SceneGenesis: 3D Scene Synthesis via Semantic Structural Priors and Mesh-Guided Video-Geometry Fusion

Yueming Zhao[1], Hongyu Yang[2,3,4,*], Di Huang[1]
[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of Artificial Intelligence, Beihang University, Beijing, China
[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China
{zhaoyueming, hongyuyang, dhuang}@buaa.edu.cn

Abstract

Generating high-quality, controllable, and structurally consistent 3D scenes in complex multi-object environments remains a fundamental challenge. We present **SceneGenesis**, a unified framework that synthesizes 3D scenes by combining semantic structural priors with mesh-guided video–geometry fusion. SceneGenesis first employs large language models to convert textual descriptions into category-aware object specifications, which are transformed into structured meshes using procedural approximations and pretrained asset generators, enabling precise layout control and scalable scene construction. To obtain rich and style-controllable appearances, SceneGenesis generates multi-view video representations conditioned on the initialized structure. A mesh-guided video–geometry fusion module then consolidates video evidence with mesh priors through *mesh-conditioned fragment initialization, progressive geometric refinement, and structure-aware optimization*, substantially improving global geometric fidelity and visual realism. Experiments demonstrate that SceneGenesis supports flexible style variation and object-level editing while achieving strong controllability, scalability, and structural quality.

Introduction

Large-scale 3D scene generation is a fundamental capability for virtual environments, simulation, robotics, and autonomous driving. Real-world applications demand diverse and semantically structured scenes with controllable layouts, scalable content generation, and consistent 3D geometry. However, achieving these goals simultaneously remains a long-standing challenge.

Current methods can be broadly categorized into three paradigms. *Rule-based approaches* (Feng et al. 2023; Hu et al. 2024; Li et al. 2024) leverage procedural modeling to generate structured layouts efficiently. These methods excel at ensuring geometric regularity and semantic plausibility but suffer from limited realism and diversity due to reliance on rigid templates and fixed asset libraries (Bucher and Armeni 2025). *Neural 3D generation methods* (Zhang et al. 2024; Yang et al. 2024b) synthesize objects or scenes directly from data, enabling high visual fidelity. However, their reliance on complex training pipelines (Zhou et al. 2024b) and dense supervision hampers scalability and interpretability in large-scale, multi-object scenes (Lu et al. 2024). *Image- and video-driven methods* (Yang et al. 2024a; Engstler et al. 2025; Yan et al. 2025) enhance scalability and reduce 3D annotation costs by generating scenes from 2D views, but the absence of strong geometric priors leads to inconsistent structures and weak semantic control (Gao et al. 2024; Yu et al. 2024; Chung et al. 2023).

To address these limitations, we propose **SceneGenesis**, a unified framework that systematically incorporates *semantic structural priors* into every stage of the generation pipeline, from initialization to synthesis to reconstruction. Our key insight is to maintain explicit object-level structure via mesh representations throughout the process, enabling fine-grained controllability, scalable content generation, and faithful 3D geometry.

We begin with a **semantic structural initialization module** that converts a 3D layout and style prompt into structured meshes. A large language model parses the layout into object-level descriptions, which are instantiated using a category-aware strategy: fine-grained objects *e.g.*, vehicles, furniture) are generated via pretrained mesh models , while large-scale elements (*e.g.*, buildings, terrain) are approximated procedurally. This produces semantically coherent and spatially organized scenes.

Next, the **geometry-conditioned video synthesis module** renders multi-view semantic and depth maps from the mesh, which are used to guide a pretrained video diffusion model. A consistency-guided latent fusion strategy further enhances temporal coherence in long sequences by adaptively blending overlapping content across views. The core contribution of our method lies in the **mesh-guided video-geometry fusion module**, which reconstructs a coherent 3D scene by aligning mesh anchors with multi-view video outputs. This is achieved through three complementary steps: *fragment initialization* aligns noisy depth clouds derived from videos to mesh structure; *progressive refinement* updates 3D Gaussians using mesh-conditioned features via cross-attention and residual MLPs to jointly refine geometry and texture; and *structure-aware optimization* enforces global geometric and photometric consistency.

Beyond high-quality reconstruction, our framework supports prompt-driven style variation and object-level editing

*Corresponding author

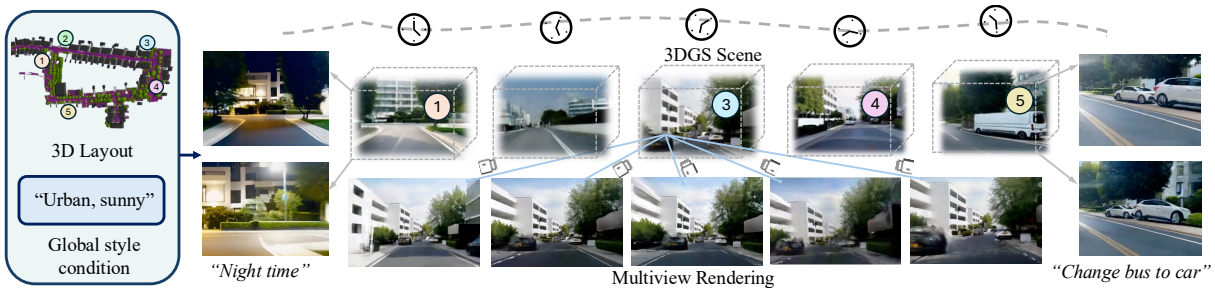

Figure 1: **Scene-level results of *SceneGenesis*.** From a 3D layout and a style prompt, our framework synthesizes structured mesh-based scenes and reconstructs coherent 3D representations via mesh-guided video-geometry fusion. Shown are multi-view renderings (center), global style variations (left), and object-level editing (right), highlighting controllability, scalability, and 3D structural consistency.

