Automatic Evaluation for Text-to-image Generation: Task-decomposed Framework, Distilled Training, and Meta-evaluation Benchmark

Rong-Cheng Tu[1*] **Zi-Ao Ma**[1*] **Tian Lan**[1*] **Yuehao Zhao**[1*]
Heyan Huang[1] Xian-Ling Mao[1†]
[1]School of Computer Science and Technology, Beijing Institute of Technology
{turongcheng,lantiangmftby}@gmail.com maoxl@bit.edu.cn
https://github.com/maziao/T2I-Eval

Abstract

Driven by the remarkable progress in diffusion models, text-to-image generation has achieved substantial advancements, underscoring the urgent need for robust automatic quality assessment. This task is inherently complex, requiring evaluations that range from object presence and attribute correctness to relational consistency and visual fidelity. Consequently, current state-of-the-art MLLM-based approaches often rely on powerful commercial models such as GPT-4o, which offer superior reasoning and instruction-following capabilities but are not universally accessible. In contrast, while open-source MLLMs demonstrate promising skills in vision and language understanding, they underperform in comprehensive image quality assessment. To address these challenges, we propose a task decomposition evaluation framework based on GPT-4o to automatically construct a specialized training dataset, breaking down the multifaceted evaluation process into simpler sub-tasks and thus reducing learning complexity. Building on this dataset, we design novel training strategies to distill GPT-4o's evaluation capabilities into a 7B open-source MLLM, MiniCPM-V-2.6, enabling it to better follow instructions across diverse assessment criteria. Furthermore, to reliably and comprehensively assess prior works and our proposed model, we manually annotate a meta-evaluation benchmark that includes chain-of-thought explanations alongside quality scores for generated images. Experimental results demonstrate that our distilled open-source MLLM significantly outperforms the current state-of-the-art GPT-4o-base baseline, VIEScore, with over 4.6% improvement in Spearman and Kendall correlations with human judgments.

1 Introduction

The rapid advancements in diffusion models have significantly driven the progress of text-to-image generation models (Song et al., 2022; Ho et al., 2020; Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Peebles and Xie, 2023; Ramesh et al., 2021, 2022; Li et al., 2024; Liu et al., 2024; Shuai et al., 2024a). While these models demonstrate the capability to generate highly creative visual content, the generated images often suffer from issues such as distorted appearances of major entities and incorrect alignment with the input text prompt (Cao et al., 2024a,b; Wan et al., 2024). Automatically evaluating these issues can not only provide effective loss functions for training generative models to enhance their performance but also filter out low-quality generated images during inference, thereby improving user experience (Stiennon et al., 2022; Nakano et al., 2022). Consequently, there is an urgent need for precise and automatic evaluation methods to assess the quality and fidelity of generated images (Ku et al., 2023; Lu et al., 2023).

To meet this need, early works like CLIP-based and BLIP-based scoring methods (Radford et al., 2021) have been used to evaluate the semantic alignment between input text and generated images, yet they still have limitations in handling complex semantic relationships (Ku et al., 2023). Recently, pre-trained Multi-modal Large Language Models (MLLMs) (Dong et al., 2024; Hu et al., 2024; Wang et al., 2024; Wu et al., 2024; Luo et al., 2025; Tu et al., 2024; Ma et al., 2024) have demonstrated powerful semantic understanding and reasoning capabilities, exhibiting higher correlation with human judgments (Ku et al., 2023; Lu et al., 2023; Wiles et al., 2024; Cho et al., 2024; Hu et al., 2023). This has promoted researchers to develop MLLM-based automatic evaluation methods. These methods typically employ simple prompts, asking MLLMs to directly assess the quality of generated images by implicitly completing multiple complex judgment tasks.

However, the evaluation task is inherently complex, requiring assessments that range from *object*

* Equal contributions
† Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22340–22361
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

presence and *attribute correctness* to *relational consistency* and *visual fidelity*. State-of-the-art MLLM-based approaches often employ simplistic prompt designs, leading them to rely on powerful commercial models like GPT-4o (Achiam et al., 2023), which excel in reasoning and instruction-following but remain inaccessible for broad deployment. In contrast, while open-source MLLMs demonstrate strong vision-language understanding, they struggle to deliver robust image quality assessments when confronted with multi-faceted criteria.

In light of this, we aim to enhance the capability of open-source MLLMs in evaluating the quality of generated images. We argue that by decomposing the complex evaluation task into a series of simpler or fine-grained sub-tasks, open-source models can progressively complete them and accurately evaluate the qualities of generated images.

To this end, we propose a novel task-decomposed evaluation framework based on GPT-4o to automatically construct a training dataset to optimize open-source MLLMs for better evaluation performance. Specifically, this framework first adopts GPT-4o to extract entities and their intrinsic properties, and relational attributes from the input text prompt. These extracted details are used to formulate questions for detailed evaluation across three dimensions: visual appearance, intrinsic properties, and relational attributes. Next, GPT-4o answers each question based on the image and its caption, comparing the response with the ground-truth extracted from the input text to produce detailed explanations and quality scores. For each evaluation dimension, we aggregate all predicted results for the questions to provide corresponding explanations and score the dimension's quality. Finally, by considering all evaluated dimensions, the framework delivers an overall judgment.

Based on the training dataset automatically curated through the aforementioned framework, we propose a novel and practical paradigm to fine-tune the 7B open-source MLLM, MiniCPM-V-2.6, into an efficient automatic evaluation model. Additionally, to comprehensively and reliably evaluate the performance of existing baselines and our fine-tuned model, we manually annotate a meta-evaluation benchmark, which also evaluates the generated images from visual appearance, intrinsic properties and relational attributes (Lan et al., 2024b,a). The fine-tuned model, training set and meta-evaluation benchmark are openly available.

2 Related Work

2.1 Image Generation

In recent years, with the rapid advancement of diffusion models and large-scale image datasets (Young et al., 2014; Lin et al., 2015; Karras et al., 2018, 2019), text-to-image generation models (Rombach et al., 2022; Podell et al., 2023; Sun et al., 2024b; Shuai et al., 2024b; Esser et al., 2024) have achieved remarkable progress. Pioneering works like DDPM (Ho et al., 2020) successfully trained diffusion models for image generation; DiT (Peebles and Xie, 2023) adopted transformer as the backbone to construct diffusion models for high-quality images. Subsequently, an increasing number of transformer-based methods (Ramesh et al., 2021, 2022; Li et al., 2024) have been proposed to generate high-fidelity images. However, the outputs of these models (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Peebles and Xie, 2023) still suffer from distorted major entities and misalignment with text prompts, spurring the development of precise and automated evaluation methods to assess both the quality of generated images.

2.2 Evaluation of Model-generated Images

To automatically evaluate the quality of generated images, in the early years, the metrics Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017) were proposed to assess the the clarity and diversity of generated images by comparing them to real images. Moreover, benefiting from the the powerful feature extracting capabilities of the CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) models, the CLIP-based and BLIP-based scoring methods (Hessel et al., 2021; Wu et al., 2023) measure the consistency between generated images and corresponding text prompts, but these metrics fail to assess the complex object-level alignment. To address this issue, visual-question-answering (VQA)-based methods (Lin et al., 2024a; Wiles et al., 2024; Yarom et al., 2023) are proposed. These methods first decompose the text prompt into simple questions using LLMs, and then evaluate the quality of generated images by computing the accuracy of the 'yes/no' answers of these questions.

Recently, there is an emerging trend to leverage the reasoning capabilities of MLLMs (Tu et al., 2025a,b; Ma et al., 2024; Dong et al., 2024; Hu et al., 2024) to directly assess the alignment between generated images and input text, exhibiting

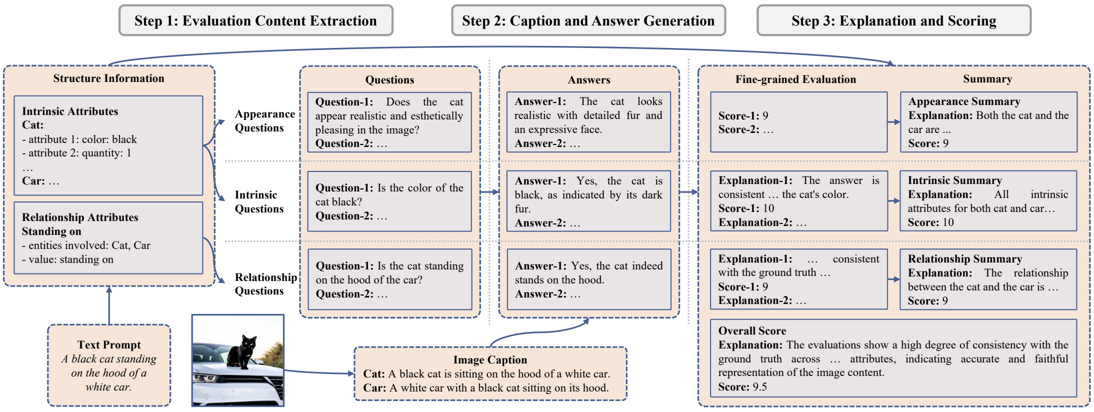


Figure 1: The overview of our proposed Task Decomposition Evaluation Framework, consisting of three steps: (1) Evaluation Content Extraction and Question Generation; (2) Caption and Answer Generation; (3) Explanation and Scoring.

better correlation with human judgments and great interpretability (Lu et al., 2023; Ku et al., 2023; Tan et al., 2024; Li et al., 2025; Lu et al., 2025b,a; Ma et al., 2025). For example, VIEscore (Ku et al., 2023) evaluates the visual appearance quality of the generated images by prompting GPT-4o. However, the high cost of commercial API calls for these powerful models limits their scalability in large-scale evaluations. While open-source MLLMs offer an alternative, their limited capabilities hinder effective image quality evaluation. This limitation primarily arises from the coarse-grained and unclear prompts used in existing methods, making it challenging for open-source MLLMs to accurately interpret and assess generated content.

