3D Space as a Scratchpad for Editable Text-to-Image Generation

Supplementary Material

Figure 9. **Additional qualitative examples from GenAIBench dataset.**

Figure 10. **Additional qualitative examples from CompoundPrompts dataset.**

7. Additional Qualitative Results

We illustrate some additional qualitative results on the GenAIBench dataset in Figure 9 and on the CompoundPrompts dataset in Figure 10. We highlight the errors made by the baselines with bold red text. These examples further show that our method excels in reasoning about counting, spatial, attribute, and comparative planning demanded by prompts for image generation.

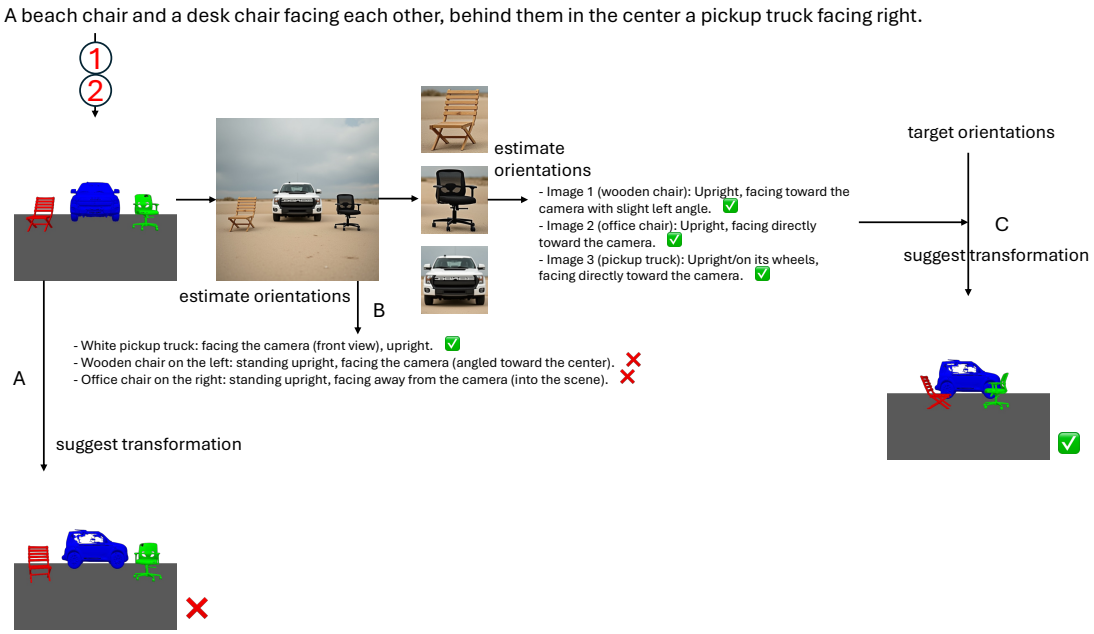

Figure 11. **Strategies for determining rotations.** We examine a scenario that requires accurate rotation planning for multiple objects described in a natural language prompt. A: Directly providing the LLM with multiview renders of the entire scene and asking it to output transformation matrices fails to produce correct rotations. B: Estimating orientations from the full synthesized image also proves unreliable: while the pickup truck is interpreted correctly, both chairs receive incorrect orientation predictions. C: Our proposed strategy isolates each object by cropping its image, enabling the LLM to infer its orientation independently; these estimated orientations, paired with the desired target orientations, are then passed to a secondary agent that suggests appropriate transformations. This crop-based decomposition yields far more accurate rotation predictions, improving robustness even though its quantitative impact is modest (Table 3) due to benchmarks containing few orientation-sensitive prompts. Nonetheless, the qualitative failures observed in (A) and (B) highlight the importance of dedicated orientation planning, motivating our explicit modeling of this capability.

8. Reasoning about Orientation using LLMs

In Figure 11 we explore a situation which requires specific rotation planning. We show that A - simply supplying the renders to the LLM and prompting to output transformation matrices does not work. For estimating orientations, B - using the full image leads to erroneous estimations. Our strategy of C - using crops to estimate individual orientations and then supplying to another LLM along with target orientations is a more robust option. As shown in Table 3 the improvement offered by adding this agent is not significant due to benchmarks not containing many samples that require specific orientation planning. However, we focus on modeling this aspect robustly upon qualitatively noticing inconsistencies.