such as insertion and replacement (see Fig. 1). These capabilities highlight the controllability and structural awareness of our approach. Our contributions include:

- We propose **SceneGenesis**, a unified 3D scene generation framework that integrates semantic structural priors throughout the pipeline, achieving controllable layouts, consistent synthesis, and high-quality reconstruction with object-level editing and style variation.
- We develop a **geometry-conditioned video synthesis module** that combines fine-grained mesh priors and a consistency-guided latent fusion strategy for scalable, temporally consistent video generation.
- We introduce a **mesh-guided video-geometry fusion module** that fuses mesh constraints and multi-view video cues through fragment initialization, progressive refinement, and structure-aware optimization, enabling accurate, high-fidelity 3D scene reconstruction.

Related work

3D Representations

Scene representation is fundamental for 3D understanding and content generation. Explicit formats such as point clouds (Aliev et al. 2020) and meshes (Gkioxari, Malik, and Johnson 2019) are widely used for their interpretability and semantic compatibility; meshes in particular support category-aware modeling and direct spatial editing. However, their discrete nature limits fine-grained detail recovery in large, complex scenes and reduces sensitivity to 2D supervision. Implicit neural representations, such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2020), enable high-fidelity rendering from sparse views but incur high computational cost and are unsuitable for real-time applications. More recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) provides an efficient alternative with a good trade-off between visual quality and rendering speed (Huang et al. 2024). Yet, most existing representations lack strong structural controllability and are hard to integrate into multi-stage generation pipelines. In contrast, our framework leverages a structured mesh representation to enforce semantic consistency and spatially aligned editing, while remaining compatible with fast 3DGS rendering.

3D Scene Generation

Early 3D scene generation methods (Merrell, Schkufza, and Koltun 2010; Zhang et al. 2021) rely on procedural modeling with manually defined rules and parametric templates. While efficient and structurally regular, these approaches lack realism and adaptability due to rigid templates and limited asset diversity. Recent efforts incorporate large language models (Zhou et al. 2024a; Zhang et al. 2025) to improve semantic controllability, yet still depend on fixed rule engines and constrained object libraries, limiting their applicability to complex, open-world scenarios. Neural generative methods synthesize 3D scenes directly from data using representations such as voxels (Li et al. 2025), NeRF (Lu et al. 2024), and 3DGS (Zhou et al. 2024b). These approaches offer high visual fidelity and realism when trained on structured inputs like scene graphs or multi-view images. However, their heavy computation, poor scalability to multi-object environments, and weak controllability remain challenges (Lu et al. 2024; Zhou et al. 2024b). To enhance scalability and reduce annotation costs, image- and video-driven methods have emerged. Image-based approaches (Yu et al. 2024; Shriram et al. 2024) enable multi-view synthesis via iterative generation, but lack explicit geometric priors, leading to limited structural coherence and semantic alignment. Video-driven methods (Gao et al. 2024; Yan et al. 2025; Wen et al. 2024) leverage temporal continuity and generative diffusion models to produce stylistically diverse sequences, yet often suffer from geometric inconsistency, weak object-level control, and reliance on extensive pretraining or fine-tuning.

To address these limitations, our method introduces a unified mesh-guided framework that integrates semantic structure across multiple stages, enabling controllable, scalable, and structurally consistent 3D scene generation.

Method

We introduce *SceneGenesis*, a unified framework for large-scale 3D scene generation that systematically integrates mesh-based structural priors across all stages of the pipeline (Figure 2). Unlike prior work that treats structure and appearance as separate stages or lacks geometric control, our framework tightly couples their modeling. The pipeline comprises three key modules: (1) a **Semantic Structural**

