A Comprehensive Survey of Interaction Techniques in 3D Scene Generation

Yuqi Li[1] , **Siwei Meng**[2] , **Chuanguang Yang**[3] , **Weilun Feng**[3] , **Junming Liu**[4] , **Zhulin An**[3] , **Yikai Wang**[5] and **Yingli Tian**[1]
[1]The City College of New York, City University of New York, USA
[2]Department of Computer Science, University of Nevada, Reno, USA
[3]Institute of Computing Technology, Chinese Academy of Sciences, China
[4]Tongji University, China
[5]Beijing Normal University, China
yikaiw@bnu.edu.cn, ytian@ccny.cuny.edu

Abstract

3D scene generation has rapidly evolved, significantly promoting the innovation of content creation. In this context, interaction techniques serve as a pivotal bridge connecting user intent with the generative models, thereby enabling precise control, real-time feedback and personalized customization of complex 3D scenes. Existing literature reviews predominantly focus on general generative paradigms, or are limited to specific subdomains such as single-object modeling, while often overlooking the systematic classification of interaction mechanisms. To bridge this gap, this work presented a comprehensive survey of interaction techniques in 3D scene generation. We proposed a unified taxonomy that categorized existing methods into three primary paradigms: Interactive Generation, Interactive Editing, and Embodied Interaction. For each category, we analyzed representative methods in terms of controllability, interaction granularity, and physical consistency, and discussed their advantages and limitations. We further summarized commonly used datasets and evaluation protocols for interactive 3D scene generation. Finally, we discussed the future directions toward more physically grounded, multi-modal, and user-centered interactive 3D scene generation systems. A curated list of the related papers mentioned in this work can be found at Awesome-Interactive-Techniques-in-3D-Scene-Generation-Lists.

1 Introduction

Recent advancements in generative models have transformed the 3D scene generation from a fully manual design task into a data-driven synthesis process. Instead of explicitly constructing every scene component, users only need to provide partial information and rely on generative models to infer the remaining structure [Lin *et al.*, 2023; Wen *et al.*, 2025a]. This transition fundamentally changes the role of the user, shifting scene generation from a static modeling task to an interactive process that requires continuous user input and feedback. Interactive 3D scene generation refers to the mechanisms in which users inject their intentions, constraints, or corrective feedback into the scene synthesis process. Unlike the traditional modeling workflows, where users explicitly specify geometry and parameters, interaction in generative systems often operates at different abstraction levels. User input involves encoding high-level semantics, partial structural information, or localized editing operations, while the model is responsible for translating these signals into coherent 3D representations.

Existing studies defines interaction according to the locus of user intervention within the workflow, and formulating interaction as the condition setting during the scene generation[Wen *et al.*, 2025a]. While a dominant paradigm approaches interaction as generative conditioning, distinct methodologies prioritize interactive editing to refine scene geometry and semantics after synthesis. Furthermore, emerging research [Fan *et al.*, 2025] extends the interaction towards action-level modeling in embodied interaction, emphasizing the generation of physically consistent human–scene and human–object interactions in evolving environments.

Despite the rapid proliferation of research across these interaction paradigms, the literature remains fragmented, lacking a unified framework to organize these diverse techniques. As summarized in Table 1, existing surveys predominantly focus on broad generative architectures [Wen *et al.*, 2025a], such as the transition from NeRF to 3D Gaussian Splatting, or are limited to specific subdomains such as 3D representation [Wang, 2024] and human motion [Xue *et al.*, 2025]. Crucially, these reviews often overlook the interaction mechanisms themselves, failing to systematically analyze how user intent is encoded, processed, and integrated into the 3D synthesis pipeline. Consequently, there is a lack of comprehensive guidance on selecting appropriate interaction techniques for different stages of the 3D creation pipeline.

This survey aims to present a comprehensive and unified view of interaction techniques in 3D scene generation. We propose a structured taxonomy that organizes existing methods into the three aforementioned paradigms: **(1) Interactive Scene Generation**, which unifies semantic-driven generation grounded in multi-modal directives with structure-guided generation constrained by explicit spatial layouts; **(2) Interactive Scene Editing**, which focuses on post-synthesis refinement

Category	Works	Focused Tasks	Date
Broad Review	[Wen *et al.*, 2025a]	3D Scene Generation	May 2025
	[Miao *et al.*, 2025]	4D Generation	Mar 2025
Subdomain Review	[Wang, 2024]	3D Representation	Oct 2024
	[Fan *et al.*, 2025]	Human Interaction	Mar 2025
	[Xue *et al.*, 2025]	Human Motion Video	July 2025
Broad Review	**Ours**	**Interaction Techniques**	**Jan 2026**

Table 1: Comparison between this work and previous surveys. Our survey is the first to systematically categorize and analyze interaction techniques across the entire 3D generation lifecycle.

through high-level semantic modification or precise geometry-aware manipulation; and **(3) Embodied Interaction**, which extends the scope to dynamic action model, emphasizing physically plausible human–scene and human-object interactions within the synthesized environment. Beyond this taxonomy, we critically analyze the technical foundations and evolutionary trajectories within each category. We also summarize essential datasets and evaluation benchmarks tailored for interactive 3D scene tasks.

Scope of This Survey. In this survey, we provide a comprehensive overview of interaction techniques for 3D scene generation, covering areas of generation, editing, and human-scene interaction. First, we introduce core concepts of common 3D representation methods, generative models, and problem formulation in interaction 3D scene (Section 2). Then, we classify recent research into three main paradigms: Interactive 3D Scene Generation (Section 3.1), Interactive 3D Scene Editing (Section 3.2), and Embodied Interaction (Section 3.3). For each paradigm, we review representative methods and analyze their interaction mechanisms, control characteristics, and technical distinctions. We then summarize commonly used datasets and evaluation metrics for interactive 3D scene generation (Section 4). Finally, we discuss open challenges and future research directions in Section 5.