9. Scratchpad rendering

We show the various design choices we explore for rendering in Figure 12. We only show the front view here for each of the choices but the same strategy is applied for other views of each corresponding choice. The choices are shown in the order they were listed in Table 4. We draw rulers up till the fixed bounds seen in the figure so as to constrain

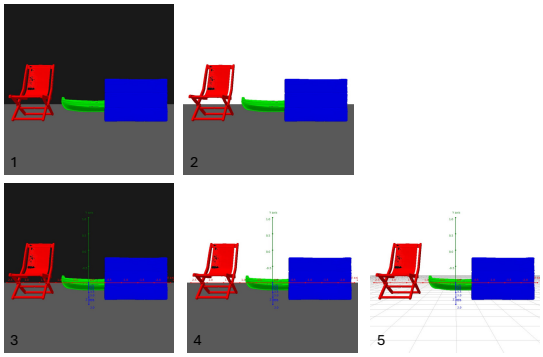

Figure 12. **Various rendering designs explored.** We illustrate the different rendering configurations explored in our design ablation, shown here only for the front-view camera but applied analogously to all viewpoints. Each variant corresponds to the ordered settings listed in Table 4, including changes to background color, ruler visibility, and grid placement. Rulers are drawn up to fixed spatial bounds to constrain object placement, and cameras are positioned with fixed viewing directions and distances chosen to ensure all meshes remain fully visible.

placement within a region. The camera in each view is created with fixed directions per view (eg. $[0.0, 0.0, -1.0]$ for front) wrt to the center point of the scene and a distance which is fixed as the minimum distance required for all meshes to be completely visible. For rendering proposal views among which the final CameraPicker agent chooses the final view, we use the following directions:

```
directions = {
    "proposal0": [0.0, 0.0, -1.0],
    "proposal1": [-0.3, 0.0, -1.0],
    "proposal2": [0.3, 0.0, -1.0],
    "proposal3": [0.0, 0.3, -1.0],
    "proposal4": [0.0, 1.0, 0.0],
}
```

These denote front, partial right, partial left, partial top, and top views in order.

10. Evaluation on T2I-CompBench

We evaluate the baselines and our method on 2D/3D spatial, complex, and numeracy tasks of T2I-CompBench. We use the same strategy as proposed by T2I-CompBench for evaluation, where for 2D/3D spatial tasks and numeracy we use UniDet [55] and for the complex category we use their proposed 3-in-1 metric. We observe that our method improves over baselines across all the reasoning domains.

Method	Reasoning modality	2D spatial	3D spatial	Complex	Numeracy
Flux	-	0.29	0.39	0.37	0.61
Idea2Img	text	0.38	0.42	0.49	0.65
RPG - Flux	2D	0.43	0.43	0.55	0.64
Ours	3D	**0.48**	**0.54**	**0.65**	**0.67**

Table 6. **Evaluation on T2I-CompBench.**

Method	Ours	MUSES	Flux.1 + camera	Idea2Img + camera	RPG + camera
VQAScore	0.83	0.68	0.67	0.80	0.72

Table 7. GenAIBench Advanced additional baselines.

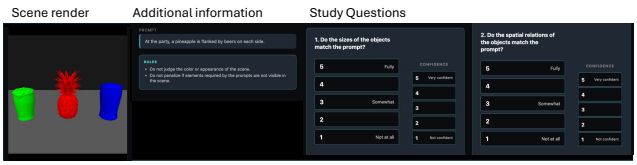

Figure 13. User study

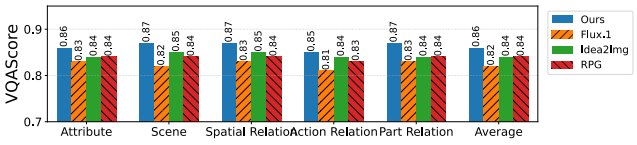

Figure 14. GenAI-Bench Basic plot

11. Comparison to MUSES

Table 7 shows that our method outperforms MUSES significantly. This could be attributed to the fact that their LLM planning stage relies on 2D bounding boxes for object placement.

12. Camera Control

We extracted cameras chosen by our method and constructed text prompts as; "front view", "viewed from left", "viewed from right", "slight top view", "top view"; and appended them to the original prompts. We show in Table 7 that using these prompts results in improvement in the Flux.1 baseline, but lower improvement in the reasoning baselines like RPG and Idea2Img.

13. GenAIBench Basic

Evaluation on GenAIBench Basic is presented in Figure 14. Our method outperforms baselines but with lower margins as compared to GenAIBench Advanced. We hypothesize this is due to most examples in the "Basic" portion referring to single subjects.

14. User Evaluation

We created a user study with 30 examples of 3D scratchpads. We show the interface in Figure 13. There were 160 evaluated examples across 9 individuals. On a scale of 1 to 5, average rating of whether sizes are consistent is 4.5 (with a confidence of 4.7). Average rating of spatial relations is 4.6 (with a confidence of 4.7).