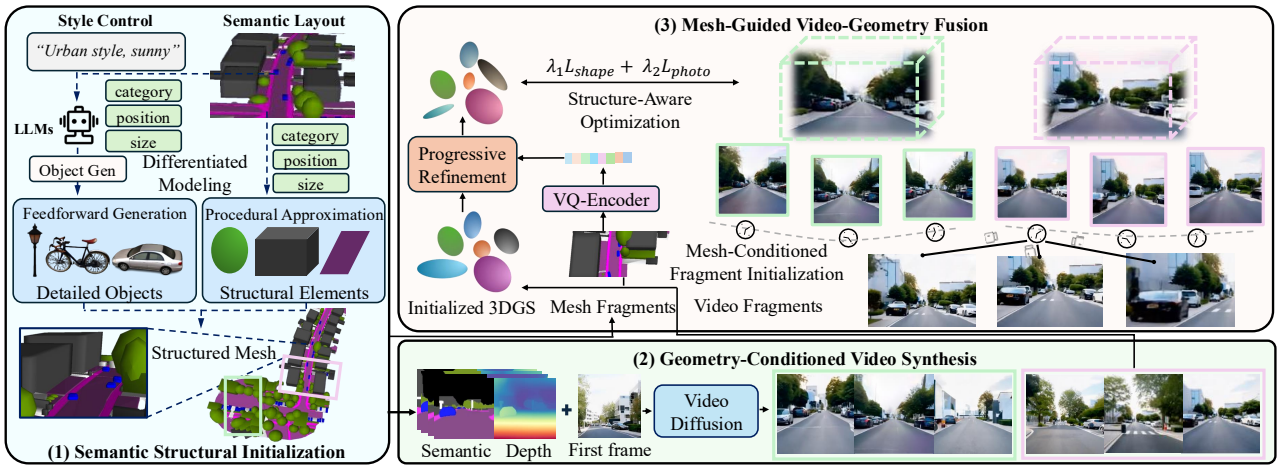

Figure 2: **Overview of the *SceneGenesis* framework.** Our method comprises three stages: (1) **Semantic Structural Initialization**, which parses a semantic layout and style prompt into object-level descriptions, then generates structured meshes via a category-aware strategy; (2) **Geometry-Conditioned Video Synthesis**, where rendered depth and semantic maps guide a pretrained video diffusion model to synthesize multi-view, style-controllable sequences; and (3) **Mesh-Guided Video-Geometry Fusion**, which aligns video-derived geometry with mesh priors through fragment initialization, cross-attentive refinement, and structure-aware optimization, enabling high-fidelity 3D reconstruction with both geometric accuracy and appearance realism.

Initialization module that transforms a 3D semantic layout and style prompt into a structured mesh scene, leveraging a category-aware modeling strategy to instantiate fine-grained assets and procedural structures based on object semantics; (2) a **Geometry-Conditioned Video Synthesis** module that renders semantic and depth maps from the mesh and conditions a pretrained video diffusion model to synthesize multi-view, style-consistent sequences without additional training; (3) a **Mesh-Guided Video-Geometry Fusion** module that reconstructs a coherent 3D Gaussian scene by aligning video-derived geometry with mesh priors via fragment initialization, progressive refinement, and structure-aware optimization.

Semantic Structural Initialization

We begin with a controllable semantic–structural initialization module that constructs semantically organized and spatially coherent 3D scenes, forming the foundation of our pipeline. Given a 3D semantic layout and a global style prompt, the module generates a mesh-based scene with object-level controllability and consistent stylistic structure. By adopting a category-aware modeling strategy, it yields high-quality object meshes and coherent overall layouts, providing strong structural priors for downstream video synthesis and 3D reconstruction.

Semantic Layout and Style Conditioning. The input to the initialization module consists of a semantic layout and a style prompt. The semantic layout is defined as a set of object-level parameters:

$$\mathcal{L} = \{(c_i, \mathbf{p}_i, \mathbf{s}_i)\}_{i=1}^{N}, \tag{1}$$

where c_i, $\mathbf{p}_i$, and $\mathbf{s}_i$ denote the object category, 3D position, and scale, respectively. A global style prompt S (*e.g.*, *urban, pastoral, snowy*) defines the overall visual theme of the scene.

To enrich object semantics, we leverage a pretrained large language model (LLM) (Achiam et al. 2023) to generate fine-grained natural language descriptions for each object:

$$d_i = \mathrm{LLM}(c_i, \mathbf{p}_i, \mathbf{s}_i, S), \tag{2}$$

where a natural language description d_i captures shape, material, color, function, and contextual cues. These descriptions serve as object-specific prompts for downstream 3D mesh modeling, enabling both semantic alignment and stylistic consistency at the object level.

Category-Aware Mesh Generation Strategy. To generate a scene that is both detailed and scalable, we introduce a category-aware modeling strategy that differentiates between fine-scale and structural elements.

For small and detail-rich objects (*e.g.*, vehicles, street lamps, furniture), we adopt a conditional feedforward generator to produce high-fidelity meshes. Specifically, we use TRELLIS (Xiang et al. 2025) as the mesh generator, which employs a structured latent representation combining sparse 3D grids and multiview visual features to decode into high-quality meshes with fine structural and appearance fidelity:

$$\mathcal{M}_i = \mathcal{G}_{\text{FF}}(d_i), \tag{3}$$

where $\mathcal{G}_{\text{FF}}$ indicates the generator, which directly maps each semantic description d_i to a high-resolution 3D mesh $\mathcal{M}_i$.

For large-scale structural elements (*e.g.*, buildings, terrain, vegetation), we apply procedural modeling to ensure regularity and spatial efficiency:

$$\mathcal{B}_i = \mathcal{G}_{\text{geo}}(c_i, \mathbf{p}_i, \mathbf{s}_i), \tag{4}$$

where $\mathcal{G}_{\text{geo}}$ derives basic geometric primitives from KITTI-360 bounding boxes, using object categories to parameterize coarse shapes via representative points and surfaces, yielding structurally plausible and semantically aligned approximations $\mathcal{B}_i$.

Finally, all generated components are spatially assembled to form the initial scene:

$$\mathcal{S}_{\text{init}} = \bigcup_{i=1}^{N} T(\mathcal{M}_i \cup \mathcal{B}_i; \mathbf{p}_i, \mathbf{s}_i), \tag{5}$$

where $T(\cdot)$ applies spatial transformations based on object position $\mathbf{p}_i$ and scale $\mathbf{s}_i$.

This structured mesh-based initialization ensures layout-level controllability, stylistic coherence, and scalability, laying a geometry-aware foundation for subsequent video synthesis and 3D reconstruction.

Geometry-Conditioned Video Synthesis

We develop a geometry-conditioned video synthesis module that generates photorealistic and spatially consistent scene videos under explicit mesh guidance. By conditioning a pretrained video diffusion model on rendered depth and semantic maps from the initialized mesh, this module enables geometry-aware, scalable video generation without additional training.