2 Background

2.1 3D Scene Representations

3D representation fundamentally determines the fidelity of generation and granularity of user interaction. We categorize existing representations into explicit, implicit, and hybrid paradigms, analyzing their modelings and applications in terms of editability and generative consistency.

Explicit Representations. Explicit representations define geometry directly in Euclidean space. **Polygonal Meshes** [Wang, 2024] allow for precise manipulation of topology and texture, making them the standard for graphics pipelines and downstream editing. However, generating meshes with coherent topology remains challenging for neural networks due to their discrete nature. **Point Clouds** [Wang, 2024] offer a flexible, unstructured format that is easy to generate via regression but lacks surface connectivity, complicating physical simulation and high-quality rendering. **Voxel Grids** [Wang, 2024; Jiang *et al.*, 2024] discretize space into regular cubic volumes , simplifying generation into a classification task. Yet, their cubic memory complexity limits resolution, making them unsuitable for large-scale detailed scene interaction.

Implicit Representations. Implicit methods represent scenes as continuous functions parameterized by neural networks. **Neural Radiance Fields (NeRF)** [Mildenhall *et al.*, 2021] map spatial coordinates and viewing directions (x, d) to density and color (σ, c), enabling photorealistic view synthesis. **Signed Distance Fields (SDF)** [Park *et al.*, 2019] map coordinates to the distance to the nearest surface $s(x)$, facilitating precise geometric operations and surface extraction. While implicit representations excel in visual fidelity, they entangle scene attributes within network weights, making localized interaction computationally expensive and difficult without explicit decoupling.

Hybrid Representations. To bridge the gap between explicit controllability and implicit rendering quality, **3D Gaussian Splatting (3DGS)** [Kerbl *et al.*, 2023] has emerged as a dominant paradigm. 3DGS represents a scene as a set of anisotropic particles with center μ, covariance Σ, color c, and opacity α, allowing for real-time selection and transformation, while the differentiable rasterization pipeline ensures NeRF-level rendering quality. This duality makes 3DGS particularly advantageous for real-time interactive generation and editing.

2.2 Deep Generative Models

Generative models serve as the computational engine for synthesizing 3D structures from user intents. Early interactive systems predominantly leveraged **Generative Adversarial Networks (GANs)** [Goodfellow *et al.*, 2014] due to the efficiency. By optimizing a minimax process between a generator and a discriminator, GANs synthesize 3D content in a single forward pass, enabling millisecond-level feedback essential for real-time exploration. However, GANs often struggle with training instability and mode collapse, limiting their ability to capture the complex and multi-modal distributions required for diverse scene generation. Generative models have shifted towards **Diffusion Models (DMs)** [Ho *et al.*, 2020], which formulate generation as an iterative denoising process. While this iterative nature incurs higher inference computational costs, it unlocks superior generation quality and training stability. Diffusion models also promote progressive interaction models [Rombach *et al.*, 2022]. User signals, whether text prompts, coarse layouts, or sparse strokes, can be injected at different noise levels to improve the global structure or local details. This granularity of control makes diffusion the standard for modern interactive systems, driving parallel research into distillation techniques to alleviate reasoning bottlenecks and approach real-time performance.

2.3 Problem Formulation

We formally define the problem of Interactive 3D Scene Generation, where $\mathcal{S}$ denotes the space of 3D scene representations. The goal is to learn a generative model Φ that synthesizes a scene $S \in \mathcal{S}$ conditioned on a multi-modal user control signal C. At step t, the user observes the current state S_{t-1} and provides an interaction signal u_t. The model updates the scene:

$$S_t = \Phi(S_{t-1}, u_t, \theta) \tag{1}$$

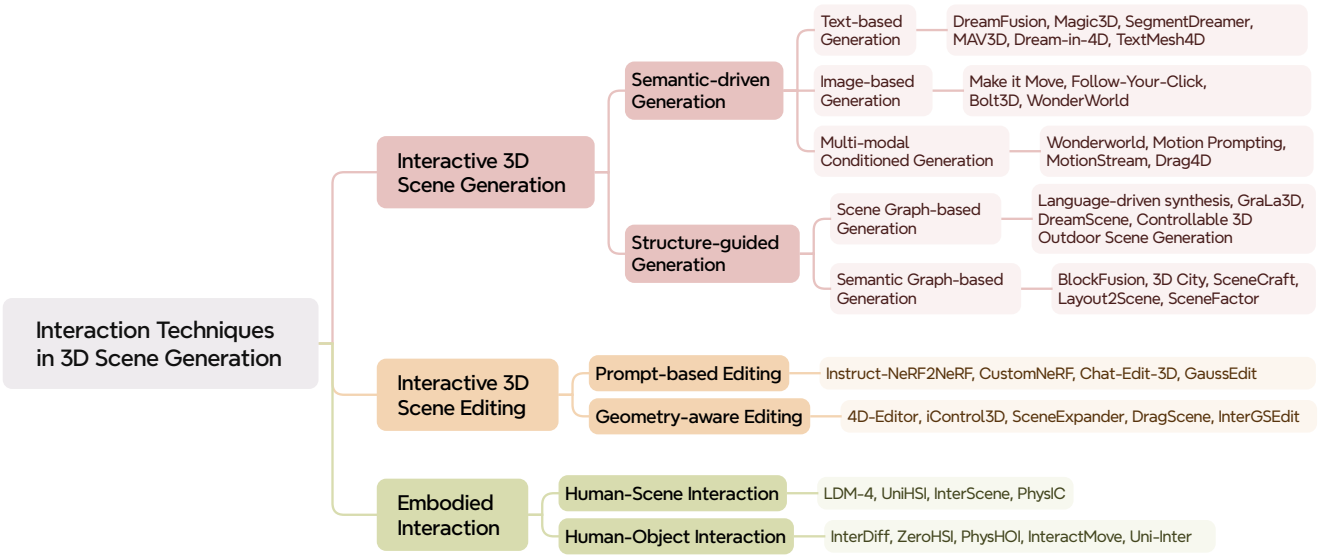

Figure 1: A taxonomy of interaction techniques in 3D dynamic scene generation. The taxonomy organizes existing methods according to user interaction contents and environments, covering interactive generation, interactive editing, and embodied interaction.

where θ represents the learned parameters. The objective is to maximize the plausibility of the generated scene while strictly adhering to the user's constraints:

$$\max_{\theta} \mathbb{E}_{S\sim P_{data}}[\log P_{\Phi}(S_t|S_{t-1}, u_t)] - \lambda \mathcal{L}_{align}(S_t, u_t) \quad (2)$$

where $\mathcal{L}_{align}$ measures the alignment between the scene and user intent (e.g., CLIP score for text, IoU for layout), and λ balances realism and controllability. This formulation unifies the three paradigms discussed in this survey: *generation* (S_0 from null), *editing* (updating S via u), and *embodied interaction* (optimizing S for physical agent dynamics).

