TSLP: Text driven and style transferable indoor furniture layout generation pipeline

Zhiguo Xu[1,*], Jian Zhang[1,*], Shenghao Yang[1], Xiaoyan Liu[1,✉], Mingbo Zhao[1]

[1] School of Information and Intelligent Science, Donghua University, Shanghai 201620, China
* Equal contribution

Abstract. With the rising demand for interior design, existing 3D generation methods face a dual dilemma. Retrieval-based frameworks lack stylistic steerability due to restricted asset diversity whereas holistic image-to-3D synthesis yields monolithic representations that hinder instance-level control. To this end, we propose TSLP, a text driven and style transferable indoor furniture layout generation pipeline. The innovation of the framework lies in unifying instance-level control with aesthetic style customization. Specifically, we initially employ structured prompts to guide a diffusion model in generating interior images, achieving control over the categories, features, and spatial layouts of individual furniture instances. Subsequently, addressing the limits of text in describing complex aesthetics, TSLP enables visual style customization using a single reference image. Finally, decoupled 3D assets are reconstructed through pose estimation, and high-fidelity textures are baked using a texture synthesis model. Experimental results demonstrate that TSLP establishes a complete workflow from customized layout to aesthetic customization, achieving customized 3D furniture layout generation while significantly lowering the design threshold for users.

Keywords: 3D Generation, furniture Layout Generation, Style Transfer, Text-driven Generation, Instance-level Control

1 Introduction

In real life, an indoor scene is physically realized by a sequential arrangement of objects within a bounded indoor space [26]. In practice, an effective text-to-3D interior furniture generation model must possess strong natural language understanding capabilities, as well as the ability to generate individual 3D assets and textures, and place them in specific coordinates, which is a highly challenging task.

There has already been some research in this field. For example, existing retrieval-based methods rely on matching input features with a model database to place furniture at calculated coordinates [35, 51]. Their limitations are evident: the diversity of the generated results is restricted, making it difficult to meet the varied demands of diverse application scenarios. With the continuous development of the 3D generation field, several very powerful general-purpose 3D generation models have recently emerged[40, 50]. While they possess strong capabilities for generating 3D models from images, they

typically can only generate holistic 3D furniture layout assets based on input images. Due to their holistic nature, users are largely unable to fine-tune the results using 3D industry software. Of course, there are also some cutting-edge research efforts in this field that combines the strengths of both approaches: they separate individual pieces of furniture from input images for 3D generation and calculate the corresponding coordinates for placement [13, 33]. However, these methods still face two major issues: first, they can only faithfully reproduce the information in the image and lack editability. Second, users cannot independently select furniture styles. Newer 3D style transfer tasks [19]generate 3DGS formats, which cannot be adjusted within industrial software.

We propose TSLP, a text-driven and style transferable indoor furniture layout generation pipeline. A 3D furniture layout is defined as a structured spatial configuration comprising the mesh geometry, three-dimensional position, orientation, and semantic category of each furniture instance. Our approach initially employs structured prompts to guide a diffusion model in generating interior images, achieving control over the categories and layouts of individual furniture instances. Addressing the limitations of language in describing complex aesthetics, TSLP integrates an image-driven style transfer module to enable global aesthetic customization via a single reference image. Subsequently, individual furniture masks are extracted through object detection and semantic segmentation, with a deduplication module ensuring the uniqueness of scene modeling. These decoupled instance images are fed into pose estimation and 3D generation modules to reconstruct texture-free assets with their respective coordinates. Finally, high-fidelity six-view textures are generated and baked onto the assets using a texture synthesis model, establishing a complete 3D generation workflow from customized layout to aesthetic stylization. Since the final output consists of multiple independent objects, users can fine-tune prompts for specific items to achieve instance-level intervention of the generated results.

In summary, our main contributions are as follows:

- We propose TSLP, a text-driven 3D furniture layout generation pipeline that converts natural language into 3D furniture layouts and completes the placement.
- We incorporate a style transfer module, allowing users to specify the style of the 3D furniture layout based on reference images to meet personalized needs.
- We adopt an approach that jointly reconstructs 3D geometry and spatial coordinates, supporting prompt adjustment for individual objects to achieve fine-grained control over the generation of each furniture piece.

2 Related Work

2.1 3D Diffusion-based Generation

Benefiting from advanced 3D Variational Autoencoders (VAEs) such as 3DShape2VecSet [45], diffusion models have achieved remarkable progress in 3D content creation. The emergence of large-scale, high-quality datasets[4] has further catalyzed a wave of state-of-the-art frameworks.

These methodologies can be broadly categorized into two classes: the former class is two-stage decoupled generation, which decomposes synthesis into sequential geometric and textural phases, typically generating untextured meshes via ShapeVAE and DiT before synthesizing high-resolution UV maps through image-conditioned modules[50]. The other class is unified latent space generation, which jointly models geometry and appearance within a shared 3D latent space to recover textured representations via differentiable decoders[40]. While these works primarily focus on object-centric synthesis, their advancements in 3D diffusion provide a robust foundation for scene-level generation.

In our pipeline, we strategically leverage texture synthesis models to apply high-fidelity textures to pre-positioned untextured assets.

2.2 3D Indoor Scene Generation

3D indoor scene Generation aims to construct interior environments that are physically plausible, practically functional, and aesthetically satisfactory, leveraging input configurations that may be explicit or implicit and encode domain-specific priors, subject to various spatial and semantic constraints[48]. Related Research in Computer Graphics has witnessed substantial progress, in content creation and understanding in particular, driven by the convergence of massive visual datasets and deep learning innovations [26].

