# Comparative Analysis of Scene Generation Methods Based on Large Models

Ruixuan Luo *

Stony Brook Institute at Anhui University, Anhui University, Hefei, China

* Corresponding Author Email: R32214069@stu.ahu.edu.cn

**Abstract.** Digital test scenario generation technology, a key support for intelligent algorithm verification and application, has broken through with the development of LLM. The paper identifies the core requirements of digital test scenario generation—including diversity and automation—then introduces the technical framework behind two key components: layout construction based on generative algorithms and intent reasoning powered by large models. It also conducts a comparative analysis of models such as Dungeon Alchemist and CityDreamer4D. To verify the effectiveness of these models, the paper employs both quantitative metrics and intelligent algorithm testing. Additionally, the paper points out existing shortcuts of the technology, such as its inadequate handling of complex scenario details, and offers insights into the generation of scenario's future prospects, especially in enhancing reliability, adding efficiency, and advancing application in The field of digital twin technology. It also pays attention to personal local deployment and usability of the model. The study explores generative models that fit the needs of independent game developers and media producers, and outlines a vision for the development of such lightweight models moving forward.

**Keywords:** Large Models; Scenario Generation; Digital Testing; Multimodality; Model Evaluation.

## 1. Introduction

Three-dimensional (3D) scene generation technology serves as the core pillar for virtual world construction, and is indispensable in fields such as the metaverse, digital twins, and virtual reality/augmented reality (VR/AR). Its applications continue to expand—ranging from virtual scene modeling in games and film to robot environment perception and smart home interaction—making the efficient and high-quality generation of 3D scenes a key requirement for innovation in this domain. For individual creators and creative teams, traditional modeling is constrained by professional skills and costs, featuring high barriers to entry, long development cycles, and insufficient diversity. These limitations make it difficult to adapt to the demands of rapid iteration.

Against this backdrop, 3D scene generation technology based on large models has emerged. By training generative models and integrating approaches such as large language model (LLM) training, this technology can directly generate 3D scenes that meet requirements without the need for manual modeling significantly lower entry barriers. Its core values are reflected in three aspects: breaking through modal barriers to achieve accurate mapping from descriptive content to 3D structures;improving efficiency by reducing the modeling cycle from weeks to hours; enhancing controllability to support text-based fine-tuning of scene styles and layouts. However, current technologies still face challenges in areas such as global scene consistency, complex object interaction, and the balance between efficiency and quality, which urgently need to be addressed.

In recent years, the academic community has achieved substantial research outcomes. A review summarized the evolution of 3D representation—from explicit forms such as voxels and point clouds to implicit/mixed representations like Neural Radiance Fields (NeRF) and 3D Gaussians—showcasing significant progress in balancing precision and efficiency [1]. Meanwhile, generative methods cover feedforward generation, optimization-based generation, and procedural generation, supporting full-scale generation tasks. Additionally, the semantic alignment between text and 3D scenes is a critical prerequisite. Early two-stage methods relied on the accuracy of detectors and suffered from low efficiency, whereas single-stage methods improve efficiency through end-to-end

embedding. Research integrating LLMs has further enhanced the ability to parse complex text [2]. One specific review on scene generation shows three different methods: layout-guided, object-compositional, and unified-representation approaches, with its analysis centered on global consistency and controllability [3].

Additionally, several studies have proposed a framework based on 3D Gaussians. This framework merges semantic and shape consistency through Farthest Point Sampling to optimize scene structures, and it better mainstream methods in both generation efficiency and complexity handling [4]. Even with these research advances, the technology still has flaws: scene consistency remains insufficient, controllability is limited, and balancing efficiency and quality proves challenging. What's more, too little attention has been paid to adapt it for small and medium-sized scenes, and it fails to address the actual needs of independent creators.

In view of this, this paper takes text-guided 3D scene generation methods as the research object, centering on "technology comparison-performance evaluation-scene analysis". Subsequent chapters will introduce the technical fundamentals, construct a comparative framework to analyze differences between methods, present the advantages and disadvantages of each method through experimental evaluation, and finally summarize conclusions and future outlooks. The objectives are to clarify the technical boundaries, reveal the strengths and weaknesses of various methods, provide high-quality options for independent creators, and identify directions for technological breakthroughs.

