minor changes made in image editing, such as small patch edits. Consequently, MLLMs might perceive two images as identical even when humans recognize the edits as successful. This issue may stem from MLLMs focusing on high-level image features while overlooking finer details like color and texture differences, as illustrated in Figure 5. This limitation is apparent in both GPT-4v and LLaVA, highlighting a challenge in synthetic image evalua-

Figure 5: Representative pairs that MLLMs misunderstood as identical images. Images in the first row are the inputs and in the second row are the edited.

tion accuracy on image editing tasks.

Both MLLMs and human evaluators display a broader range of views regarding perceptual quality compared to semantic consistency. From Table 2 and Table 3, we can observe that the correlation scores of PQ are generally lower than the correlation scores of SC and Overall, even on human raters. This suggests the human raters' perspective on evaluating perceptual quality is more diverse. Possible impacting factors include the rater's eyesight condition, screen resolution, rating leniency, etc. In the context of MLLMs, we found that MLLMs while being able to correctly recognize the naturalness and artifacts of the image, the rating scores are as diverse as human rating scores even though we have provided a marking rubric.

5.3 VIESCORE and Auto-metrics vs Human

We report the human correlations in Table 2 and 3 to compare the performance between our VIESCORE and popular auto-metrics. To ensure a fair comparison, we only included automatic metrics that have been previously reported in related research for the specific tasks under consideration.

DINO is an effective metric in subject-driven tasks. The DINO metric demonstrates sensitivity to variations within the same class of subjects, making it an effective metric for measuring whether the subject in the synthesized image aligns with the token subject. Our correlation result shows that DINO outperforms GPT-4v and CLIP-I on subject-driven image generation and editing tasks, suggesting that DINO highly aligns with human's perspective on semantic consistency where subject fidelity is considered.

LPIPS metric proves to be effective in control-guided tasks, but less effective in image editing tasks. As discussed in section 3.3, LPIPS has great ability in detecting distortions. Since the control-guided task is a less mature research direction compared to image editing tasks, distortions are often found in the synthetic images from the control-guided task. On the other hand, current image editing models can synthesize images with less distortions. This explains the high correlation in the control-guided task.

CLIP-Score has a much weaker correlation with human ratings in the text-to-image task than GPT-4v. We also noticed none of the synthetic images achieved higher than 0.3 CLIP-Score, even if they are regarded as having high semantic consistency by human raters. This can be due to different evaluation focuses, as humans tend to grab the abstract idea from the prompt to access the image, but CLIP-Score considers the whole text prompt.

GPT-4v outperforms other auto-metrics on ImagenHub leaderboard rankings. The correlation of model rankings on ImagenHub was evaluated against CLIP Score and LPIPS metrics, as shown in Table 5, and compared with MLLMs in the VIESCORE. We found that GPT-4v can achieve a positive correlation with the model rankings on every task. This shows the sign of capability for MLLMs as evaluators for image synthesis research.

6 Conclusion

In this paper, we propose the VIESCORE for synthetic image evaluation across seven popular image synthesis tasks and comprehensively access the efficacy using human ratings from ImagenHub. Our experiment reported that VIESCORE with GPT-4v backbone is significantly more effective than other open-source MLLMs in assessing synthetic images, achieving a correlation of over 0.3 to human ratings on a portion of the tasks, especially on par with humans on the Text-To-Image task. However, it notes a lower correlation in image editing tasks for MLLMs, including GPT-4v. Comparing our VIESCORE to existing auto-metrics, we found that GPT-4v is more effective than auto-metrics in most tasks, while DINO is more effective in subject-driven image generation and editing tasks. GPT-4v also shows a higher ranking correlation with the ImagenHub leaderboard than other automatic metrics. This marked a milestone towards explainable metrics for conditional image synthesis evaluation. Our future research will focus on investigating the use of distillation models to replicate human-like performance in evaluating synthetic images.

7 Limitations

OpenAI Security and Privacy Policy. Due to ChatGPT's security and privacy policy, AI-generated images that resemble a real person or photograph will be refused by GPT-4v for evaluation. The model will return results similar to "I am sorry, but I cannot process these images as they contain real people.". We simply drop those results by keyword matching.