Existing approaches generally fall into four paradigms based on their underlying scene representations[39]: rule-driven procedural synthesis[35], native 3D neural generation[1, 25], 2D-to-3D reconstruction via image synthesis [9, 13, 33], and spatiotemporal video generation[17]. Within the context of 2D-to-3D generation, it is worth noting that several preliminary works [5] focus specifically on generating interior image from text. Furthermore, while some recent pioneering works[8]have achieved end-to-end text-to-3D scene generation, they inherently lack the capability to incorporate reference images for style transfer. Scene stylization pipelines such as [49] accept style images as guidance, but stylize objects through cascaded per-object texture optimization without text-driven instance-level layout control.Such coherent layouts also benefit embodied downstream tasks such as vision-and-language navigation [10, 46].

Although our proposed pipeline employs a 2D-to-3D methodology as an intermediate bridge, the framework as a whole represents a text-driven, style-transferable system. While 3DGS excels in rendering[37, 41], its incompatibility with standard industrial software limits its practical utility in design workflows. Consequently, we commit to a mesh-based framework that successfully accomplishes the end-to-end generation task from textual descriptions to stylized 3D furniture layouts, addressing a gap in existing methodologies.

2.3 Style Transfer

Evolution of 2D Neural Style Transfer. 2D style transfer evolved from intensive iterative optimization [7]to efficient feed-forward networks[15]. Feature alignment techniques later enabled zero-shot arbitrary stylization [11, 18].With diffusion models, the

paradigm shifted toward conditional generation. Moving away from time-consuming fine-tuning[32], recent works favor tuning-free adapters for image prompt injection [43]or explicit feature decoupling[38].

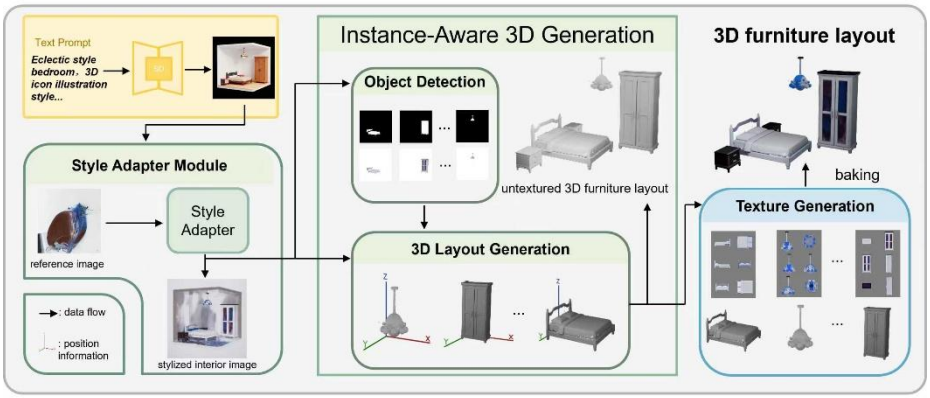

Fig. 1. Overview of the proposed framework. Given a text prompt and a reference image, a Style Adapter first generates a stylized interior image. The Instance-Aware 3D Generation (I3DG) module first identifies individual furniture entities through object detection, and subsequently leverages 3D Layout Generation Module to decompose the layout into precise object-level representations. Finally, a Texture Generation Module synthesizes and bakes the textures to produce the complete 3D furniture layout.

Extension to 3D and Dynamic Scenes. Benefiting from mature 2D generative priors, 3D stylization has rapidly advanced from single objects to complex scenes[3], primarily falling into two categories.

Joint optimization of explicit/implicit representations. Early methods optimized explicit meshes [23]or point clouds [2]. Beyond mere geometry editing, 3DStyleNet[44] introduced a joint optimization framework that achieves dual transfer of 3D geometric deformation and texture features via part-aware affine transformations. The rise of NeRFs introduced per-scene optimization [12]and feed-forward models [21]. Recently, 3D Gaussian Splatting improved both rendering quality and real-time performance [16, 20].

Multimodal-driven stylization. Enabled by multi-view diffusion [14] and SDS [29], text-driven 3D stylization has gained traction.

Despite progress in direct 3D stylization, the complexity of 3D spaces makes stable, high-quality generation challenging. Therefore, our method leverages mature 2D style transfer. By executing high-quality style translation in the 2D domain first and effectively mapping it to the subsequent 3D modeling stage, we stably produce the desired 3D stylization results.

3 Method

We propose TSLP: Text driven and style transferable indoor furniture layout generation pipeline, shown in Fig .1. Our TSLP framework takes two primary inputs: a text prompt describing furniture attributes and spatial relationships, and a stylistic feature extracted from a reference image. Based on these, it synthesizes a collection of high-fidelity 3D furniture layouts.

3.1 Style and Textual Feature Processing Module

Currently, mainstream text-to-2D-image models predominantly adopt SD-like architectures. We adopt a diffusion-based framework building upon Seedream 4.0[34], which comprises an efficient and scalable DiT backbone and a robust VAE with a high compression ratio. This architecture effectively transforms textual information into high-fidelity interior images. This design considerably reduces the number of image tokens in the latent space, resulting in a significant reduction in both training and inference FLOPs.

Such exceptional generative capability effectively translates textual information into the image domain, thereby achieving dimensional alignment with RGB style images. Consequently, this alignment facilitates the fusion of style information and textual information via an SD-Adapter-like architecture.