Structure-Aware Video Generation. To synthesize spatially aligned and stylistically coherent videos, we build upon Cosmos-Transfer (Alhaija et al. 2025), a pretrained diffusion model designed for generic video generation. Cosmos-Transfer adopts a Transformer-based denoising architecture and supports flexible conditioning at inference.

Given the initialized mesh scene $\mathcal{S}_{\text{init}}$, we render depth and semantic maps under specified camera views t (intrinsics and extrinsics) for each frame:

$$D_t, S_t = \mathcal{R}(\mathcal{S}_{\text{init}}, t), \tag{6}$$

where $\mathcal{R}(\cdot)$ denotes the rendering operation. These maps encode explicit structural and semantic information of the scene. The rendered maps are encoded into a conditioning token $\mathbf{c}_t$:

$$\mathbf{c}_t = \text{Encode}(D_t, S_t), \tag{7}$$

which is injected into the latent denoising process of Cosmos-Transfer via:

$$z_t' = \mathcal{T}(z_t, \mathbf{c}_t), \tag{8}$$

where z_t denotes the noisy latent and $\mathcal{T}(\cdot)$ represents the conditional fusion operator. This structure-aware conditioning ensures that the generated video respects the geometric layout and semantic attributes specified by the mesh while maintaining the style prompt.

Consistency-Guided Latent Fusion. Despite structural conditioning, frame-wise video generation with sliding windows may still introduce local flickering or style drift, especially at overlapping segments. To address this, we propose a *consistency-guided latent fusion* strategy that adaptively merges latent features based on temporal coherence, thereby ensuring smooth transitions and globally consistent appearance across long sequences.

Given n overlapping frames between the k-th and $k+1$-th sliding window, with latent codes $z_i^{(k)}$, $i = 1, ..., n$ and z_{target} denotes the latent code of the target frame in overlapping region, we compute similarity-based attention weights:

$$\alpha_i = \frac{\exp(-\lambda \| z_i^{(k)} - z_{\text{target}} \|)}{\sum_{j=1}^{n} \exp(-\lambda \| z_j^{(k)} - z_{\text{target}} \|)}, \tag{9}$$

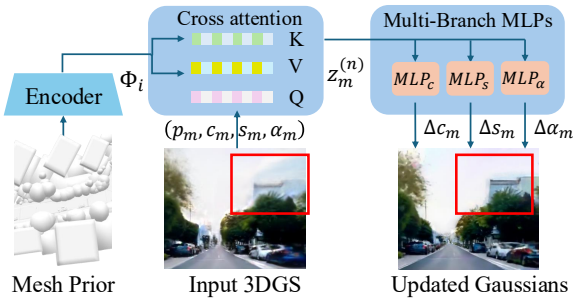

Figure 3: **Mesh-Guided Progressive Refinement.** Given initial 3D Gaussians and mesh priors, we extract mesh features using a VQ-encoder and fuse them with Gaussian point positions via a cross-attention mechanism. The fused features are processed by multi-branch MLPs to predict residuals for color, scale, and opacity.

and obtain the fused latent representation:

$$z_{\text{fused}} = \sum_{i=1}^{n} \alpha_i \cdot z_i^{(k)}, \tag{10}$$

where λ controls the sharpness of the similarity weighting. Higher weights are assigned to latents that are more temporally consistent with the reference, while inconsistent latents are downweighted. This inference-time mechanism requires no retraining or architectural modification of the pretrained video diffusion model. By integrating this fusion step, we significantly reduce visual artifacts at segment boundaries and enhance both the spatial coherence and stylistic continuity of long-range video generation.

Mesh-Guided Video-Geometry Fusion

We introduce a mesh-guided video-geometry fusion module that extends VideoLifter (Cong et al. 2025) by explicitly incorporating mesh priors into reconstruction. Unlike VideoLifter, which relies primarily on photometric cues from video frames, our approach integrates mesh information at three levels: (1) mesh-conditioned fragment initialization that anchors depth fragments to semantic surfaces; (2) refinement with VQ-encoded mesh features fused into Gaussian attributes via cross-attention; and (3) a structure-aware optimization that regularizes Gaussians toward mesh geometry. This enhances geometric guidance and yields more stable and consistent 3D reconstructions.

Mesh-Conditioned Fragment Initialization. We first generate an initial 3D Gaussian representation by aligning depth-derived point clouds with the coarse mesh structure. For each frame i, a camera-space point cloud $\mathbf{X}_i^{\text{cam}} \in \mathbb{R}^{N_i \times 3}$ is derived from its depth map and transformed into the global coordinate system $\mathbf{X}_i^{\text{world}}$ using camera extrinsics. To enforce spatial alignment, we uniformly sample semantic anchor points $\mathbf{X}_{\text{semantic}} \in \mathbb{R}^{N_s \times 3}$ from the surface of the initialized mesh. We then filter $\mathbf{X}_i^{\text{world}}$ by retaining only points within a fixed distance ϵ from the mesh surface:

$$\mathbf{X}_i^{\text{clip}} = \left\{ \mathbf{x}_k \in \mathbf{X}_i^{\text{world}} \;\middle|\; \min_{\mathbf{y} \in \mathbf{X}_{\text{semantic}}} \|\mathbf{x}_k - \mathbf{y}\|_2 < \epsilon \right\}. \tag{11}$$

This produces mesh-aligned 3D fragments $\mathbf{X}_i^{\text{clip}}$ that serve as robust initialization for the subsequent Gaussian representation. Unlike prior methods that rely on learned filtering or heuristics, our initialization directly anchors reconstruction to structured geometry.