OpenAI GPT-4v Playground vs API. While GPT-4v Playground allows the user to keep a session, the OpenAI API does not provide such a function. While we believe using GPT-4v playground might yield better performance, especially in an In-Context learning setting, we can only rely on API due to the large scale of the experiment.

8 Potential Risks

Multimodal models can inadvertently perpetuate or amplify biases present in their training data. The interpretation and evaluation of synthetic images depend heavily on context. A multimodal model might not fully grasp certain images' nuances or cultural sensitivities, leading to inappropriate or offensive outputs.

9 Artifacts

All datasets and models are publicly accessible for academic use, and the official OpenAI API is available for academic purposes.

10 Computational Experiments

All open-source model experiments were conducted on an NVIDIA RTX A6000 GPU. Approximately 250 US dollars were spent on an OpenAI API call for GPT-4v experiments.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2023. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

deep floyd.ai. 2023. If by deepfloyd lab at stabilityai.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, 28.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023a. Gptscore: Evaluate as you desire. *ArXiv*, abs/2302.04166.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023b. Dreamsim: Learning new dimensions of human visual similarity using synthetic data.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT Press.

Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. 2023. Photoswap: Personalized subject swapping in images. *arXiv preprint arXiv:2305.18286*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *ArXiv*, abs/2307.06350.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*.

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. 2022. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35:35072–35086.

Younghyun Kim, Sangwoo Mo, Min-Kyung Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. 2023. Bias-to-text: Debiasing unknown visual biases through language interpretation. *ArXiv*.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2023. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. 2023. Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dongxu Li, Junnan Li, and Steven CH Hoi. 2023a. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. 2023c. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023b. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*.

Lingxiao Lu, Bo Zhang, and Li Niu. 2023a. Dreamcom: Finetuning text-guided inpainting model for image composition. *ArXiv*, abs/2309.15508.

Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. 2023b. Vim: Probing multimodal large language models for visual embedded instruction following. *ArXiv*, abs/2311.17647.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023c. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *arXiv preprint arXiv:2305.11116*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023d. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *ArXiv*, abs/2305.11116.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047.

OpenAI. 2023. Gpt-4 technical report.

openjourney.ai. 2023. Openjourney is an open source stable diffusion fine tuned model on midjourney images.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11.

Archit Parnami and Minwoo Lee. 2022. Learning from few examples: A summary of approaches to few-shot learning. *ArXiv*, abs/2203.04291.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

runwayml. 2023. Stable diffusion inpainting.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep

language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David Forsyth, and Anand Bhattad. 2023. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now. *ArXiv*, abs/2311.17138.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2023. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*.

stability.ai. 2023. Stable diffusion xl.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079.

Chen Henry Wu and Fernando De la Torre. 2023. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. *ArXiv*, abs/2305.14282.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv*, abs/2306.13549.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS dataset and benchmark track*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023b. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *ArXiv*, abs/2311.01361.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Prompt Templates

Prompt Engineering. We found that not all MLLMs can fully understand our prompt to give a desired output format consistently. Thus we required MLLMs to output a JSON format, which is supposed to be capable for most MLLMs.

Prompt Design. The prompt is divided into two segments: the 'context prompt' and the 'rating prompt'. The ultimate prompt provided to the model is a combination of these two segments.

Context

You are a professional digital artist. You will have to evaluate the effectiveness of the AI-generated image(s) based on the given rules. You will have to give your output in this way (Keep your reasoning concise and short.):
{
"score" : [...],
"reasoning" : "..."
}

PQ Rating Prompt Template (for all tasks)

RULES:
The image is an AI-generated image. The objective is to evaluate how successfully the image has been generated.
On a scale 0 to 10:
A score from 0 to 10 will be given based on image naturalness.
(0 indicates that the scene in the image does not look natural at all or gives an unnatural feeling such as a wrong sense of distance, wrong shadow, or wrong lighting. 10 indicates that the image looks natural.)
A second score from 0 to 10 will rate the image artifacts.
(0 indicates that the image contains a large portion of distortion, watermarks, scratches, blurred faces, unusual body parts, or subjects not harmonized. 10 indicates the image has no artifacts.)
Put the score in a list such that output score = [naturalness, artifacts]

