Supplementary Material

1. Differences between ours and prior work.

Our work focuses on generating compositional 3D scenes from single images, whereas the prior Layout Learning derives 3D compositions from textual descriptions. Text-to-3D generation places a strong emphasis on the diversity of the generated content, along with the proper structure and fine textures of the generated assets, while image-to-3D content generation emphasizes the fidelity and consistency of the generated assets with respect to the input images. This fundamental difference in input types presents unique challenges. Thus, our method must accurately interpret and render the complex visual details of an image into a coherent 3D layout. Technically, while Layout Learning employs Score Distillation Sampling (SDS) for text-based supervision, our approach utilizes the differentiable rendering, where we propose a novel combination of optimal transport-based appearance loss and semantic loss using the DINO-v2 to optimize the layout of 3D assets generated from the first stage under the supervision of input single images. Our framework allows users to choose whether to optimize rotation (r) as a variable. While end-to-end generation via large-scale multi-object 3D datasets is a promising direction, it is limited by the high cost of data construction and model retraining. In contrast, our modular two-stage approach enables low-cost composition without the need for large-scale finetuning. Moreover, our method can facilitate the creation of such datasets by providing structured multi-object scenes, forming a cycle.

2. More implementation details

Throughout the experiment, both the reference and rendered images are maintained at a resolution of 256×256. In the long-range appearance loss function L_a, we estimate the depth of the images using DPT-DINOv2-base [7]. For pixel matching, we adopt Sinkhorn divergences [2] as an efficient approximation of the optimal transport algorithm and use a GPU implementation provided by GemoLoss [4] with the parameter ϵ set to 0.01. In the high-level semantic loss function L_s, we extract image features using DINOv2-base [7] and compute the loss using features of the last hidden state. L_s is subjected to a warm-up period of 200 iterations. In Eq.5, we set λ to 0.8.

3. User study

We conducted a user study to compare our method with others. We gathered 280 responses from 40 human participants. Each participant was shown a reference image alongside four 3D assets (including our model and baseline model DreamGaussian [9], Wonder3D [6] and LRM [5]) simultaneously and asked to select the most realistic assets based on geometry, texture quality, and accurate placement. All options were presented in a randomized order with no time constraints. Figure 1 illustrates that our approach significantly outperformed previous methods in terms of human preference.

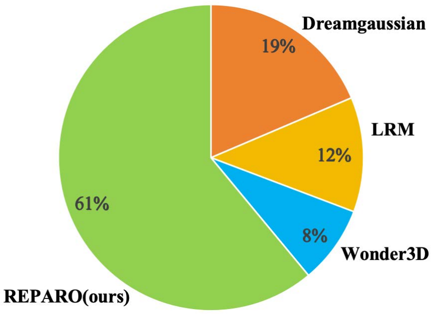

Figure 1. User study of different models.

4. Additional ablation study

Rendering resolution. In our ablation study, we analyzed the impact of different renderer resolutions on layout alignment, as shown in the Table 1. The default resolution of 256×256 in the REPARO framework achieves the highest CLIP score (0.833 in DreamGaussian-based REPARO) and provides excellent semantic consistency and spatial arrangement. While a resolution of 512×512 shows comparable performance, the marginal gains do not justify the significantly higher computational costs and longer optimization time. Given that 256×256 strikes a balance between performance and optimization efficiency, it is selected as the default resolution for REPARO, offering strong semantic alignment and consistent results with manageable resource requirements.

Effectiveness of differentiable method. As shown in Tab. 2, we implement a baseline using grid search optimization for individual asset layout adjustments. In this grid search, we explore a range of translations and scales defined by predetermined intervals, ensuring that each object's position and size were individually optimized. In both DreamGaussian-based and TripoSR-based setups, our differentiable method outperformed the grid search baseline, demonstrating its superior ability to align objects contextually within the scene. We also conduct a user study with 25 participants, gathering 350 responses that demonstrate our method's significant superiority over the grid search approach. Furthermore, the differentiable method optimizes 20 layouts in 1.5 hours, whereas the grid search method requires 4.5 hours and demonstrates lower accuracy.

Table 1. Comparison of different resolutions of renderer used in layout alignment.

(a) DreamGaussian-based REPARO.

Resolution	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
64 × 64	0.827	17.008	0.822	0.243
128 × 128	0.828	17.315	0.829	0.231
256 × 256	0.833	17.279	0.826	0.234
512 × 512	0.828	17.189	0.825	0.237

(b) TripoSR-based REPARO.