3 Categories of Interactive Techniques

We categorize interactive techniques based on the granularity of user control and the stage of intervention within the generative pipeline. As illustrated in Figure 1, we structure existing methodologies into three primary paradigms: (1) Interactive Scene Generation, dealing with the synthesis of 3D scenes driven by high-level semantics and explicit structural constraints; (2) Interactive Scene Editing, focusing on the refinement of geometry and appearance; and (3) Embodied Interaction, which extends generation to dynamic and physics-aware interactions between agents and environments. The following content critically analyzes the technical foundations and representative frameworks within each paradigm.

3.1 Interactive 3D Scene Generation

Within interactive generation, existing interactive generation methods can be broadly divided according to the precision with which user intent constrains the scene structure, as illustrated in Figure 2. Semantic-driven interaction provides a high-level and low-effort control interface, where users express intent through abstract signals such as text, images, or audio. In contrast, structure-guided interaction emphasizes precise control by allowing users to explicitly specify spatial constraints, such as object layouts, scene graphs, or geometric proxies. This contrast reflects a fundamental trade-off in interactive generation between flexible but ambiguous semantic control and precise but effort-intensive structural control, which underlies the design choices of most interactive 3D scene generation systems.

Semantic-driven Generation

Semantic-driven generation allows users to synthesize 3D environments using high-level abstract prompts, such as text descriptions, reference images, or motion trajectories. This paradigm relies on distilling priors from large-scale 2D generative models into consistent 3D representations.

Text-based Generation. Several works focused on effectively mapping to target 3D representations from user-provided text prompts, which usually utilize text-to-image models as the prior information for generating 3D scenes. Early text-to-3D approaches [Poole *et al.*, 2023; Lin *et al.*, 2023] leverage pre-trained text-to-image diffusion models as semantic priors, optimizing 3D scene representations to align with text-conditioned visual distributions. This strategy enables data-efficient 3D generation without large-scale 3D-text datasets and has become a dominant paradigm for semantic-driven interaction. However, text-driven interaction also exposes fundamental limitations. Natural language descriptions rarely encode precise spatial layouts or physical constraints, leading to challenges in multi-object arrangement and geometric consistency. While recent methods [Zhu *et al.*, 2025] introduce multi-stage optimization or hierarchical prompting to mitigate these issues, the core ambiguity of text-based control remains an inherent property of this interaction paradigm.

Beyond static 3D scenes, users increasingly expect to model temporal and spatial evolution in dynamic environments, which further amplifies the challenges of semantic-driven inter-

action. MAV3D [Singer *et al.*, 2023] appears as a pioneering effort that enables direct generation of 4D scenes from text descriptions. It optimizes a dynamic NeRF representation augmented with motion fields and leverages a pretrained text-to-video diffusion model to guide scene attributes, producing dynamic scenes observable from arbitrary viewpoints. However, coupling static appearance and motion within a single optimization process often leads to artifacts such as temporal inconsistency and view-dependent distortions. Dream-in-4D [Zheng *et al.*, 2024] addresses this by further introducing a two-stage framework that explicitly decouples static scene generation from motion learning. Similarly, TextMesh4D [Dai *et al.*, 2025] proposes a Jacobian Deformation Field combined with local–global semantic regularization, enabling direct generation of dynamic mesh sequences while preserving spatiotemporal consistency. These text-to-4D methods expose a fundamental limitation of semantic-driven interaction: maintaining geometric and temporal consistency across space and time under weak semantic constraints.

Image-based Generation. To mitigate the semantic ambiguity inherent in text prompts, Image-based Generation leverages visual exemplars to provide explicit spatial cues. This interaction modality significantly constrains the solution space, making it effective for tasks requiring high fidelity. The evolution of this paradigm has been characterized by a shift from slow optimization-based synthesis to efficient and interactive frameworks. Early approaches primarily utilized GANs to synthesize panoramas from image priors [Hu *et al.*, 2022], however they often lacked flexibility in local control. To enhance interactivity, Follow-Your-Click [Ma *et al.*, 2024] introduced regional selection mechanisms, allowing users to animate specific scene areas via sparse clicks. However, a major bottleneck for interactive applications has been generation latency. Addressing this, Bolt3D [Szymanowicz *et al.*, 2025] proposes a feed-forward latent diffusion model that bypasses lengthy per-scene optimization, capable of outputting 3D representations from unposed images at interactive frame rates. Similarly, to facilitate rapid online creation, WonderWorld [Yu *et al.*, 2025] introduces Fast Layered Gaussian Surfaces, that employs geometric initialization and directed deep diffusion, enabling users to generate consistent scenes from a single image and layout prompts in seconds. These advancements are pivotal, as they transition image-to-3D generation from an offline processing task to a real-time creative workflow.

Multi-modal Conditioned Generation. Standard text or image prompts often struggle to articulate dynamic behaviors, complex temporal evolution, or non-visual attributes. Recent research extends semantic interaction into the spatiotemporal domain by incorporating additional modalities, such as audio cues [Zhang *et al.*, 2024b] and motion trajectories. These signals enrich control by explicitly encoding motion intent and temporal patterns, bridging the gap between static scene generation and dynamic storytelling.

