A Background

A.1 Text-conditioned Image Synthesis

The popular large-scale image generation models such as Imagen [20], DALL-E2 [15] and Stable Diffusion [17] demonstrate extraordinary generation quality, but a subtle difference in input prompt could lead to a dramatic change of semantic style, thus not directly suitable for image editing [2, 18]. Besides, the prompt-to-prompt [7] achieves image editing with cross-attention control on the observation of interaction between the pixels to the text embedding. InstructPix2Pix [3] leverages the complementary abilities of a pre-trained language model GPT-3 [4], and a pre-trained text-to-image model Stable Diffusion [16] to generate large pairs of multi-modal training data and perform image editing following human instructions. ControlNet [25] learns task-specific conditions in an end-to-end way and achieves robust control effects. Recently, growing interests [13] in computer vision focus on aligning text-to-image synthesis [9] or visual editing [26] by using human feedback. Typically, a reward model is expected to evaluate images by training on task rewards using proximal policy optimization (PPO [21]).

A.2 Image-level Evaluation Metric

CLIP-based CLIP [14] was pre-trained on large-scale image-caption pairs through contrastive learning, making it a highly versatile tool for natural language processing and computer vision applications. CLIPScore [8] measures the cosine similarity value of image and caption representations extracted from the CLIP feature extractors. Formally, $CLIPScore = max(cos(f_I(v_i), f_C(c_i)), 0)$, where f_I, f_C is the image and caption feature extractor. CLIPScore is a reference-free metric and outperforms previous reference-based metrics like CIDEr [23] and SPICE [1]. Within the text-to-image domain, previous work also relies on the same approach to measure the alignment between the text prompt and the generated image. However, vision-and-language models exhibit deficiencies in compositional understanding and are insensitive to word orders. In [24], they show BLIP [11] and CLIP [14] only achieve random chance level understanding ability on attribution, relation, and order understanding. They furthermore propose NegCLIP to improve the original CLIP model by generating additional hard negative captions and optimizing the same contrastive objective. They show that obtaining specific and low-cost negative examples can result in significant enhancements in compositional tasks without losing existing ability.

BLIP-based BLIP [11] filters out noisy synthetic captions to effectively make use of the noisy web data through bootstrapping based on a novel multimodal mixture of Encoder-Decoder. Beyond that, BLIPv2 [10] introduces Query Transformer that bootstraps image and text representation learning and then bootstraps large language model for image-to-text generations. BLIPv2 achieves state-of-the-art performance on a wide range of understanding-based and generation-based vision-language tasks, including image-text retrieval, image captioning, and visual question answering. We suppose the grounding objective in BLIPv2's pre-training can bootstrap its performance in evaluating text-to-image synthesis. Specifically, we utilize BLIPv2 to compute the image-text matching score using "ITM" head and "ITC" head. BLIP-ITC uses a simple cosine similarity function over the extracted image and text features. In contrast, BLIP-ITM uses cross-attention to fuse multimodal features to capture fine-grained similarity.

A.3 Evaluation Datasets

The MSCOCO [12] has been widely used for object segmentation, although there is a dearth of varied prompts, indicating a lack of diversity. Winoground [22] is designed for evaluating the ability of vision and language models to conduct visio-linguistic compositional reasoning. For a pair of two distinct images, their captions are composed of identical sets of words, but in a different order. Many state-of-the-art vision and language models only achieve random chance performance, making it a good testbed for evaluation. DrawBench [19] tackles prompt diversity issues by collecting challenging descriptions for image generation. There are a set of 11 prompt categories that test various capabilities of models, including the ability to accurately depict colors, numbers of objects, spatial relations, and text, as well as more complex prompts such as long textual descriptions, rare words, and misspelled

prompts. In [5], they evaluate the visual reasoning of text-to-image models and propose PaintSkills, a diagnostic dataset and evaluation toolkit designed to measure object recognition, counting, and spatial understanding. Recent studies in compositional text-to-image synthesis[6] collect Concept Conjunction prompt dataset which focuses on two objects with different colors in the text prompt, and Attribute Binding prompt dataset that is sampled from COCO captions.