SC Rating Prompt Template (Text-Guided Image Generation)

RULES:
The image is an AI-generated image according to the text prompt. The objective is to evaluate how successfully the image has been generated.
On a scale 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the AI-generated image does not follow the prompt at all. 10 indicates the AI-generated image follows the prompt perfectly.)
Put the score in a list such that output score = [score].
Text Prompt: <prompt>

SC Rating Prompt Template (Text/Mask-Guided Image Editing)

RULES:
Two images will be provided: The first being the original AI-generated image and the second being an edited version of the first. The objective is to evaluate how successfully the editing instruction has been executed in the second image. Note that sometimes the two images might look identical due to the failure of the image edit.
On scale of 0 to 10:
A score from 0 to 10 will be given based on the success of the editing. (0 indicates that the scene in the edited image does not follow the editing instructions at all. 10 indicates that the scene in the edited image follows the editing instruction text perfectly.)
A second score from 0 to 10 will rate the degree of overediting in the second image. (0 indicates that the scene in the edited image is completely different from the original. 10 indicates that the edited image can be recognized as a minimally edited yet effective version of the original.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the editing success and 'score2' evaluates the degree of overediting.
Editing instruction: <instruction>

SC Rating Prompt Template (Control-Guided Image Generation)

RULES:
Two images will be provided: The first being a processed image (e.g. Canny edges, openpose, grayscale, etc.) and the second being an AI-generated image using the first image as guidance. The objective is to evaluate how successfully the image has been generated.
On scale 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the second image does not follow the prompt at all. 10 indicates the second image follows the prompt perfectly.)
A second score from 0 to 10 will rate how well the generated image is following the guidance image. (0 indicates that the second image does not follow the guidance at all. 10 indicates that the second image is following the guidance image.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the prompt and 'score2' evaluates the guidance.
Text Prompt: <prompt>

SC Rating Prompt Template (Subject-Driven Image Generation)

RULES:
Two images will be provided: The first is a token subject image and the second is an AI-generated image using the first image as guidance. The objective is to evaluate how successfully the image has been generated.
On a scale of 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the second image does not follow the prompt at all. 10 indicates the second image follows the prompt perfectly.)
A second score from 0 to 10 will rate how well the subject in the generated image resembles the token subject in the first image. (0 indicates that the subject in the second image does not look like the token subject at all. 10 indicates the subject in the second image looks exactly like the token subject.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the prompt and 'score2' evaluates the resemblance.
Text Prompt: <prompt>

SC Rating Prompt Template (Subject-Guided Image Editing)

RULES:
Three images will be provided: The first image is an input image to be edited. The second image is a token subject image. The third image is an AI-edited image from the first image. it should contain a subject that looks like the subject in the second image. The objective is to evaluate how successfully the image has been edited.
On a scale 0 to 10:
A score from 0 to 10 will rate how well the subject in the generated image resembles the token subject in the second image. (0 indicates that the subject in the third image does not look like the token subject at all. 10 indicates the subject in the third image looks exactly like the token subject.)
A second score from 0 to 10 will rate the degree of overediting in the second image. (0 indicates that the scene in the edited image is completely different from the first image. 10 indicates that the edited image can be recognized as a minimally edited yet effective version of the original.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the resemblance and 'score2' evaluates the degree of overediting.
Subject: <subject>

Task (Total number of Models)	$d_{SF}(r_{\text{Human}}, r_{\text{Method}})\downarrow$ GPT4v	LLaVA	LPIPS	CLIP	$\rho_S(r_{\text{Human}}, r_{\text{Method}})\uparrow$ GPT4v	LLaVA	LPIPS	CLIP
Text-guided Image Generation (5)	2	6	N/A	8	0.90	0.50	N/A	-0.20
Mask-guided Image Editing (4)	2	8	2	0	0.80	-1.00	0.80	1.00
Text-guided Image Editing (8)	12	16	20	16	0.67	0.48	0.17	0.48
Subject-driven Image Generation (4)	4	6	0	6	0.20	-0.40	1.00	-0.20
Subject-driven Image Editing (3)	2	2	4	4	0.50	0.50	-0.50	-1.00
Multi-concept Image Composition (3)	0	0	2	2	1.00	1.00	0.50	0.50
Control-guided Image Generation (2)	0	0	0	2	1.00	1.00	1.00	-1.00

Table 5: Ranking judgment from each metric method. $d_{SF}(r, r_*)$ is the Spearman's footrule and $\rho_S(r, r_*)$ is the Spearman's rho (correlation). LPIPS is not available for the first task because there are no reference images.