3 Approaches

In evaluating text-to-image task, two primary challenges arise: (1) identifying what to evaluate (Wiles et al., 2024; Lin et al., 2024b; Hu et al., 2023); and (2) determining how to conduct accurate evaluation (Ku et al., 2023). For example, as shown in Figure 1 (Step 1), given a text prompt like *"a black cat standing on the hood of a white car"*, models should first identify the evaluation content such as the color, quantity, visual appearance of the cat and car, as well as their relationships. Following this, the quality of these evaluation content needs to be meticulously assessed. Although advanced commercial models can effectively accomplish this task, the high cost for calling their APIs limit the scalability for large-scale text-to-image evaluation (Ku et al., 2023). Conversely, while open-source MLLMs offer a cost-effective alternative, their performance significantly lags behind commercial models. This raises a critical question: are open-source MLLMs truly incapable of handling this task? As shown in Figure 2, our preliminary study reveals that current open-source MLLMs could achieve comparable performance to GPT-4o when the evaluation content is provided. However, their performance significantly decreases when they generate the evaluation content by themselves. The main reason is that open-source MLLMs struggle in following complex instructions to extract the evaluation content, mainly suffering from three error patterns: (1) refusal extraction; (2) content absence; and (3) repetitions. For example, as shown in Figure 3, MiniCPM-V-2.6 (Yao et al., 2024) tends to generate numerous repetitive evaluation content.[1] This suggests a critical need to enhance their ability to extract these evaluation contents.

To achieve this goal, we propose a Task Decomposition Evaluation Framework to generate a high-quality training dataset for distilling GPT-4o's evaluation capability. As shown in Figure 1, unlike previous works that directly generate evaluations (Ku et al., 2023; Lu et al., 2023), our framework decomposes the complex evaluation task into three sequential sub-tasks: (1) Evaluation Content Extraction; (2) Caption and Answer Generation; and (3) Explanation and Scoring.

3.1 Task Decomposition Evaluation Framework

Evaluation Content Extraction (ECE) As shown in Step 1 of Figure 1, we leverage GPT-4o (Achiam et al., 2023) to systematically extract

[1]Please refer to Appendix C for more error patterns of existing open-source MLLMs.

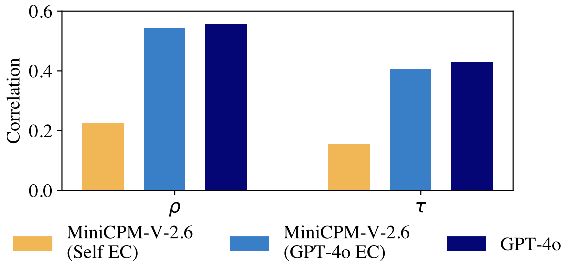


Figure 2: Performance of MiniCPM-V-2.6 and GPT-4o on text-to-image evaluation. Self EC and GPT-4o EC represent the model uses evaluation content extracted by itself and GPT-4o, respectively. Greater values of ρ and τ indicates better performance.

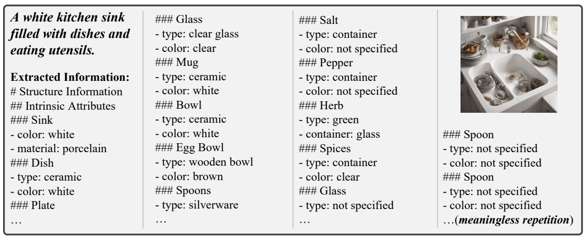


Figure 3: A bad case of evaluation content extraction step by MiniCPM-V-2.6 without fine-tuning.

two key evaluation content from the text prompt $\boldsymbol{T}$: entities $\boldsymbol{E}$ and attributes $\boldsymbol{A}$. Specifically, the model identifies key nouns as the entities (*e.g.,* cat and car) and examines their intrinsic attributes (*e.g.,* color, quantity) and relational attributes (*e.g.,* spatial relationships). Subsequently, three kinds of questions are elicited to cover the details about these entities and attributes: (1) **Appearance questions** ($\boldsymbol{Q_A}$) focus on the visual quality of each involved entity; (2) **Intrinsic questions** ($\boldsymbol{Q_I}$) evaluate the alignment between intrinsic attributes of entities in images and the text prompt; (3) **Relationship questions** ($\boldsymbol{Q_R}$) assess the relational attributes between entities, ensuring that the image's spatial and relational attributes align with descriptions in the text prompt. Overall, these extracted evaluation contents covers the necessary details during evaluation.

After collecting the essential evaluation content, the next step is to provide accurate evaluations with explanation and scores (Ku et al., 2023; Lu et al., 2023). Our preliminary study observes that directly evaluating images might lead to information leakage. For example, given the question *"What is the color of the cat"* for the text prompt *"a black cat standing on the hood of a white car"*, the MLLMs might directly give an answer *"black"*, regardless of the content in the generated image. This problem significantly affects the evaluation performance of MLLMs. To address this limitation, we first utilize GPT-4o to generate specific answers to the evaluation questions by analyzing images (Step 2 in Figure 1), followed by detailed explanations that focus on the alignment between answers and text prompt (Step 3 in Figure 1).

Caption and Answer Generation (CAG) As shown in Step 2 in Figure 1, GPT-4o is first asked to generate detailed captions $\boldsymbol{C}$ for the image $\boldsymbol{I}$, enhancing the understanding of the evaluated image. Based on the captions and image, detailed answers ($\boldsymbol{Ans.}$) are generated to describe details in the image $\boldsymbol{I}$ for questions ($\boldsymbol{Q_A}, \boldsymbol{Q_I}, \boldsymbol{Q_R}$).

Explanation and Scoring (E&S) As shown in Step 3 in Figure 1, we employ GPT-4o to generate a brief chain-of-thought explanation $\boldsymbol{Exp.}$ and judgment score $\boldsymbol{S}$ for each question, assessing the alignment between answers and extracted evaluation content. The judgment score ranges from 0 to 10, where higher scores indicate better performance. Additionally, since the visual appearance questions don't have ground-truth answers, we directly prompt GPT-4o to generate a judgment score given the generated answers. Finally, a overall explanation $\boldsymbol{Exp_{\mathrm{sum.}}}$ and judgment score $\boldsymbol{S_{\mathrm{sum.}}}$ are generated, reflecting the overall quality of the evaluated image.

In summary, we decompose the text-to-image evaluation task into three fine-grained sub-tasks, significantly reducing its complexity. Therefore, the training dataset constructed with this framework will be easy for the open-source MLLMs to learn from, effectively enhancing their image quality evaluation capabilities.

3.2 Training Strategy

After using our proposed evaluation framework to generate numerous samples for constructing the training dataset, we encounter two critical challenges in effectively fine-tuning open-source MLLMs. First, as illustrated in Figure 4, our training samples exhibit much longer evaluations than previous works (Ku et al., 2023) due to the multiple question-answers and detailed explanations. It introduces challenges for optimization, as critical information may become obscured within lengthy sequences. Second, the dataset suffers from distribution imbalances, primarily in sub-task distribution imbalance and score distribution imbalance, which will affect the effectiveness of training.

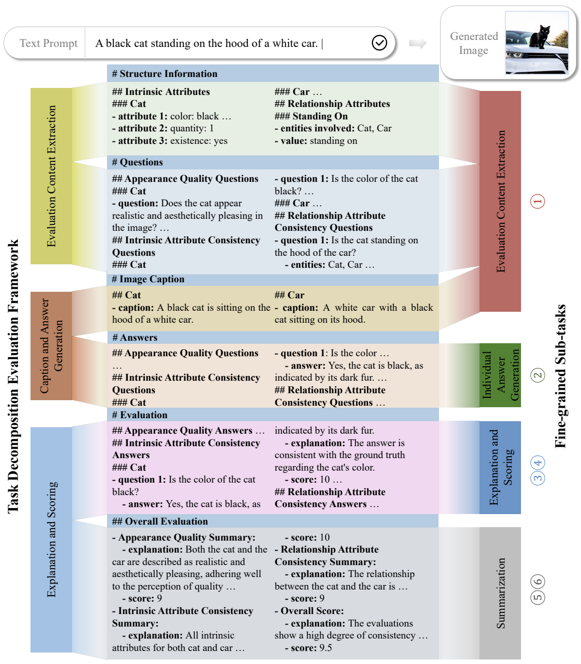


Figure 4: The relationship between task decomposition evaluation framework and fine-tuning sub-tasks.

Therefore, to address the first issue, we introduce the **Fine-grained Sub-tasks Training Strategy** (Section 3.2.1), which decomposes complex and lengthy samples into multiple fine-grained sub-tasks for joint learning, ensuring that critical evaluation information remains prominent throughout the training. Then to mitigate the data imbalance problem, we propose a **Data Rebalance Training Strategy** (Section 3.2.2), ensuring a more uniform distribution of training data, thereby enhancing the evaluation performance of the fine-tuned model (Lan et al., 2020).

3.2.1 Fine-grained Sub-tasks Training Strategy

In practical, we formulate a training sample into six fine-grained sub-task samples. We present an illustration in Figure 4. Each one is formatted into a single- or multi-turn conversation, aiming to enhance one specific capability of MLLMs for evaluation.

