EvalMuse-40K: A Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Alignment Evaluation

Shuhao Han[1,2], Haotian Fan[2], Jiachen Fu[1], Liang Li[2], Tao Li[2], Junhui Cui[2], Yunqiu Wang[2], Yang Tai[2], Jingwei Sun[2], Chunle Guo[1,3]*, Chongyi Li[1,3]

[1]VCIP, CS, Nankai University
[2]ByteDance Inc
[3]NKIARI, Shenzhen Futian

hansh@mail.nankai.edu.cn, fujiachen2005@gmail.com, {guochunle, lichongyi}@nankai.edu.cn,
{fanhaotian, liliang.58, litao.walker, cuijunhui, wangyunqiu.qqq, taiyang, sunjingwei.jerry}bytedance.com

Abstract

Text-to-Image (T2I) generation models have achieved significant advancements. Correspondingly, many automated methods emerge to evaluate the image-text alignment capabilities of generative models. However, the performance comparison among these automated methods is constrained by the limited scale of existing datasets. Additionally, existing datasets lack the capacity to assess the performance of automated methods at a fine-grained level. In this study, we contribute an **EvalMuse-40K** dataset, gathering 40K image-text pairs with fine-grained human annotations for image-text alignment-related tasks. In the construction process, we employ various strategies such as balanced prompt sampling and data re-annotation to ensure the diversity and reliability of our dataset. This allows us to comprehensively evaluate the performance of image-text alignment methods for T2I models. Based on this dataset, we introduce an efficient automated evaluation method termed **FGA-BLIP2**, which enables **F**ine-**G**rained **A**lignment evaluation solely by inputting images and text leveraging **BLIP2**, without visual question answering for each fine-grained element. Experimental results show the proposed FGA-BLIP2 efficiently achieves good performance on multiple image-text alignment datasets. Meanwhile, benefiting from the high efficiency and fine-grained evaluation capability of FGA-BLIP2, we apply it as a reward model to improve text-to-image models, which effectively enhances the image-text alignment ability of text-to-image models.

Code — https://github.com/DYEvaLab/EvalMuse

Introduction

Recently, advanced Text-to-Image (T2I) models (Li et al. 2024d; Peebles and Xie 2023; Esser et al. 2024; Podell et al. 2024; Rombach et al. 2022; Saharia et al. 2022; Yu et al. 2022) are capable of generating numerous impressive images. However, these models may still generate images that fail to accurately match the input text, such as inconsistency in quantities (Kirstain et al. 2023; Xu et al. 2023; Wu et al. 2023). Given the high cost and inefficiency of manual evaluation, developing a reliable automatic evaluation metric and corresponding benchmark is vital. They can effectively evaluate the performance of existing models and provide guidance for improvements in future models.

Since traditional evaluation metrics, such as FID (Heusel et al. 2017) and CLIPScore (Hessel et al. 2021), are not well-suited for assessing the consistency of T2I models. Recent works (Kirstain et al. 2023; Hu et al. 2023; Wiles et al. 2024) have explored the construction of new evaluation metrics. PickScore (Kirstain et al. 2023) fine-tunes CLIP with annotated preference data to enhance image-text alignment evaluation. VQAScore (Li et al. 2024b) queries VQA models to obtain alignment scores based on logit values. However, these methods are failed when we need to evaluate whether a specific fine-grained element within the text is accurately depicted in the corresponding image. TIFA (Hu et al. 2023), VQ2 (Yarom et al. 2023) and Gecko (Wiles et al. 2024) decompose the prompt into multiple elements, formulate related questions for each, and then average the answers to generate a final alignment score. These methods allow for fine-grained evaluation. However, due to the lack of fine-grained annotation data, the performance comparison of these methods is still conducted at the overall level of image-text pairs, with little consideration of accuracy at the element level. Therefore, to better explore the performance of existing T2I evaluation methods, we contribute a new benchmark, **EvalMuse-40K**, featuring fine-grained human annotations of image-text pairs.

EvalMuse-40K includes 4K prompts, 40K image-text pairs, and more than 1M fine-grained human annotations. To ensure the diversity of prompts, EvalMuse-40K includes 2K real prompts and 2K synthetic prompts. To evaluate various skills, current benchmarks typically adopt relatively simple prompt designs when assessing specific skills, such as "a photo of three bags" for counting skill evaluation. For prompts that involve multiple skills, categories like "complex" are usually used for representation. With the development of T2I models, they have achieved good performance on simple single-skill prompts. Therefore, we aim to explore the performance of generative models in terms of each skill when using prompts that contain multiple skills. When prompts have multiple categories, simple downsampling of majority-class samples is insufficient to achieve sample balance. So, we propose a MILP-based sampling

*Corresponding author.