Mesh-Conditioned Progressive Refinement. To enhance fidelity and coherence, we propose a progressive refinement that injects mesh-derived spatial priors into the optimization of 3D Gaussian parameters via multi-branch residual MLPs as Figure 3. Specifically, given the mesh $\mathcal{M}_i$ of a video fragment, we extract structured spatial features using the VQ-encoder f_{mesh}, capturing rich spatial context:

$$\Phi_i = f_{\text{mesh}}(\mathcal{M}_i) \in \mathbf{R}^{H \times W \times D}. \tag{12}$$

At the n-th refinement step, the current state of a Gaussian point is:

$$\mathbf{g}_m^{(n)} = (\mathbf{p}_m, c_m, s_m, \alpha_m), \tag{13}$$

where $\mathbf{p}_m$, c_m, s_m, and α_m denote the position, color, scale, and opacity, respectively. We use $\mathbf{p}_m$ to construct a Gaussian query vector $\mathbf{q}_m^{(n)}$, which is fused with mesh features via cross-attention:

$$\mathbf{z}_m^{(n)} = \text{CrossAttn}(\mathbf{q}_m^{(n)}, \Phi_i). \tag{14}$$

To improve specialization and representation power, we design three independent MLP branches to predict residuals for color, scale, and opacity:

$$\Delta c_m, \Delta s_m, \Delta \alpha_m = \text{MLP}(\mathbf{z}_m^{(n)}) \tag{15}$$

The Gaussian point attributes are updated as follows:

$$\begin{aligned} c_m^{(n+1)} &= c_m^{(n)} + \Delta c_m, \\ s_m^{(n+1)} &= s_m^{(n)} \cdot \exp(\Delta s_m), \\ \alpha_m^{(n+1)} &= \alpha_m^{(n)} + \Delta \alpha_m. \end{aligned} \tag{16}$$

This multi-stage refinement progressively enhances geometric detail and appearance realism while preserving the mesh-constrained structural layout. Detailed hyperparameter settings for the refinement module are provided in the appendix.

Structure-Aware Optimization. To further enforce geometric alignment and appearance consistency, we design a joint optimization objective that includes both structural and photometric terms.

Global Shape Consistency. We encourage the learned Gaussian distribution $\mathbf{G}$ to adhere to the mesh geometry by minimizing the distance to sampled mesh surface points $\mathbf{X}_{\text{mesh}}$, where $\mathbf{g}_k$ denotes the 3D position of the k-th Gaussian center:

$$\mathcal{L}_{\text{shape}} = \frac{1}{|\mathbf{G}|} \sum_{k=1}^{|\mathbf{G}|} \min_{\mathbf{x}_j \in \mathbf{X}_{\text{mesh}}} \|\mathbf{g}_k - \mathbf{x}_j\|_2^2. \tag{17}$$

Photometric Consistency. To fully leverage image supervision from the synthesized video, we optimize the learnable parameters of Gaussians and MLPs using a combination of L_1 loss and SSIM loss (Wang et al. 2004). For each camera view v, the rendered image $\hat{I}_v$ is supervised against the ground-truth I_v as:

$$\mathcal{L}_{\text{photo}} = \sum_{v=1}^{V} \left(\|\hat{I}_v - I_v\|_1 + \lambda_{\text{ssim}} \cdot \mathcal{L}_{\text{SSIM}}(\hat{I}_v, I_v) \right), \tag{18}$$

where λ_{ssim} balances the perceptual similarity constraint.

Final Objective. The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{shape}} + \lambda_2 \mathcal{L}_{\text{photo}}, \tag{19}$$

where λ_1 and λ_2 balance structural alignment and appearance fidelity.

This fusion module tightly integrates symbolic structure (from mesh priors) and visual realism (from video diffusion) in a joint optimization loop. Following the video fusion strategy of VideoLifter (Cong et al. 2025), all fragments are reconstructed and optimized independently, then merged using global registration. Spatial consistency and clustering in attribute space (*e.g.*, color, normals) are applied to align overlapping regions, yielding a coherent and redundancy-free scene.

Experiments

Experimental Setup

We conduct experiments on the KITTI-360 dataset (Liao, Xie, and Geiger 2022) for quantitative comparisons. KITTI-360 provides 3D bounding box annotations for multiple classes (*e.g.*, building, car, road, vegetation), forming extensive 3D scene layouts. Each layout is represented as a triangle mesh, enabling fast rendering of semantic and depth maps. For comprehensive evaluation, we use diverse scene layouts from KITTI-360 paired with various text prompts describing style and conditions. While most prompts depict urban scenes, we also include rural and natural settings as well as time-specific cues (*e.g.*, "sunny morning", "night city"), covering a broad range of appearances and scales to validate controllability, scalability, and generality. Our pipeline incorporates several pretrained modules: TRELLIS (Xiang et al. 2025) for category-aware 3D object mesh generation, ChatGPT-4 (Achiam et al. 2023) for fine-grained style prompt construction, and Cosmos-Transfer (Alhaija et al. 2025) as the base video diffusion model. All experiments are conducted on a single NVIDIA A800 GPU.

We compare SceneGenesis with three representative baselines: (1) Urban Architect (Lu et al. 2024), a NeRF-based generator trained with SDS (Poole et al. 2022); (2) GALA3D (Zhou et al. 2024b), which combines procedural layout generation with neural rendering; (3) Vista (Gao et al. 2024), a diffusion-based video synthesis model conditioned on layout and style.