The style transfer module is built upon the SDXL [28] framework, achieving tuning-free style transfer by integrating IP-Adapter [43] and ControlNet[47]. Specifically, the reference image is first encoded into global features, and the style and content are then explicitly decoupled by subtracting the content textual features. Based on the fine-grained analysis of U-Net attention layers introduced by InstantStyle[38], these decoupled style features are exclusively injected into the transformer block to strictly govern the color, texture, and atmosphere.

By feeding both the generated interior image t and the reference style image s into the style transfer adapter, we obtain the stylized interior image L_{style} . Overall, this module presents an effective mechanism for fusing stylistic and textual information, which can be formulated as

$$L_{style} = \text{Adapter}\big(s, SD(t)\big) \tag{1}$$

where SD denotes the SD-like diffusion model and $Adapter(\cdot)$ represents the style transfer adapter. Through this proposed mechanism, textual features and visual information are successfully coupled.

3.2 Instance-Aware 3D Generation (I3DG) module

Reconstruction Object Detection Module. To specify the furniture objects $\mathcal{M}$ for downstream modules, the pipeline first employs an object detection stage to identify targeted furniture assets within the stylized interior image t_s. To enhance cross-scene adaptability and framework flexibility, we optionally employ Qwen3.5[54] to perform

category generalization z_{label} based on the provided furniture types. When enabled, this approach eliminates the need for predefined static labels, allowing the generalized semantic information to be fed into the detector as dynamic target anchors.

Given the inherent challenges of furniture occlusion and spatial overlap in the interior image, we utilize the Intersection over Union (IoU) metric to prune redundant bounding boxes. When the overlap ratio between detected objects exceeds a predefined threshold, the redundant candidate is discarded to maintain spatial integrity. Subsequently, a semantic segmentation module identifies the target objects based on the detection priors, generating binary mask z_{mask}. The segmentation results are then used to extract the corresponding RGB images of the furniture z_{rgb}, which serve as inputs for the subsequent generative modules.

$$z_{rgb}, z_{mask} = \text{seg}(\text{detect}(t_s, z_{label}), z_{label}) \tag{2}$$

where $detect(\cdot)$ denotes the object detection module and $\text{seg}(\cdot)$ denotes the semantic segmentation module.

3D Layout Generation Module. The module, which incorporates the Multi-Instance Diffusion (MIDI[13]) framework, extends pre-trained single-object generative models to compositional 3D scene generation, enabling an end-to-end mapping from a single image to multi-instance 3D scenes.

Conditioning Encoding. For each instance z, we construct a 7-channel composite conditional input $y_i \in R^{h \times w \times 7}$ by concatenating z_{rgb}, z_{mask}, and L_{style} along the channel dimension. A DINO v2 encoder is utilized to process these composite inputs, effectively fusing global context with local details. The resulting features are then injected into the denoising network via a cross-attention mechanism.

Parallel Denoising and Spatial Relationship Modeling. The geometric representations of all instances are encoded into low-dimensional latent tokens $\{z_i\}_{i=1}^{N}$ via a VAE. Distinct from conventional object-by-object generation, this paradigm employs a Rectified Flow architecture to perform parallel denoising on N instances within a shared Diffusion Transformer (DiT) module[27]. We incorporate a Multi-Instance Attention mechanism that adapts the standard self-attention into a cross-instance variant, facilitating spatial reasoning across different furniture entities. This allows the Query of each instance to aggregate Keys and Values from all instances within the scene

$$f_i^{out} = \text{Multi-Instance-Attention}\left(f_i, \{f_j\}_{j=1}^{N}\right) \tag{3}$$

where f_i is the feature of instance i. Consequently, spatial dependencies between objects are implicitly modeled during the denoising process, eliminating the need for explicit post-processing pose optimization.

Global Coordinate Generation. Upon completion of the denoising process, the latent tokens z_i are reconstructed into Signed Distance Function (SDF) values or triplane features via the VAE decoder, followed by mesh extraction using Marching Cubes. Crucially, the decoded 3D instances are positioned within a global scene coordinate system,

specifically a normalized space $[-1,1]^3$, ensuring consistent relative spatial configurations. This design ensures that the decoded 3D instances are directly output with precise relative positions, maintaining rigorous scene-level geometric consistency

$$\{\mathcal{M}_i\}_{i=1}^{N} = \{\text{MIDI}\left(\text{Enc}\left(\text{Cat}\left(z_{rgb}, z_{mask}, t_s\right)\right)\right)\}_{i=1}^{N} \tag{4}$$

where M_i denotes the untextured mesh of instance i, $\text{Enc}(\cdot)$ denotes the encoder module and $\text{MIDI}(\cdot)$ denotes the Multi-Instance Diffusion (MIDI[13]) framework.

3.3 Texture Generation Module

We adopt Hunyuan3D-Paint[50] as our texture generation module, which follows a three-stage framework for texture synthesis: preprocessing, multi-view generation, and texture baking. Specifically, the segmented RGB image z_{rgb} and the generated untextured mesh $\mathcal{M}_i$are fed into this module to synthesize the corresponding texture for each instance.

First, an Image Delighting module is employed to convert the input image into an unlit state. This preprocessing step effectively eliminates the interference of illumination and shadows on the intrinsic texture. Subsequently, a geometry-aware greedy strategy is utilized to select 8 to 12 optimal viewpoints. This process initializes with 4 fixed orthogonal viewpoints and iteratively supplements them via greedy selection to maximize the UV coverage area on the mesh surface.