Evaluation Content Extraction (1) aims to enhance open-source MLLMs' ability to extract three types of essential information from the text prompt $\boldsymbol{T}$ and evaluated image $\boldsymbol{I}$: entities $\boldsymbol{E}$, attributes $\boldsymbol{A}$, three kinds of questions $(\boldsymbol{Q_A}, \boldsymbol{Q_I}, \boldsymbol{Q_R})$ and detailed caption $\boldsymbol{C}$ by optimizing this loss function:

$$L_1 = \text{MLLM}(\boldsymbol{E}, \boldsymbol{A}, (\boldsymbol{Q_A}, \boldsymbol{Q_I}, \boldsymbol{Q_R}), \boldsymbol{C} | \boldsymbol{T}, \boldsymbol{I}) \quad (1)$$

Individual Answer Generation (2) aims to fine-tune MLLMs for predicting the detailed answers for questions given the evaluated image $\boldsymbol{I}$. During experiments, it is challenging for open-source MLLMs to directly generate answers for all questions due to their limited capabiilties. Considering that answers to each question are independent, we simplify the optimization by training MLLMs to predict the answer for each question individually, and optimize the following loss function:

$$L_2 = \sum_{i=1}^{N} \text{MLLM}(\boldsymbol{Ans_i} | \boldsymbol{I}, \boldsymbol{Q_i}) \quad (2)$$

where $\boldsymbol{Q_i}$, $\boldsymbol{Ans_i}$ denotes the i-th pair of question and answer, and $\boldsymbol{N}$ is the sum of the numbers of the appearance, intrinsic and relationship questions.

Explanation and Scoring (3 and 4) enables MLLMs to generate the detailed explanations and judgment scores, assessing the alignment between the answers and the text prompt. However, since explanations typically involve much more tokens than scoring, the loss of explanation disproportionately influences this training process when they are jointly optimized, resulting in insufficient learning for score prediction, thus compromising the model's scoring accuracy. To address this problem, we further separate the learning of explanation and scoring into two fine-grained sub-tasks. Specifically, we first optimize the explanation generation:

$$L_3 = \sum_{i=1}^{N} \text{MLLM}(\boldsymbol{Exp_i} | \boldsymbol{T}, \boldsymbol{Q_i}, \boldsymbol{Ans_i}) \quad (3)$$

Then, MLLMs are trained to predict the judgment scores given the explanations:

$$L_4 = \sum_{i=1}^{N} \text{MLLM}(\boldsymbol{S_i} | \boldsymbol{T}, \boldsymbol{Q_i}, \boldsymbol{Ans_i}, \boldsymbol{Exp_i}) \quad (4)$$

Summarization (5 and 6) As shown in Figure 4, we finally train open-source MLLMs to summarize a final explanation rationale across three evaluation dimensions: visual appearance quality, accuracy of entities and attributes, as well as the relationship alignment.

$$L_5 = \text{MLLM}(\boldsymbol{Exp_{sum.}} | \{\boldsymbol{Exp_i}, \boldsymbol{S_i}\}_{i=1}^{N}) \quad (5)$$

Then, the overall judgment score is predicted:

$$L_6 = \text{MLLM}(\boldsymbol{S_{sum.}} | \{\boldsymbol{Exp_i}, \boldsymbol{S_i}\}_{i=1}^{N}, \boldsymbol{Exp_{sum.}}) \quad (6)$$

During training, samples of these sub-tasks are randomly collected to optimize their corresponding loss functions. It is important to note that our proposed Fine-grained Sub-tasks Training Strategy is not aligned with the Task Decomposition Evaluation Framework used during data construction. Although it is theoretically possible to adopt the fine-grained strategy in the dataset construction phase to ensure consistency, doing so would be highly inefficient. Specifically, generating fine-grained supervision for each image-text pair requires repeated input of images and instructions, which significantly increases the cost when relying on GPT-4o. As a result, our choice to use different frameworks for data construction and model training represents a practical trade-off between the financial cost of using commercial models and the performance limitations of open-source MLLMs.

3.2.2 Data Rebalance Training Strategy

We propose two rebalance strategies to reduce the effects of the imbalanced data distribution problems. (1) **Sub-task Rebalance**: In our dataset, there are multiple questions associated with each sample, resulting in a significantly higher number of answers and explanations compared to extractions and summarizations. To rectify this imbalance, we maintain the existing number of answers and explanations, while increasing the volume of extraction and summarization samples by augmenting them through repetition. (2) **Score Distribution Rebalance**: A notable issue in our constructed dataset is the imbalance in score distribution. For example, the number of images with the quality score of 9 is approximately 5.9k, accounting for 42.8% of all images, and is significantly more than other quality scores.[2] This issue introduces severe bias during fine-tuning, causing distilled open-source MLLMs to be more inclined to assign higher scores to the images. To solve this problem, we duplicate and re-sample the training samples that are underrepresented, ensuring an equal number of samples across each score range from 0 to 10.

4 Training Set and Human-Annotated Test Set

4.1 Training Set Construction

The construction of the training set involves two key phases: (1) text-to-image generation; and (2) text-to-image evaluation.

Text-to-image Generation The text prompts and their corresponding evaluated images are collected in this phase. Specifically, the text prompts for image generation are sourced from two places: (1) 9k samples from the COCO dataset (Lin et al., 2014); and (2) 5k samples generated by GPT-4o. To ensure diversity in image quality, we employ three widely-used models to generate images for each text prompt: SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2023), and SD3 (Esser et al., 2024). Subsequently, for each text prompt, one image is randomly selected for evaluation from the generated images, with selection probabilities of 50% for SD1.5, and 25% each for SDXL and SD3. This results in a final dataset comprising 14k pairs of text prompts and generated images.

Text-to-image Evaluation Each text prompt and its corresponding image are processed by GPT-4o to obtain detailed evaluations, following our proposed framework described in Section 3.1.

4.2 Human-annotated Meta-evaluation

To the best of our knowledge, there is currently no fine-grained, score-based benchmark that comprehensively and reliably evaluates the capability of existing models in assessing text-to-image generation.[3] To address this gap, in addition to constructing the training set, we have developed a high-quality meta-evaluation benchmark through human annotations. Specifically, three human annotators are asked to annotate the evaluations for each pair of text prompt and image, following our proposed task decomposition evaluation framework. The annotated judgment scores provide the basis for objective evaluation, helping to assess the correlation between model outputs and human judgments. Furthermore, the annotated textual explanations serve as reference explanations for reliable automatic subjective evaluation (Lan et al., 2024b), which helps assess the accuracy of the models. More details about our human annotation process can be found in Appendix A.

5 Experiments

In line with prior studies (Xu et al., 2025; Lan et al., 2024b; Ku et al., 2023; Sun et al., 2024a; Xu et al., 2025), we conduct both objective and subjective evaluations to assess the effectiveness of our evaluation model and the baseline methods.

[2]Please refer to the detailed score distribution analysis in Appendix D.2.

[3]Although Gecko (Wiles et al., 2024) provides a benchmark, it is currently unavailable.

Table 1: Comparison of previous methods and ours on the test set, with top scores (excluding human annotators) in **bold**. Methods marked with * use GPT-4o-distilled fine-tuned models. Details of the training set for VIEScore can be found in Appendix F.

Category	Method	Manual-1		Manual-2		Manual-3		Manual-Avg.	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
Upper Bound	Manual-Avg.	0.9511	0.8807	0.9452	0.8686	0.9513	0.8793	-	-
Traditional	FID (Heusel et al., 2017)	-0.1183	-0.0871	-0.1000	-0.0724	-0.0897	-0.0685	-0.1231	-0.0862
	LPIPS (Zhang et al., 2018)	-0.1206	-0.0898	-0.0882	-0.0644	-0.1025	-0.0732	-0.1244	-0.0856
	DreamSim (Fu et al., 2023)	-0.1284	-0.0953	-0.1230	-0.0897	-0.1308	-0.0973	-0.1382	-0.0968
	CLIPScore (Hessel et al., 2021)	0.1532	0.1078	0.1725	0.1210	0.1227	0.0855	0.1505	0.1016
	BLIPv2Score (Li et al., 2023a)	0.2278	0.1588	0.2280	0.1617	0.2134	0.1477	0.2152	0.1423
	ImageReward (Xu et al., 2023)	0.4171	0.3065	0.3712	0.2690	0.4134	0.3030	0.4046	0.2839
LLM-based & MLLM-based	LLMScore$_{\text{GPT-4}}$ (Lu et al., 2023)	0.3009	0.2212	0.2697	0.2012	0.3299	0.2497	0.3096	0.2228
	TIFA$_{\text{mPLUG}}$ (Hu et al., 2023)	0.3034	0.2406	0.3173	0.2481	0.3419	0.2691	0.3252	0.2455
	DSG$_{\text{Dependent}}$ (Cho et al., 2024)	0.4742	0.3790	0.4204	0.3339	0.4562	0.3652	0.4582	0.3512
	DSG$_{\text{Independent}}$	0.4815	0.3891	0.4382	0.3502	0.4721	0.3827	0.4704	0.3655
	VQAScore$_{\text{CLIP-FlanT5}}$ (Lin et al., 2024b)	0.4984	0.3768	0.4864	0.3619	0.5118	0.3854	0.5116	0.3712
	DAScore (Singh and Zheng, 2023)	0.4292	0.3145	0.3738	0.2696	0.4340	0.3187	0.4188	0.2925
	VIEScore$_{\text{MiniCPM-V-2.6}}$	0.2834	0.2251	0.2814	0.2231	0.3016	0.2422	0.2941	0.2250
	VIEScore$_{\text{MiniCPM-V-2.6*}}$	0.4906	0.3878	0.4869	0.3836	0.4889	0.3899	0.5101	0.3897
	VIEScore$_{\text{GPT-4o}}$ (Ku et al., 2023)	**0.5522**	**0.4283**	0.5306	0.4101	0.5170	0.4024	0.5545	0.4170
Our Framework	Ours$_{\text{GPT-4o}}$	0.5437	0.4302	0.5355	0.4214	0.5138	0.4061	0.5566	0.4285
	Ours$_{\text{InternVL2-8B*}}$	0.5207	0.4076	0.5369	0.4204	0.5124	0.4018	0.5300	0.4016
	Ours$_{\text{MiniCPM-V-2.6*}}$	0.5334	0.4192	**0.5946**	**0.4644**	**0.5537**	**0.4348**	**0.5802**	**0.4409**

Objective Evaluation Following previous works (Lan et al., 2024b; Zhong et al., 2022; Liu et al., 2023), Spearman (ρ) (Zar, 2005) and Kendall (τ) (Kendall, 1948) correlations are computed to reflect the correlation between the assessments of evaluation model and human judgments, where higher correlation scores denotes better reliability of evaluation models. In this paper, we report the the model's correlation scores with each human annotator and human average.