Benchmark	Dataset Size			Alignment			Faithfulness	T2I Models
	Prompt	Image	Ann.	Likert	Ele.-Ann.	Ele.-Category	Structure	Num.
PartiPrompt (Yu et al. 2022)	1.6K	-	-	✗	✗	✗	✗	-
DrawBench (Saharia et al. 2022)	200	-	-	✗	✗	✗	✗	-
T2I-CompBench (Huang et al. 2023)	6K	-	-	✗	✗	✓	✗	-
TIFA160 (Hu et al. 2023)	160	800	1.6K	✓	✗	✓	✗	5
GenAI-Bench (Li et al. 2024a)	1.6K	9.6K	28.8K	✓	✗	✗	✗	6
Gecko (Wiles et al. 2024)	2K	8K	108K	✓	✓	✗	✗	4
EVALALIGN (Tan et al. 2024)	3K	21K	132K	✓	✗	✓	✓	8
EvalMuse-40K (ours)	**4K**	**40K**	**1M**	✓	✓	✓	✓	**20**

Table 1: In comparison to existing T2I model evaluation benchmarks, EvalMuse-40K collects a large number of human annotations (Ann.). Furthermore, EvalMuse-40K offers fine-grained annotations at the element (Ele.) level and categorizes these elements into different skills in image-text alignment. Additionally, EvalMuse-40K includes annotations for structural problems in generated images. To ensure reliable evaluation of automated metrics, we also randomly generate image-text pairs from 20 different T2I models.

method (Vonikakis, Subramanian, and Winkler 2016) that minimizes the difference between the sampled distribution and a uniform distribution, ensuring a more balanced distribution of prompts after sampling. This strategy allows sampling prompts with multiple dimensions and categories. Using this method, we sampled 2,000 real prompts. And the synthesized prompts are then constructed for specific skills in image-text alignment, such as quantity and location. By synthesizing specific prompts and sampling from real prompts, EvalMuse-40K can be used to compare the evaluation capabilities of models on specific skills, while also providing a more comprehensive comparison of evaluation model in real-world scenarios. For fine-grained evaluation, we use a large language model for the elemental splitting of prompts and question generation, and to increase the diversity of the generated images, we generate images using a variety of T2I models. Compared with the previous benchmark (see Tab. 1 for details), EvalMuse-40K not only scores image-text alignment but also performs more fine-grained annotations for elements split from the prompts. Using EvalMuse-40K, we evaluated the performance of current alignment evaluation methods.

For fine-grained evaluation of image-text alignment, most of the current methods are implemented using Multi-modal Large Language Models (MLLMs) with visual question-answering for each element. This makes it very redundant that each image-text pair needs to be evaluated multiple times to achieve a complete evaluation. Therefore to achieve efficient fine-grained evaluation, we introduce a new evaluation method termed FGA-BLIP2, which uses a visual language pre-training model to jointly fine-tune image-text alignment scores and element-level annotations. Given a single image-text pair, FGA-BLIP2 enables both overall and fine-grained image-text alignment evaluation at once without the need for additional visual question-answering. Benefiting from the efficiency and fine-grained evaluation capability of FGA-BLIP2, we utilize it as a reward model to fine-tune the T2I model, thereby improving the image-text alignment ability of the T2I model. Our data, models, code, and results serve as valuable resources to support further research.

To summarize, our contributions are listed as follows.

- EvalMuse-40K adopts a MILP-based sampling strategy to ensure the balance and diversity of multi-category complex prompts.
- EvalMuse-40K collects a large-scale dataset with categorical fine-grained annotations, enabling a more comprehensive evaluation of existing fine-grained alignment evaluation methods.
- We propose an efficient fine-grained alignment evaluation method termed FGA-BLIP2, which can perform both overall and fine-grained image-text alignment evaluation at once.
- Experiments show FGA-BLIP2 is efficient and performs well in image-text alignment evaluation, while offering better guidance for T2I model improvement.

Related Work

Image-Text Alignment Benchmarks. Many different benchmarks have been proposed to evaluate the image-text alignment of T2I models. Early benchmarks are small-scale and mostly rely on captions from existing datasets like COCO (Cho, Zala, and Bansal 2023; Hu et al. 2023; Lin et al. 2014; Ramesh et al. 2022), focusing on limited sample skills. Other benchmarks (e.g., HPDv2 (Wu et al. 2023) and Pick-a-pic (Kirstain et al. 2023)) use side-by-side model comparisons to evaluate the quality of the generated images. Recently, benchmarks like DrawBench (Saharia et al. 2022), PartiPrompt (Yu et al. 2022), and T2I-CompBench (Huang et al. 2023) have introduced a set of prompts and focused on evaluating specific skills of generative models, including counting, spatial relationships, and attribute binding. Moreover, some benchmarks (e.g., GenAI-Bench (Li et al. 2024a) and RichHF-18K (Liang et al. 2024)) provide human annotations on image-text align scores to validate the relevance of automated metrics with human preference. For fine-grained evaluation, benchmarks like TIFA (Hu et al. 2023), SeeTRUE (Yarom et al. 2023) and DSG (Cho et al. 2024) extract elements from prompts and generate corresponding questions. However, prompts in current benchmarks are mostly designed for specific skills or are simply sampled from multiple prompt sets, which leads to an imbalanced distribution. Moreover, when comparing fine-