During the generation phase, a Dual-stream Conditional Diffusion architecture is adopted. The reference branch injects reference image details using frozen Stable Diffusion (SD) 2.1[31] weights and original VAE features at timestep $\mathrm{t} = 0$, serving as a soft regularization to prevent distribution shift. Concurrently, the generation branch concatenates the geometric features, which are derived from canonical Normal Maps and Canonical Coordinate Maps (CCM), with latent noise, and injects learnable camera embeddings to facilitate multi-view diffusion. A Multi-task Attention mechanism is then employed to execute reference attention and multi-view attention in parallel, achieving a joint constraint that preserves fine image details while ensuring cross-view consistency.

Finally, multi-view images are generated through Dense-view Inference. Following single-image super-resolution enhancement, these images are projected onto the UV space. For unassigned regions on the mesh, a KD-Tree is constructed to query the textures of nearest neighbor vertices, which are then interpolated using Inverse Distance Weighting (IDW) to yield a seamless texture map. The entire texture mapping process for the scene instances can be formulated as

$$\{\mathcal{M}_{i^{tex}}\}_{i=1}^{N} = \{\Phi_{texture}\left(z_{rgb}^{i}, \mathcal{M}_i\right)\}_{i=1}^{N} \tag{5}$$

where $\mathcal{M}_{i^{tex}}$ denotes the textured mesh of instance i, $\Phi_{texture}$ denotes the texture generation module.

4 Experimental Setup

4.1 Implementation Details

TSLP is implemented based on two critical functional modules, 3D generation and style transfer. We leverage a SD-like diffusion model [34] to generate 2D images from the text descriptions. These images are resized to 512×512 before being fed into the image encoder, DINOv2[24]. For the 3D generation backbone, we adopt the Rectified Flow Architecture, which is centered around a denoising Transformer network consisting of 21 attention modules, which is a design previously validated in various 3D generation tasks [13]. The style transfer module is built upon the SDXL architecture, achieving training-free style transfer by integrating IP-Adapter and ControlNet. We adopt Grounding DINO[22] for object detection to generate text-driven bounding box proposals. For both our pipeline and the MIDI[13] baseline, object category labels are generated using Qwen3.5, ensuring consistent semantic grounding across methods. The number of generated instances per scene is likewise unrestricted for both methods.

Additionally, to achieve more accurate extraction of individual furniture items, we introduce SAM 2.1[30] and an IoU-based anchor box optimization strategy, which effectively reduces the overlap rate of anchor boxes and enhances the precision of semantic segmentation. Our task is strictly defined as indoor furniture layout generation, and planar background structures (e.g., walls and floors) are deliberately avoided, since such static architectural elements only introduce unnecessary geometric ambiguity. Any wall instances produced are identified via a fixed bounding box volume ratio threshold, applied uniformly to all scenes and both methods, and removed prior to computing the evaluation metrics. The entire pipeline can be executed on a single RTX 4090 GPU with 24GB of VRAM for typical scenes in our evaluation.

Table 1. Statistical summary of the indoor scene dataset across various room categories.

Room Type	Rooms	Furniture	Avg. Furniture
Bedroom	35	176	5.0
Dining Room	43	336	7.8
Living Room	43	548	12.7
Total	121	1060	8.8

Data. We performed additional processing on the 3D-FRONT dataset [6] as curated by DiffuScene[36] to construct the dataset for our method. The dataset provides detailed descriptions of individual furniture pieces and precise spatial coordinate distributions. Specifically, we constructed structured prompt templates for image generation and employed Qwen3.5[54] to populate them with furniture attributes such as style and relative positions, forming complete text inputs primarily in Chinese, with a small number of examples in English. All prompts shown in the figures are presented in English.This approach significantly facilitates the generation of unique and specific natural language descriptions for interior image.

All quantitative and qualitative results in this paper are based on 121 scenes used exclusively for evaluation, comprising 35 bedrooms, 43 dining rooms, and 43 living rooms, with statistics summarized in Table 1. A cross-check of our scene IDs against MIDI's official split [13], in which the last 1,000 rooms of the official scene list form the test set, shows that 34 of these scenes, approximately 28%, lie in its training portion. Since both systems share identical frozen MIDI weights, this overlap affects both equally, but absolute scores may be optimistically biased, as further discussed in Sec. 5.1. Style reference images are curated from WikiArt [55], covering Realism, Abstract Expressionism, Magic Realism, and Abstract Art.

Baselines. Previous works predominantly focused on generating furniture placement parameters and populating layouts using existing 3D assets. In contrast, our method not only optimizes furniture placement but also possesses the capability to generate 3D furniture assets autonomously. Given the unique challenges of furniture layout generation, we selected representative State-of-the-Art (SOTA) models in the current 3D generation field as baselines: TRELLIS[40], which utilizes structured latent representations. Hunyuan3D 3.1[53], which adopts a two-stage generation paradigm. And MIDI[13], which focuses on multi-instance attention.

To the best of our knowledge, TSLP is the first pipeline unifying text-driven instance-level layout control, reference-image-based style transfer, and decoupled 3D asset generation in a single workflow. Consequently, we show its effectiveness primarily through highly faithful furniture layouts and highly consistent style transfer results.

Metrics. Since our framework focuses on instance synthesis without explicitly generating structural elements like floors or walls, we argue that Object Overlapping Rate (OOR) is a more appropriate metric than Object Out-of-Boundary (OOB) for evaluating our layout quality. We adopt two distinct formulations: (1) OOR_{IL3D}[52] , defined as the ratio of intersection volume to total volume, and (2) $OOR_{LLplace}$[42] , which normalizes the intersection by the volume of the smaller object to more strictly penalize inter-object nesting. As presented in Table 2, although the overlapping rate naturally increases in denser scenes with more furniture, our method consistently achieves a lower OOR compared to the MIDI baseline on this evaluation set. Given the same frozen MIDI backbone, this indicates that our added modules improve spatial separation even in highly cluttered scenes, rather than establishing absolute generalization performance.