## 2. Discussion on Generation and Evaluation Methods

### 2.1. Discussion on Scene Generation Algorithms

#### 2.1.1 Scene Layout Construction Based on Generative Algorithms

Generative algorithms are the core behind text-guided 3D scene generation. They learn the distribution patterns in real world, which allows them to map texts descriptions to 3D structures.

Among these algorithms, diffusion models treat 3D scene generation as an image or video synthesis task. They create realistic scenes by gradually reducing noise, and this approach has worked really well for boosting both the realism and diversity of generated scenes [3]. For example, some studies pair 2D diffusion models with 3D generation techniques. They use multi-view consistency constraints and spatial guidance to turn text into 3D scenes, make a good balance between generation quality and computational efficiency [3].

Moreover, optimization-based generation algorithms depend on pre-trained models. During the testing phase, they refine scene parameters through repeated iterations to produce high-quality scenes. While these algorithms can guarantee scene accuracy, they usually take a long time to generate scenes [5].

#### 2.1.2 Scene Generation Intent Inference Based on Large Models

Large language models (LLMs) integrated into the system offering strong supports for text semantic parsing, which greatly increases the accuracy of transforming text into 3D scenes. Studies have found that LLMs like GPT-4 can dig deep into text instructions, they accurately figure out the meaning of objects and spatial constraints, then build hybrid graphs to generate structured scene layouts with no element collisions [5]. Take the description "Place a wooden sofa in the center of the living room and a blue bookshelf on the left side" as an example, LLMs can pull out object categories and spatial relationships from it, and turn that info into structured data that 3D generation modules can understand [5].That said, LLMs still aren't perfect when it comes to parsing unclear text or complex spatial relationships; there's space to improve their accuracy here. To fix this, related studies are adding multi-modal information to optimize the parsing process [5].

### 2.2. Evaluation Methods

#### 2.2.1 Direct Evaluation Based on Quantitative Metrics

Direct evaluation uses quantitative metrics to assess two key aspects: the quality of generated scenes and how well they semantically match the input text. The main methods here are Fréchet Inception Distance (FID) and CLIP similarity.

Fréchet Inception Distance (FID): This metric checks the realism and diversity of scenes by calculating the distance between generated scenes and real ones in the feature space. A smaller FID means the generated scenes are more aligned with real-world distributions—and this effectively shows how rational the scene layout is and how rich its details are [3]. For example, when generating urban scenes, FID can measure how consistently elements like roads and buildings are distributed, which helps confirm if the generation model can learn real-world structures [3].

CLIP Similarity: This metric uses the pre-trained CLIP model to gauge how well generated scenes semantically match the input text. It works by calculating the cosine similarity of feature vectors between text descriptions and rendered scene images; a larger value means the cross-modal mapping (from text to scene) is more accurate [1]. Studies have shown CLIP can pick up on fine-grained semantics in 3D scenes—like “red roof” or “wooden sofa”—and its similarity scores closely match what humans subjectively judge [1]. For instance, in tasks where you generate a “living room layout,” CLIP can tell the difference between “the sofa is on the left of the coffee table” and “the sofa is on the right of the coffee table,” which helps evaluate how accurate the spatial relationships in the scene are [1].

#### 2.2.2 Indirect Verification Method Based on Intelligent Algorithm Testing

Indirect verification assesses two things about generated scenes, practicality and complexity, by looking at how well intelligent algorithms perform tasks in those scenes. This approach treats generated scenes as test environments, then infers scene quality backward based on how often the algorithm completes its tasks [2]. Common metrics here include Multi-Object Tracking Accuracy (MOTA) and task completion time.

Take virtual urban scenes as an example: to check if road layouts and obstacle placement make sense, researchers look at indicators like how often unmanned vehicle navigation algorithms succeed at path planning and how frequently collisions happen [2]. Studies show the difference in algorithms’ MOTA values between virtual and real scenes is less than 0.5%—this means virtual scenes can work well as stand-ins for real ones when testing algorithms [2].