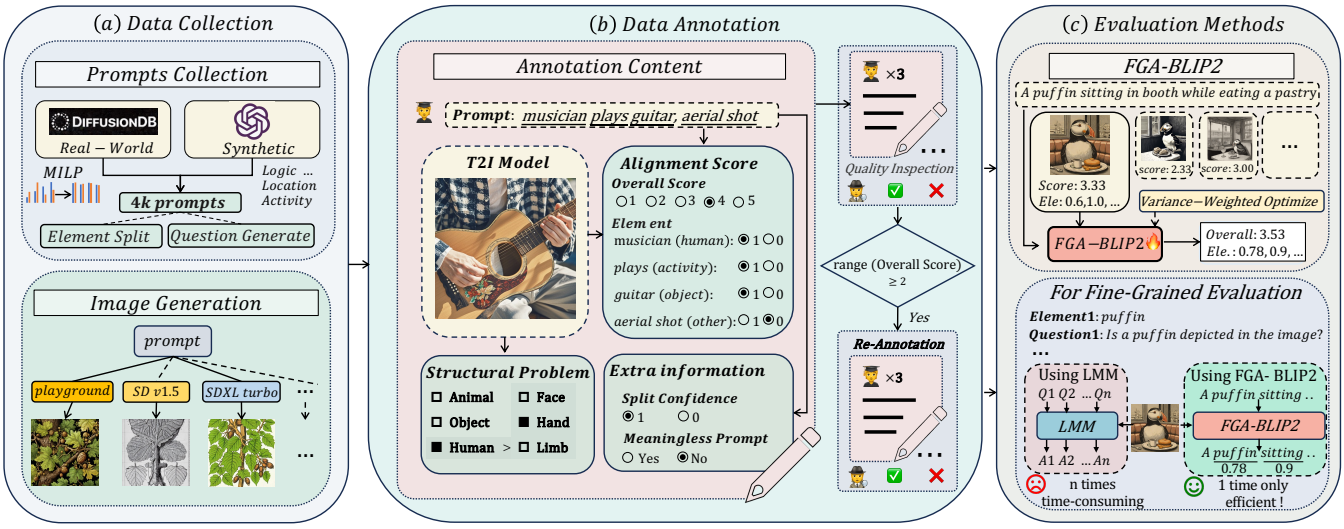

Figure 1: Overview construction of EvalMuse-40K, consisting of (a) data collection, (b) data annotation, and (c) evaluation methods. (a) We collect 2K real-world prompts and 2K synthetic prompts, using the MILP (Vonikakis, Subramanian, and Winkler 2016) sampling strategy to ensure category balance and diversity. Prompts are further divided into elements, and corresponding questions are generated. We also use various T2I models to generate images. (b) During data annotation, we label the fine-grained alignment levels of image-text pairs, structural problems of the generated images, and some extra information. For annotated data with large differences in overall alignment scores, we perform re-annotation to ensure the reliability of the benchmark. (c) Leveraging annotated data, we propose FGA-BLIP2 for fine-grained alignment evaluation, which is more efficient compared to using Large Multimodal Models (LMMs) for visual question answering.

grained evaluation methods, these benchmarks still average the scores into a single overall score for comparison, lacking a detailed assessment of fine-grained evaluation capabilities. EvalMuse-40K adopts a new prompt collection process to ensure diversity and balance. Additionally, it provides large-scale human annotations on both overall and fine-grained aspects of image-text alignment, allowing deeper analysis of how well evaluation metrics align with human preference.

Automated Metrics for Image-Text Alignment. Early metrics such as FID (Heusel et al. 2017), IS (Salimans et al. 2016), and LPIPS (Zhang et al. 2018) focus on image quality or distribution similarity but fail to assess image-text alignment. Recent works primarily adopt CLIP-Score (Hessel et al. 2021) (measuring text-image feature cosine similarity) and BLIP2Score (leveraging BLIP2's (Li et al. 2023b) strong performance) as alignment metrics. Human preference models (e.g., ImageReward (Xu et al. 2023), PickScore (Kirstain et al. 2023), HPSv2 (Wu et al. 2023)) fine-tune vision-language models like CLIP on large-scale human ratings, but their reliance on side-by-side image comparisons hinders accurate alignment scoring and fine-grained evaluation. For fine-grained tasks, TIFA (Hu et al. 2023), VQ2 (Yarom et al. 2023), and Gecko (Wiles et al. 2024) split prompts into elements and use VQA models for evaluation, which increases the evaluation time of image-text pairs.. To address these limitations, we directly predict both overall and fine-grained alignment scores via a vision-language model, achieving one-step output of the overall and fine-grained element scores.

EvalMuse-40K Benchmark

In this section, we detail EvalMuse-40K, a reliable and fine-grained benchmark with comprehensive human annotations for T2I evaluation. We present the overview construction of EvalMuse-40K in Fig. 1. In the construction process, we employ various strategies to ensure the diversity and reliability of our benchmark. Next, we introduce the construction process of EvalMuse-40K from two aspects: data construction and data annotation followed by a statistical analysis.

Data Construction

To better evaluate the T2I task, we collect 2K real prompts for broader category coverage and 2K synthetic prompts for specific skills. Below, we outline the process of collecting real and synthetic prompts, generating images, and performing element splitting and question generation.

Real Prompts Collection. Prompts in current benchmark typically belong to a single category, but real user prompts are often more complex, involving multiple attributes such as quantity and color. However, existing benchmarks usually label such complex prompts uniformly with tags like "complex", resulting in an unbalanced distribution of these complex prompts. Prompts in current benchmarks (Hu et al. 2023; Cho et al. 2024) are typically imbalanced across categories, even when Gecko (Wiles et al. 2024) assigns higher sampling weights to under-represented skills during resampling. To address the above issues, we collect a large number of prompts to ensure sufficient quantities across different categories and design a comprehensive categorization system. Each prompt is classified across three dimensions: sub-