B More Results

In Table B, we show the full table that includes variants of our LLMScore (CapCLIP, CapMETEOR, DescCLIP, DescMETEOR) on General Bench.

Table 1: The correlation between automatic evaluation metrics and human rankings on text-to-image synthesis. Our devised metrics LLMScore significantly surpass existing metrics in terms of Kendall's τ and Spearmanr's ρ with $p < 0.001$.

Human	Metric	COCO2014		COCO2017		DrawBench		PaintSkills	
		$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
Overall	CLIP	0.1971	0.2655	0.2227	0.2771	0.1530	0.2143	0.4715	0.5869
	NegCLIP	0.2164	0.2905	0.2793	0.3523	0.1463	0.1999	0.4911	0.6313
	BLIP-ITM	0.3252	0.4255	0.0928	0.1155	0.1044	0.1455	0.4755	0.6214
	BLIP-ITC	0.3465	0.4535	0.1703	0.2121	0.1569	0.2171	0.4743	0.5864
	CapCLIP	0.0263	0.0335	-0.0274	-0.0315	0.0056	0.0072	0.3035	0.3751
	CapMETEOR	0.0710	0.0960	0.0512	0.0650	0.0951	0.1312	0.2315	0.3056
	DescCLIP	0.1377	0.1799	0.1130	0.1424	0.1136	0.1557	0.3825	0.4917
	DescMETEOR	0.1175	0.1549	0.0567	0.0702	0.0028	0.0048	0.0678	0.0827
	LLMScore	**0.3629**	**0.4612**	**0.3357**	**0.4275**	**0.2230**	**0.3023**	**0.5600**	**0.6853**
Error Counting	CLIP	0.1464	0.2142	0.1888	0.2677	0.1360	0.1910	0.3052	0.2891
	NegCLIP	0.2116	0.3061	0.1795	0.2581	0.1179	0.1596	0.4563	0.4908
	BLIP-ITM	0.2251	0.3289	0.1137	0.1635	0.0871	0.1189	0.4622	0.4997
	BLIP-ITC	0.2636	0.3739	0.1849	0.2620	0.1506	0.2029	0.6178	0.6511
	CapCLIP	0.0266	0.0362	-0.0068	-0.0085	0.0544	0.0704	0.4963	0.5332
	CapMETEOR	0.0822	0.1197	0.0004	0.0013	0.0173	0.0192	0.3274	0.3636
	DescCLIP	0.1433	0.2145	0.0338	0.0477	-0.0039	-0.0022	0.2978	0.3289
	DescMETEOR	0.1398	0.2010	-0.0829	-0.1198	-0.1348	-0.1791	0.0881	0.0924
	LLMScore	**0.2792**	**0.4006**	**0.2138**	**0.3125**	**0.2125**	**0.2839**	**0.6444**	**0.7066**

C Human Annotation

For each image-text pair, we ask 2 annotators to rate the overall and error counting. We will show the details of annotation interface in Section C.1.

C.1 Human Ratings Interface

In Figure 1, we show the interface for human ratings over the image quality from two objectives, overall and error counting. Human annotators are required to rate the overall quality of the image on a scale of 1-10 and count the errors in the image on a scale of 0-9.

Question

Given the Text Prompt **"A gold chair and a red clock."**:

[Overall Quality]According to the Text Prompt, verify the Overall and Compositional quality of the Generated Images below by rating on a scale of 10.

Text Prompt **"A gold chair and a red clock."**
Rate the overall quality of the Generated Images in terms of matching the Text Prompt:

- 1 - Poor.
- 2 - Very Bad.
- 3 - Bad: Low quality, merely aligned with the text prompt.
- 4 - Not okay.
- 5 - Neutral (Leaning Negative)
- 6 - Neutral (Leaning Positive)
- 7 - Okay
- 8 - Good: High quality, aligned with the text prompt.
- 9 - Very Good
- 10 - Perfect.