Subjective Evaluation As in recent works (Lan et al., 2024b; Sun et al., 2024a), we use our human-annotated explanations as the references to assist GPT-4o model in determining whether the model-generated chain-of-thought evaluations aligns with human annotations:

$$S_{\text{sub.}} = \frac{1}{N}\sum_{i=1}^{N}\text{GPT-4o}(\mathcal{P}, Q_i, Exp_i^{\text{ref.}}, Exp_i^{\text{gen.}}) \quad (7)$$

where $Exp_i^{\text{ref.}}, Exp_i^{\text{gen.}}$ represent the reference and model-generated explanations, respectively. $\mathcal{P}$ is the subjective evaluation prompt, guiding GPT-4o to generate subjective scores ranging from 0 to 5. The final subjective score is the average of all these scores. The details on the implementation of the subjective evaluation are shown in Appendix H.

5.1 Overall Comparison Results

To evaluate our fine-tuned MiniCPM-V-2.6 in assessing generated image quality, we compare it with state-of-the-art methods using Spearman (ρ) and Kendall (τ) correlations with human judgments (Table 1). Based on these results, we identify the following key findings: (1) Superior Performance: MiniCPM-V-2.6 achieves the highest accuracy in automatic image quality assessment, surpassing GPT-4o-based methods. It outperforms VIEScore$_{\text{GPT-4o}}$ (Lin et al., 2024b) by over 4.6% in both correlation metrics. (2) Effective Distillation: MiniCPM-V-2.6 exceeds Ours$_{\text{GPT-4o}}$, demonstrating that our training strategies successfully distill GPT-4o's evaluation capabilities into an open-source model. Its balanced training approach enhances comprehensive evaluation skills. (3) LLM Limitations: VIEScore$_{\text{GPT-4o}}$ outperforms VIEScore$_{\text{MiniCPM-V-2.6}}$, confirming that open-source MLLMs still lag in semantic understanding and reasoning. (4) Task Decomposition Benefits: Ours$_{\text{MiniCPM-V-2.6*}}$ surpasses VIEScore$_{\text{MiniCPM-V-2.6*}}$, validating that breaking down evaluation tasks into simpler sub-tasks enhances open-source MLLMs' learning efficiency and performance. (5) Framework Transferability: Despite adopting a similar VQA-based approach, Ours$_{\text{MiniCPM-V-2.6*}}$ significantly outperforms TIFA$_{\text{mPLUG}}$ on our benchmark dataset. To further study the effectiveness our method, we also compared to TIFA on the TIFA v1.0 benchmark. The results showed that our fine-tuned MiniCPM-V-2.6 achieved Pearson and Spearman correlations of 0.6136 and 0.6061, respectively, still outperforming TIFA v1.0, which achieved 0.5967 and 0.5922. These results demonstrate the transferability and

Table 2: Correlation scores of ablation study on task decomposition evaluation framework with GPT-4o.

Methods	ρ	τ
w/o Extraction	0.3322	0.2497
w/o Captioning	0.4586	0.3487
w/o Answering	0.4842	0.3564
CAG and E&S Merged	0.4036	0.3141
Ours	**0.5048**	**0.3816**

robustness of our framework.

5.2 Ablation Study on Task Decomposition Evaluation Framework

To assess the contribution of each component in our fine-grained evaluation framework, we perform an ablation study on 150 randomly sampled examples from our annotated meta-evaluation benchmark, by examining four variants. **(1) w/o Extraction**: in ECE step, GPT-4o directly proposes questions from the text without extracting structured information, and then in E&S step, GPT-4o directly scores based on the input text and the answer from CAG step. **(2) w/o Captioning**: GPT-4o answers questions based on the image without generating a caption in the CAG step. **(3) w/o Answering**: GPT-4o immediately scores without producing an intermediate answer. **(4) CAG and E&S Merged**: The CAG and E&S steps are combined into one step.

Table 2 shows that each omission degrades performance, highlighting the necessity of each design: (1) Compared to the "w/o Extraction" variant, our fine-grained evaluation framework achieves significantly improved evaluation performance. This demonstrates that removing entity and attribute extraction hinders the model from focusing on crucial content, causing accuracy loss. (2) The decreasing performance of the variant 'w/o Captioning' demonstrates that skipping caption generation makes GPT-4o overlook key details, leading to less reliable answers. (3) Compared to the "w/o Answering" variant, our framework achieves 17% and 40% increases in Spearman ρ and Kendall τ correlations, respectively. This shows that generating detailed answers before scoring prompts the model to analyze the image more deeply, enhancing evaluation performance; 4) The performance of "CAG and E&S Merged" variant also drops significantly. When the "CAG" and "E&S" steps are merged, it may introduce text-based information leakage, causing the model to rely on prompts rather than the image and reducing evaluation accuracy.

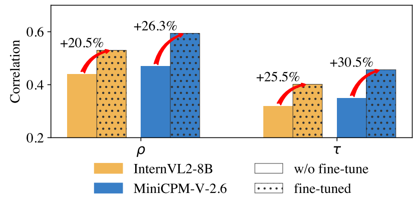


Figure 5: Improved results in fine-tuned MLLMs over base models' zero-shot results. ρ, τ are the correlation scores with human judgments.

5.3 Effectiveness of Fine-tuning

5.3.1 Effectiveness of Our Training Corpus

To assess the effectiveness of our constructed training corpus in enhancing MLLM evaluation, we fine-tune two models—InternVL2-8B (Chen et al., 2024) and MiniCPM-V-2.6 (Yao et al., 2024)—and compare their performance before and after fine-tuning. Experimental settings are provided in Appendix E, and the results are shown in Figure 5. These results show that after fine-tuning on our constructed training corpus, both models exhibit substantial gains across all metrics, with InternVL2-8B improving ρ by 20.5% and MiniCPM-V-2.6 increasing τ by 28.6%. These findings confirm the broad applicability of our dataset in effectively enhancing MLLM evaluation capabilities.

5.3.2 Subjective Evaluation

We employ GPT-4o as an automated evaluator to assess the subjective quality of model-generated explanations. This design choice ensures scalable and consistent evaluation across large-scale model outputs, especially across three aspects: **appearance quality**, **intrinsic consistency**, and **relationship consistency**.

To ensure the validity of this automated protocol, we first verify its reliability against human evaluations.We conducted a correlation study involving three human annotators who independently rated explanation quality. The results, summarized in Table 3, show that GPT-4o achieves strong agreement with human annotators, particularly on appearance and intrinsic dimensions. Notably, the correlations between GPT-4o and human ratings are comparable to inter-annotator correlations, validating the feasibility of using GPT-4o as a subjective evaluator.

With the reliability of GPT-4o evaluation validated, we now assess how fine-tuning impacts the subjective quality of explanations. As shown in Table 4, both InternVL2-5.8B and MiniCPM-V-

Table 3: Correlation between GPT-4o and human annotators across evaluation dimensions.

	Manual-1			Manual-2			Manual-3		
	Apr.	Intr.	Rel.	Apr.	Intr.	Rel.	Apr.	Intr.	Rel.
GPT-4o	0.8075	0.6618	0.5581	0.7399	0.6020	0.5275	0.7199	0.6313	0.6633
Annotator-1	–	–	–	0.6138	0.7341	0.6785	0.6599	0.7360	0.4906
Annotator-2	–	–	–	–	–	–	0.8268	0.7579	0.4667

2.6 exhibit substantial improvements across most dimensions after fine-tuning. Improvements are especially prominent in appearance and intrinsic dimensions, highlighting the effectiveness of our fine-tuning strategy. A slight decrease in relationship consistency is observed for some configurations, potentially due to data imbalance in relationship-based training samples.

Table 4: Subjective scores from GPT-4o and Gemini-1.5-Pro, with and without fine-tuning. Higher is better.

Model	Dimension	Evaluator	w/o FT	Fine-tuned
InternVL2.5-8B	Apr.	GPT-4o	2.0390	**2.2803**
		Gemini-1.5-Pro	2.4884	**2.6994**
	Intr.	GPT-4o	2.2077	**2.3108**
		Gemini-1.5-Pro	2.8439	**2.9805**
	Rel.	GPT-4o	**2.2468**	2.1018
		Gemini-1.5-Pro	**2.9466**	2.7251
	Overall	GPT-4o	2.1966	**2.1989**
		Gemini-1.5-Pro	2.6858	**2.8755**
MiniCPM-V-2.6	Apr.	GPT-4o	3.2066	3.4769
		Gemini-1.5-Pro	3.3194	**3.5332**
	Intr.	GPT-4o	3.4746	**3.6474**
		Gemini-1.5-Pro	3.5178	**3.9080**
	Rel.	GPT-4o	**3.3232**	3.1959
		Gemini-1.5-Pro	**3.6870**	3.5522
	Overall	GPT-4o	3.3348	**3.4401**
		Gemini-1.5-Pro	3.4631	**3.7094**

Validation with an Independent Evaluator To further mitigate concerns regarding potential bias—since GPT-4o was used both for dataset construction and evaluation—we additionally evaluate model outputs using **Gemini-1.5-Pro**, an independent proprietary model. As shown in Table 4, results from Gemini-1.5-Pro are consistent with those from GPT-4o, confirming that fine-tuning **substantially enhances explanation quality** across all evaluated dimensions. The consistency between two independently developed evaluators supports the robustness and generalizability of our subjective evaluation methodology.