Quantitative metrics include CLIP similarity (Radford et al. 2021) for text–image alignment. We further conduct a user study with 94 participants, who rated randomized multi-view results on a 5-point Likert scale for perceptual quality and prompt alignment.

Quantitative Results

Table 1 reports CLIP similarity and FID scores, averaged over 20 rendered views per scene. Our method achieves the highest CLIP score of 67.02, demonstrating strong semantic alignment with input prompts, and the lowest FID of 50.19, indicating superior visual fidelity. In comparison, Vista achieves a CLIP score of 65.23 and FID of 52.47, benefiting from pretrained diffusion priors but lacking explicit

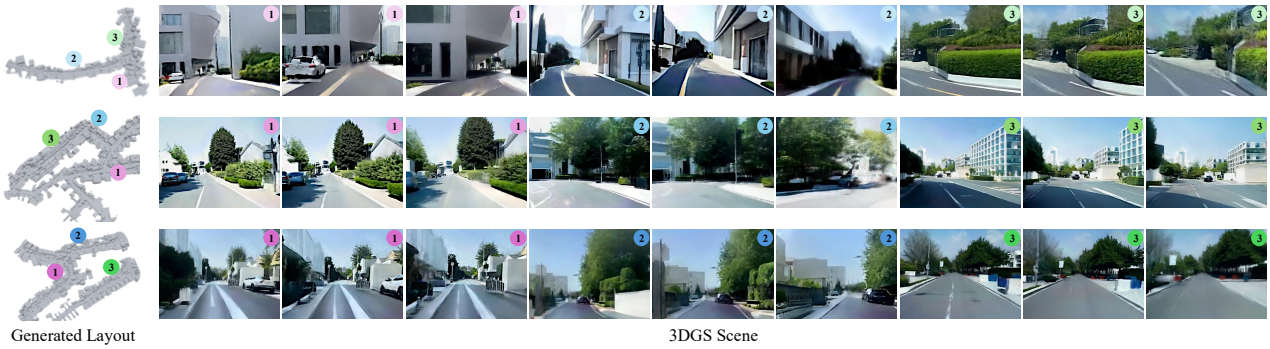

Figure 4: **Controllable and Scalable 3D Scene Generation.** We visualize our model's results on complex urban scenes. **Left:** Generated 3D layouts representing object-level controllable structures with semantic guidance. **Right:** Corresponding synthesized 3DGS scenes rendered from long video trajectories.

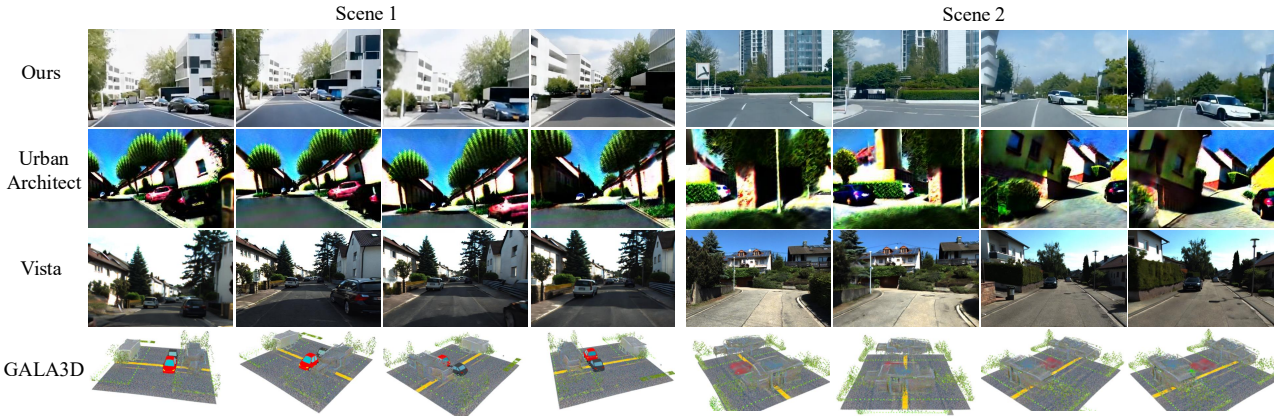

Figure 5: **Qualitative Comparison.** Visual results across two scenes comparing our method with Urban Architect, GALA3D, and Vista. Our approach achieves superior 3D consistency, semantic fidelity, and object-level controllability.

Method	CLIP Score ↑	FID (Kitti-360) ↓
Ours	**67.02**	**50.19**
Vista	65.23	52.47
Urban Architect	55.76	86.12
GALA3D	42.16	143.74

Table 1: **Quantitative Comparison.** CLIP measures text-image alignment; FID evaluates visual realism.

Method	Quality	Consistency
Vista	4.01	4.27
Urban Architect	3.26	3.76
GALA3D	2.67	3.05
Ours	**4.12**	**4.61**

Table 2: **User Study Results.** Mean scores from 94 participants on visual quality and prompt alignment.

3D structure modeling, leading to view inconsistency. Urban Architect yields a CLIP score of 55.76 and FID of 86.12, with oversaturated textures and blurred geometry due to weak structural priors. GALA3D performs the worst (CLIP 42.16, FID 143.74) as its rule-based modeling lacks flexibility and struggles with complex layouts.