B Supplementary Information

B.1 Human Correlation Study

In our paper content, we only reported the Spearman correlations. Here we included the Pearson and Kendall correlation in Table 7 for comparative analysis of Human-to-Human (H-H) correlation.

B.2 Zero-shot vs One-shot on VIE

We applied only zero-shot and one-shot experiments in this paper because not even GPT-4v can produce anything with few-shot setting in our context. We report full table of GPT-4v performance in Table 6 for zero-shot vs one-shot results.

B.3 Autometrics vs Human Detail results

See Table 8, 9, 10, 11 for detail statistics of CLIP-Score, LPIPS, DINO, and CLIP-I correlation with human ratings.

B.4 ImagenHub Models used

- Text-guided Image Generation: Stable Diffusion (SD) (Rombach et al., 2022), SDXL (stability.ai, 2023), DALLE-2 (Ramesh et al., 2022), DeepFloydIF (deep floyd.ai, 2023), OpenJourney (openjourney.ai, 2023).
- Mask-guided Image Editing: SD (runwayml, 2023), SDXL (stability.ai, 2023), GLIDE, BlendedDiffusion (Avrahami et al., 2022)
- Text-guided Image Editing: MagicBrush (Zhang et al., 2023a), InstructPix2Pix (Brooks et al., 2023), Prompt-to-Prompt (Mokady et al., 2023), CycleDiffusion (Wu and la Torre, 2023), SDEdit (Meng et al., 2021), Text2Live (Bar-Tal et al., 2022), DiffEdit (Couairon et al., 2022), Pix2PixZero (Parmar et al., 2023).
- Subject-driven Image Generation: DreamBooth (Ruiz et al., 2023), DreamBooth-Lora (Hu et al., 2021), BLIP-Diffusion (Li et al., 2023a), TextualInversion (Gal et al., 2022).
- Subject-driven Image Editing: PhotoSwap (Gu et al., 2023), DreamEdit (Li et al., 2023c), BLIP-Diffusion.
- Multi-concept Image Composition: CustomDiffusion (Kumari et al., 2023), DreamBooth, TextualInversion.
- Control-guided Image Generation: ControlNet (Zhang and Agrawala, 2023), UniControl (Qin et al., 2023).

B.5 ImagenHub Human data information

We showed the total human rating data we used for each task in Table 12.