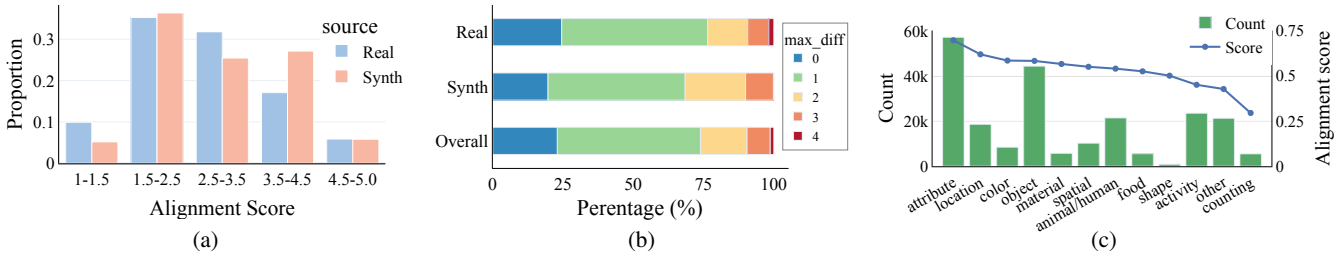

Figure 2: Statistical Charts. (a) Distribution of annotated alignment scores in real prompt samples and synthetic prompt samples. (b) Distribution of maximum score differences between different annotators for the same image-text pair. (c) Distribution of element counts and scores across different skills in fine-grained annotations.

ject category, logical relationship, and image style. All dimensions allow multiple labels, with logical relationship and image style including a "None" label. To better capture the semantic characteristics of prompts, we added a fourth dimension by calculating BERT (Kenton and Toutanova 2019) embeddings and clustering prompts based on their semantic features. Then using the MILP strategy to sample the data ensures uniformity of each category across the four dimensions.

Specifically, we first randomly sample 100K prompts from DiffusionDB (Wang et al. 2023b), and then use GPT-4 (Achiam et al. 2023) to label each prompt across the three dimensions mentioned above. The three dimensions cover 24 categories, including most common types. Specific classifications and examples are in the supplementary materials. We then calculate the semantic features using BERT and cluster them into seven categories using the K-means algorithm. Finally, we employ the MILP strategy to sample 2,000 prompts according to a specific distribution $D = \{D_m\}_{m=1}^{M}$, where $D_m \in \mathbb{R}^{H_m}$ represents the distribution ratio of each category in the m-th dimension of the sampled data and H_m denotes the total number of categories in dimension m. Our optimization goal is to set $D_m = R_m/H_m$,, where R_m represents the average number of categories per prompt in dimension m. Compared to (Vonikakis, Subramanian, and Winkler 2016), by setting R_m, we allow each dimension to have multiple category labels. The specific sampling strategy and the distribution of the sampled data are shown in the supplementary material.

Synthetic Prompts for Various Skills. To evaluate generative models comprehensively, we use GPT-4 with specific templates and a rich corpus to generate 2K synthetic prompts, divided into six categories: Object Count; Object Color and Material; Environment and Time Setting; Object Activity and Perspective Attributes; Text Rendering; and Spatial Composition Attributes, with different templates for each. Details are in the supplementary material.

Images Generation. To ensure image diversity, we select 20 diffusion-based generative models (given diffusion models' strong performance), classified into four groups: basic stable diffusion models; advanced open-source models; efficient models; and proprietary models. For each prompt, a subset of models is randomly sampled for image generation with default parameters, resulting in 40K image-text pairs from 4K prompts, ensuring a diverse dataset for annotation.

Element Splitting and Question Generation. For fine-grained annotation and evaluation, we split 4K collected prompts into fine-grained elements (unlike word-level annotation in RichHF and Gecko). In addition, categorizing each element allows us to examine the model's capabilities in various aspects at a fine-grained level. We then generate yes/no judgment questions, enhancing controllability and interpretability. Generated questions are filtered to ensure each element has a corresponding one, with templates in the supplementary material.

Data Annotation

In this section, we describe how we perform the data annotation. We first define the content and templates of the annotations and then detail the entire annotation process.

Annotation Content and Templates. The annotation includes image-text alignment, structural problems, and extra information (see Fig. 1(b)). In terms of image-text alignment, annotators score the alignment using a 5-point Likert scale and label whether fine-grained elements in the prompt are aligned with the image. Structural issues in generated images are marked and categorized into humans, animals, and objects, with human figures further subdivided by regions like face, hands, and limbs. To address potential inaccuracies, we introduce a splitting confidence label to flag incorrect element splitting by GPT-4. Moreover, we add a label to indicate whether a prompt is meaningful, especially for prompts originating from real users that may contain unclear meanings.

Annotation Process. To improve annotation quality, our process has three stages. *1) Pre-annotation:* We train annotators using clear standards and a small pre-annotation dataset, then refine evaluation criteria based on review feedback. *2) Formal annotation:* During formal annotation, each image-text pair is labeled by three annotators, with an additional annotator assigned for quality control. Annotators also identify and flag any NSFW content in the generated images. *3) Re-annotation:* For image-text pairs where alignment scores from the three annotators show significant discrepancies (range ≥ 2), we conduct re-annotation to reduce subjective bias. Ultimately, each pair is annotated by 3 to 6 annotators, and the average score is used as the final label.

Data Statistics and Reliability Analysis

We statistics the image-text alignment scores from annotations and the histogram of the alignment scores is shown in Fig. 2a. The alignment scores are distributed across all ranges, with a higher concentration in the middle range. This distribution provides a sufficient number of both positive and negative samples, enabling a robust evaluation of the consistency between existing image alignment metrics and human preferences and facilitating the training of a scoring model aligned with human preferences.

To analyze annotator agreement on scoring data, we calculate the maximum score difference for each image-text pair and plot it as a histogram in Fig. 2b. 75% samples show a score difference < 2. For score differences ≥ 2, we obtain double annotations (from 3 to 6) by re-annotating, further reducing the inter-annotator disagreement.