Table 2. Quantitative comparison with MIDI on the bedroom (B), living room (L), and dining room (D) subsets. Both systems share identical frozen MIDI weights, as detailed in Sec. 4.1.

Method	OOR_{IL3D} ↓	$OOR_{LLplace}$ ↓
MIDI(B)	3.01	4.00
Ours(B)	**0.47**	**1.76**
MIDI(D)	7.23	3.46
Ours(D)	**4.89**	**2.57**
MIDI(L)	8.43	3.23
Ours(L)	**7.10**	**3.10**

4.2 Stylized 3D Generation

As observed from the comparison in Fig. 2, the texture features and color distributions of the reference style images are successfully integrated into the target scenes. This transformation process does not compromise the structural integrity of the images because the geometric boundaries between objects and the background remain distinct. This preservation ensures that objects in the stylized scenes can still be accurately identified by object detection and semantic segmentation algorithms, which further validates the effectiveness of our style transfer module.

The visual results are illustrated in Fig. 2. To evaluate the robustness of TSLP, we selected four artistic styles for testing, including Realism, Abstract Expressionism, Magic Realism, and Abstract Art.

4.3 Non-stylized 3D Generation

Qualitative comparisons are shown in Fig. 3. As demonstrated in the visual results, scene-level 3D generation such as TRELLIS and Hunyuan3D 3.1 produce monolithic, non-decomposable scenes. Despite their holistic appearance, these models often exhibit noticeable topological distortions (e.g., merged furniture boundaries), color deviations and rendering artifacts that fail to faithfully represent the input. While object-level 3D generation like MIDI facilitate independent asset manipulation, the synthesized objects often exhibit fragmented geometries, noticeable spatial overlaps and inconsistent coloring. These issues are primarily rooted in imprecise semantic segmentation and the inherent limitations of their rendering modules.

In contrast, TSLP employs a high-precision object detection and semantic segmentation module specifically to ensure the geometric integrity of each furniture instance. As demonstrated in Fig. 3, these modules effectively reduce the boundary fusion and fragmented meshes prevalent in baseline methods. Simultaneously, a dedicated rendering pipeline Hunyuan3D-Paint is utilized to enhance color fidelity, ensuring that the synthesized furniture layout exhibit realistic texture.

4.4 Text-driven 3D generation

Our system features a text-driven pipeline for indoor furniture layout generation. Benefiting from the object-level generation mechanism, we can achieve relatively fine manipulation of individual objects by fine-tuning specific descriptions within the input prompts. As shown in Fig .4, by modifying only the clauses pertinent to a specific item, such as its spatial prepositions or material attributes, our model manages to maintain relatively stable spatial consistency of the remaining $N-1$ objects while achieving controlled variations in the target object's position, category, or topological structure. This demonstrates the method's certain robustness in handling complex instructions and provides some reference for interactive scene design.

4.5 Ablation Study

We conducted further ablation experiments on our dataset to verify the effectiveness of the key improvements made over the MIDI baseline. Specifically, we evaluated the following two core components: 1) the object detection and semantic segmentation module, and 2) the texture rendering module.

Fig. 2. 3D furniture layout generation results of TSLP under different art styles.

The object detection and semantic segmentation module. We compared our enhanced detection and segmentation module against the baseline to verify its impact on visual quality. As shown in Fig. 5, the precision of this module is a critical determinant of the final 3D asset generation.