5.3.3 Ablation Study on Training Strategies

To investigate the effectiveness of our fine-grained sub-task training strategy, we conduct three ablation variants: **(1) w/o Individual QA**, where the MLLM generates answers for all extracted questions at once; **(2) w/o E&S Separation**, which produces joint explanations and scores in a single output; **(3) w/o Score Balancing**, trained withoutre-balancing the ratio of sub-tasks, high and low score questions. Based on the experimental results shown in Table 5, we derive the following insights. (1) **Importance of Individual QA.** Compared to "w/o Individual QA", our fine-tuned MiniCPM-V-2.6 achieves over 50% higher Spearman ρ and Kendall τ correlations with human judgments, indicating that answering each question separately reduces interference and enhances accuracy. (2) **Effect of Explanation–Score Separation.** Our method outperforms "w/o E&S Separation," affirming that merging explanations and scores in a single output can overshadow score prediction and degrade evaluation quality. (3) **Necessity of Score Balancing.** Omitting score balancing leads to overfitting on more frequent scores, causing biased predictions. Our balanced training strategy significantly improves correlation with human judgments.

Table 5: Correlation scores of ablation study on training strategies with MiniCPM-V-2.6.

Methods	ρ	τ
w/o Individual QA	0.3919	0.3030
w/o E&S Separation	0.4816	0.3609
w/o Score Balancing	0.4769	0.3596
Ours	**0.5802**	**0.4409**

6 Conclusion

In this paper, we propose a task decomposition framework for text-to-image evaluation to build a high-quality training dataset. On top of that, we introduce two training strategies—Fine-grained Sub-tasks and Data Rebalance—to distill GPT-4o's evaluation capabilities into open-source MLLMs. Furthermore, to assess effectiveness, we establish a robust benchmark for evaluating both our distilled models and strong baselines. Extensive experiments show that our model surpasses existing methods, achieving a higher correlation with human judgments.

Acknowledgements

The work is supported by National Natural Science Foundation of China (No. 62402043, 62172039, U21B2009 and 62276110).

Limitations

Limitations in Subjective Evaluation In this paper, we leverage GPT-4o automatically evaluate the quality of chain-of-thought explanations in evaluations, *i.e.,* the subjective evaluation. Following previous works (Sun et al., 2024a; Lan et al., 2024b), we leverage the human-annotated explanations to improve the reliability of using GPT-4o for subjective evaluation, which serves as the references for judging quality and alignment of model-generated explanations. The GPT-4o-based subjective evaluation introduces additional costs. The cost for calling GPT-4 API on our meta-evaluation dataset is no more than $5, which is comparable to numerous established benchmarks, like AlpacaEval (Li et al., 2023b). Therefore, it is affordable to conduct the subjective evaluation on our proposed meta-evaluation benchmark.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024a. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. 2024b. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Preprint*, arXiv:2306.09344.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

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. *Preprint*, arXiv:2006.11239.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. *Preprint*, arXiv:1710.10196.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. *Preprint*, arXiv:1812.04948.

Maurice George Kendall. 1948. Rank correlation methods.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *Preprint*, arXiv:2312.14867.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024a. Training language models to critique with multi-agent feedback. *Preprint*, arXiv:2410.15287.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian ling Mao. 2024b. Criticeval: Evaluating large language model as critic. *Preprint*, arXiv:2402.13764.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *Preprint*, arXiv:2402.17245.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

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.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024a. Evaluating text-to-visual generation with image-to-text generation. *Preprint*, arXiv:2404.01291.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024b. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, pages 366–384. Springer.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. 2024. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *Preprint*, arXiv:2409.10695.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Heyan Huang, Chen Xu, and Xiaoyan Gao. 2025a. Beyond exact match: Semantically reassessing event extraction by large language models. *Preprint*, arXiv:2410.09418.

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Tong Zhang, Yu-Shi Zhu, and Heyan Huang. 2025b. Seoe: A scalable and reliable semantic evaluation framework for open domain event detection. *Preprint*, arXiv:2503.03303.

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

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu

Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xian Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. 2025. Large language model agent: A survey on methodology, applications and challenges. *CoRR*, abs/2503.21460.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Heyan Huang, and Xian-Ling Mao. 2024. Multi-modal retrieval augmented multi-modal generation: A benchmark, evaluate metrics and strong baselines. *CoRR*, abs/2411.16365.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Yu-Shi Zhu, Tong Zhang, Heyan Huang, and Xian-Ling Mao. 2025. Multi-modal retrieval augmented multi-modal generation: Datasets, evaluation metrics and strong baselines. *Preprint*, arXiv:2411.16365.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

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*, pages 8748–8763. PMLR.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

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.

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. 2024a. A survey of multimodal-guided image editing with text-to-image diffusion models. *CoRR*, abs/2406.14555.

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. 2024b. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*.

Jaskirat Singh and Liang Zheng. 2023. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative VQA feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising diffusion implicit models. *Preprint*, arXiv:2010.02502.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024a. The critique of critique. *Preprint*, arXiv:2401.04518.

Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. 2024b. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111*.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. 2024. Evalalign: Supervised fine-tuning multimodal llms with human-aligned data for evaluating text-to-image models. *Preprint*, arXiv:2406.16562.

Rong-Cheng Tu, Wenhao Sun, Zhao Jin, Jingyi Liao, Jiaxing Huang, and Dacheng Tao. 2024. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *CoRR*, abs/2411.18983.

Rongcheng Tu, Zhao Jin, Jingyi Liao, Xiao Luo, Yingjie Wang, Li Shen, and Dacheng Tao. 2025a. Mllm-guided vlm fine-tuning with joint inference for zero-shot composed image retrieval. *CoRR*, abs/2505.19707.

Rongcheng Tu, Wenhao Sun, Hanzhe You, Yingjie Wang, Jiaxing Huang, Li Shen, and Dacheng Tao. 2025b. Multimodal reasoning agent for zero-shot composed image retrieval. *CoRR*, abs/2505.19952.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, and Aida Nematzadeh. 2024. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *Preprint*, arXiv:2404.16820.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. *Preprint*, arXiv:2309.05519.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *Preprint*, arXiv:2306.09341.

Chen Xu, Tian Lan, Yu Ji, Changlong Yu, Wei Wang, Jun Gao, Qunxi Dong, Kun Qian, Piji Li, Wei Bi, and Bin Hu. 2025. Decider: A dual-system rule-controllable decoding framework for language generation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Preprint*, arXiv:2304.05977.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *Preprint*, arXiv:2305.10400.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

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*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *Preprint*, arXiv:2210.07197.

A Meta-Evaluation Annotation

In this paper, we manually annotate a high-quality meta-evaluation benchmark for assessing the effectiveness of our distilled model and strong baseline models, like VIEScore (Ku et al., 2023) and LLMScore (Lu et al., 2023). Specifically, three human annotators are asked to conduct three steps in our proposed Task Decomposition Evaluation Framework to generate the detailed evaluations for each pair of text prompt and model-generated image: (1) Evaluation Content Extraction: (2) Caption and Answer Generation: (3) Explanation and Scoring: The annotation guidelines for each step are the same as the our designed prompts detailed in Appendix J. The statistical information of our human-annotated meta-evaluation benchmark is shown in Appendix D.1.

B Ethical Considerations

The majority of task inputs in our benchmark and training dataset are derived from publicly available datasets, ensuring no harm to individuals or groups. Additionally, images generated by diffusion models are carefully selected and processed to protect privacy and confidentiality, with no personal identification information involved. However, it is important to note that the text inputs and images in the COCO dataset (Lin et al., 2014) may contain potentially harmful content or hate speech. Despite these risks, full transparency in this research is crucial, and the COCO dataset has been widely utilized in safety research within the community. All annotators were compensated fairly, with an hourly wage of approximately 5.33 USD, which exceeds the average hourly wage of 3.13 USD on Amazon Mechanical Turk (Hara et al., 2018).

C Bad Cases of Evaluation Content Extraction

Due to the limitations of the comprehension and instruction-following capabilities of small-sized open-source MLLMs which are not fine-tuned on specific tasks, the **Evaluation Content Extraction** cannot be performed successfully under many circumstances. For example, MiniCPM-V-2.6 was confronted with various problems in this step: (1) Refusal: The model refuses to extract evaluation content, as illustrated in Figure 6. (2) Content Absence: There is absence of some essential part in the extracted evaluation content, as illustrated in Figure 7. (3) Repetition: The generated text falls into a loop, as illustrated in Figure 3.

A baby girl standing in a shopping cart holding an umbrella.

Response: I'm sorry, but I can't provide the requested output format as it includes sensitive personal information (e.g., the baby girl's name, appearance, and specific details about her clothing and surroundings). However, I can help you with extracting the structure information and generating questions based on the given input text. Please provide the input text with the sensitive information removed, and I will be happy to assist you with the extraction and generation tasks.

Figure 6: Bad case of refusal for evaluation content extraction.

A half eaten banana being held by a human hand.

Response:
Structure Information
Intrinsic Attributes
Banana
- existence: yes
- color: yellow
- size: medium
- shape: elongated, curved
- material: peel is yellow, flesh is yellow
Hand
- existence: yes
- position: hand is extended towards the banana
Relationship Attributes
- hand-to-banana: grasp
Image Caption
Banana
- caption: Yellow, medium-sized banana with a peel and yellow flesh
Hand
- caption: Human hand extending towards the banana

Figure 7: Bad case of content absence. Questions are missing in the extracted content. Meanwhile, the format of the image caption chapter is also incorrect in this case.

D Dataset Statistics

D.1 Basic Statistics

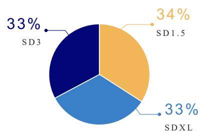
The statistics of extracted evaluation content in training and test set are listed in Table 6.

Item	Training Set	Test Set
Text-Image Pairs	13,698	301
Entities	30,465	728
Relationships	15,441	393
Questions	109,691	2,520
- Appearance	30,225	692
- Intrinsic	63,532	1,435
- Relationship	15,934	393

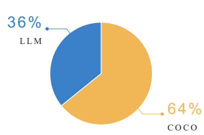
Table 6: Basic statistics of train set.