Image Model	Backbone: GPT-4v					
	Zero-Shot			One-Shot		
	M-H$^{SC}_{0shot}$	M-H$^{PQ}_{0shot}$	M-H$^{O}_{0shot}$	M-H$^{SC}_{1shot}$	M-H$^{PQ}_{1shot}$	M-H$^{O}_{1shot}$
Text-guided Image Generation						
DeepFloydIF	0.5182	0.3509	0.5479	0.4272	0.2048	0.3849
Stable Diffusion XL	0.5684	0.2823	0.5301	0.5136	0.1522	0.3735
Dalle-2	0.5046	0.2192	0.4871	0.4469	0.1822	0.5364
OpenJourney	0.4835	0.1624	0.4648	0.4563	0.2730	0.3750
Stable Diffusion 2.1	0.5957	0.1981	0.4658	0.5988	0.0820	0.3311
Mask-guided Image Editing						
SDXL-Inpainting	0.5461	0.2331	0.4772	0.5308	0.3460	0.5261
SD-Inpainting	0.5607	0.4253	0.544	0.3759	0.3446	0.3969
GLIDE	0.4663	0.2816	0.4499	0.4247	0.1056	0.3536
BlendedDiffusion	0.3695	0.2363	0.2563	0.4054	0.1624	0.3283
Text-guided Image Editing						
MagicBrush	0.3273	0.3696	0.3395	0.3613	0.5135	0.4727
InstructPix2Pix	0.3094	0.4461	0.3363	0.4423	0.3106	0.3921
Prompt-to-prompt	0.3094	0.3696	0.3395	0.2514	0.2057	0.2068
CycleDiffusion	0.4488	0.6124	0.3927	0.3522	0.3374	0.1578
SDEdit	0.1607	0.3944	0.1570	0.1754	0.3837	0.2814
Text2Live	0.1875	0.4158	0.1964	0.2817	0.2357	0.2753
DiffEdit	0.1803	0.5957	0.0247	0.1761	0.4874	0.1281
Pix2PixZero	0.2144	0.4502	0.2193	-0.0588	0.3609	-0.0588
Subject-driven Image Generation						
DreamBooth	0.4975	0.2199	0.4787	0.5409	0.1930	0.5848
BLIP-Diffusion	0.3367	0.0663	0.2845	0.1176	0.3402	0.1194
TextualInversion	0.5564	0.2398	0.4795	0.3882	0.0010	0.3035
DreamBooth-Lora	0.2938	0.2448	0.3285	0.0856	0.3860	0.1225
Subject-driven Image Editing						
PhotoSwap	0.3711	0.1246	0.1598	-0.0782	0.0385	-0.1063
DreamEdit	0.3817	0.4419	0.1580	0.1384	0.3037	0.0954
BLIP-Diffusion	0.2671	0.3488	0.1379	-0.1368	0.1333	-0.0309
Multi-concept Image Composition						
CustomDiffusion	0.4781	0.431	0.4263	0.5064	0.0194	0.4867
DreamBooth	0.1494	0.2367	0.232	0.0396	0.0633	0.0694
TextualInversion	0.3703	0.269	0.3857	0.0183	0.2745	0.0266
Control-guided Image Generation						
ControlNet	0.4270	0.4827	0.4753	0.3561	0.4052	0.4055
UniControl	0.5797	0.4173	0.3972	0.4655	0.4737	0.4988

Table 6: Comprehensive study on the Spearman correlation between GPT4v-to-Human (GPT4v-H) ratings across various models, in zero-shot (0shot) and one-shot (1shot) settings, across different test categories: Semantic Consistency (SC), Perceptual Quality (PQ), and Overall (O).