Trajectory-based interaction has emerged as a critical direction for defining spatiotemporal dynamics. Motion Prompting [Geng *et al.*, 2025] enables fine-grained control by conditioning video generation on user-drawn sparse or dense trajectories, permitting the specification of both object-specific paths and global camera motion. Moving towards real-time feedback, MotionStream [Shin *et al.*, 2025] proposes an interactive video world model that synthesizes content in a streaming manner, allowing users to manipulate trajectories or transfer motion styles while viewing results instantaneously. In the 3D domain, Drag4D [Kang *et al.*, 2025] establishes a unified framework that integrates object motion control with background synthesis, enabling users to define 3D trajectories for objects generated from single images. By allowing users to choreograph the movement within generated scenes, these methods unlock significant potential for interactive 3D animation and virtual cinematography.

Structure-guided Generation

Structure-guided interactive generation introduces explicit spatial constraints into the generation process, enabling users to precisely control scene structure, object relationships, and layout configurations. Unlike semantic-driven interaction, which relies on the model to infer spatial organization from abstract cues, structure-guided methods directly constrain where and how scene elements are arranged. This paradigm shifts part of the generation responsibility from the model to the user, resulting in improved determinism and controllability at the cost of increased interaction effort.

Scene Graph-based Generation. Scene graphs provide a symbolic representation of scene structure by encoding objects as nodes and spatial or semantic relationships as edges. Early scene graph–based methods [Ma *et al.*, 2018] primarily relied on data-driven retrieval pipelines, where scene graphs served as blueprints for selecting and placing object assets from databases. While effective for enforcing relational consistency, these approaches often suffer from limited geometric diversity and coarse spatial control. Recent advances aim to bridge symbolic structure and continuous geometry. GraLa3D [Huang *et al.*, 2024] jointly leverages scene graphs and layouts by using large language models to infer non-overlapping spatial arrangements from text, combining relational constraints with geometric guidance. DreamScene [Li *et al.*, 2025] further integrates LLM-based agents to infer hybrid graphs encoding both object semantics and spatial constraints, enabling end-to-end scene synthesis. Similarly, Liu et al. [Liu *et al.*, 2025b] transform sparse scene graphs into dense embedding maps that condition diffusion-based generators, improving spatial fidelity while preserving user-specified relationships. These methods demonstrate that scene graphs are most effective when coupled with continuous spatial representations, highlighting the trend toward symbolic–geometric hybrid control.

Semantic Graph-based Generation. Unlike the scene graph, semantic-based layout serves as an intermediate representation for encoding the structure and semantic organization of 3D scenes. It provides users with advanced guidance for 3D scene generation, ensuring the controllability and consistency of scene element placements. One aspect of the research focuses on 2D semantic layouts. BlockFusion [Wu *et al.*, 2024] generates 3D scenes in the form of cubic blocks. It introduced a two-dimensional layout adjustment mechanism, which enables users to precisely determine the placement and arrangement of elements by manipulating the bounding boxes of two-dimensional objects, thereby providing users

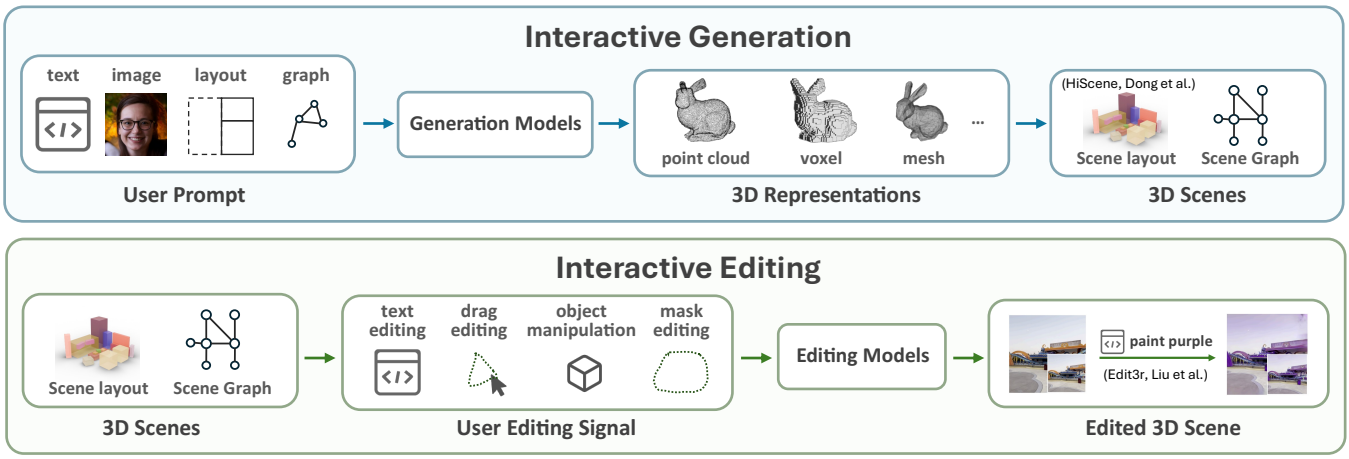

Figure 2: Comparison of Interactive Generation and Interactive Editing Paradigms. The framework compares how user prompts and editing signals are mapped through generation or editing models into 3D representations and edited scenes.

with more control over the generation process. 3D City [Feng *et al.*, 2025] integrated 2D layout, 3D terrain, and procedural modeling techniques to create a fully editable 3D environment, allowing users to efficiently customize and modify scene components. They utilize the structured data extraction process to separate model and layout features from textual descriptions to create 2D layouts to guide 3D terrain generation.