For fine-grained annotation ,we conduct statistics on the quantity and alignment scores of elements by category. As shown in Fig. 2c, the overall alignment scores for most categories are around 50%, ensuring a balanced distribution of positive and negative samples. Additionally, the images generated by T2I models exhibit relatively poor consistency with the text in aspects of counting, shape, and activity.

Methods for Alignment Evaluation

Existing fine-grained image-text alignment evaluation methods predominantly rely on element-wise prompt decomposition and question generation, requiring multiple rounds of visual question answering (VQA) via Multi-Modal Large Language Models (MLLMs) for each element in the image-text pair. This iterative VQA process leads to substantial computational overhead when conducting comprehensive fine-grained evaluation for a single image-text pair. To address this inefficiency, we propose a novel fine-grained evaluation method termed **FGA-BLIP2**, which enables one-step inference to simultaneously output both the overall alignment score of the image-text pair and the fine-grained alignment scores for each token in the text.

A straightforward approach to achieving fine-grained evaluation using **V**ision-**L**anguage pre-training **M**odels (VLMs) is to adopt the Image-Text Contrastive (ITC) learning setup from models like CLIP or BLIP2. After aligning image and text features, cosine similarity calculations between image features and features of specific tokens from text and are directly used to obtain fine-grained alignment scores. However, this method fails to fuse image and text features, making it difficult to output more accurate alignment scores. Thus, in our FGA-BLIP2, we leverage cross-attention operation to fuse image information into text features, ensuring that the text features are enriched with visual details from the image. Based on the positions of elements in the text, we extract the features of specific elements. By feeding these position-specific, image-fused elements features into the classification head, we can accurately output fine-grained alignment scores for each specific element, thereby achieving fine-grained evaluation.

Specifically,, we adopt the setting of Image-Text Matching (ITM) in BLIP2 (Li et al. 2023b), where the query and

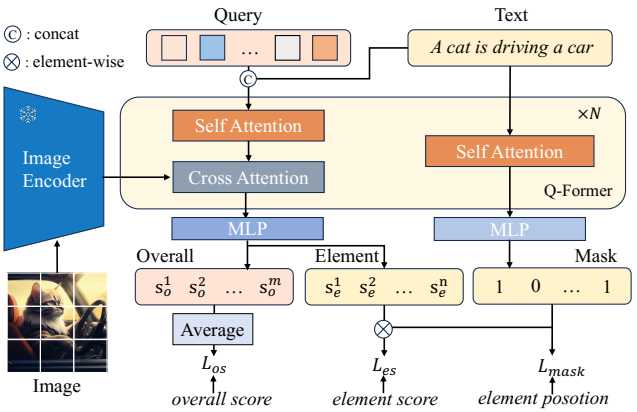

Figure 3: Structure of FGA-BLIP2. m and n denote query and text token lengths, respectively; s_o and s_e represent overall and element scores, respectively; L_{os} represents the overall alignment score loss, L_{es} is the element alignment score loss, and L_{mask} is the loss for predicting valid elements. FGA-BLIP2 jointly optimizes these losses to achieve fine-grained evaluation.

text after embedding are concatenated and then employed self-attention to aggregate the information. After that, these embeddings are processed through cross-attention with the image and yields query embedding and text embedding outputs. The final alignment score is obtained by a two-class linear classifier, where the query part is averaged to produce the overall alignment score, while the text part at each corresponding position provides the element-specific alignment scores. Since not all tokens in the text are highly relevant to the alignment task, we introduce an additional operation to predict valid tokens. The text is first fed to a self-attention layer to extract text features, which then are passed to an MLP to predict a mask that represents the validity of each text token. **In summary**, when given an image-text pair as input, FGA-BLIP2 not only outputs an **overall alignment score** (os) but also provides **element-level alignment scores** (es) and generates a **mask** to judge the validity of text tokens.

We observe that some prompts in the dataset are either too simple or overly complex, resulting in minimal differences in alignment scores across images generated by different models. Such data may lead the evaluation model to focus more on prompt complexity than on the actual alignment level of the image-text pairs during training. We therefore design a variance-weighted optimization strategy for the image-text alignment task. We calculate the variance $\sigma(p)$ of the alignment scores across different images generated using the same prompt and use this to adjust the loss weights of different prompts during training.

The final loss objective function is as follows:

$$L_{total} = e^{\sigma(p)} \cdot (L_{os} + \lambda L_{es} + \eta L_{mask}), \qquad (1)$$

where weighting parameters are set to $\lambda = 1$ and $\eta = 0.1$. L_{os} denotes the L1 loss between the predicted overall alignment score and human annotation. L_{es} denotes the L1 loss