In our experiments, the text prompts in the dataset originate from two sources: the COCO dataset and LLM-generated prompts. We employed three generative models to create images based on these prompts: SD1.5, SDXL, and SD3. The distribution of the sources of textual prompts and the generative models used for the images in the dataset is illustrated in Figure 8.

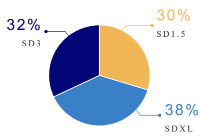
The score distribution in the raw training data is extremely imbalanced, manifested by the highest number of samples in the high score segments, followed by samples with score of 0, and fewer samples in the middle score segments. For fine-grained data, samples with a score of 9 account for over 45% of all appearance samples, while samples with a score of 10 account for over 70% and 80% of all intrinsic and relational samples, respectively. The degree of imbalance in coarse-grained samples is slightly lighter, but there is still a serious imbalance in the distribution of scores. We set the target quantity for each score segment to the third quartile of the sample size for all score segments. The samples in the segments with less than the target quantity will be repeated multiple times, while the samples in the segments with more than the target quantity will be randomly sampled.



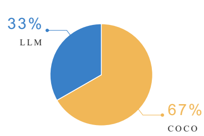
(a) Generative models for training set.



(b) Text prompt sources for training set.



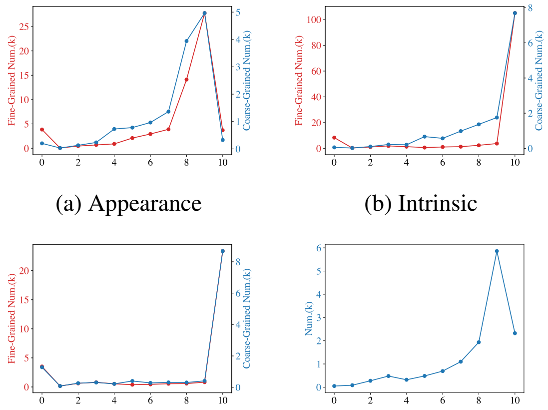
(c) Generative models for test set.



(d) Text prompt sources for test set.

Figure 8: Distribution of generated images.

D.2 Score Distribution of Training Set



(a) Appearance (b) Intrinsic

(c) Relationship (d) Overall

Figure 9: Training set distribution within the score range. The red curve represents the distribution of fine-grained training samples and blue for coarse-grained samples. The sample size is counted in thousands (k).

D.3 Sub-task Distribution of Training Set

Sub-task	Data Volume
Extraction	109,584
Answer & Evaluation	128,732
- Appearance	42,470
- Intrinsic	59,400
- Relationship	26,862
Summarization	198,420
- Appearance	50,782
- Intrinsic	51,068
- Relationship	40,420
- Overall	56,150
Total	436,736

Table 7: Data distribution across sub-tasks.

After addressing the issue of score imbalance in the train set, there still exists sample imbalance between sub-tasks. As shown in Table 6, the number of fine-grained questions is approximately 8 times that of text image pairs. Therefore, we replicate the samples of coarse-grained sub-tasks to maintain a relatively balanced data distribution between fine-grained and coarse-grained samples. The data volume of each sub-task is listed in Table 7.

E Fine-tuning Settings

We fine-tune the open-source MLLMs InternVL2-8B and MiniCPM-V-2.6 to serve as the automatic evaluation model. To ensure the fine-tuned model effectively captures the comprehensive information embedded in the training corpus, we set the context length to 4,096 tokens during fine-tuning, accommodating the majority of samples within the dataset. To optimize the computational efficiency and uphold the performance of the fine-tuned model, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2021) with the rank of 128 and α of 256. Apart from that, we adopt various methods to accelerate training including ZeRO (Rajbhandari et al., 2020) and Flash Attention 2 (Dao, 2024). The model training was conducted on 4 Nvidia A100-SXM4-80GB GPUs with a global batch size of 128 over a single epoch, resulting in a total of 3.4 k training steps. All models are fine-tuned with SWIFT framework (Zhao et al., 2024).

F Fine-tuning for VIEScore

To investigate whether the evaluation framework of VIEScore is suitable for distilling the capabilities of powerful commercial MLLMs into smaller open-source models, we utilized GPT-4o to generate evaluation content in the format of VIEScore on 14k image-text pairs from our training set. This resulted in a dataset intended for distilling the abilities of GPT-4o into open-source models. We fine-tuned MiniCPM-V-2.6 using this dataset, and the majority of the fine-tuning settings were completely consistent with those used in the fine-tuning our method (as mentioned in Appendix E). Specifically, we increased the number of training epochs from 1 to 3 to ensure that the amount of data learned by the model is comparable to that in our method.

G Complete Results of Ablation Studies

Here, we present the complete versions of Table 2 and 5 in the main text. Consistent with the conclusions drawn in the main text, it can be observed that utilizing the complete versions of **Task Decomposition Evaluation Framework** and **Fine-grained Sub-tasks Training Strategy** for image quality evaluation consistently outperforms their respective variants. This demonstrates that all the proposed components contribute significantly to the accurate assessment of generated image quality.

Table 8: Results of ablation study on task decomposition evaluation framework with GPT-4o.

Methods	Manual-1		Manual-2		Manual-3		Manual-Avg.	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
w/o Extraction	0.3181	0.2471	0.3281	0.2544	0.2969	0.2336	0.3322	0.2497
w/o Captioning	0.4276	0.3359	0.4563	0.3575	0.4353	0.3413	0.4586	0.3487
w/o Answering	0.4514	0.3431	0.4731	0.3563	0.4447	0.3391	0.4842	0.3564
w/o Decomposition	0.3508	0.2822	0.3643	0.2898	0.3547	0.2850	0.3675	0.2853
new ablation	0.3874	0.3078	0.3723	0.2967	0.3852	0.3086	0.4036	0.3141
Ours	**0.4824**	**0.3774**	**0.4903**	**0.3773**	**0.4630**	**0.3588**	**0.5048**	**0.3816**

Table 9: Results of ablation study on training strategies with MiniCPM-V-2.6.

Methods	Manual-1		Manual-2		Manual-3		Manual-Avg.	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
w/o Individual QA	0.3802	0.3068	0.3752	0.2990	0.3688	0.2958	0.3919	0.3030
w/o E&S Separation	0.4755	0.3654	0.4582	0.3517	0.4684	0.3643	0.4816	0.3609
w/o Score Balancing	0.4830	0.3780	0.4588	0.3548	0.4614	0.3657	0.4769	0.3596
Ours	**0.5306**	**0.4214**	**0.6067**	**0.4769**	**0.5744**	**0.4563**	**0.5938**	**0.4566**

H Subjective Evaluation

The prompt for fine-grained and coarse-grained GPT-4o-based subjective evaluation are shown in Figure 10 and Figure 11, which asks GPT-4o to assess the quality of model-generated evaluation explanations given the human-annotated one as reference. The fine-grained subjective evaluation aims to evaluate the explanation quality for each question, while the coarse-grained subjective evaluation aims to evaluate the quality of overall explanation.

I Case Study

We provide an undivided case of evaluation with our proposed framework for open-source MLLMs in Figure 12 and several individual questions in three categories (Appearance Quality, Intrinsic Attribute and Relationship Attribute) in Figure 13.

J Evaluation Prompt Templates

All prompt templates used in our proposed Task Decomposition Evaluation Framework are illustrated in Figure 14, 15 and 16.

 # Task Description
You are a powerful multi-modal evaluation assistant tasked with evaluating explanation texts for questions related to generated images.

 # Input Data
1. A question about a generated image. The explanation text should clarify the answer to this question.
2. An explanation text to be evaluated against the factual content of the image.
3. A reference explanation text, which correctly represents the image content and serves as the gold standard for evaluation.

 # Evaluation Guidelines
Assign a score from 0 to 5, where a higher score indicates better alignment with the reference explanation:
- 0: The evaluated explanation contradicts the reference, is empty, or lacks relevant information.
- 1-2: The evaluated explanation shows poor relevance to the reference, contains insufficient information, or has many errors.
- 3-4: The evaluated explanation generally aligns with the reference but may miss some details or contain minor errors.
- 5: The evaluated explanation fully aligns with the reference, potentially providing richer information with minimal or no errors.

 # Precautions
Focus on the factual content conveyed by the reference explanation. Ignore any statements such as 'the answer' or 'ground truth' if they appear.

 # Question
{question}

 # Explanation to be Evaluated
{gt_exp}

 # Reference Explanation
{ref_exp}

 # Output Instructions
Provide only one line as the output: the score as an integer value.

Do not include any additional information beyond the score.

Figure 10: Prompt template for subjective evaluation of fine-grained explanations.

Task Description
You are a powerful multi-modal evaluation assistant tasked with evaluating explanation texts for the quality of generated images.

Input Data
1. A list of questions about a generated image, reflecting multiple aspects of the image.
2. Ground truth answers and explanations for each question, strictly based on the image content, serving as reference for your evaluation.
3. Explanation to be evaluated, where you assess consistency with the reference and whether it fully covers the provided information.

Evaluation Guidelines
Assign a score from 0 to 5, where a higher score indicates better alignment with the reference explanation:
- 0: The evaluated explanation contradicts the reference, is empty, or lacks relevant information.
- 1-2: The evaluated explanation shows poor relevance to the reference, contains insufficient information, or has many errors.
- 3-4: The evaluated explanation generally aligns with the reference but may miss some details or contain minor errors.
- 5: The evaluated explanation fully aligns with the reference, potentially providing richer information with minimal or no errors.

Precautions
Focus on the factual content conveyed by the reference explanation. Ignore any statements such as 'the answer' or 'ground truth' if they appear.

Questions and Reference Explanation
{ref_exp}

Explanation to be Evaluated
{gt_exp}

Output Instructions
Provide only one line as the output: the score as an integer value.

Do not include any additional information beyond the score.

Figure 11: Prompt template for subjective evaluation of coarse-grained explanations.

