SpatialDiff: 3D-Aware Object Movement via Implicit Spatial Modeling

Supplementary Material

In this supplementary material, we first present quantitative and qualitative results on OBJect-3DIT, comparing SpatialDiff against various baseline methods. This is followed by additional visual results generated by our approach. Furthermore, we provide comparisons with a commercial baseline and present additional ablation studies. Finally, we detail the setup of our user study and conclude with a discussion on the limitations of SpatialDiff alongside potential directions for future work.

1. Results on OBJect-3DIT

We obtained the quantitative results on the OBJect-3DIT test dataset, as shown in Table 4. Our method consistently achieves the best overall performance on this dataset. Notably, OBJect-3DIT features relatively simpler scenes. Compared with the results on SpatialBench (Table 1), most existing methods perform better in instruction-following under both GPT-SC and Qwen-SC evaluations. The visualization results on this dataset are shown in Figure 8.

2. Additional Qualitative Results

Figure 9 shows more results of complex scene object movement. We observe that in the clock example (Row 1, Column 1), SpatialDiff not only moves the clock onto the clothes but also generates a natural indentation, indicating that the movement operation aligns with realistic 3D physical behavior. In the apple example (Row 2, Column 1), the apple is accurately placed between the watermelon and the banana, an instance involving significant front–back occlusion, which demonstrates that SpatialDiff has a strong understanding of 3D spatial relationships. Moreover, in other examples involving depth inconsistencies or occlusions, SpatialDiff is still able to achieve superior performance in both instruction following and object consistency.

3. Comparison with Commercial Baseline

Our method continues to outperform the commercial baseline Nano Banana 2 in Table 5. Specifically, while Nano Banana 2 maintains competitive performance in certain generation quality metrics such as GPT-PQ, SpatialDiff achieves significant margins in spatial instruction adherence (Qwen-SC) and overall accuracy (Qwen-O).

4. Additional Ablation

We evaluated the variant, *Full model w/o ISM*, by replacing VGGT features with learnable queries while retaining LDS. Tab. 6 shows a large drop in instruction following, validating the critical role of ISM. Compared to "w/o LDS", LDS acts as *linker* that couples depth supervision with VGGT priors, enforcing spatial consistency.

5. User Study Details

Figure 6 shows the user study interface for evaluating Semantic Consistency (H-SC) and Perceptual Quality (H-PQ). Each case includes two questions, one on Semantic Consistency and the other on Perceptual Qualit. H-SC primarily evaluates whether the generated image follows the editing instruction, regardless of image quality; H-PQ primarily evaluates whether the attributes and shape of the edited object have changed and whether the unedited regions remain consistent, without considering whether the instruction was completed. Participants rate their responses on a 5-point scale: (1) "Very inconsistent", (2) "Somewhat inconsistent", (3) "Fair", (4) "Quite consistent", and (5) "Very consistent".

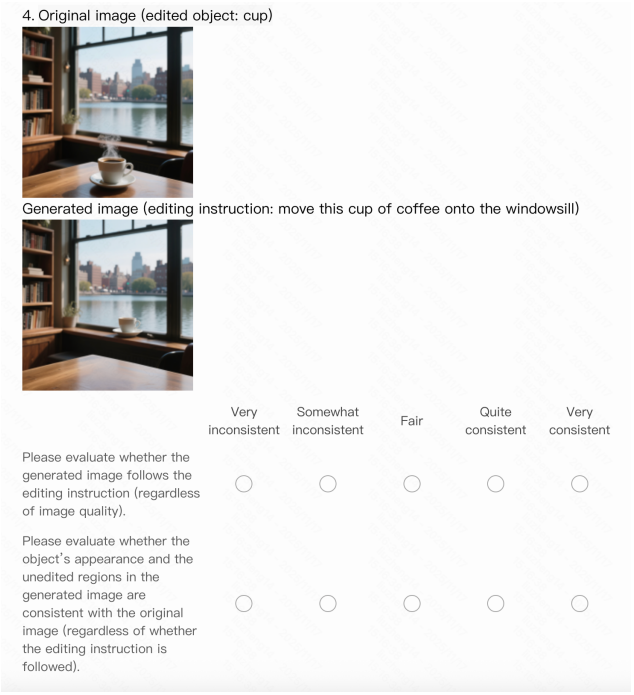

Figure 6. **Screenshot of our user study rating interface.**

6. Limitations and Discussion

In Fig. 7, incorrect perspective scaling and incomplete shadow removal are still observed in SpatialDiff. This reflects a limitation in physical reasoning required to fully handle complex object interactions. Precise distance control also remains challenging due to natural language ambiguity.