Method	EvalMuse-40K (ours)		TIFA		AIGCIQA		AGIQA3K		CompBench	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
♢ CLIPScore (Hessel et al. 2021)	0.2993	0.2933	0.3003	0.3086	0.2337	0.6839	0.5972	0.6839	0.2044	0.1944
♢ BLIPv2Score (Hessel et al. 2021)	0.3583	0.3348	0.4287	0.4543	0.3784	0.2576	0.6230	0.7380	0.3967	0.3940
♢ ImageReward (Xu et al. 2023)	0.4655	0.4585	0.6211	0.6336	0.5870	0.5911	0.7298	0.7298	0.4367	0.4307
♢ PickScore (Kirstain et al. 2023)	0.4399	0.4328	0.4279	0.4342	0.5045	0.4998	0.6977	0.7633	0.1115	0.0955
♢ HPSv2 (Wu et al. 2023)	0.3745	0.3657	0.3647	0.3804	0.6068	0.5989	0.6349	0.7000	0.2844	0.2761
♢ VQAScore (Li et al. 2024b)	0.4877	0.4841	0.6951	0.6585	0.6394	0.5869	0.6273	0.6677	0.5832	0.5328
♣ InternVL2.5 (1B) (Chen et al. 2024)	0.3489	0.3456	0.5765	0.5681	0.5383	0.5308	0.5568	0.5858	0.4960	0.4756
♣ InternVL2.5 (2B) (Chen et al. 2024)	0.4392	0.4347	0.6210	0.6262	0.5346	0.5307	0.5480	0.6159	0.5194	0.5011
♣ InternVL2.5 (4B) (Chen et al. 2024)	0.4249	0.4347	0.6746	0.6744	0.5587	0.5628	0.6797	0.7232	0.4004	0.4554
♣ InternVL2.5 (8B) (Chen et al. 2024)	0.5057	0.4659	0.7141	0.7012	0.5682	0.5655	0.7171	0.7885	0.5034	0.5196
♣ LLaVA-NeXT (8B) (Li et al. 2024c)	0.4554	0.4732	0.6045	0.6244	0.5046	0.5254	0.6797	0.6877	0.5001	0.5333
♣ InternVL2 (8B) (Chen et al. 2024)	0.4165	0.4014	0.5717	0.5817	0.4681	0.4309	0.6332	0.7006	0.4960	0.5272
♣ Qwen2-VL (7B) (Wang et al. 2024)	0.4715	0.4417	0.6391	0.6672	0.6263	0.5646	0.6920	0.7815	0.4776	0.4872
♣ InternVL2.5* (1B) (Chen et al. 2024)	0.7679	0.7588	0.7296	0.7154	0.6590	0.6474	0.7338	0.8024	0.5713	0.5545
♣ InternVL2* (2B) (Chen et al. 2024)	0.7591	0.7507	0.7109	0.7034	0.6227	0.6166	0.7538	0.8427	0.5611	0.5641
♣ InternVL2.5* (2B) (Chen et al. 2024)	**0.7712**	0.7594	0.7285	0.6810	**0.6840**	**0.6736**	0.7651	**0.8429**	0.5757	0.5622
♣ Qwen2.5-VL* (3B) (Bai et al. 2025)	0.7710	**0.7624**	**0.7311**	**0.7273**	0.6541	0.6568	**0.7706**	0.8427	**0.5995**	**0.5813**
FGA-BLIP2 (1B) (Ours)	**0.7789**	**0.7727**	**0.7577**	**0.7523**	**0.7257**	**0.7172**	**0.7713**	**0.8668**	**0.6164**	**0.6063**

Table 2: Quantitative comparison between our methods and the state-of-the-art methods which output overall alignment score on multiple benchmarks. ♢ vision-language pre-training models, ♣ LMM-based models. *Refers to finetuned models.

between the predicted element score and human fine-grained annotation. L_{mask} denotes the L1 loss between the predicted valid text token and the real elements.

To our knowledge, few existing Vision-Language Pre-trained Models (VLMs) are used for fine-grained alignment evaluation, as they typically only consider overall prompt-image consistency during pre-training, neglecting the consistency of fine-grained elements within prompts. Benefiting from the extensive fine-grained annotation data in EvalMuse-40K, our proposed FGA-BLIP2 fine-tunes existing VLMs, significantly enhancing their fine-grained evaluation capabilities. Meanwhile, compared to LMM-based methods, FGA-BLIP2 eliminates the need for element-wise decomposition of prompts and multiple rounds of visual question answering, thereby improving the efficiency of fine-grained evaluation. Additionally, FGA-BLIP2 achieves favorable performance when compared to LMM-based methods with the same scale of parameters.

Experiments

Experimental Setup

Datasets. We use one-quarter of the samples from EvalMuse-40K as test set, ensuring no overlap in prompts between the training and test sets. The test set includes 500 real prompts and 500 synthetic prompts. We train FGA-BLIP2 and other methods on the training set and then test all evaluation methods on the test set.

Training Settings. We initialize FGA-BLIP2 using BLIP-2 (Li et al. 2023b) fine-tuned on COCO (Lin et al. 2014). We train it for 5 epochs on an A100 GPU with a max learning rate of $1e^{-5}$ and a min learning rate of 0, following BLIP-2's setup. For LMM-based models, we transform both holistic alignment evaluation and fine-grained alignment evaluation into visual question answering (VQA) formats, and perform fine-tuning via LoRA (Hu et al. 2022). For different LMM-based models, we use the same set of questions.

Evaluation Settings. To evaluate the correlation between predicted alignment scores and annotated scores, we employ two metrics: Spearman's Rank Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC). For assessing the performance of fine-grained methods, we use F1-Score and Accuracy (ACC) to measure the accuracy of each fine-grained element. Additionally, SRCC is utilized to quantify the correlation between the average score of fine-grained elements for each image-text pair and the annotated alignment score.