Resolution	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
64 × 64	0.821	17.769	0.865	0.214
128 × 128	0.821	17.758	0.865	0.217
256 × 256	0.822	17.751	0.865	0.216
512 × 512	0.823	17.824	0.865	0.215

Table 2. Comparison of optimization methods. HP† is the mean score of human preference.

(a) DreamGaussian-based REPARO.

Method	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓	HP† ↑
Grid search	0.820	17.974	0.850	0.206	21.15%
Differentiable	0.833	17.279	0.826	0.234	78.85%

(b) TripoSR-based REPARO.

Method	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓	HP† ↑
Grid search	0.818	17.911	0.866	0.210	20.58%
Differentiable	0.822	17.751	0.865	0.216	79.42%

Feature matching for layout initialization. We implement a simple feature matching initialization via DINO feature map grid search between GT and prediction and compare it with our learned layout alignment. As shown in Table 3, the performance is slightly worse than the original method without initialization, as feature matching introduces unnecessary bias.

Table 3. Comparison of the effect of initialization based on feature matching.

(a) DreamGaussian-based REPARO.

Init.	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
✓	0.818	17.262	0.824	0.235
	0.833	**17.279**	**0.826**	**0.234**

(b) TripoSR-based REPARO.

Init.	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
✓	0.818	17.329	0.863	0.221
	0.822	**17.751**	**0.865**	**0.216**

Ablation of semantic loss term. We conduct an ablation study for the L_s loss with a VGG [8] backbone using both DreamGaussian-based and TripoSR-based methods. As shown in Tab. 4, the results demonstrate that the DINO-v2 backbone outperforms the VGG backbone in both configurations. This advantage may come from DINO-v2's self-supervised learning method and large amounts of training data.

Impact of loss weight λ on performance. Fig. 2 include results adjusting the weight λ of the semantic loss.

Table 4. Comparison of different L_s backbone.

(a) DreamGaussian-based REPARO.

L_s Backbone	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DINOv2	**0.833**	**17.279**	**0.826**	**0.234**
VGG	0.831	17.268	0.825	**0.234**

(b) TripoSR-based REPARO.

L_s Backbone	CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DINOv2 [7]	**0.822**	**17.751**	**0.865**	**0.216**
VGG [8]	0.820	17.748	**0.865**	0.217

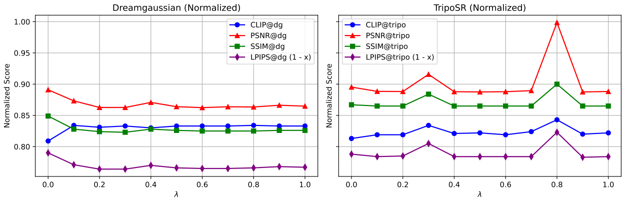

Figure 2. Ablation results of hyperparameter λ.

Effectiveness of semantic loss. Fig. 3 shows qualitative results comparing $\mathcal{L}_a$, $\mathcal{L}_s$, and $\mathcal{L}_a+\mathcal{L}_s$. $\mathcal{L}_s$ consistently improves spatial alignment and semantic plausibility although quantitative gains in Table 2 appear marginal. These qualitative examples support the necessity of including $\mathcal{L}_s$.

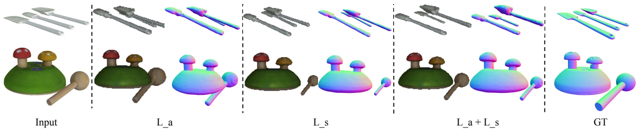

Figure 3. Qualitative visualization of appearance loss and semantic loss.

5. More qualitative results

We present additional qualitative results that are not provided by the GSO [3] dataset. As illustrated in the Figure 4, these qualitative examples are selected from the benchmark provided by ComboVerse [1]. During layout alignment, the translation and rotation parameters of the assets are optimized. The final results demonstrate that REPARO achieves excellent spatial arrangement and semantic consistency.

Real-world photograph inputs. As shown in Figure 5, we test our method on realistic cases from the ScanNet++ dataset [10], including complex photographs like "*a desktop with two remote controls*" and "*an office desk with two monitors and two keyboards*". The results show that our method effectively handles real-world inputs and accurately generates 3D layouts from photographs.