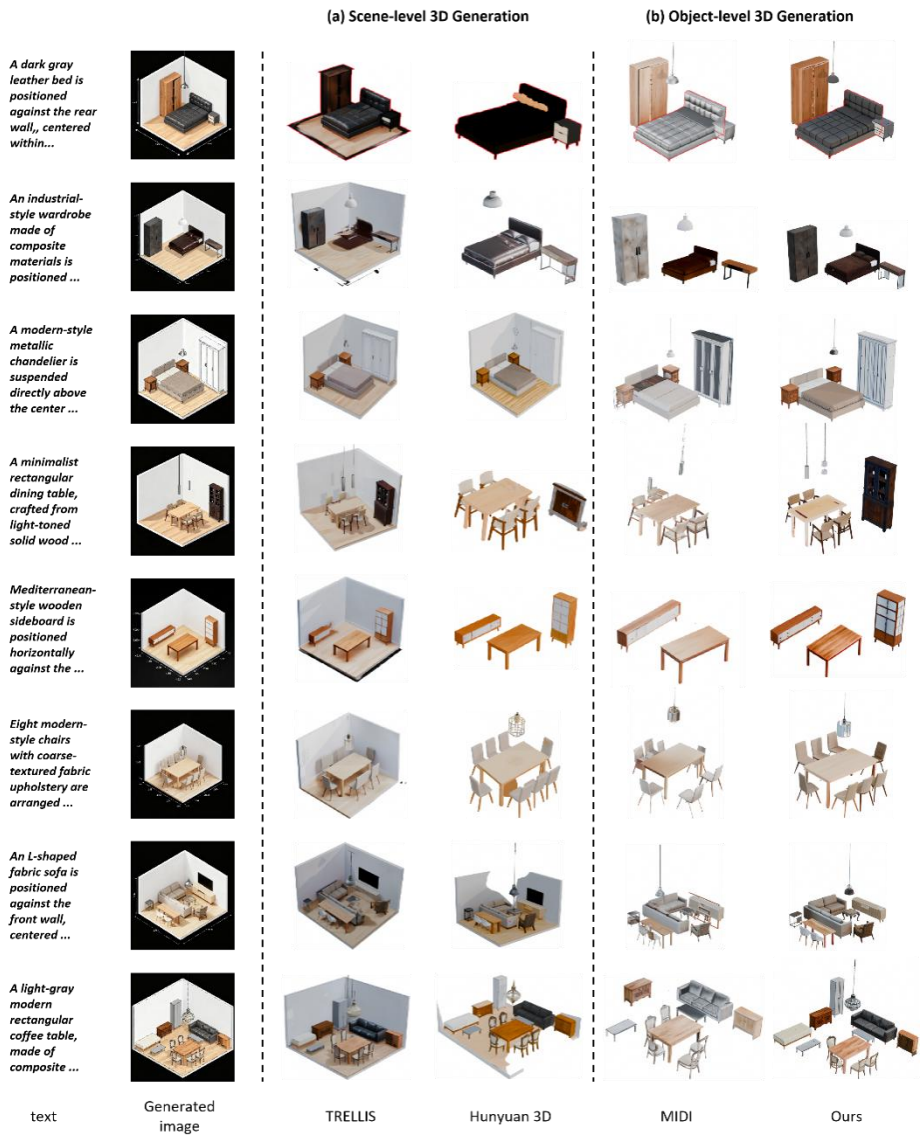

Fig. 3. Qualitative comparisons of different 3D furniture layout generation methods. (a)Scene-level 3D Generation, which outputs a holistic but non-decomposable scene. (b)Object-level 3D Generation, which facilitates independent asset generation and flexible asset placement.

The baseline approach treats each mask independently and handles overlapping regions through simple coverage. This often leads to redundant detections of a single object, resulting in fragmented or point-like masks that severely degrade the 3D synthesis. Furthermore, the baseline lacks a filtering mechanism for noise, causing the model to erroneously generate non-existent objects.

Fig. 4. Qualitative results of text-driven 3D layout generation and fine-grained control.

The texture rendering module. We reconstructed the original texture pipeline used in MIDI to evaluate the fidelity of our integrated rendering module. The fundamental difference between the two approaches lies in their architectural design.

As demonstrated in Fig. 3, the visual results generated by our method exhibit significantly higher fidelity to the original source images. By utilizing a coupled generation approach, our pipeline effectively reduces the misalignment and seam artifacts that often occur during the independent projection phase of decoupled methods. Consequently, the synthesized 3D assets maintain better chromatic consistency and finer textural details.

5 Conclusion

5.1 Limitations and Future Work

Due to computational constraints, evaluation is limited to 121 scenes, a portion of which overlaps with MIDI's public training split as disclosed in Sec. 4.1. This overlap was unintentional and was identified only after the completion of our experiments, and we acknowledge the oversight. Validation on fully held-out scenes, together with larger-scale and cross-dataset evaluation, is left to future work. Scalability is another concern, as inference time and memory footprint increase with more furniture instances. Generation quality may degrade in densely populated layouts, and such scenes may also require hardware with larger memory capacity.