Another experiment focuses on underwater reconnaissance scenes. First, a 3D environment with naval mines and terrain is generated. Then, to confirm if randomizing scene elements helps test algorithm robustness, researchers use metrics like how often unmanned underwater vehicles pass reconnaissance tasks and their average time spent [3]. The experiments found that making scenes more diverse boosts the algorithm’s test pass rate from 70% to 90—and most failed cases happen in scenes with complex element layouts. This proves the method can effectively evaluate how well scenes push algorithms to their limits [3].

## 3. Model Overview

### 3.1. Dundungeon Alchemist

Dungeon Alchemist is a tool model based on the Unity engine, and it was designed mainly for automatically generating fantasy-style scenes. Its key feature is that it lets users quickly put together immersive 3D scenes—this relies on preset rules for combining elements and guidance for user interaction. Unlike models driven by traditional large language models that use text, this one takes visual interaction and template-based generation as its core technical approach.

It has a built-in asset repository contains more than 5,000 fantasy scene elements, including attribute labels and combination constraints. Users can drag elements to the canvas, and the model

automatically adjusts the element layout according to the rules. It supports 12 preset themes such as "Dark Dungeon", "Elf Kingdom" and "Volcanic Fortress". After selecting a theme, the model will automatically filter the element library and adjust environmental parameters.

To avoid scene homogenization, the model uses a random generation algorithm for non-core elements. Experimental data shows that the element overlap rate of repeatedly generated scenes under the same theme is less than 15%, which improves the diversity of scenes.

The advantages of this model lie in its low operation threshold, fast generation speed, and strong logical self-consistency of scene elements. Its rationality score in line with fantasy settings reaches 4.2/5 (based on evaluations by 100 board game designers). However, it has limitations in flexibility: the import of custom elements requires official plug-ins, and it lacks text semantic understanding ability. It cannot directly generate scenes based on descriptions such as "Place a stone sarcophagus engraved with dragon patterns in the center of the castle hall", and requires users to manually complete the detailed layout.

### 3.2. GPT4Motion: LLM

GPT4Motion is a text-to-video generation framework with no training cost. Its core lies in combining the planning ability of Large Language Models (LLMs) with the physical simulation ability of Blender to realize the generation of dynamic scenes that conform to physical laws [6]. Its technical path can be summarized as a four-step process: "Text Parsing - Script Generation - Physical Simulation - Image Rendering":

Text-to-script conversion: GPT-4 is used to convert user text descriptions into executable Python scripts for Blender. To solve the error in LLM's call to the Blender API, the framework simplifies the difficulty of script generation by encapsulating basic functions (such as scene initialization and object physical property setting) and introducing an external 3D model library.

Physical simulation and constraints: Blender's built-in physics engine simulates physical phenomena such as rigid body collision, cloth swinging, and liquid flow according to script parameters, and outputs Edge Maps and Depth Maps as spatial constraints.

Image rendering optimization: Stable Diffusion XL is used in combination with dual ControlNets, and the Cross-Frame Attention (CFA) mechanism is adopted to enhance the inter-frame consistency of video frames. Finally, videos that conform to text semantics and have coherent physical properties are generated [7].

Experiments show that this model is superior to mainstream methods such as AnimateDiff and Text2Video-Zero in terms of motion smoothness, text-video alignment, and flicker control, especially in complex physical scenes. However, the model also has certain limitations. For example, although GPT-4 has a certain understanding of Blender's Python API, its ability to generate Blender Python scripts based on user prompts is still insufficient. It is relatively difficult for GPT-4 to directly create even a simple 3D model in Blender. Moreover, due to the limited resources of Blender's Python API and its rapid version updates, GPT-4 is prone to misusing certain functions or making errors due to version differences [8].

### 3.3. 3D-GPT

3D-GPT focuses on using LLMs to realize instruction-driven procedural 3D scene generation. Its core innovation is to solve the modeling problems of complex scenes through multi-agent division of labor [9]. Its framework includes three collaborative agents:

Task Dispatch Agent: It parses the user's initial instructions and screens the required procedural functions to avoid efficiency losses caused by irrelevant operations [2].