Overall Alignment Evaluation

In Tab. 2, we report the results of FGA-BLIP2, along with existing VLM-based and LLM-based methods, on our EvalMuse-40K and four other benchmarks that contain annotated image-text overall alignment scores for generative models, namely TIFA (Hu et al. 2023), AIGCIQA (Wang et al. 2023a), AGIQA3K (Li et al. 2023a), and CompBench (Huang et al. 2023). It can be observed that FGA-BLIP2 achieves superior performance across all datasets. Additionally, we find that for LLM-based methods, as the number of model parameters increases, their performance on multiple datasets improves. Meanwhile, using EvalMuse-40K, we fine-tuned some LLM-based models with similar numbers of parameters to FGA-BLIP2. It is evident that compared to the original model, the fine-tuned version achieves significant performance gains across all benchmarks. Notably, prompts in EvalMuse-40K don't overlap with other benchmarks. This good generalization demonstrates the balance and diversity of our sampled prompts and the reliability of our annotation data.

Fine-Grained Evaluation

Current fine-grained evaluation methods for image-text alignment primarily rely on LLM-based models to perform visual question answering (VQA) for each element in an image-text pair. They then derive alignment scores from the

Method	Time Complexity	Overall			Real			Synth		
		SRCC	F1	ACC	SRCC	F1	ACC	SRCC	F1	ACC
Blipv2Score (ITC) (Li et al. 2023b)	$o(1)$	0.3666	0.7417	0.6163	0.3547	0.7480	0.6197	0.4000	0.7209	0.6059
InternVL2.5 (1B) (Chen et al. 2024)	$o(n)$	0.3937	0.7221	0.6381	0.3597	0.7221	0.6307	0.4659	0.7220	0.6609
InternVL2.5 (2B) (Chen et al. 2024)	$o(n)$	0.3916	0.7220	0.6373	0.3307	0.7222	0.6284	0.5287	0.7210	0.6649
InternVL2.5 (4B) (Chen et al. 2024)	$o(n)$	0.4182	0.7366	0.6589	0.3733	0.7335	0.6467	0.5187	0.7470	0.6964
InternVL2.5 (8B) (Chen et al. 2024)	$o(n)$	0.4702	0.7377	0.6668	0.4249	0.7384	0.6595	0.5727	0.7357	0.6890
LLaVA-NeXT (8B) (Li et al. 2024c)	$o(n)$	0.4842	0.7478	0.6669	0.4413	0.7513	0.6659	0.5673	0.7359	0.6702
Qwen2-VL (7B) (Wang et al. 2024)	$o(n)$	0.4473	0.7495	0.6621	0.3768	0.7506	0.6565	0.5885	0.7458	0.6793
Qwen2.5-VL (7B) (Bai et al. 2025)	$o(n)$	0.5159	0.7399	0.6817	0.4643	0.7369	0.6711	0.6143	0.7500	0.7143
Ovis2 (8B) (Lu et al. 2024)	$o(n)$	0.4276	0.7550	0.6669	0.3879	0.7560	0.6624	0.5084	0.7519	0.6809
InternVL2 (8B) (Chen et al. 2024)	$o(n)$	0.4301	0.7210	0.6474	0.3783	0.7220	0.6409	0.5447	0.7174	0.6672
InternVL2.5* (1B) (Chen et al. 2024)	$o(n)$	0.6164	0.8080	0.7626	0.6166	0.8053	0.7537	0.6082	0.8172	0.7900
InternVL2* (2B) (Chen et al. 2024)	$o(n)$	0.6246	0.7477	0.7664	0.6087	0.7328	0.7563	0.6501	0.7896	0.7973
InternVL2.5* (2B) (Chen et al. 2024)	$o(n)$	0.6190	0.8093	0.7642	0.6154	0.8051	0.7538	0.6180	0.8236	0.7962
Qwen2.5-VL* (3B) (Chen et al. 2024)	$o(n)$	**0.6503**	**0.8128**	**0.7689**	**0.6402**	**0.8089**	**0.7591**	**0.6646**	**0.8257**	**0.7992**
Blipv2Score* (ITC) (Li et al. 2023b)	$o(1)$	0.5593	0.7751	0.7164	0.5328	0.7768	0.7127	0.5968	0.7693	0.7277
FGA-BLIP2 (1B) (Ours)	$o(1)$	**0.6876**	**0.8190**	**0.7699**	**0.6355**	**0.8162**	**0.7598**	**0.7767**	**0.8287**	**0.8008**

Table 3: Quantitative comparison between FGA-BLIP2 and other methods for fine-grained evaluation on EvalMuse-40K. Time complexity denotes the number of times required to perform a complete fine-grained evaluation on an image-text pair, where n is the number of fine-grained elements in the prompt. Here, SRCC represents the correlation between the average of fine-grained evaluation scores in the image-text pair and the overall annotated alignment score.

Setting	TIFA			AGIQA3K		
	SRCC	PLCC	KRCC	PLCC	SRCC	KRCC
es_avg, w/o mask	0.630	0.621	0.464	0.545	0.591	0.386
es_avg, with mask	**0.717**	**0.711**	**0.553**	**0.612**	**0.664**	**0.436**
os, w/o var	0.751	0.742	0.584	0.769	0.864	0.581
os, with var	**0.754**	**0.748**	**0.589**	**0.771**	**0.865**	**0.583**

Table 4: Ablation study for FGA-BLIP2.

