The Fabrication of Reality and Fantasy: Scene Generation with LLM-Assisted Prompt Interpretation Supplementary Materials

Anonymous ECCV 2024 Submission

Paper ID #3294

A LLM-Driven Detail Synthesis

In this work, as described in the Sec. 4.1 of the main paper, we emphasized that by leveraging LLMs, we have significantly enriched responses to encompass additional information, such as *layout, detailed descriptions, background scenes,* and *negative prompts*. To achieve this, we facilitated an interaction with a LLM as shown in Fig. 1. The input given to the LLM, depicted on the left side of the figure, includes detailed task specifications and in-context learning examples to enhance the LLM's comprehension. The response from the LLM, shown on the right, is rich with details extracted from the prompt. Notably, the descriptions are particularly crucial for our work, serving as indispensable information for the later image generation stage.

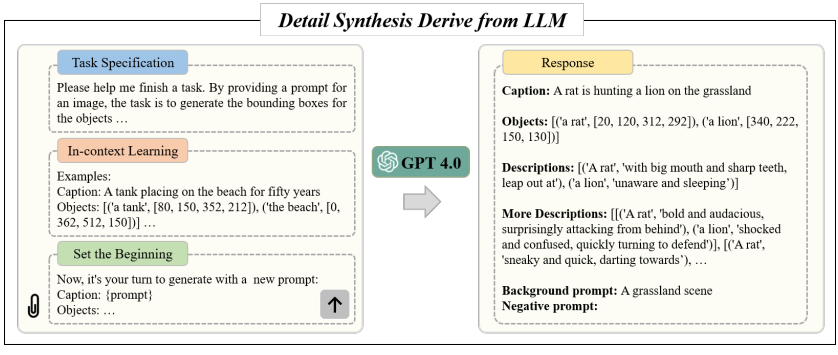

Fig. 1: Detail Synthesis. The illustration of the interaction with a LLM in our work.

B Qualitative Comparison on RFBench

In Fig. 2 and Fig. 3, we present additional qualitative examples to showcase the exceptional outcomes of our work. Fig. 2 shows the results under the category ***Realistic and Analytical***, while Fig. 3 shows the category ***Creativity and Imagination***. Both figures demonstrate that our method achieves more accurate editing results compared to other approaches.

