MLLMSplat: A 2D MLLM-Powered Framework for 3D Gaussian Splatting Understanding, Generation, and Editing

Supplementary Material

A. Evaluation of Geometric Consistency

To assess the geometric fidelity of the generated 3DGS, we employ the 3D consistency metric from WorldScore [1]. This metric leverages DROID-SLAM [2] to estimate pixel-wise depth across frames and quantifies the reprojection error between co-visible pixels in consecutive views. As shown in Table 1, our method significantly outperforms existing baselines, achieving superior 3D consistency across both the RealEstate10K and DL3DV-10K datasets.

Table 1. Comparison on the WorldScore 3D Consistency metric.

Method	WorldScore ↓	
	RealEstate10K	DL3DV-10K
Director3D	122.91	112.40
SplatFlow	108.44	86.39
Prometheus	51.06	46.32
Ours	**11.89**	**8.55**

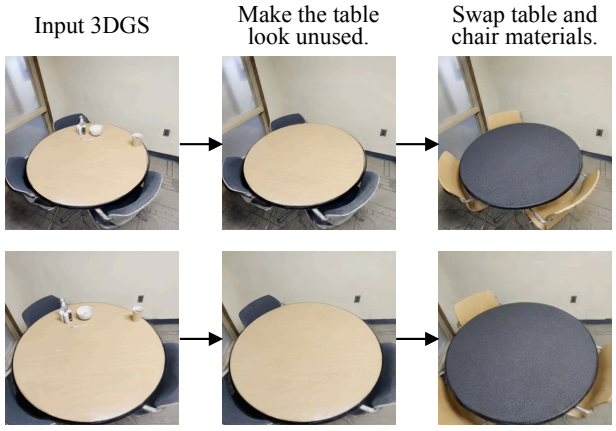

Figure 1. Visualization of our model in complex editing scenarios.

B. Evaluation of Complex Editing

In Figure 1, we demonstrate our framework's capacity for sequential editing involving multiple objects and logical reasoning. These results suggest that our approach successfully propagates advanced MLLM reasoning capabilities into the 3DGS domain. Despite these strengths, we observe that our model struggles with instructions requiring intricate reasoning. We attribute this bottleneck to the underlying MLLM backbone, which exhibits inherent limitations in handling such complex reasoning tasks even in the 2D image domain.

C. Analysis of Contextual Capacity

Within our framework, Gaussian tokens are processed analogously to visual tokens in the backbone Multimodal Large Language Model (MLLM), thereby inheriting the visual token budget. Consistent with established performance profiles in MLLMs, context window saturation precipitates a degradation in instruction-following efficacy and induces the "lost-in-the-middle" phenomenon. These behavioral shifts are primarily attributable to attention dilution as the sequence length approaches context limits.

D. Equivalence of RoPE and GaPE

For queries and keys originating from the same view, the dot product of the unit-level Positional Encoding (PE) components in both GaPE and RoPE simplifies to an identity matrix. Given that the remaining intra-unit components are identical by design, GaPE and RoPE are mathematically equivalent under intra-view attention.

References

[1] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *IEEE/CVF International Conference on Computer Vision*, pages 27713–27724, 2025. 1

[2] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, pages 16558–16569, 2021. 1