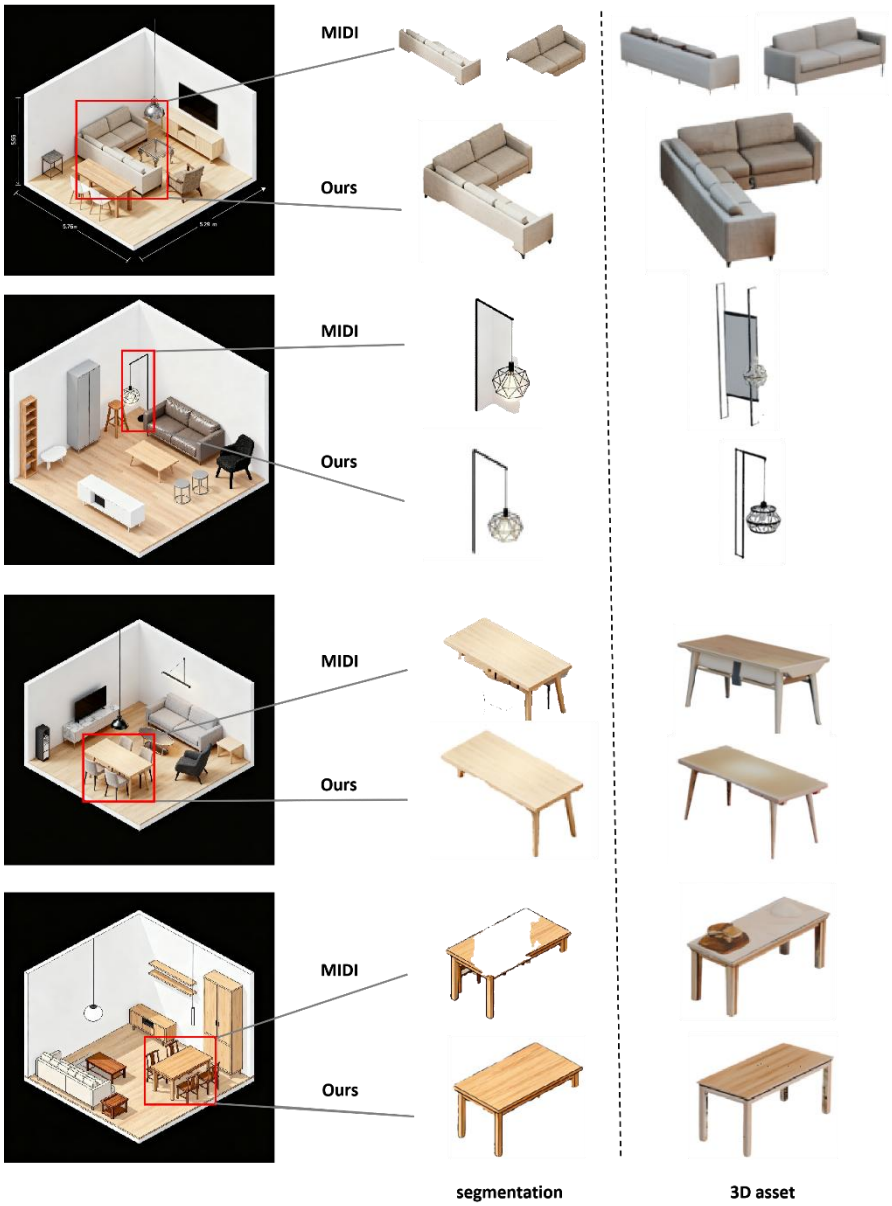

Fig. 5. Qualitative comparisons on the detection and segmentation modules for the final 3D asset generation.

Positional editing of individual furniture items is also limited, as the text-to-image model does not always follow spatial instructions faithfully. The method may also produce spurious walls and suffer from orientation misalignment. Moreover, as generation is not reconstruction, no paired ground truth exists for metrics such as CD or LPIPS,

and our preliminary user studies remained inconclusive due to uncontrolled biases in the study design. Large-scale quantitative comparisons (e.g., FID, CD, LPIPS) and rigorous user studies are thus left to future work.

5.2 Conclusion

In this paper, we propose TSLP, a text driven and style transferable pipeline for indoor 3D furniture layout generation. By bridging text-to-image synthesis, style transfer, and instance-level 3D reconstruction, the proposed framework enables the generation of decomposable 3D furniture layouts with controllable per-instance furniture geometry, spatial arrangement, and visual style. We note that TSLP is primarily a system-level integration of existing modules. Its contribution lies in workflow design rather than a single novel method.

Experimental results show that TSLP can produce spatially coherent layouts with reduced object overlap relative to the baseline on our evaluation set, while maintaining consistency with the reference style. The modular design of the pipeline further allows flexible integration of existing generative components, providing a practical solution for controllable 3D content creation. Overall, this work highlights the feasibility of incorporating visual style into structured 3D scene generation and provides a step toward more controllable and adaptable 3D layout synthesis frameworks.

References

1. Bucher, M.J., Armeni, I.: ReSpace: Text-Driven Autoregressive 3D Indoor Scene Synthesis and Editing, http://arxiv.org/abs/2506.02459, (2026).
2. Cao, X. et al.: PSNet: A Style Transfer Network for Point Cloud Stylization on Geometry and Color. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3326–3334 IEEE, Snowmass Village, CO, USA (2020). https://doi.org/10.1109/WACV45572.2020.9093513.
3. Chen, Y. et al.: Advances in 3D Neural Stylization: A Survey, http://arxiv.org/abs/2311.18328, (2024).
4. Deitke, M. et al.: Objaverse-XL: A Universe of 10M+ 3D Objects. In: Oh, A. et al. (eds.) Advances in Neural Information Processing Systems. pp. 35799–35813 Curran Associates, Inc. (2023).
5. Eldesokey, A., Wonka, P.: BUILD-A-SCENE: INTERACTIVE 3D LAYOUT CON- TROL FOR DIFFUSION-BASED IMAGE GENERATION.
6. Fu, H. et al.: 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics.
7. Gatys, L.A. et al.: Image Style Transfer Using Convolutional Neural Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.265.

8. Gu, Z. et al.: ArtiScene: Language-Driven Artistic 3D Scene Generation Through Image Intermediary. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2891–2901 IEEE, Nashville, TN, USA (2025). https://doi.org/10.1109/CVPR52734.2025.00275.
9. Han, H. et al.: REPARO: Compositional 3D Assets Generation with Differentiable 3D Layout Alignment.
10. Huang, J. et al.: Instance-Aware Visual Language Grounding for Consumer Robot Navigation. IEEE Trans. Consum. Electron. 71, 4, 12519–12526 (2025). https://doi.org/10.1109/TCE.2025.3601582.
11. Huang, X., Belongie, S.: Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1510–1519 IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.167.
12. Huang, Y.-H. et al.: StylizedNeRF: Consistent 3D Scene Stylization as Stylized NeRF via 2D-3D Mutual Learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18321–18331 IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01780.
13. Huang, Z., et al.: MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23646–23657 (2025).
14. Huang, Z. et al.: MV-Adapter: Multi-View Consistent Image Generation Made Easy.
15. Johnson, J. et al.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B. et al. (eds.) Computer Vision – ECCV 2016. pp. 694–711 Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43.
16. Kim, G. et al.: FPGS: Feed-Forward Semantic-aware Photorealistic Style Transfer of Large-Scale Gaussian Splatting, http://arxiv.org/abs/2503.09635, (2025).
17. Li, R. et al.: 4K4DGen: Panoramic 4D Generation at 4K Resolution, http://arxiv.org/abs/2406.13527, (2024).
18. Li, Y. et al.: Universal Style Transfer via Feature Transforms.
19. Liu, H. et al.: Stylos: Multi-View 3D Stylization with Single-Forward Gaussian Splatting, http://arxiv.org/abs/2509.26455, (2026).
20. Liu, K. et al.: StyleGaussian: Instant 3D Style Transfer with Gaussian Splatting. In: SIGGRAPH Asia 2024 Technical Communications. pp. 1–4 ACM, Tokyo Japan (2024). https://doi.org/10.1145/3681758.3698002.
21. Liu, K. et al.: StyleRF: Zero-Shot 3D Style Transfer of Neural Radiance Fields. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8338–8348 IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.00806.
22. Liu, S. et al.: Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In: Leonardis, A. et al. (eds.) Computer Vision – ECCV 2024. pp. 38–55 Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72970-6_3.