Reward Model	ImageReward	PickScore	Qwen2.5-VL (8B)	InternVL2.5 (8B)
None	0.036	0.2211	56.07	51.26
Blipv2Score	1.321	0.2264	71.91	70.09
LLlaVA-Bert	1.541	0.2234	78.58	78.18
FGA-BLIP2	**1.643**	**0.2375**	**81.42**	**81.55**

Table 5: Using the DDPO setting optimized for image-text alignment, with Stable Diffusion v1.4 as the baseline.

answers and calculate the correlation between these scores and the overall annotated scores. However, due to limitations in annotation data, these methods cannot directly measure the accuracy of each VQA response from LLM-based models. Benefiting from fine-grained annotation in EvalMuse-40K, we are able to evaluate the VQA capabilities of LLM-based models. We also report the fine-grained evaluation performance of the ITC setting in BLIP2.

From Tab. 3, the results show that, when compared to LLM-based methods with a comparable number of parameters, FGA-BLIP2 achieves superior performance in fine-grained alignment evaluation. Furthermore, as it eliminates the need for multiple rounds of visual question answering, FGA-BLIP2 enables more efficient fine-grained assessment.

Ablation Study

The ablation study for FGA-BLIP2 is shown in Tab. 4. When evaluating using the overall alignment scores (os) output by FGA-BLIP2, our model achieves better correlation across four datasets compared to the method without the variance optimization strategy (w/o var). We also report the fine-grained element scores (es) output by FGA-BLIP2 and take the average (avg) as the image-text alignment score for evaluating the model's performance on the datasets. Averaging valid tokens with masks during output can significantly enhance the performance of fine-grained evaluation.

Application as A Reward Model

We applied FGA-BLIP2 as a reward model to the DDPO (Black et al. 2024) algorithm (see Fig. **??** and Tab. 5). Compared with the original SD 1.4 (Luo et al. 2023) and the LLaVA-Bert method used in DDPO for optimizing image-text alignment, using FGA-BLIP2 achieved favorable results in multiple metrics.

Conclusion

In this work, we contribute EvalMuse-40K, which contains a large number of manually annotated alignment scores and fine-grained element scores, enabling a comprehensive evaluation of the correlation between automated metrics and human judgments in image-text alignment-related tasks. We also propose FGA-BLIP2, an efficient fine-grained alignment evaluation method. Unlike LLM-based approaches needing multiple VQA rounds, it does overall and fine-grained alignment evaluation in one pass by joint fine-tuning. Extensive experiments demonstrate that FGA-BLIP2 achieves state-of-the-art performance on EvalMuse-40K and manifests strong zero-shot generalization ability on the other benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62306153, 62225604), the Natural Science Foundation of Tianjin, China (24JCJQJC00020),

the Young Elite Scientists Sponsorship Program by CAST (YESS20240686), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63243143), and Shenzhen Science and Technology Program (JCYJ20240813114237048). This work was also funded by ByteDance Inc. The computational devices are partly supported by the Supercomputing Center of Nankai University (NKSC).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning. In *ICLR*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cho, J.; Hu, Y.; Baldridge, J. M.; Garg, R.; Anderson, P.; Krishna, R.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *ICLR*.

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 3043–3054.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 12606–12633.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 6629–6640.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 20406–20417.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 78723–78747.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 36652–36663.

Li, B.; Lin, Z.; Pathak, D.; Li, J.; Fei, Y.; Wu, K.; Ling, T.; Xia, X.; Zhang, P.; Neubig, G.; et al. 2024a. GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation. *arXiv preprint arXiv:2406.13743*.

Li, B.; Lin, Z.; Pathak, D.; Li, J.; Fei, Y.; Wu, K.; Xia, X.; Zhang, P.; Neubig, G.; and Ramanan, D. 2024b. Evaluating and Improving Compositional Text-to-Visual Generation. In *CVPR*, 5290–5301.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2023a. Agiqa-3k: An open database for ai-generated image quality assessment. *TCSVT*, 6833–6846.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024c. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. 2024d. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv preprint arXiv:2405.08748*.

Liang, Y.; He, J.; Li, G.; Li, P.; Klimovskiy, A.; Carolan, N.; Sun, J.; Pont-Tuset, J.; Young, S.; Yang, F.; et al. 2024. Rich human feedback for text-to-image generation. In *CVPR*, 19401–19411.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv:2405.20797*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 36479–36494.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurIPS*, 2234–2242.

Tan, Z.; Yang, X.; Qin, L.; Yang, M.; Zhang, C.; and Li, H. 2024. EvalAlign: Evaluating Text-to-Image Models through Precision Alignment of Multimodal Large Models with Supervised Fine-Tuning to Human Annotations. *arXiv preprint arXiv:2406.16562*.

Vonikakis, V.; Subramanian, R.; and Winkler, S. 2016. Shaping datasets: Optimal data selection for specific target distributions across dimensions. In *ICIP*, 3753–3757.

Wang, J.; Duan, H.; Liu, J.; Chen, S.; Min, X.; and Zhai, G. 2023a. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CICAI*, 46–57.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2023b. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *ACL*, 893–911.

Wiles, O.; Zhang, C.; Albuquerque, I.; Kajić, I.; Wang, S.; Bugliarello, E.; Onoe, Y.; Knutsen, C.; Rashtchian, C.; Pont-Tuset, J.; et al. 2024. Revisiting Text-to-Image Evaluation with Gecko: On Metrics, Prompts, and Human Ratings. *arXiv preprint arXiv:2404.16820*.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 15903–15935.

Yarom, M.; Bitton, Y.; Changpinyo, S.; Aharoni, R.; Herzig, J.; Lang, O.; Ofek, E.; and Szpektor, I. 2023. What you see is what you read? improving text-image alignment evaluation. In *NeurIPS*, 1601–1619.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. In *TMLR*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.