In SpatialDiff, the tokens extracted from the input image by the VGGT backbone are first fused into a learnable query via cross-attention, and then concatenated with other information along the token sequence dimension within MMDiT. While this design introduces useful 3D priors and implicitly models 3D spatial positional information within the network, the additional conditioning tokens increase the computational burden during both training and inference, thereby reducing efficiency. With 1024 spatial tokens, the inference time increased from 29s to 59s. A potential direction for future work is to integrate these 3D priors directly into the model module parameters, so that the extra conditioning computation can be removed during inference—that is, concatenating the 3D prior tokens only during training, while relying solely on the 2D image tokens as conditioning information at inference time.

Figure 7. Limitations of the approach.

Table 4. **Quantitative comparison** on OBJect-3DIT. GPT-SC, GPT-PQ, and GPT-O refer to the metrics evaluated by GPT-5, while Qwen-SC, Qwen-PQ, and Qwen-O refer to the metrics evaluated by Qwen3-VL-32B-Instruct. Scores range from 0 to 1. ↑: higher is better.

Method	GPT-SC↑	GPT-PQ↑	GPT-O↑	Qwen-SC↑	Qwen-PQ↑	Qwen-O↑
Flux-Kontext [2]	0.424	0.831	0.594	0.253	**0.858**	0.466
OmniGen2 [40]	0.416	0.587	0.494	0.407	0.691	0.530
Step1X-Edit [21]	0.528	0.637	0.580	0.552	0.765	0.650
BAGEL [47]	0.697	0.518	0.601	0.643	0.700	0.671
Qwen-Image-Editing [39]	0.783	0.807	0.795	0.746	0.839	0.791
SpatialDiff (Ours)	**0.891**	**0.833**	**0.862**	**0.776**	0.847	**0.811**

Table 5. Quantitative results compared with a commercial baseline.

Method	GPT-SC↑	GPT-PQ↑	GPT-O↑	Qwen-SC↑	Qwen-PQ↑	Qwen-O↑
Nano Banana 2	0.711	**0.892**	0.796	0.594	**0.840**	0.706
SpatialDiff (Ours)	**0.803**	0.886	**0.843**	**0.778**	0.838	**0.807**

Table 6. Additional Ablation Studies on the ISM Module.

Method	GPT-SC↑	GPT-PQ↑	GPT-O↑	Qwen-SC↑	Qwen-PQ↑	Qwen-O↑
Full model w/o LDS	0.518	0.743	0.620	0.513	0.684	0.592
Full model w/o ISM	0.656	**0.882**	0.761	0.635	0.775	0.702
Full model (SpatialDiff)	**0.804**	0.871	**0.837**	**0.796**	**0.835**	**0.815**