The user study results in Table 2 show that our method attains the highest scores in both visual quality (4.12) and prompt consistency (4.61), confirming its ability to generate photorealistic and semantically faithful scenes. Vista achieves moderately strong ratings but lacks 3D alignment and fine-grained object-level control, while Urban Architect and GALA3D exhibit visual artifacts and slower synthesis performance.

Efficiency. For a typical scene with ∼1,000 objects and a 200-frame trajectory, semantic–structural initialization and mesh generation take about 20 minutes, geometry-conditioned video synthesis about 80 minutes, and 3D reconstruction about 50 minutes. The full pipeline completes in roughly 2.5 hours with peak memory usage below 30 GB, offering a favorable balance between quality and scalability relative to existing approaches.

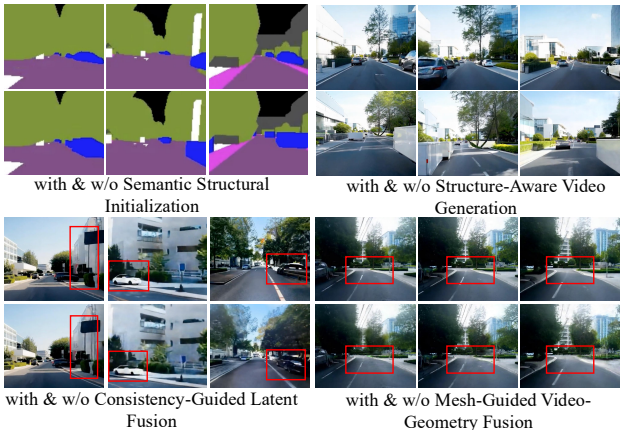

Figure 6: **Qualitative Results of Ablation Study.** We visualize the effects of ablating four key modules: Semantic Structural Initialization, Structure-Aware Video Generation, Consistency-Guided Latent Fusion, and Mesh-Guided Video-Geometry Fusion.

Configuration	CLIP Score
w/o Semantic Structural Initialization	25.73
w/o Structure-Aware Video Generation	25.81
w/o Consistency-Guided Latent Fusion	29.98
w/o Mesh-Guided Video-Geometry Fusion	29.52
Full model (Ours)	**30.23**

Table 3: **Quantitative Results of Ablation Study.** CLIP score comparison across ablated variants.

Qualitative Results

Figure 1 illustrates results on complex scenes, highlighting our framework's geometric consistency, semantic richness, and scalability. Figure 5 compares methods on two sequences (50 and 120 frames). Urban Architect produces blurry, unstable reconstructions; Vista offers consistent styles but shows spatial inconsistency and weak object control; GALA3D fails on complex layouts with fragmented geometry. In contrast, our method leverages mesh priors and multi-view video conditioning to preserve object count, structural integrity, and style coherence across views and time, enabling faithful, dynamic 3D scene synthesis.

Ablation Study

We conduct ablation studies to evaluate the contribution of each core module by systematically replacing them with corresponding base models. Results are shown in Figure 6, where the top row shows our full model and the bottom row displays the ablated variants. Replacing the *Semantic Structural Initialization* and *Geometry-Conditioned Video Synthesis* with raw KITTI-360 bounding boxes degrades object fidelity and layout control, *e.g.*, cars are incorrectly generated as cubes. Removing the *Consistency-Guided Latent Fusion* reverts to unmodified Cosmos-Transfer (Alhaija et al. 2025), introducing flickering and style drift across frames.

Figure 7: **Editing capabilities of SceneGenesis.** (a) Global style transfer (*e.g.*, urban to pastoral) and (b) object-level editing (*e.g.*, vehicle removal). These examples demonstrate fine-grained semantic control and structural consistency.

Without the *Mesh-Guided Video-Geometry Fusion*, the system reduces to VideoLifter (Cong et al. 2025), causing geometry misalignments and surface artifacts, *e.g.*, uneven road regions lacking coherence.

Quantitative results in Table 3 further support these findings: removing any single module leads to a clear performance drop, confirming that each component is essential for the controllability, consistency, and fidelity of our unified mesh-guided framework.

Scene and Object Editing

To demonstrate the controllability of our pipeline, we conduct editing experiments in two scenarios: global style manipulation and object-level replacement. For global editing, we modify the style token (*e.g.*, "urban" $\rightarrow$ "pastoral") during initialization. The entire scene adjusts accordingly, including vegetation, buildings, and lighting, while maintaining structural layout. For object-level editing, we selectively modify or replace the semantic description of individual objects (*e.g.*, delete "car"). The generated scene reflects the updated object attributes accurately, without affecting surrounding elements. These results validate our system's capacity for flexible and targeted control over both global and local aspects of the 3D scene, enabling scalable content authoring and scenario customization.

Conclusion

We present **SceneGenesis**, a unified framework for controllable, scalable 3D scene generation that combines semantic-driven initialization, geometry-conditioned video synthesis, and structure-aware reconstruction. Guided by a mesh-centric prior, SceneGenesis balances efficiency, 3D consistency, and fine-grained control. Experiments on diverse layouts show clear gains over baselines and support coherent scene generation with flexible editing. Despite improved 3D consistency, the reconstruction quality still depends on the reliability of synthesized videos and rendered depth, especially in dense scenes with heavy occlusions. Future work will explore tighter coupling between video generation and geometric regularization.