Image Model	H-H$_{\text{pear}}^{SC}$	H-H$_{\text{pear}}^{PQ}$	H-H$_{\text{pear}}^{O}$	H-H$_{\text{spea}}^{SC}$	H-H$_{\text{spea}}^{PQ}$	H-H$_{\text{spea}}^{O}$	H-H$_{\text{kend}}^{SC}$	H-H$_{\text{kend}}^{PQ}$	H-H$_{\text{kend}}^{O}$
Text-guided Image Generation									
DeepFloydIF	0.5933	0.3086	0.5595	0.5635	0.3029	0.5131	0.5360	0.2878	0.4581
Stable Diffusion XL	0.5990	0.4957	0.5945	0.5807	0.4992	0.5896	0.5468	0.4719	0.5289
Dalle-2	0.5208	0.5024	0.4630	0.5019	0.4680	0.4348	0.4654	0.4459	0.3820
OpenJourney	0.5678	0.3853	0.5513	0.5321	0.3600	0.4861	0.5017	0.3442	0.4347
Stable Diffusion 2.1	0.6202	0.3227	0.5397	0.5979	0.2772	0.4962	0.5707	0.2636	0.4557
Mask-guided Image Editing									
SDXL-Inpainting	0.6550	0.5929	0.6578	0.6574	0.5928	0.6556	0.6160	0.5382	0.6040
SD-Inpainting	0.6606	0.5197	0.5716	0.6590	0.5166	0.5394	0.6222	0.4728	0.5039
GLIDE	0.6253	0.5496	0.6144	0.5894	0.5530	0.5695	0.5573	0.4984	0.5357
BlendedDiffusion	0.5863	0.5873	0.5879	0.5051	0.5511	0.4224	0.4911	0.5346	0.4157
Text-guided Image Editing									
MagicBrush	0.6217	0.5251	0.6288	0.6219	0.5190	0.6289	0.5740	0.4740	0.5651
InstructPix2Pix	0.6573	0.6158	0.6632	0.6600	0.5955	0.6561	0.6250	0.5502	0.6157
Prompt-to-prompt	0.5954	0.5084	0.5699	0.5880	0.5028	0.5811	0.5611	0.4537	0.5470
CycleDiffusion	0.5908	0.5848	0.6101	0.5482	0.5887	0.5891	0.5228	0.5378	0.5600
SDEdit	0.2303	0.4717	0.1674	0.2657	0.4705	0.1991	0.2618	0.4211	0.1957
Text2Live	0.3167	0.6013	0.2890	0.2675	0.5757	0.1524	0.2648	0.5440	0.1503
DiffEdit	0.2513	0.6331	0.3570	0.3286	0.6214	0.4265	0.3268	0.5924	0.4247
Pix2PixZero	0.4747	0.5763	0.5247	0.3311	0.5770	0.3327	0.3305	0.5299	0.3312
Subject-driven Image Generation									
DreamBooth	0.6337	0.3988	0.5834	0.6452	0.3871	0.6208	0.6010	0.3787	0.5625
BLIP-Diffusion	0.4970	0.2663	0.4394	0.4458	0.3263	0.4390	0.4090	0.3180	0.3993
TextualInversion	0.5987	0.3219	0.5533	0.6000	0.3351	0.5686	0.5683	0.3078	0.5226
DreamBooth-Lora	0.5014	0.4571	0.4278	0.3903	0.4430	0.3878	0.3831	0.4169	0.3756
Subject-driven Image Editing									
PhotoSwap	0.4685	0.3213	0.5025	0.4805	0.2961	0.4973	0.4412	0.2768	0.4368
DreamEdit	0.5684	0.2319	0.5520	0.5867	0.2245	0.5460	0.5485	0.2130	0.4892
BLIP-Diffusion	0.5411	0.4086	0.5074	0.5359	0.4033	0.5051	0.5221	0.3779	0.4857
Multi-concept Image Composition									
CustomDiffusion	0.7257	0.4889	0.7215	0.7256	0.4838	0.7217	0.7101	0.4665	0.6963
DreamBooth	0.6560	0.6583	0.6575	0.6209	0.6423	0.6222	0.6068	0.6228	0.6022
TextualInversion	0.6833	0.6009	0.6799	0.6990	0.5803	0.6980	0.6935	0.5563	0.6898
Control-guided Image Generation									
ControlNet	0.6166	0.5730	0.5830	0.6144	0.5682	0.5868	0.5585	0.5408	0.5429
UniControl	0.6014	0.6194	0.6131	0.6060	0.6062	0.5954	0.5577	0.5741	0.5533

Table 7: Comparative analysis of Human-to-Human (H-H) correlation ratings across various models. Metrics used include Pearson's (pear), Spearman's (spea), and Kendall's (kend) correlation coefficients, across different test categories: Semantic Consistency (SC), Perceptual Quality (PQ), and Overall (O).

Image Model	Metric: CLIP-Score M-H^{SC}_{corr}	M-H^{PQ}_{corr}	M-H^{O}_{corr}
Text-guided Image Generation			
DeepFloydIF	0.0272	0.1025	0.0332
OpenJourney	-0.1628	-0.0875	-0.1907
DALLE2	-0.0946	0.1469	-0.0381
SD	-0.0528	-0.0691	-0.0632
SDXL	-0.1265	-0.1500	-0.1830

Table 8: CLIP-Score vs Human correlation on Text-guided image generation task.

Image Model	Metric: LPIPS M-H^{SC}_{corr}	M-H^{PQ}_{corr}	M-H^{O}_{corr}
Text-guided Image Editing			
InstructPix2Pix	0.1652	0.4717	0.2045
CycleDiffusion	-0.0936	0.3193	-0.0211
MagicBrush	0.2146	0.3722	0.2667
Text2Live	-0.0812	0.2906	-0.0787
DiffEdit	0.0943	0.3299	0.1440
Pix2PixZero	0.1379	0.0256	0.1370
Prompt2prompt	0.1918	0.1929	0.1798
SDEdit	0.1381	0.0441	0.0857
Mask-guided Image Editing			
Glide	-0.1098	0.0647	-0.0662
BlendedDiffusion	0.0980	0.1371	0.0598
SDInpaint	-0.2447	-0.0749	-0.2110
SDXLInpaint	-0.1496	0.1318	-0.0607
Control-guided Image Generation			
ControlNet	0.3447	0.3916	0.3888
UniControl	0.4319	0.5048	0.4904

Table 9: LPIPS (signs inverted) vs Human correlation on several tasks.