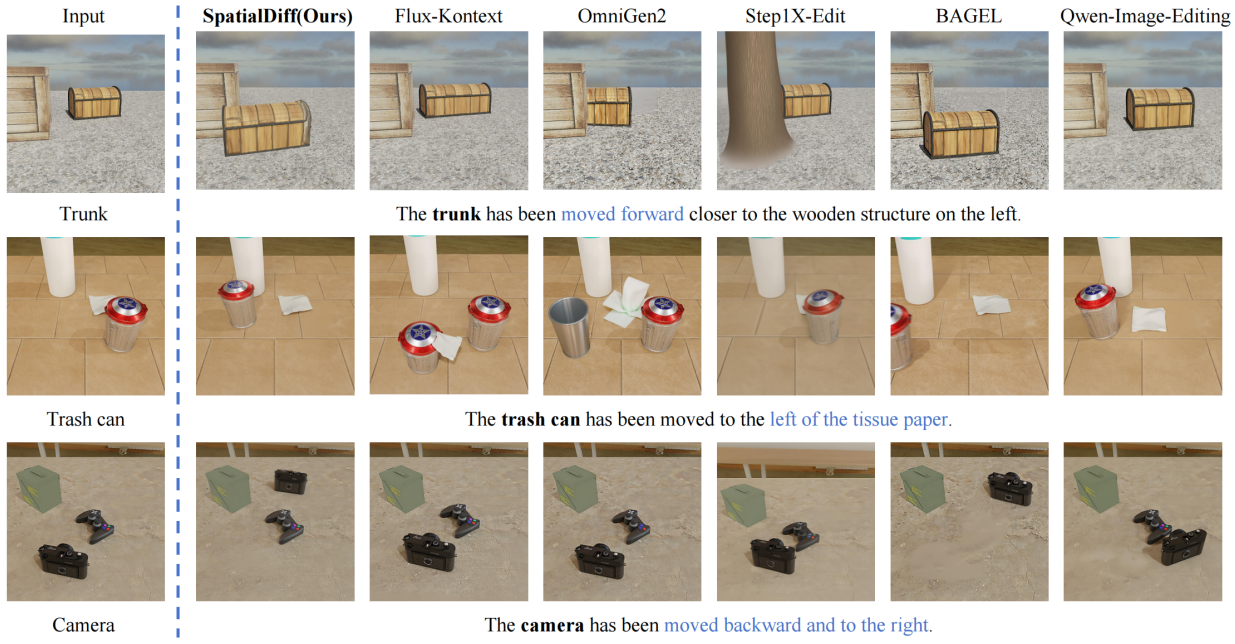

Figure 8. **Qualitative comparison** on OBJect-3DIT. The leftmost image represents the input image, with the target object to be edited shown below it. Each example is accompanied by the corresponding editing instruction, where the blue text indicates the spatial position involved in the editing operation.