One the other aspect of research focuses on more complex 3D semantic layouts, which have better interactivity and controllability. SceneCraft [Yang *et al.*, 2024] proposed a detailed method for indoor scenes generating, the core of which lies in converting 3D semantic layout into multi-view 2D proxy maps, and designing a semantic and depth-conditioned diffusion model to generate multi-view images, which are used to learn NeRF as the final scene representation. Layout2Scene [Chen *et al.*, 2025] guides the scene generation through 3D semantic layout based on geometric and appearance diffusion priors. It combines semantic layout with text description, where the text description is used to specify the scene type, and the 3D semantic layout provides fine-grained control to eliminate the ambiguity in the global text description. SceneFactor [Bokhovkin *et al.*, 2025] maps the textual descriptions of the scene area to a 3D semantic layout map, and guide the high-fidelity geometric synthesis of scene geometry corresponding to the proxy semantics, enabling users to achieve more controllable scene generation and editing.

Structure-guided interaction enables precise and deterministic control by explicitly encoding spatial structure, making it particularly suitable for complex multi-object scenes. However, it requires users to invest more effort in specifying layouts or structural proxies, highlighting a fundamental trade-off between interaction burden and controllability.

3.2 Interactive 3D Scene Editing

Interactive 3D scene editing focuses on post-generation refinement, allowing users to modify geometry, appearance, semantics, or spatial layout of an existing scene. Unlike interactive generation, where user input conditions synthesis from scratch, editing-based interaction decouples interaction from initial generation and emphasizes local controllability, consistency preservation, and iterative correction. This paradigm is essential for correcting model errors, enforcing fine-grained constraints, and supporting creative workflows that require incremental adjustments.

Prompt-based Editing. Prompt-based editing enables users to specify editing intent through high-level instructions, such as text or reference images. These methods typically rely on pretrained vision–language models to interpret user commands and propagate semantic changes to 3D representations. Previous approaches [Haque *et al.*, 2023] extend text-to-image diffusion models to 3D domains, enabling instruction-based modification of scene attributes. CustomNeRF [He *et al.*, 2024] introduces a unified text and image-driven editing framework that uses 2D masks to localize edits and adopts a local–global optimization strategy to maintain multi-view consistency. Dialogue-based interaction further enhances controllability by allowing users to iteratively refine instructions. Chat-Edit-3D [Fang *et al.*, 2024] slices 3D scenes into 2D texture atlases and employs LLMs to parse user instructions, enabling conversational editing through 2D operations. GaussEdit [Shu *et al.*, 2025] extends this paradigm by allowing users to define regions of interest via bounding boxes and applying category-aware diffusion regularization to ensure physically plausible edits. Prompt-based editing offers intuitive semantic control but remains sensitive to region localization accuracy and cross-view consistency, particularly for complex geometric changes.

Geometry-aware Editing. Geometry-aware editing provides direct and precise control by allowing users to manipulate scene geometry through clicks, strokes, dragging, or selection. These interactions explicitly operate on spatial structures, making them more deterministic than prompt-based approaches. 4D-Editor [Jiang *et al.*, 2023] enables object-level editing in dynamic NeRFs using user strokes on single frames, maintaining spatiotemporal consistency via hybrid semantic features. iControl3D [Li *et al.*, 2024b] allows users to specify semantic segmentations, scribbles, or depth cues, which are fused with diffusion-generated images to construct complete

scenes. SceneExpander [Zhang *et al.*, 2024c] supports group-aware editing, where dragging one object triggers automatic re-layout of surrounding objects to preserve spatial coherence. DragScene [Gu *et al.*, 2024] and InterGSEdit [Wen *et al.*, 2025b] further exploit 3D Gaussian Splatting to enable real-time, multi-view consistent geometric manipulation. While geometry-aware editing excels in precision and determinism, it often requires more complex interfaces and careful consistency handling, especially for large-scale or dynamic scenes.

3.3 Embodied Interaction

Embodied interaction extends interactive 3D scene generation beyond manipulation to action execution within generated environments. In this paradigm, interaction is realized through dynamic behaviors of humans and objects, requiring joint reasoning over geometry, motion, contact, and physical constraints. Unlike generation or editing, embodied interaction inherently involves dynamics and physical feasibility.

Human-Scene Interaction (HSI). Early studies explored possibilities of human interaction with static scenes based on simple scene conditions and initial human states [Wang *et al.*, 2021]. Later works incorporated action labels and text descriptions as semantic conditions to enhance user control. Advancements in LLMs have significantly enhanced applications in this area. UniHSI [Xiao *et al.*, 2023] proposed a unified HSI framework, supporting users to control various scene physical interactions uniformly via language commands, characterized by using LLMs to generate interaction plans, making training interaction-annotation-free. Current progress lies in generating realistic and physically plausible human interaction in 3D scenes. InterScene [Pan *et al.*, 2024] proposes dividing character-scene interaction into "interaction" and "navigation" phases, using interaction controllers (e.g., sit/stand actions) and navigation controllers (obstacle avoidance path following) respectively, to enable physics-simulated characters to execute long-term interaction actions in multi-object, complex 3D scenes. PhySIC [Yalandur Muralidhar *et al.*, 2025] reconstructs physically plausible 3D human-scene interactions from a single RGB image. By jointly optimizing human, scene, and contact information, it resolves depth ambiguities and physical inconsistencies in invisible parts.

Human-Object Interaction (HOI). HOI is considered a specialized subset of HSI, emphasizing the accurate execution of approaching, grasping, and manipulating objects. Some studies focus on the interaction generation of whole-body motion. InterDiff [Xu *et al.*, 2023] uses a diffusion network to generate human-object interaction sequences and employs a physics-aware interaction correction module using relative motion patterns around contact points to ensure physical perceptibility. ZeroHSI [Li *et al.*, 2024a] reconstructs realistic human motion by extracting motion priors from video generation models and using differentiable neural rendering, achieving human-scene and human-object interaction in static and dynamic 3D environments without paired motion capture data.