Acknowledgments

This work is partly supported by the National Key R&D Program of China (2022ZD0161902), and the Fundamental Research Funds for the Central Universities.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alhaija, H. A.; Alvarez, J.; Bala, M.; Cai, T.; Cao, T.; Cha, L.; Chen, J.; Chen, M.; Ferroni, F.; Fidler, S.; Fox, D.; Ge, Y.; Gu, J.; Hassani, A.; Isaev, M.; Jannaty, P.; Lan, S.; Lasser, T.; Ling, H.; Liu, M.; Liu, X.; Lu, Y.; Luo, A.; Ma, Q.; Mao, H.; Ramos, F.; Ren, X.; Shen, T.; Sun, X.; Tang, S.; Wang, T.; Wu, J.; Xu, J.; Xu, S.; Xie, K.; Ye, Y.; Yang, X.; Zeng, X.; and Zeng, Y. 2025. Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control. *arXiv preprint arXiv:2503.14492*.

Aliev, K. M.; et al. 2020. Neural Point-Based Graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 696–712.

Bucher, M. J.; and Armeni, I. 2025. ReSpace: Text-Driven 3D Scene Synthesis and Editing with Preference Alignment. *arXiv preprint arXiv:2506.02459*.

Chung, J.; Lee, S.; Nam, H.; Lee, J.; and Lee, K. M. 2023. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*.

Cong, W.; Zhu, H.; Wang, K.; Lei, J.; Stearns, C.; Cai, Y.; Wang, D.; Ranjan, R.; Feiszli, M.; Guibas, L.; et al. 2025. Videolifter: Lifting videos to 3d with fast hierarchical stereo alignment. *arXiv preprint arXiv:2501.01949*.

Engstler, P.; Shtedritski, A.; Laina, I.; Rupprecht, C.; and Vedaldi, A. 2025. Syncity: Training-free generation of 3d worlds. *arXiv preprint arXiv:2503.16420*.

Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36: 18225–18250.

Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37: 91560–91596.

Gkioxari, G.; Malik, J.; and Johnson, J. 2019. Mesh R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9785–9795.

Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D. A.; Schmid, C.; and Fathi, A. 2024. Scenecraft: An llm agent for synthesizing 3d scenes as blender code. In *Forty-first International Conference on Machine Learning*, 19252–19282.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference papers*, 1–11.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (SIGGRAPH)*, 42: 1–14.

Li, B.; Guo, J.; Liu, H.; Zou, Y.; Ding, Y.; Chen, X.; Zhu, H.; Tan, F.; Zhang, C.; Wang, T.; et al. 2025. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11971–11981.

Li, X.-L.; Li, H.; Chen, H.-X.; Mu, T.-J.; and Hu, S.-M. 2024. Discene: Object decoupling and interaction modeling for complex scene generation. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 1–12.

Liao, Y.; Xie, J.; and Geiger, A. 2022. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36: 3292–3310.

Lu, F.; Lin, K.-Y.; Xu, Y.; Li, H.; Chen, G.; and Jiang, C. 2024. Urban architect: Steerable 3d urban scene generation with layout prior. *arXiv preprint arXiv:2404.06780*.

Merrell, P.; Schkufza, E.; and Koltun, V. 2010. Computer-generated residential building layouts. In *ACM SIGGRAPH Asia 2010 Conference Papers*, 1–12.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, volume 65, 99–106.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763. PmLR.

Shriram, J.; Trevithick, A.; Liu, L.; and Ramamoorthi, R. 2024. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; and Zhang, X. 2024. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6902–6912.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21469–21480.

Yan, Y.; Xu, Z.; Lin, H.; Jin, H.; Guo, H.; Wang, Y.; Zhan, K.; Lang, X.; Bao, H.; Zhou, X.; et al. 2025. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 822–832.

Yang, S.; Tan, J.; Zhang, M.; Wu, T.; Li, Y.; Wetzstein, G.; Liu, Z.; and Lin, D. 2024a. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv preprint arXiv:2408.13252*.

Yang, X.; Man, Y.; Chen, J.; and Wang, Y.-X. 2024b. SceneCraft: Layout-guided 3D scene generation. *Advances in Neural Information Processing Systems*, 37: 82060–82084.

Yu, H.-X.; Duan, H.; Hur, J.; Sargent, K.; Rubinstein, M.; Freeman, W. T.; Cole, F.; Sun, D.; Snavely, N.; Wu, J.; et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6658–6667.

Zhang, Q.; Wang, C.; Siarohin, A.; Zhuang, P.; Xu, Y.; Yang, C.; Lin, D.; Zhou, B.; Tulyakov, S.; and Lee, H.-Y. 2024. Towards text-guided 3d scene composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6829–6838.

Zhang, S.-K.; Li, Y.-X.; He, Y.; Yang, Y.-L.; and Zhang, S.-H. 2021. Mageadd: Real-time interaction simulation for scene synthesis. In *Proceedings of the 29th ACM international conference on multimedia*, 965–973.

Zhang, Y.; Li, Z.; Zhou, M.; Wu, S.; and Wu, J. 2025. The scene language: Representing scenes with programs, words, and embeddings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24625–24634.

Zhou, M.; Wang, Y.; Hou, J.; Zhang, S.; Li, Y.; Luo, C.; Peng, J.; and Zhang, Z. 2024a. SceneX: Procedural Controllable Large-scale Scene Generation. *arXiv preprint arXiv:2403.15698*.

Zhou, X.; Ran, X.; Xiong, Y.; He, J.; Lin, Z.; Wang, Y.; Sun, D.; and Yang, M.-H. 2024b. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. arXiv:2402.07207.