Given the Text Prompt **"A gold chair and a red clock."**:

Text Prompt **" gold chair and a red clock."**
[Error Counting] Provide the number of composition errors Y (scale: 0-9) in the Generated Images compared to the Text Prompt. One error should be counted for each incorrect color, spatial position, shape, size, material, or relationship among objects. If an object category mentioned in the Text Prompt is missing in the caption, count it as 4 errors. The over-specifications in the image caption should be counted as only one error::

- No composition error found.
- 1 composition error.
- 2 composition error.
- 3 composition error.
- 4 composition error.
- 5 composition error.
- 6 composition error.
- 7 composition error.
- 8 composition error.
- 9 or above 9 composition error.

Figure 1: Amazon Mechanical Turk Platform. Questions Layout for Human Raters for Overall and Compositional ratings of the generated image given the text prompt.

D Visual Descriptions

Our approach involves utilizing both global and local descriptions of an image. Initially, a general caption is generated for the image. Then, a dense caption model is employed to describe the objects in detail. This technique enables the extraction of both the overall context of the image and the specific attributes of individual objects, thereby providing a comprehensive description of the image.

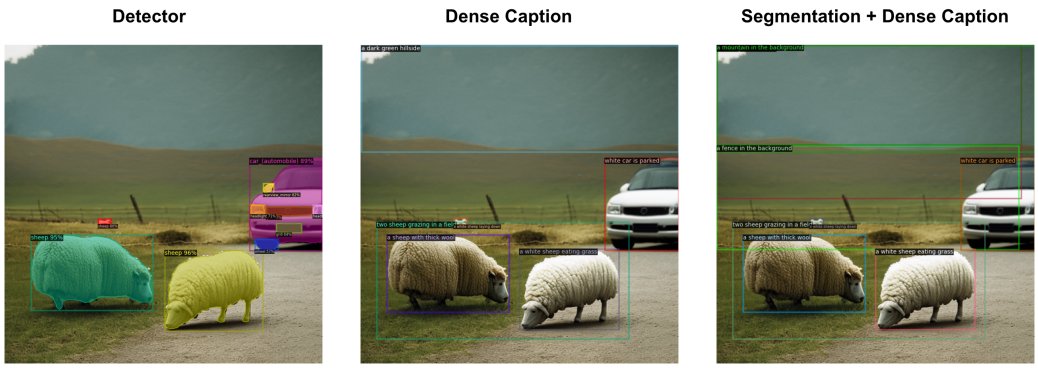

Figure 2: Comparison between, detector, dense caption, and segmentation results over the synthesized image. Alternatively, an object segmentation model can be applied to identify the objects in the image more accurately.

E Example of GPT Text-to-Image Evaluation.

Stage	Example
Prompt	a light brown bathroom with a white sink and mirror.
Description	This image showcases a clean and modern bathroom with a white theme. The main focus is the white sink, which is placed on a wooden bathroom counter. The counter is brown and extends across the entire width of the image, covering the lower part of the white wall behind it. Above the sink, there is a large mirror on the wall that reflects the shower curtain as well as a silver light fixture with two light bulbs, which are switched on and provide ample light to the room. The silver faucet complements the white sink and is slightly towards the left side of the sink. A white soap dish is positioned on the right side of the counter, next to an additional sink. A silver handle can be seen on a drawer beneath the counter, near the sink. On the left side of the picture, a brown wooden door can be partially seen, while on the right side, a small part of a white toilet is visible. The overall setup of the bathroom, with its white walls, bright lights, and well-arranged counter, gives off a spacious and organized vibe.
Rating Rationale	Overall Semantics: 0.8/1.0 The alignment between the image caption and the text prompt is mostly good. The generated image caption describes a white sink, wooden counter, mirror, and other elements like a soap dish, faucet, and lights that complement the room. This is not far from the text prompt of a light brown bathroom with a white sink and mirror.
Rating Rationale	Error Counting: 1 (number of errors) The composition error mainly lies in the description of the color of the bathroom. The text prompt mentions a light brown bathroom, but the image caption emphasizes a white theme, with only the wooden counter and door being brown.