Compared to kinematics-based methods, physics-based methods are more advantageous for HOI because physics simulators can explicitly constrain humanoid and object dynamics, creating more realistic human-machine interactions in 3D scenes. PhysHOI [Wang *et al.*, 2023] is a representative work introducing a physics-based whole-body HOI imitation framework, making HOI imitation feasible in diverse scenes. InteractMove [Cai *et al.*, 2025] first incorporated text-conditioned 3D interaction actions (including grasping, manipulation) into a scene generation framework, emphasizing physical feasibility, enabling users to generate the physical process of humans interacting with movable objects in 3D scenes (including grasping, moving, and collision avoidance) via text control. Uni-Inter [Liu *et al.*, 2025a] attempts to unify multiple interaction types (including object-object, human-scene) beyond single interaction types, proposing a unified framework to model volumetric representations of humans and various interaction entities, and generating physically plausible actions via joint-level probability prediction to support composite interaction scenes.

4 Datasets and Evaluation

In this section, we review commonly used datasets and evaluation metrics from an interaction-centric perspective, highlighting how current dataset support different interaction paradigms and significant gaps in this field.

4.1 Datasets

Existing datasets can be broadly categorized into generation and editing, human–scene interaction, and human–object interaction, each reflecting different assumptions about interaction complexity and physical grounding. In Table 2, **generation and editing datasets,** such as 3D-FRONT [Fu *et al.*, 2021] and Objaverse-XL [Deitke *et al.*, 2023], emphasize scene diversity and structural annotations, supporting controllable synthesis and post-generation modification. However, they often lack interaction trajectories or physical annotations. **Human–scene interaction datasets,** including PROX [Hassan *et al.*, 2019] and HUMANISE [Wang *et al.*, 2022], focus on capturing realistic human behaviors within scenes, but are limited in scale due to data collection complexity. Human–object interaction datasets, such as GRAB [Taheri *et al.*, 2020] and BEHAVE [Bhatnagar *et al.*, 2022], provide fine-grained contact and manipulation annotations, enabling physics-aware modeling, but typically cover fewer object categories. However, a dataset that simultaneously supports interactive generation, editing, and embodied interaction is still lacking, motivating future efforts toward unified, interaction-centric benchmarks.

4.2 Evaluation

Controllability is used to evaluate the ability to respond to user input. CLIP Score[Hessel *et al.*, 2021] is utilized to measure the alignment between the generated images and the conditioned text, which can reflect whether the generated scene follows the user-specified prompt. **Motion quality** focuses on evaluating the authenticity, accuracy, and diversity of the interactive movements in the generated scenarios. The indicators usually include Diversity, Multi-modality, R Precision, and MPJPE. **Physical scalability** is used to evaluate how the generated interactions comply with the physical constraints of the environment or objects, which is typically measured through Non-Collision, Contact, and Penetration. **User research** mainly measures the perceived naturalness and overall

Category	Dataset	Type	Representation	Volume	Source	Venue
Generation and Editing	SceneNN [Hua *et al.*, 2016]	Indoor scenes	Mesh	502K	Synthetic	3DV 2016
	DAVIS [Perazzi *et al.*, 2016]	Real-world scenes	Video	50 sequences	Camera capture	CVPR 2016
	3D-FRONT [Fu *et al.*, 2021]	Indoor scenes	CAD	6.8K	Synthetic	ICCV 2021
	Objaverse-XL [Deitke *et al.*, 2023]	3D objects	CAD	10M+	Scan	NeurIPS 2023
	SG3D [Zhang *et al.*, 2024e]	Real-world scenes	RGB-D	4.9K	Scan	arXiv 2024
	SpatialGen [Fang *et al.*, 2025]	Indoor scenes	Panoramic images	4.7M	Synthetic	arXiv 2025
Human-Scene Interaction	PROX [Hassan *et al.*, 2019]	Interaction sequences	Mesh	0.1M	Camera capture	ICCV 2019
	GTA-IM [Cao *et al.*, 2020]	Interaction sequences	RGB images	1M	Synthetic	ECCV 2020
	SAMP [Hassan *et al.*, 2021]	Interaction sequences	Voxel	185K	Camera capture	ICCV 2021
	HUMANISE [Wang *et al.*, 2022]	Interaction sequences	Point cloud	19.6K	Synthetic	NeurIPS 2022
	LINGO [Jiang *et al.*, 2024]	Interaction sequences	Voxel	16 hours	Camera capture	SIGGRAPH Asia 2024
Human-Object Interaction	GRAB [Taheri *et al.*, 2020]	Interaction sequences	Mesh	1,334 seq. (1.6M frames)	Camera capture	ECCV 2020
	BEHAVE [Bhatnagar *et al.*, 2022]	Interaction sequences	Mesh	15.2K	Camera capture	CVPR 2022
	HOI-M^3 [Zhang *et al.*, 2024a]	Interaction sequences	Mesh	199 seq. (181M frames)	Hybrid	CVPR 2024
	FORCE [Zhang *et al.*, 2024d]	Interaction sequences	Mesh	450 seq. (192K frames)	Hybrid	3DV 2025
	CORE4D [Liu *et al.*, 2025c]	Interaction sequences	Mesh	11K	Hybrid	arXiv 2024

Table 2: Overview of datasets for interactive generation, editing and embodied human interaction. We highlight current datasets in representation, scale and interaction types.

quality of the generated interaction through the subjective assessment by evaluators. Participants are usually required to rank or rate the generated interaction sequences based on multiple aspects, typically including authenticity, naturalness, and physical acceptability. Platforms such as Prolific and Amazon Mechanical Turk facilitate the recruitment of diverse participants and enable the effective expansion of user research.

5 Challenges and Future Directions

Interactive 3D scene generation has progressed rapidly; however, current systems still fall short of practical requirements for reliable control, consistent feedback, and physically plausible interaction. The core difficulties are not only model capacity, but also the lack of interaction-centric data, evaluation protocols, and system-level design that jointly support generation, editing, and embodied behaviors.

5.1 Challenges

Interaction-centric data and evaluation. A primary bottleneck is the scarcity of datasets that explicitly capture *interactive signals* and their corresponding *target outcomes*. Existing resources are often designed for static generation or isolated interaction tasks, and rarely provide unified annotations across semantic intent, structural constraints, editing operations, and embodied trajectories. On the evaluation side, commonly used metrics are still dominated by visual fidelity or geometric reconstruction quality, which are insufficient to assess interactive systems. What remains under-specified is an interaction-aware evaluation protocol that measures (i) controllability under user constraints, (ii) consistency across viewpoints and time, and (iii) response stability under iterative user feedback.