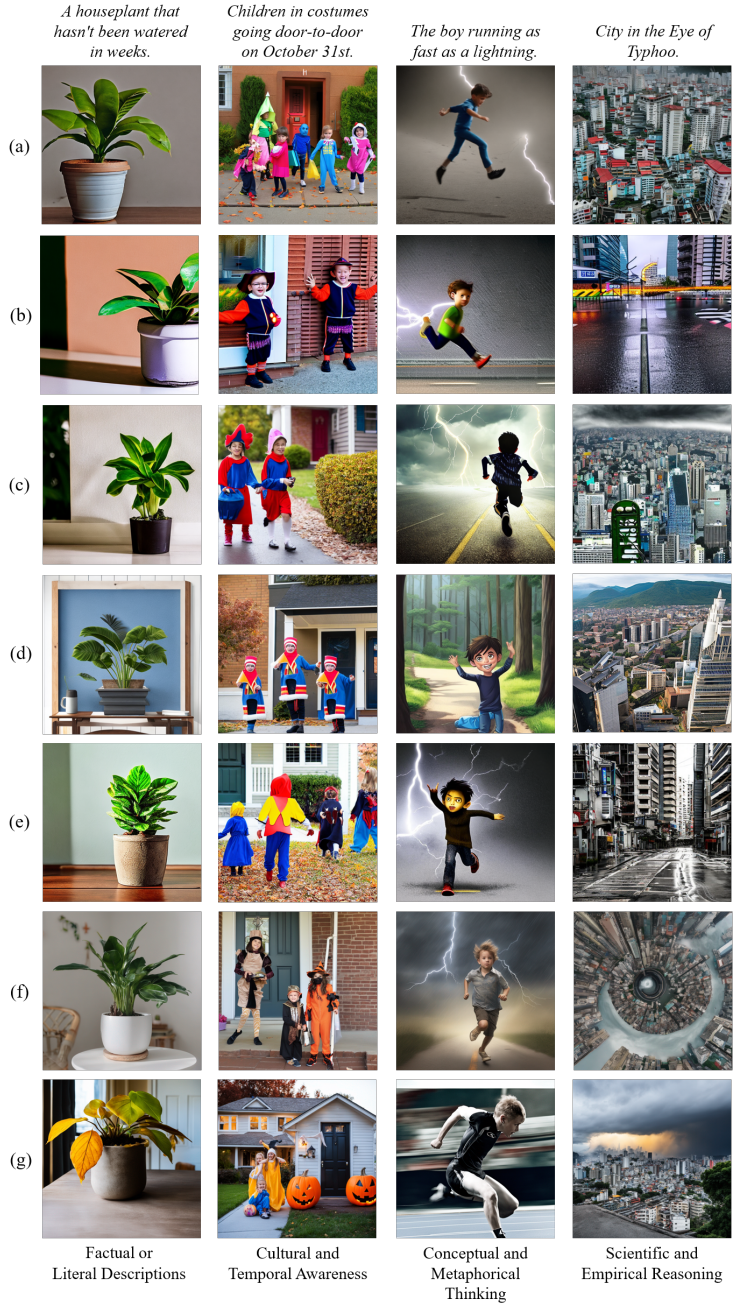

Fig. 2: More results on ***Realistic and Analytical.*** The compared models include (a) Stable Diffusion, (b) MultiDiffusion, (c) AttendandExcite, (d) LMD, (e) BoxDiff, (f) SDXL, (g) Ours