Table 2: Examples of GPT Text-to-Image Evaluation: the prompt is used for image generation, while the description is generated by GPT-4

We showcase examples for text-to-image evaluation in Table 2 and Table 3 on synthesized images in Figure 3. The Prompt section illustrates the initial input provided to the text-to-image model, which in this instance is a short description of a bathroom. The subsequent Description section is the multi-granularity object-centric description for the image. Here, the model describes a light-filled, modern bathroom with various elements like a white sink, mirror, soap dish, and a wooden counter. Then there are two rating settings, the first is *Overall* (Overall Semantics), and the second is *Error Counting*. Both settings evaluate the text-to-image synthesis with a score and a rationale. Here, a score of 0.8 out of 1.0 reflects a high level of correspondence between the synthesized image and the prompt, despite a few discrepancies. And 1 represents there are only one errors found in the synthesized image compare to the prompt.

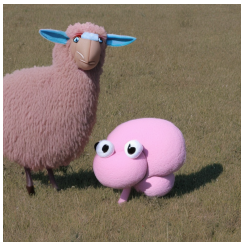

Figure 3: *Left*: Synthesized by Stable-diffusion-2 with text prompt "A light brown bathroom with a white sink and mirror" sampled from Attribute Binding Contrast dataset. *Right*: Synthesized by Stable-diffusion-2 with text prompt "A brown bird and a pink sheep." sampled from COCO.

Stage	Example
Prompt	a brown bird and a pink sheep.
Description	In this vivid image, there are two unique pink animals standing on a large patch of green grass that covers the entire ground. On the left side, there is a pink sheep with a blue tag attached to its ear. The sheep has distinct blue ears and a prominent head, showcasing a contrast in colors against its pink body. It also has two sturdy black legs standing firmly on the grass. On the right side of the image, there is a pink pig with black eyes. The pig's eyes are embedded on a pink and white stuffed animal-like face, giving it a cute and charming look. The pink sheep and the pig are casting shadows towards the bottom of the image, blending with the green grass beneath. Overall, the colorful and playful scene provides a fantastical and captivating atmosphere, with the pink sheep and pink pig standing as the central focus.
Rating Rationale	Overall Semantics: 0.8/1.0 There is a clear focus on the pink sheep in the image caption, but the prompt's mention of a brown bird is entirely missing. The overall composition is good and somewhat aligned, but it could be better.
Rating Rationale	Error Counting: 4 (number of errors) The composition error comes from the missing brown bird (4 errors). The color, size, and material of the pink sheep, as well as the relationship between the two animals are well-described, making this error count relatively low.

Table 3: Examples of GPT Text-to-Image Evaluation: the prompt is used for image generation, while the description is generated by GPT-4

Broader Impact

The framework proposed in this paper first integrates GPT-4 for text-to-image evaluation and showcases how to take advantage of the existing large-scale pre-trained models (GPT-4) for measuring the alignment between the generated images and text, we also propose a new metric, LLMScore which provides interpretable rating and well aligns with the human scores on several datasets. This work sheds light on the value of large language models on the evaluation of text-to-image synthesis, we hope it can help the future text-to-image synthesis work on improving the groundedness and compositionality, either as a reward signal or evaluation metric; our preliminary work on the interpretability of LLMScore, may have the potential to be used for explanation, controllable generation, and image editing.

Limitations

One limitation of our work is that it relies on GPT, which is not free for the public, and may limit its fast plug-in capability, future work may consider replacing this component with a publicly available LLM model (e.g., LLaMA) or our in-house finetuned image captioning model. Another potential issue for this work is, since it incorporates the exsiting large language models, it may inherit its own biases that could propagate to the metric. The future work who considers adopting our LLMScore metric should be cautious on the specific domains to make sure no harmful biases get propagated.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 1

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 1

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models, 2022. 2

[6] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. 2

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 1

[9] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback, 2023. 1

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014. 1

[13] André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. Tuning computer vision models with task rewards. *arXiv preprint arXiv:2302.08242*, 2023. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[16] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[18] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1

[22] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[23] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[24] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv e-prints*, pages arXiv–2210, 2022. 1

[25] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[26] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 1