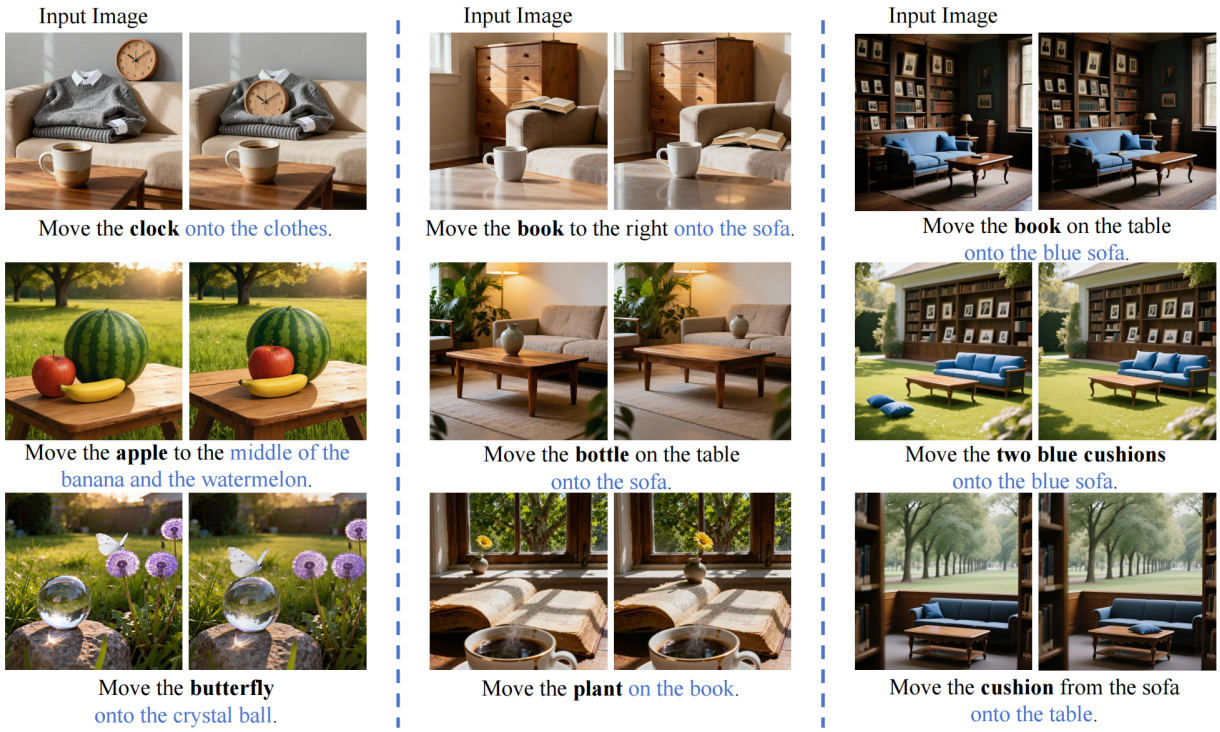

Figure 9. **More results of SpatialDiff**. Each example is accompanied by the corresponding editing instruction, where the blue text indicates the spatial position involved in the editing operation.

Prompt for PQ Score

Task Definition:
You will be provided with two images: one is the reference image before editing, and the other is the result image after editing. The description {`prompt`} specifies a spatial adjustment operation applied to the target object {`edit_object`} in the reference image. Although the description indicates the intended operation, your evaluation should **not** focus on whether the editing instruction is successfully completed. Instead, focus on the overall visual quality, naturalness, and spatial consistency of the edited image.

Scoring Criteria:
- **Stability:** Whether the non-edited regions remain stable, with continuous background structure, texture, and lighting, free from noticeable damage or blurring.
- **Artifacts:** Whether there are artifacts, ghosting, edge misalignment, or residual traces around the edited object.
- **Consistency:** Whether lighting, shadow, perspective, and depth relationships are consistent and reasonable after editing, ensuring an overall natural and realistic appearance.

Scoring Range:
- **0.0 - 0.2 (Severe distortion):** The edited area is highly unnatural; non-edited areas are damaged or contain significant artifacts; the overall image looks inconsistent.
- **0.2 - 0.4 (Poor):** Noticeable flaws or artifacts in the edited area; object edges or background appear incoherent; low overall naturalness.
- **0.4 - 0.6 (Fair):** The editing operation is roughly achieved but with some visual inconsistencies such as local blurriness, incorrect shadows, or slight artifacts.
- **0.6 - 0.8 (Good):** The editing appears natural; the object blends well with the background; only minor inconsistencies are present.
- **0.8 - 1.0 (Excellent):** The editing is smooth and seamless; non-edited regions remain completely consistent; the image looks natural, harmonious, and free of visible artifacts or discontinuities.

Input Format: You will receive two images: the first is the pre-edit image, and the second is the post-edit image.
Output Format: Provide **only a single numerical score** between 0.0 and 1.0 to assess the overall visual naturalness and consistency.

Figure 10. Prompt for PQ Score.

Prompt for SC Score

Task Definition:
You will be provided with two images: one is the reference image before editing, and the other is the result image after editing. The description {`prompt`} specifies a spatial adjustment operation applied to the target object {`edit_object`} in the reference image. The editing operation may involve movement of the object in two-dimensional directions (up, down, left, right) or in three-dimensional space (near, far, depth). Your task is to evaluate, based on the before-and-after images, whether the editing result accurately follows the operation described in the instruction. **Do not** evaluate the visual quality of the image; only assess whether the editing operation has been completed as instructed.

Scoring Range:
- **0.0 - 0.2 (Not completed at all):** The edited image is almost unrelated to the instruction; the target object did not move as expected, or the direction/position is clearly wrong.
- **0.2 - 0.4 (Partially incorrect):** Some changes are visible, but the movement direction or magnitude does not match the instruction, or the editing effect is only partially reflected.
- **0.4 - 0.6 (Partially correct):** The object roughly moves according to the instruction, but noticeable deviations exist, such as inaccurate position, wrong depth direction, or incorrectly updated background.
- **0.6 - 0.8 (Mostly correct):** The editing result largely matches the instruction; the object's movement direction and position are generally correct, with only minor errors or unnatural aspects.
- **0.8 - 1.0 (Fully correct):** The edited image accurately executes the spatial movement described in the instruction; direction, distance, and depth relationships are natural and reasonable.

Input Format: You will receive two images: the first is the pre-edit image, and the second is the post-edit image.
Output Format: Provide **only a single numerical score** between 0.0 and 1.0 to assess the accuracy of the spatial editing operation.

Figure 11. Prompt for SC Score.