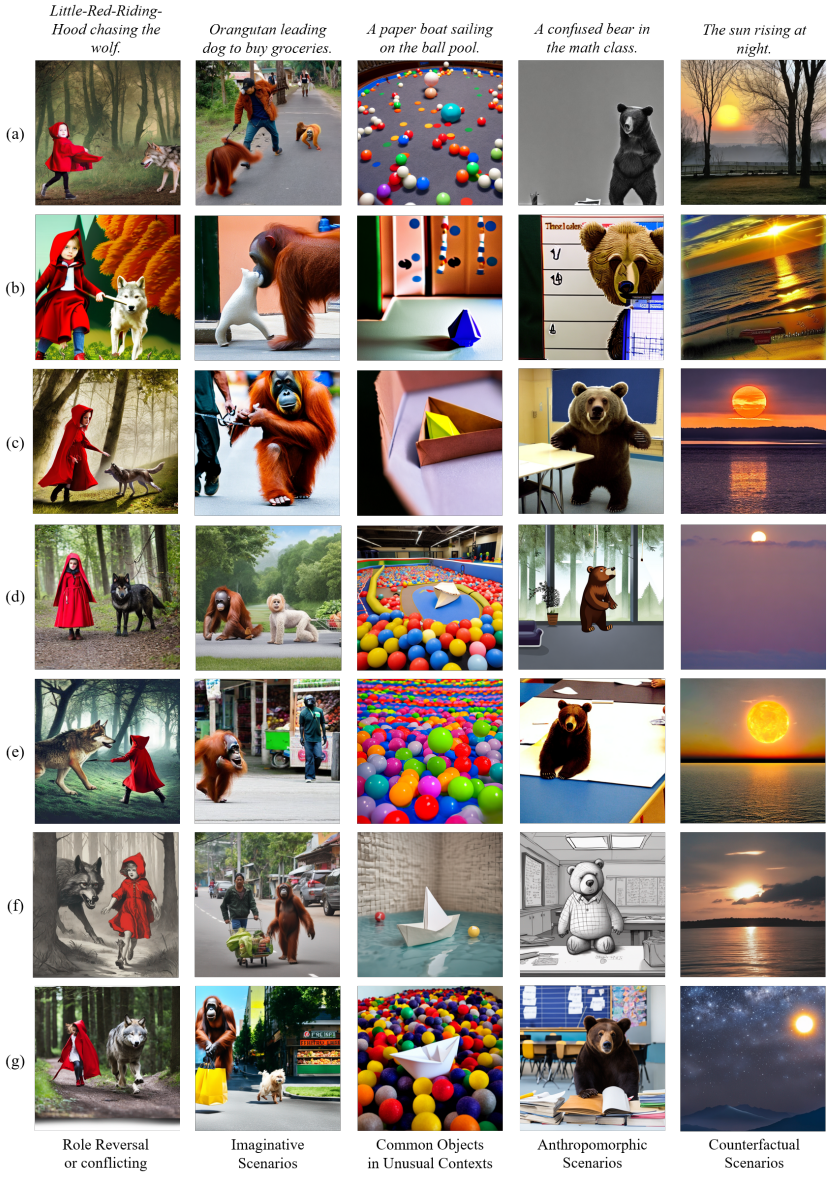

Fig. 3: More results on ***Creativity and Imagination.*** The compared models include (a) Stable Diffusion, (b) MultiDiffusion, (c) AttendandExcite, (d) LMD, (e) BoxDiff, (f) SDXL, (g) Ours

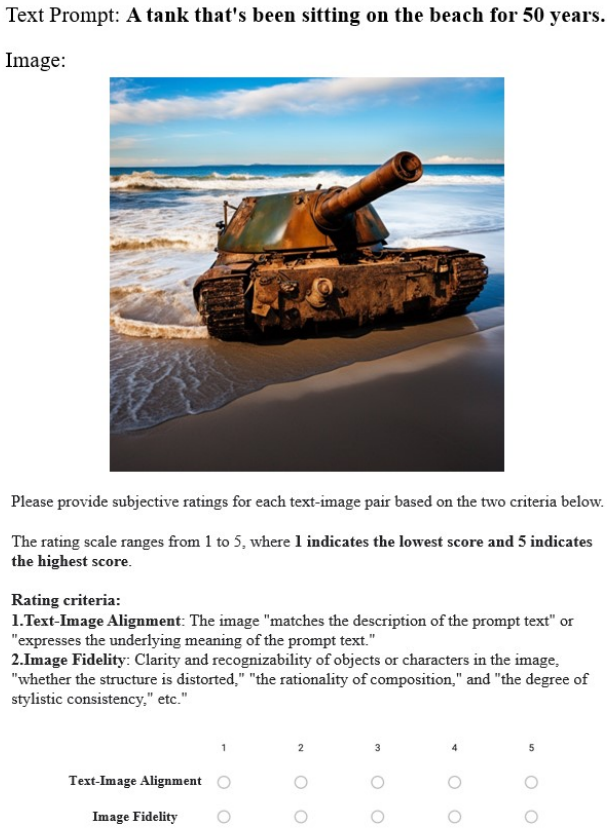

Fig. 4: Survey on Image-Text Alignment and Image Fidelity

C GPT4Score

We follow the approach of T2I-Compbench, using Multimodal LLM (MLLM) to measure the similarity between generated images and input prompts. The key deviation lies in our observation that MiniGPT4, employed in T2I-Compbench, struggles to comprehend the surreal aspects of the images effectively. Therefore, we employ GPT4, a more powerful MLLM, as our new benchmarking model for evaluation, as mentioned in the Sec. 5.1 of the main paper.

Specifically, given a generated image and its prompt, we input both the image and prompt into GPT4. Subsequently, we pose two questions to the model: "*Describe the image*" and "*Predict the image-text alignment score*", the generated image is then assigned the final output score predicted by GPT4. For detailed prompts, please refer to the appendix of T2I-Compbench.