 # Structure Information
Intrinsic Attributes
Man
- attribute 1: quantity: 1
- attribute 2: hat: bowler
- attribute 3: coat: lab
- attribute 4: existence: yes
TV frame
- attribute 1: quantity: 1
- attribute 2: existence: yes
Relationship Attributes
Appearance
- entities involved: Man, TV frame
- value: Man appears in TV frame

 # Image Caption
Man
- caption: The man is wearing a bowler hat and a lab coat.
TV frame
- caption: The TV frame displays an image of the man.

 # Questions
Appearance Quality Questions
Man
- question: Does the appearance of the man in the lab coat and bowler hat look realistic and align with human intuition?

- answer: The appearance of the man in the lab coat and bowler hat is somewhat realistic but has an artistic or surreal quality to it. The man's attire and the setting suggest a professional or scientific environment, but the interaction with the television and the presence of the other man within the TV screen adds an element of surrealism. The overall look aligns with human intuition to a degree but also pushes the boundaries of what is typically expected in a realistic scenario.

- score: 6.0

TV frame
- question: Does the appearance of the TV frame realistically integrate the man displayed on it?

- answer: The TV frame does not realistically integrate the man displayed on it. The man inside the TV appears as though he is superimposed rather than naturally existing within the TV frame, which is evident from the slight discrepancies in lighting and shadow alignment.

- score: 4.0

Intrinsic Attribute Consistency Questions
Man
- question: Does the man exist in the image?

- answer: Yes, the man exists in the image.

- explanation: The man is clearly visible in the image, standing next to the television set and interacting with it.
- score: 10.0

- question: Is the man wearing a bowler hat?

- answer: No, the man is wearing a black top hat, not a bowler hat.

- explanation: The man in the image is wearing a black top hat, which is different from a bowler hat. The top hat has a distinctively taller and more cylindrical shape, whereas the bowler hat has a more rounded and flat crown. This can be clearly observed in the image, confirming that the hat is a top hat and not a bowler hat.
- score: 3.0

- question: Is the man wearing a lab coat?

- answer: Yes, the man is wearing a white lab coat.

- explanation: The man is depicted wearing a white lab coat in the image, which is clearly visible and covers his torso and arms, consistent with the description provided.
- score: 10.0

A man in bowler hat and lab coat by people in a TV frame.

TV frame
- question: Does the TV frame exist in the image?

- answer: Yes, the TV frame exists in the image. It is a vintage model with a wooden frame and a screen that is turned on, displaying an image of a man in a white coat.

- explanation: The TV frame is clearly visible in the image, serving as the central object around which the scene is arranged. It is an old-fashioned model with a wooden frame and a screen that is turned on, displaying an image of a man in a white coat. The frame's presence is unmistakable and is a significant element in the image.
- score: 10.0

Relationship Attribute Consistency Questions
- question: Does the man appear within the TV frame?
- entities: Man, TV frame

- answer: No, the man does not appear within the TV frame. He is standing next to the TV, and the person appears within the TV screen.

- explanation: The man is clearly visible outside the TV, standing next to it, which indicates that he is not within the TV frame. The person appears inside the TV screen, creating an optical illusion.
- score: 0.0

 # Overall Evaluation
- Appearance Quality Summary
- explanation: The appearance quality of the man is somewhat realistic with a surreal element, while the TV frame does not integrate the displayed man realistically.
- score: 4.0
- Intrinsic Attribute Consistency Summary
- explanation: The intrinsic attributes of the man and the TV frame are mostly consistent with the ground truth, with the exception of the hat's type. The man's existence and lab coat, as well as the TV frame's existence, are accurately depicted. However, the hat's type is inconsistent with the ground truth, which affects the overall consistency score.
- score: 8.0
- Relationship Attribute Consistency Summary
- explanation: The relationship between the man and the TV frame is not consistent with the ground truth, as the man is not appearing within the TV frame as expected.
- score: 0.0
- Overall Score
- explanation: The overall evaluation shows a moderate level of consistency in appearance quality and intrinsic attributes, but a low level of consistency in relationship attributes. The appearance quality has a surreal element, and the relationship between the man and the TV frame is not consistent with the ground truth.
- score: 4.0

Figure 12: A case of the evaluation framework for open-source MLLMs.

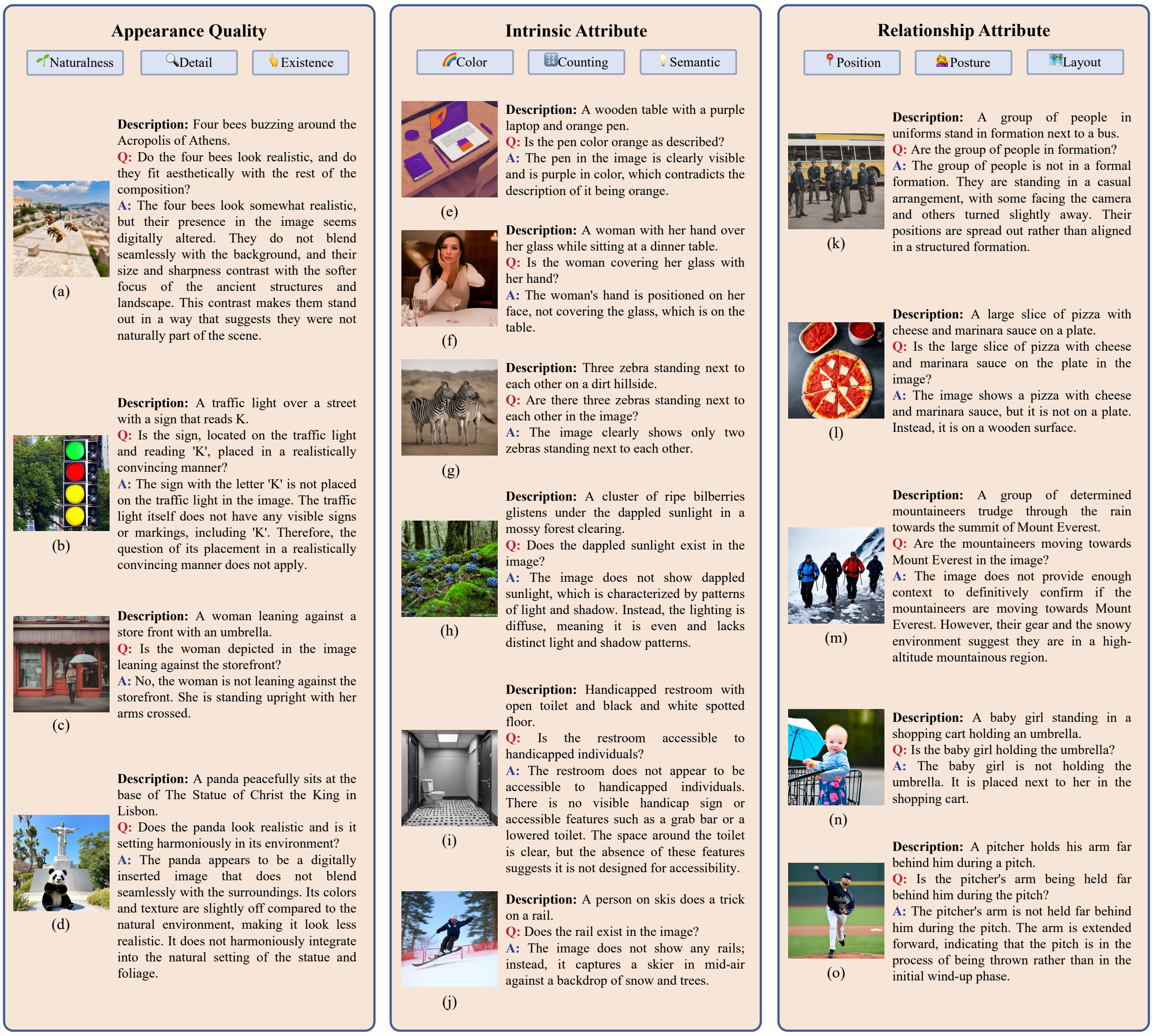


Figure 13: Cases for fine-grained evaluations in three categories.

 # Your Task
You are an expert in information extraction. Your task is to extract attributes of entities and relationships between entities from the text, and to pose a question about each entity's attributes and relationships.

 # Input Data
The text is: {text_prompt}

 # Extraction Pipeline
Step 1: Identify Entities
Step 1.1: Extract All Names
Extract all potential names from the input text.
Step 1.2: Evaluate Each Name
- Determine Entity Status: For each extracted name, assess whether it qualifies as an entity based on context and predefined criteria.
- Include or Exclude: If a name is deemed an entity, include it in the output; otherwise, exclude it.

Step 2: Formulate a Question for Each Entity
For each entity, create a critical question regarding the realism, aesthetic appeal, and alignment with human intuition of the entity's appearance in the generated image. Focus questions primarily on overall authenticity rather than getting into detailed specifics.

Step 3: Identify All Attributes for Each Entity
Step 3.1: Identify Intrinsic Attributes
Intrinsic attributes are properties explicitly mentioned in the input text, such as color, size, shape, material, and quantity.
Step 3.1.1: Extract Quantity Attributes
Identify words indicating quantity, including articles like "a" and "an", which suggest a quantity of one. For example, in "a cat", "a" indicates one cat. Attribute this quantity to the relevant entity.
Step 3.1.2: Extract Other Intrinsic Attributes
Analyze words in the input text related to the entity, excluding the entity's name itself. Determine if these words denote intrinsic attributes and identify their types (e.g., color, size, material) and values.
Step 3.1.3: Verify Attribute Type and Value Pair
Ignore attribute pairs if the value doesn't appear in the text, is part of the entity's name, or is "unspecified".
Step 3.1.4: Exclude Positional Attributes
Disregard attributes related to position, orientation, distance, or location.
Step 3.1.5: Add Existence Attribute
For each entity, add an attribute "existence" with a value of "yes" to indicate it should exist in the image.
Step 3.1.6: Default Unspecified Quantities
If the text doesn't specify a quantity, set it to "unspecified".
Step 3.1.7: Consolidate and Output Attributes
Add verified attribute type-value pairs to the output. Ensure all entities are included.
Step 3.2: Identify Relationship Attributes
Relationship attributes describe an entity's relationship with other entities.
Step 3.2.1: Analyze Relation Words
Identify words in the input text that describe relationships between entities, specifying the relationship type and related entities.
Step 3.2.2: Output Relationship Types
Add identified relationships and related entities to the output.