23. Michel, O. et al.: Text2Mesh: Text-Driven Neural Stylization for Meshes. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13482–13492 IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01313.
24. Oquab, M. et al.: DINOv2: Learning Robust Visual Features without Supervision, http://arxiv.org/abs/2304.07193, (2024).
25. Paschalidou, D. et al.: ATISS: Autoregressive Transformers for Indoor Scene Synthesis.
26. Patil, A.G. et al.: Advances in Data-Driven Analysis and Synthesis of 3D Indoor Scenes. Comput. Graph. Forum. 43, 1, e14927 (2024). https://doi.org/10.1111/cgf.14927.
27. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182 IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.00387.
28. Podell, D. et al.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, http://arxiv.org/abs/2307.01952, (2023). https://doi.org/10.48550/arXiv.2307.01952.
29. Poole, B. et al.: DreamFusion: Text-to-3D using 2D Diffusion, http://arxiv.org/abs/2209.14988, (2022).
30. Ravi, N. et al.: SAM 2: Segment Anything in Images and Videos, http://arxiv.org/abs/2408.00714, (2024).
31. Rombach, R. et al.: High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01042.
32. Ruiz, N. et al.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.02155.
33. Sautter, T. et al.: 3D-RE-GEN: 3D Reconstruction of Indoor Scenes with a Generative Framework, http://arxiv.org/abs/2512.17459, (2025). https://doi.org/10.48550/arXiv.2512.17459.
34. Seedream, T. et al.: Seedream 4.0: Toward Next-generation Multimodal Image Generation, http://arxiv.org/abs/2509.20427, (2025).
35. Sun, F.-Y. et al.: LayoutVLM: Differentiable Optimization of 3D Layout via Vision-Language Models.
36. Tang, J. et al.: DiffuScene: Denoising Diffusion Models for Generative Indoor Scene Synthesis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20507–20518 IEEE, Seattle, WA, USA (2024). https://doi.org/10.1109/CVPR52733.2024.01938.
37. Tang, Z. et al.: GarmentGS: Point-Cloud Guided Gaussian Splatting for High-Fidelity Non-Watertight 3D Garment Reconstruction. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. pp. 2048–2052 ACM, Chicago IL USA (2025). https://doi.org/10.1145/3731715.3733478.

38. Wang, H. et al.: InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation, http://arxiv.org/abs/2404.02733, (2024).
39. Wen, B. et al.: 3D Scene Generation: A Survey, http://arxiv.org/abs/2505.05474, (2025).
40. Xiang, J. et al.: Structured 3D Latents for Scalable and Versatile 3D Generation.
41. Yang, S. et al.: Zero-1-to-3DGS: a Single Image to 3D Gaussian by Consistent Multi-view Generation. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 IEEE, Nantes, France (2025). https://doi.org/10.1109/ICME59968.2025.11209455.
42. Yang, Y. et al.: LLplace: The 3D Indoor Scene Layout Generation and Editing via Large Language Model, http://arxiv.org/abs/2406.03866, (2024).
43. Ye, H. et al.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, http://arxiv.org/abs/2308.06721, (2023).
44. Yin, K. et al.: 3DStyleNet: Creating 3D Shapes with Geometric and Texture Style Variations. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12436–12445 IEEE, Montreal, QC, Canada (2021). https://doi.org/10.1109/ICCV48922.2021.01223.
45. Zhang, B. et al.: 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. ACM Trans. Graph. 42, 4, 1–16 (2023). https://doi.org/10.1145/3592442.
46. Zhang, H. et al.: Indoor Scene Recognition in Vision-and-Language Navigation. IEEE Trans. Consum. Electron. 72, 1, 1216–1222 (2026). https://doi.org/10.1109/TCE.2025.3645249.
47. Zhang, L. et al.: Adding Conditional Control to Text-to-Image Diffusion Models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3813–3824 IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.00355.
48. Zhang, S.-H. et al.: A Survey of 3D Indoor Scene Synthesis. J. Comput. Sci. Technol. 34, 3, 594–608 (2019). https://doi.org/10.1007/s11390-019-1929-5.
49. Zhang, Y. et al.: Style-Consistent 3D Indoor Scene Synthesis with Decoupled Objects. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 IEEE, Rome, Italy (2025). https://doi.org/10.1109/IJCNN64981.2025.11228381.
50. Zhao, Z. et al.: Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation, http://arxiv.org/abs/2501.12202, (2025). https://doi.org/10.48550/arXiv.2501.12202.
51. Zhou, M. et al.: RoomCraft: Controllable and Complete 3D Indoor Scene Generation, http://arxiv.org/abs/2506.22291, (2025).
52. Zhou, W. et al.: IL3D: A Large-Scale Indoor Layout Dataset for LLM-Driven 3D Scene Generation, http://arxiv.org/abs/2510.12095, (2025).
53. https://3d.hunyuan.tencent.com/.
54. https://qwen.ai/blog?id=qwen3.5.
55. https://www.wikiart.org/.