D Human Evaluation

In the human evaluation process, as introduced in the Sec. 5.4 of the main paper, we request annotators to assess the correspondence between a produced image and the textual prompt employed to create the image. Fig. 4 show the interfaces for human evaluation. The participants can choose a score from {1, 2, 3, 4, 5} and we normalize the scores by dividing them by 5. We then compute the average score across all images and all participants.

Table 1: The correlation between automatic evaluation metrics and human evaluation

Metrics	CLIPScore		GPT4Score	
	$\tau(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$	$\rho(\uparrow)$
Realistic & Analytical				
Scientific and Empirical Reasoning	-0.4880	-0.5946	0.6351	0.7157
Cultural and Temporal Awareness	-0.0476	-0.1429	0.3273	0.3780
Factual or Literal Descriptions	0.2333	0.3656	0.7620	0.8909
Conceptual and Metaphorical Thinking	-0.1952	-0.1982	0.9234	0.9633
Creativity & Imagination				
Common Objects in Unusual Contexts	-0.2381	-0.2857	-0.5345	-0.6124
Imaginative Scenarios	0.3752	0.6335	0.7265	0.8432
Role Reversal or Conflicting	0.0476	0.1429	0.5040	0.5774
Anthropomorphic Scenarios	-0.1429	-0.1429	-0.5345	-0.6124

E Human Correlation of the Evaluation Metrics

We adopt the methodology from T2I-Compbench, calculating Kendall's tau (τ) and Spearman's rho (ρ) to evaluate the ranking correlation between CLIPScore, GPT4Score, and human evaluation. For better comparison, the scores predicted by each evaluation metric are normalized to a 0-1 scale. The human correlation results are presented in Tab. 1. These results indicate that CLIP underperforms in both categories, as discussed in Section 5.1 of the main paper. This underperformance may be due to CLIP's approach to image understanding, which is often too simplistic. Nevertheless, both metrics encounter challenges with ***Creativity and Imagination***, highlighting that although GPT4Score offers a broader understanding of images, accurately assessing creativity remains a difficult task for both.

F Visualization of Ablation Study

In addition to the quantitative results presented in our ablation study, we have also included visual examples to showcase the impact of different components in our work. As shown in Fig. 5, the removal of guidance constraint and suppression constraint both causes the diffusion model to become muddled when dealing with multiple objects. Besides, eliminating the SAA module leads to unclear outcomes with the generated objects.

Fig. 5: Ablation study on various components in our work.

F.1 Effect of the hyperparameter β of guidance constraint

In our paper, we emphasize the critical role of the guidance constraint in integrating multiple objects into the background. To underscore its significance, we performed an additional ablation study focusing on the hyperparameter β, which influences the strength of guidance constraint. As shown in Fig. 6, we varied β from 0.1 to 30 to observe the effects on the generated results. The findings reveal that an optimal β value (e.g., setting it to 15) ensures objects are accurately aligned with the layout and are of high quality. However, extreme β values, such as 0.1 or 30, disrupt the layout and diminish the overall quality of the generated images.

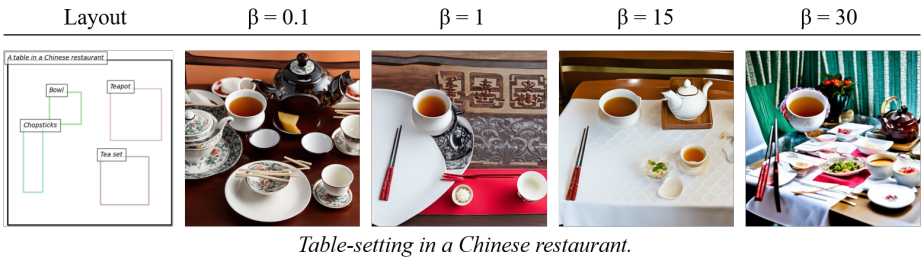

Fig. 6: Effect of the hyperparameter β of guidance constraint.