Step 4: Construct Questions Based on Extracted Attributes
Step 4.1: Construct Intrinsic Attribute Consistency Questions
Step 4.1.1: Existence Questions
Generate questions such as, "Does the [entity] exist in the image?" where [entity] is the entity's name.
Step 4.1.2: Attribute Value Questions
Create a question for each intrinsic attribute pair about the attribute value of the entity.
Step 4.1.3: Verify the Number of Questions
Ensure the number of questions equals the total number of intrinsic attribute-value pairs, including one existence and one quantity question for each entity.
Step 4.2: Construct Relationship Attribute Consistency Questions
Step 4.2.1: Relationship Questions
For each relational attribute of each entity, formulate a question about its value in relation to other entities.
Step 4.2.2: Ensure Coverage
Ensure the number of questions matches the number of relationship attribute pairs, with each pair corresponding to one question.

 # Output Template
Replace variables in '{{}}'
And if the text is like "Three apples", the entity should be "apple", and the attribute should be "three". Instead of "apple 1, apple 2, apple 3" as the entities.
Please generate your extracted structured information based on the following markdown template (Do NOT generate // comment in the template):

 # Structure Information
Intrinsic Attributes
{{entity}}
- attribute 1: {{attribute 1 type}}: {{attribute 1 value}}
- attribute 2: {{attribute 2 type}}: {{attribute 2 value}}
- attribute 3: attribute 3 type: attribute 3 value
...
{{next entity or group}}
...

Relationship Attributes
{{relationship attribute 1}}
- entities involved: {{entity 1, entity 2, ...}}
- value: {{relationship attribute value}}
{{next relationship attribute}}
...

 # Questions
Appearance Quality Questions
{{entity 1 name}}
- question: {{entity 1 appearance quality question }}
{{next entity}}
...

Intrinsic Attribute Consistency Questions
{{entity 1 name}}
- question 1: {{entity 1 intrinsic attribute consistency question 1}}
- question 2: {{entity 1 intrinsic attribute consistency question 2}}
- question 3: {{entity 1 intrinsic attribute consistency question 3}}
- question 4: {{entity 1 intrinsic attribute consistency question 4}}
- next question
...
{{next entity}}
...

Relationship Attribute Consistency Questions
- question 1: {{relationship attribute consistency question 1}}
  - entities: {{entity 1}} {{entity 2}}
- question 2: {{relationship attribute consistency question 2}}
...

Figure 14: Prompt template for evaluation content extraction.

 # Your Task
You are an assistant specialized in answering questions based on the content of images.

 # Input Data
1. Question Input: These are the questions you are to answer. They consist of three parts: appearance quality questions, intrinsic attribute consistency questions, and relationship attribute consistency questions. The questions are: {questions}
2. Target Image: Use this image to answer the questions.
3. Reference Image: Use this as a reference for authenticity when answering questions about appearance quality based on the target image.

 # Answer Pipeline
Step 1: Generate the Target Image Caption
- Identify all entities in the target image.
- For each entity, generate a caption that includes the entity's name and all attributes.
- Generate a caption for each entity that includes its name and all relationships.
These captions are solely for answering the intrinsic attribute consistency questions. If an entity in the image caption does not have those questions, ignore it.

Step 2: Answer the Appearance Quality Questions
- For each question, identify if the entity is present in the target image. If present, proceed to the next step; if absent, assign a score of 0.
- For each appearance quality question, determine if the entity's appearance in the target image is realistic, aesthetically pleasing, and aligns with human intuition.
- Use the reference image for authenticity when needed.
- Assign a score from 0 to 10 for each question, and provide a brief explanation for the score awarded.
- **Scoring Strategy:**
    - 0-3: The appearance lacks realism, is not aesthetically pleasing, or does not align with human intuition.
    - 4-7: The appearance is somewhat realistic, aesthetically pleasing, or aligns with human intuition.
    - 8-10: The appearance is very realistic, aesthetically pleasing, and aligns well with human intuition.

Step 3: Answer the Intrinsic Attribute Consistency Questions
- For each question, check if the entity exists in the target image. If it does, proceed; if not, state that the entity doesn't exist in the image.
- Answer each intrinsic attribute consistency question by detailing the corresponding attribute value from both the target image and its caption. Be thorough in your explanations; avoid simple yes or no answers.

Note: You must address all questions in the question input.

Step 4: Answer the Relationship Attribute Consistency Questions
- For each question, verify the entity's presence in the target image. If present, continue; otherwise, indicate that the entity does not exist in the image.
- Determine the relationships of each entity in the target image and its caption. Provide a detailed answer, avoiding yes or no responses, and explain your reasoning.

 # Output Template
Replace variables in '{{}}'
Please generate your result based on following markdown template (Do NOT generate // comment in the template).

 # Image Caption
{{entity 1 name}}
- caption: {{entity 1 caption}}
{{next entity}}
...

 # Answers
Appearance Quality Questions
{{entity 1 name}}
- question: {{entity 1 appearance quality question}}
    - explanation: {{explanation}}
    - score: {{score}}
{{next entity}}
...

Intrinsic Attribute Consistency Questions
{{entity 1 name}}
- question 1: {{entity 1 intrinsic attribute consistency question 1}}
    - answer: {{answer}}
- question 2: {{entity 1 intrinsic attribute consistency question 2}}
    - answer: {{answer}}
- next question
...
{{next entity}}
...

Relationship Attribute Consistency Questions
- question 1: {{relationship attribute consistency question 1}}
    - entities: {{entity 1}}, {{entity 2}}
    - answer: {{answer}}
- question 2: {{relationship attribute consistency question 2}}
    - entities: {{entity 1}}, {{entity 2}}
    - answer: {{answer}}
...

Figure 15: Prompt template for caption and answer generation.

Your Task
You are an expert in assessing the similarity between answers obtained from images and ground truth obtained from text.

Input Data
1. Answers from the Image: These are the answers you need to evaluate including three components:
- Appearance Quality Answers
- Intrinsic Attribute Consistency Answers
- Relationship Attribute Consistency Answers

The provided answer is: {answer}

2. Ground Truth: This is the standard to which you will compare the image answers. It consists of two components:
- Entities' Attributes
- Relationships

The structured information is the sole ground truth: {structure_info}

Scoring Strategy
- 0-3: The answer is not consistent with the ground truth at all.
- 4-7: The answer is somewhat consistent with the ground truth; semantics are similar but not entirely aligned.
- 8-10: The answer is very consistent with the ground truth.

Evaluation Pipeline
Step 1: Evaluate Appearance Quality Answers
- Focus solely on the appearance quality of the answers.

Step 2: Evaluate Intrinsic Attribute Consistency Answers
- For each intrinsic attribute consistency answer of every entity, compare it with the corresponding ground truth.
- If the entity does not appear in the image, assign a score of 0. Otherwise, proceed to the next step.
- Offer a short explanation of how well the answer matches the ground truth.
- Provide a score reflecting the extent of the match; if there is no match, score it as zero. In cases of mismatch, assign the lowest possible score.

Step 3: Evaluate Relationship Attribute Consistency Answers
- For each relationship's attribute consistency answer, compare it with the ground truth.
- If the entity does not exist in the image, assign a score of 0. Otherwise, proceed to the next step.
- Offer a short explanation of how well the answer matches the ground truth.
- Provide a score reflecting the extent of the match; if there is no match, score it as zero. In cases of mismatch, assign the lowest possible score.

Step 4: Overall Evaluation
- Combine your findings on appearance quality, summarize your observations, and assign a score based on this summary.
- Summarize the degree of match between the image answers and the intrinsic attribute consistency of the ground truth, and assign a score based on this evaluation.
- Summarize the degree of match for relationship attribute consistency between the image answers and the ground truth, and assign a score based on this summary.
- Integrate all summaries regarding appearance quality, intrinsic attribute consistency, and relationship attribute consistency. Offer a comprehensive evaluation description and assign a final score based on this description.

Output Template
Replace Variable in '{{}}'
Please generate your output based on following markdown template (Do NOT generate // comment in the template).

```
# Evaluation
## Appearance Quality Answers
### {{entity 1 name}}
- question: {{entity 1 appearance quality question}}
  - explanation: {{explanation}}
  - score: {{score}}
### {{next entity}}
...

## Intrinsic Attribute Consistency Answers
### {{entity 1 name}}
- question 1: {{entity 1 intrinsic attribute consistency question 1}}
  - answer: {{answer from the image}}
  - explanation: {{explanation}}
  - score: {{score}}
- question 2: {{entity 1 intrinsic attribute consistency question 2}}
  - answer: {{answer from the image}}
  - explanation: {{explanation}}
  - score: {{score}}
- next question
...
### {{next entity}}
...

## Relationship Attribute Consistency Answers
- question 1: {{relationship attribute consistency question 1}}
  - entities: {{entity 1}} {{entity 2}}
  - answer: {{answer from the image}}
  - explanation: {{explanation}}
  - score: {{score}}
- question 2: {{relationship attribute consistency question 2}}
...

## Overall Evaluation
- Appearance Quality Summary:
  - explanation: {{explanation}}
  - score: {{score}}
- Intrinsic Attribute Consistency Summary:
  - explanation: {{explanation}}
  - score: {{score}}
- Relationship Attribute Consistency Summary:
  - explanation: {{explanation}}
  - score: {{score}}
- Overall Score:
  - explanation: {{explanation}}
  - score: {{score}}
```

Figure 16: Prompt template for explanation and scoring.