Controllability under heterogeneous user inputs. Modern interactive systems face heterogeneous inputs, such as text, images, layouts, sketches, and direct manipulations. A persistent limitation is the inability to robustly resolve conflicts across modalities and across interaction stages. For example, semantic prompts may imply goals that contradict a structural layout, or geometry-aware edits may break semantic alignment. Existing methods often rely on manual weighting, leading to unstable behavior when user instructions are ambiguous. Achieving predictable controllability therefore requires principled mechanisms for constraint prioritization, uncertainty handling, and iterative correction.

5.2 Future Directions

Physics-grounded interaction and physically consistent feedback. A key direction is to incorporate stronger physical grounding into interactive generation, not only as correction but as part of the interaction loop [Meng *et al.*, 2025]. This includes learning or integrating physical priors to constrain generation and editing outcomes, and supporting embodied interaction with motion and force consistency. A practical challenge is that high-fidelity simulation can be costly, so future systems may require hybrid designs that combine fast learned dynamics with selective simulation for critical events. Physics-grounded feedback can serve as an additional interaction channel, enabling systems to explain failures and maintain consistency during iterative edits.

Unified multi-modal interaction with conflict-aware control. Another important direction is a unified interaction framework that treats multi-modal inputs as complementary constraints rather than independent conditions. This requires methods that can (i) align semantic and structural signals into a shared control space, (ii) detect and resolve conflicts across modalities, and (iii) support incremental updates without regenerating the entire scene from scratch. A promising system design is to maintain explicit intermediate states, such as structured scene graphs, editable layouts, or factorized representations. Therefore, user actions can be applied locally with predictable global effects. Ultimately, multi-modal interaction should enable more natural and efficient workflows, while providing reliable controllability in complex scenes.

6 Conclusion

This survey provides an interaction-centered perspective on 3D scene generation, systematically reviewing how users intervene in the generation pipeline through different interaction mechanisms. Rather than organizing prior work by representations or model architectures, we frame existing methods according to where user input enters the workflow and what

aspects of the scene it directly controls. We categorize interactive techniques into three major paradigms: Interactive Generation, Interactive Editing, and Embodied Interaction, which together cover intent specification during synthesis, post-generation refinement, and action execution within generated environments. This survey also summarizes commonly used datasets and evaluation practices, providing a clear fragmentation across generation, editing, and embodied interaction scenarios. We hope this survey serves as a foundation for understanding interaction mechanisms in 3D scene generation and provides guidance for developing more controllable and physically grounded interactive 3D systems.

Contribution Statement

Yuqi Li and Siwei Meng contributed equally to this work. All authors contributed to the survey design, literature review, manuscript writing, and revision.

References

[Bhatnagar *et al.*, 2022] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, 2022.

[Bokhovkin *et al.*, 2025] Aleksey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation. In *CVPR*, 2025.

[Cai *et al.*, 2025] Xinhao Cai, Minghang Zheng, Xin Jin, and Yang Liu. Interactmove: Text-controlled human-object interaction generation in 3d scenes with movable objects. In *MM*, 2025.

[Cao *et al.*, 2020] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020.

[Chen *et al.*, 2025] Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025.

[Dai *et al.*, 2025] Sisi Dai, Xinxin Su, Boyan Wan, Ruizhen Hu, and Kai Xu. Textmesh4d: High-quality text-to-4d mesh generation. *arXiv preprint arXiv:2506.24121*, 2025.

[Deitke *et al.*, 2023] Matt Deitke, Dustin Schwenk, et al. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023.

[Fan *et al.*, 2025] Siyuan Fan, Wenke Huang, Xiantao Cai, and Bo Du. 3d human interaction generation: A survey. *arXiv preprint arXiv:2503.13120*, 2025.

[Fang *et al.*, 2024] Shuangkang Fang, Yufeng Wang, Yi-Hsuan Tsai, Yi Yang, Wenrui Ding, Shuchang Zhou, and Ming-Hsuan Yang. Chat-edit-3d: Interactive 3d scene editing via text prompts. In *ECCV*, 2024.

[Fang *et al.*, 2025] Chuan Fang, Heng Li, et al. Spatialgen: Layout-guided 3d indoor scene generation. *arXiv preprint arXiv:2509.14981*, 2025.

[Feng *et al.*, 2025] Yuchuan Feng, Jihang Jiang, Jie Ren, Wenrui Li, Ruotong Li, and Xiaopeng Fan. Text-guided editable 3d city scene generation. In *ICASSP*, 2025.

[Fu *et al.*, 2021] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021.

[Geng *et al.*, 2025] Daniel Geng, Charles Herrmann, Junhwa Hur, et al. Motion prompting: Controlling video generation with motion trajectories. In *CVPR*, 2025.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. *NeurIPS*, 2014.

[Gu *et al.*, 2024] Chenghao Gu, Zhenzhe Li, Zhengqi Zhang, Yunpeng Bai, Shuzhao Xie, and Zhi Wang. Dragscene: Interactive 3d scene editing with single-view drag instructions. *arXiv preprint arXiv:2412.13552*, 2024.

[Haque *et al.*, 2023] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *ICCV*, 2023.

[Hassan *et al.*, 2019] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019.

[Hassan *et al.*, 2021] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021.

[He *et al.*, 2024] Runze He, Shaofei Huang, Xuecheng Nie, Tianrui Hui, Luoqi Liu, Jiao Dai, Jizhong Han, Guanbin Li, and Si Liu. Customize your nerf: Adaptive source driven 3d scene editing via local-global iterative training. In *CVPR*, 2024.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[Hu *et al.*, 2022] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022.

[Hua *et al.*, 2016] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016.

[Huang *et al.*, 2024] Yu-Hsiang Huang, Wei Wang, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Toward scene graph and layout guided complex 3d scene generation. *arXiv preprint arXiv:2412.20473*, 2024.