Model	Metric: DINO M-H^{SC}_{corr}	M-H^{PQ}_{corr}	M-H^{O}_{corr}
Multi-Concept Image Composition			
TextualInversion	0.0759	-0.2746	0.0754
DreamBooth	0.1027	-0.0761	0.1054
CustomDiffusion	0.1159	-0.1466	0.1074
Subject-Driven Image Generation			
DreamBoothLora	0.2335	0.0684	0.2535
BLIPDiffusion (Gen)	0.4718	0.0798	0.4751
TextualInversion	0.6508	-0.0169	0.6450
DreamBooth	0.4153	0.3535	0.4396
Subject-Driven Image Editing			
BLIPDiffusion (Edit)	0.4063	-0.1081	0.4000
DreamEdit	0.1994	-0.0877	0.1878
PhotoSwap	0.3300	0.0814	0.3424

Table 10: DINO vs Human correlation on several tasks.

Image Model	Metric: CLIP-I M-H^{SC}_{corr}	M-H^{PQ}_{corr}	M-H^{O}_{corr}
Multi-Concept Image Composition			
TextualInversion	0.1511	-0.1847	0.1523
DreamBooth	0.0741	-0.1166	0.0758
CustomDiffusion	0.2319	0.0116	0.2246
Subject-Driven Image Generation			
DreamBoothLora	0.2499	0.1801	0.2615
BLIPDiffusion (Gen)	0.2611	0.1031	0.2660
TextualInversion	0.5776	0.0362	0.5775
DreamBooth	0.1324	0.3648	0.1587
Subject-Driven Image Editing			
BLIPDiffusion (Edit)	0.4202	-0.0798	0.3844
DreamEdit	0.1927	0.1535	0.1955
PhotoSwap	0.2613	0.3028	0.2874

Table 11: CLIP-I vs Human correlation on several tasks.

Data amount per model	Total Human rating data
Task: Text-guided Image Generation	
197	2955
Task: Mask-guided Image Editing	
179	2148
Task: Text-guided Image Editing	
179	4296
Task: Subject-driven Image Generation	
150	1800
Task: Subject-driven Image Editing	
154	1386
Task: Multi-concept Image Composition	
102	918
Task: Control-guided Image Generation	
150	900
	Sum of 7 tasks
	14403

Table 12: Number of human ratings from ImagenHub used in this paper.

C Backbone Performances

C.1 Parsing MLLM outputs

We tried to parse the output using Regex and modify the format requirement if the MLLM fail to do so. If the output failed to pass our parsing rules, we fill random value as output to penalize the correlation.

C.2 Observations of GPT-4v.

GPT-4-vision-preview tends to be the best MLLM in this context. It can understand every task instruction in VIESCORE and produce reasonable scores and rationale.

C.3 Observations of LLaVA.

LLaVA-1.5-7B can also understand every task instruction in VIESCORE and produce reasonable rationale. However, the scores produced tend to be concentrated toward certain numbers.

C.4 Observations of Qwen-VL.

Qwen-VL-7B does not understand the meaning of delimiter. However, it was able to output a JSON-like dictionary following the instructions on both SC and PQ. The rationale produced is often not reasonable.

C.5 Observations of BLIP2.

BLIP-2 FLAN-T5-XXL often failed to produce the result formats according to the instructions, especially in PQ. It also tend to give 0 score in SC. Prompt engineering in our context does not solve the issue.

C.6 Observations of InstructBLIP.

InstructBLIP-T5-XL shares same observation as BLIP-2 FLAN-T5-XXL.

C.7 Observations of Fuyu.

Fuyu-8B always output 0 and failed to follow in our instruction.

C.8 Observations of CogVLM.

CogVLM tends to output numbers fall off the range [0, 10] and often failed to follow the required format. Prompt engineering in our context does not solve the issue.

C.9 Observations of OpenFlamingo.

OpenFlamingo simply printing blank as the output in our context.