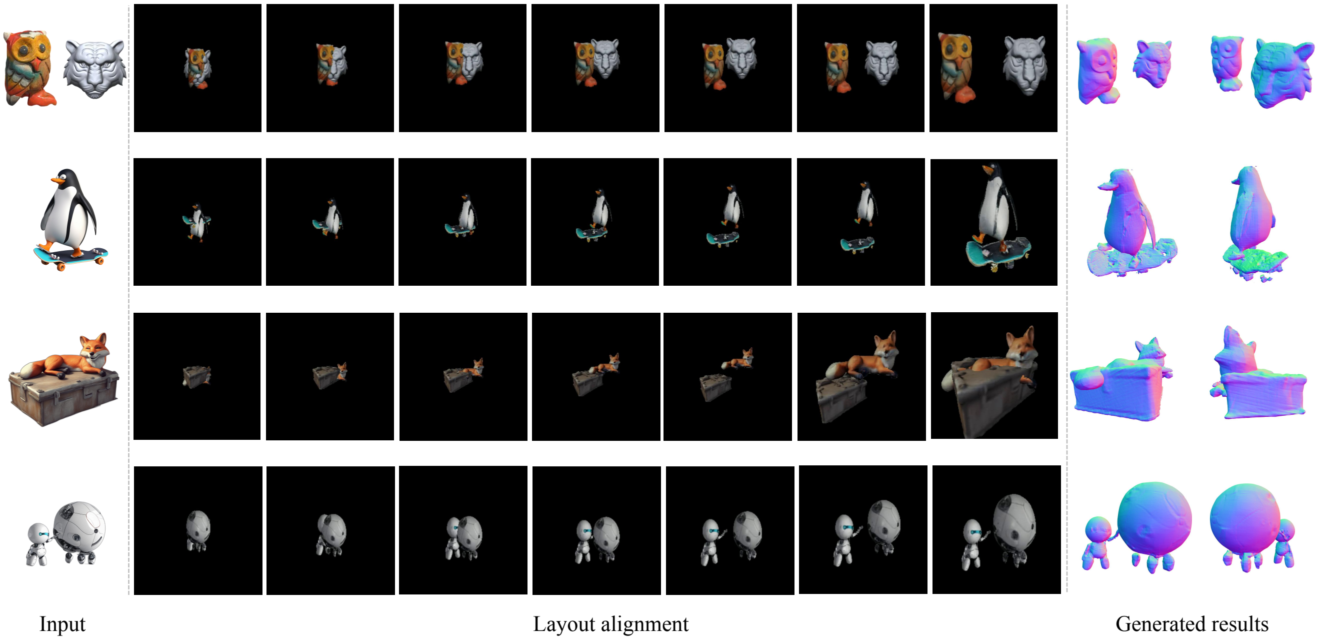

Figure 4. Selected qualitative results from ComboVerse [1] benchmark.

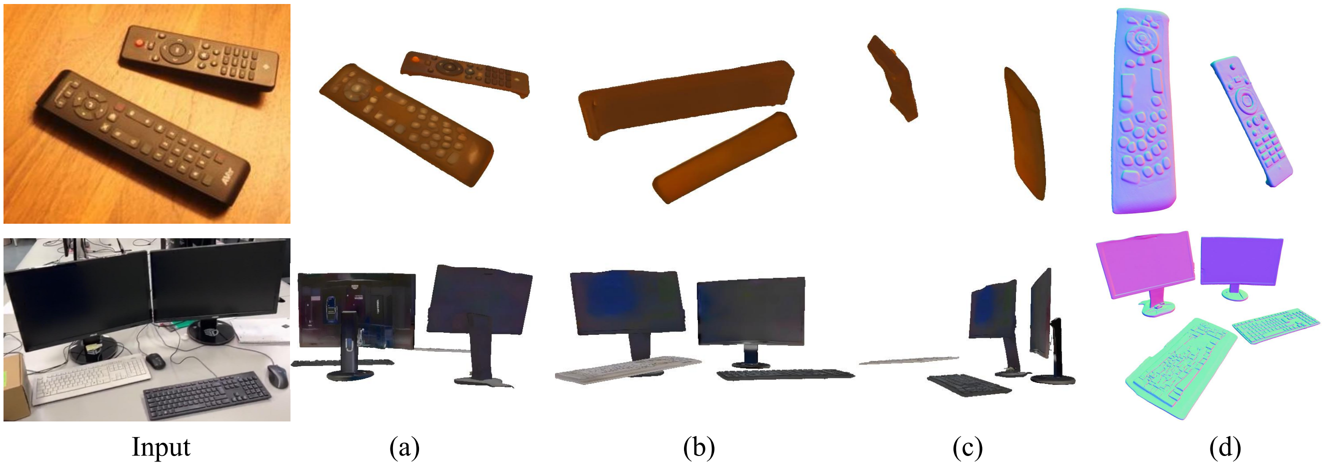

Figure 5. Results of REPARO on the Scannet++ dataset.

References

[1] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. *arXiv preprint arXiv:2403.12409*, 2024. 2, 3

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 1

[3] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2

[4] Jean Feydy, Joan Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[5] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1

[6] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 1

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[9] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1

[10] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2