[Jiang *et al.*, 2023] Dadong Jiang, Zhihui Ke, Xiaobo Zhou, and Xidong Shi. 4d-editor: Interactive object-level editing in dynamic neural radiance fields via 4d semantic segmentation. *CoRR*, 2023.

[Jiang *et al.*, 2024] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, et al. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia*, 2024.

[Kang *et al.*, 2025] Minjun Kang, Inkyu Shin, Taeyeop Lee, In So Kweon, and Kuk-Jin Yoon. Drag4d: Align your motion with text-driven 3d scene generation. *arXiv preprint arXiv:2509.21888*, 2025.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.

[Li *et al.*, 2024a] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. *arXiv preprint arXiv:2412.18600*, 2024.

[Li *et al.*, 2024b] Xingyi Li, Yizheng Wu, Jun Cen, Juewen Peng, Kewei Wang, Ke Xian, Zhe Wang, Zhiguo Cao, and Guosheng Lin. icontrol3d: An interactive system for controllable 3d scene generation. In *MM*, 2024.

[Li *et al.*, 2025] Haoran Li, Yuli Tian, Kun Lan, Yong Liao, Lin Wang, Pan Hui, and Peng Yuan Zhou. Dreamscene: 3d gaussian-based end-to-end text-to-3d scene generation. *T-PAMI*, 2025.

[Lin *et al.*, 2023] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.

[Liu *et al.*, 2025a] Sheng Liu, Yuanzhi Liang, Jiepeng Wang, Sidan Du, Chi Zhang, and Xuelong Li. Uni-inter: Unifying 3d human motion synthesis across diverse interaction contexts. In *SIGGRAPH Asia*, 2025.

[Liu *et al.*, 2025b] Yuheng Liu, Xinke Li, Yuning Zhang, Lu Qi, Xin Li, Wenping Wang, Chongshou Li, Xueting Li, and Ming-Hsuan Yang. Controllable 3d outdoor scene generation via scene graphs. *arXiv preprint arXiv:2503.07152*, 2025.

[Liu *et al.*, 2025c] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *CVPR*, 2025.

[Ma *et al.*, 2018] Rui Ma, Akshay Gadi Patil, et al. Language-driven synthesis of 3d scenes from scene databases. *ACM TOG*, 2018.

[Ma *et al.*, 2024] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024.

[Meng *et al.*, 2025] Siwei Meng, Yawei Luo, and Ping Liu. Grounding creativity in physics: A brief survey of physical priors in aigc. *IJCAI*, 2025.

[Miao *et al.*, 2025] Qiaowei Miao, Kehan Li, Jinsheng Quan, Zhiyuan Min, Shaojie Ma, Yichao Xu, Yi Yang, Ping Liu, and Yawei Luo. Advances in 4d generation: A survey. *arXiv preprint arXiv:2503.14501*, 2025.

[Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

[Pan *et al.*, 2024] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024.

[Park *et al.*, 2019] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[Perazzi *et al.*, 2016] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[Poole *et al.*, 2023] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[Shin *et al.*, 2025] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

[Shu *et al.*, 2025] Zhenyu Shu, Junlong Yu, Kai Chao, Shiqing Xin, and Ligang Liu. Gaussedit: Adaptive 3d scene editing with text and image prompts. *TVCG*, 2025.

[Singer *et al.*, 2023] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *ICML*, 2023.

[Szymanowicz *et al.*, 2025] Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. In *ICCV*, 2025.

[Taheri *et al.*, 2020] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

[Wang *et al.*, 2021] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021.

[Wang *et al.*, 2022] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *NeurIPS*, 2022.

[Wang *et al.*, 2023] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

[Wang, 2024] Zhengren Wang. 3d representation methods: A survey. *arXiv preprint arXiv:2410.06475*, 2024.

[Wen *et al.*, 2025a] Beichen Wen, Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. 3d scene generation: A survey. *arXiv preprint arXiv:2505.05474*, 2025.

[Wen *et al.*, 2025b] Minghao Wen, Shengjie Wu, Kangkan Wang, and Dong Liang. Intergsedit: Interactive 3d gaussian splatting editing with 3d geometry-consistent attention prior. In *ICCV*, 2025.

[Wu *et al.*, 2024] Zhennan Wu, Yang Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM ToG*, 2024.

[Xiao *et al.*, 2023] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023.

[Xu *et al.*, 2023] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

[Xue *et al.*, 2025] Haiwei Xue, Xiangyang Luo, et al. Human motion video generation: A survey. *PAMI*, 2025.

[Yalandur Muralidhar *et al.*, 2025] Pradyumna Yalandur Muralidhar, Yuxuan Xue, et al. Physic: Physically plausible 3d human-scene interaction and contact from a single image. In *SIGGRAPH Asia*, 2025.

[Yang *et al.*, 2024] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *NeurIPS*, 2024.

[Yu *et al.*, 2025] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, 2025.

[Zhang *et al.*, 2024a] Juze Zhang, Jingyan Zhang, et al. Hoiˆ 3: Capture multiple humans and objects interaction within contextual environment. In *CVPR*, 2024.

[Zhang *et al.*, 2024b] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. In *ECCV*, 2024.

[Zhang *et al.*, 2024c] Shao-Kui Zhang, Junkai Huang, et al. Sceneexpander: Real-time scene synthesis for interactive floor plan editing. In *MM*, 2024.

[Zhang *et al.*, 2024d] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. *CoRR*, 2024.

[Zhang *et al.*, 2024e] Zhuofan Zhang, Ziyu Zhu, Junhao Li, Pengxiang Li, , et al. Task-oriented sequential grounding and navigation in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024.

[Zheng *et al.*, 2024] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. In *CVPR*, 2024.

[Zhu *et al.*, 2025] Jiahao Zhu, Zixuan Chen, Guangcong Wang, Xiaohua Xie, and Yi Zhou. Segmentdreamer: Towards high-fidelity text-to-3d synthesis with segmented consistency trajectory distillation. In *ICCV*, 2025.