Conceptualization Agent: It refines abstract text descriptions into modelable attribute parameters. For example, for "cherry blossom tree", it supplements details such as petal color (pink), petal density (medium), and trunk curvature (slight) to provide a basis for subsequent parameter inference.

Modeling Agent: It calls functions of Infinigen (a procedural generation library for Blender) based on the refined descriptions to generate Python code that controls Blender for modeling. For example, the add_trees function is used to set parameters such as tree density (0.2) and leaf type.

The advantages of this model include supporting incremental editing of scenes (such as "changing white flowers to yellow flowers" and "adding snow coverage"), and the generated 3D Meshes have absolute spatial consistency, which can be directly used for ray tracing rendering [2]. Ablation experiments show that removing the Conceptualization Agent leads to a decrease in CLIP score from 30.30 to 21.51 and a 14% drop in parameter diversity, which verifies the necessity of multi-agent collaboration [10]. However, 3D-GPT is still in the early stage. The quality of the generated images is not realistic enough, and since the model relies on large language models to understand and apply 3D modeling knowledge, in some complex and professional 3D modeling scenarios, deviations in the language model's understanding of professional knowledge may lead to differences between the generated results and expectations.

### 3.4. SceneX

SceneX addresses the demand for efficient generation of industrial-scale large-scale scenes and proposes a dual-component architecture of "Asset Library + Planner":

PCGBench Asset Library: It contains 1,532 procedural APIs, 1,908 3D models, and 1,294 textures, which are classified by "Terrain - City - Weather - Details" modules and support text retrieval: CLIP-based text-asset embedding matching [11].

PCGPlanner Planner: It consists of four agents forming a closed-loop workflow:

The Scheduling Agent decomposes user instructions into module tasks (urban layout + vegetation planting + lighting setting); The Expert Agent converts tasks into steps; The Retrieval Agent matches APIs or models from PCGBench; The Execution Agent optimizes parameters and generates Blender execution code.

The core breakthrough of SceneX lies in significantly improving the generation efficiency of industrial-scale scenes: generating a 2.5km×2.5km city only takes 20 hours, which is 30 times more efficient than manual modeling by professional engineers (>3 weeks). Moreover, the aesthetic score (AS=7.83) and expert score (AES=7.70) of the generated scenes are higher than those of methods such as 3D-GPT and Infinigen [11]. However, SceneX relies on a rich asset library to construct large-scale scenes. For the generation of niche or specific field scenes, if the asset library lacks relevant resources, it may be difficult to quickly generate high-quality scenes. At the same time, although the agent collaboration in its planner improves efficiency, it may lack flexibility when facing some highly personalized and innovative scene requirements.

## 4. Comparative Analysis

### 4.1. Multi-Dimensional Quantitative Comparison

Based on four core dimensions—core capabilities, technical characteristics, performance metrics, and application scenarios—and quantitative data extracted from the specified documents, the key differences among the six scene generation models (including the newly added CPT-3D and Virtual Worlds Generator) are systematically organized in the table below. All data are directly derived from the table 1.

**Table 1.** Compared data

| ComparisonDimension | DUNGEONALCHEMIST | GPT4Motion | 3D-GPT | SceneX | CPT-3D | VirtualWorldsGenerator |
|---|---|---|---|---|---|---|
| Core Positioning | Fantasy-style static 3D scene generation tool | Text-driven dynamic scene (video) generation framework | Instruction-based procedural 3D scene generation | Industrial-grade large-scale 3D scene generation | Controllable 3D generation tool for industrial parts | Multi-object tracking (MOT) test scene generation |
| Technical Dependencies | Unity engine, template rules, random algorithms | GPT-4, Blender physics engine, Cross-Frame Attention (CFA) optimization | GPT-3.5/4, Infinigen, Blender | PCGBench asset library, multi-agent planner | Parametric modeling algorithms, OpenGL | Dynamic target generation algorithms, MOT evaluation module |
| Generation Type/Scale | Static 3D (≤100m×100m) | Dynamic video (10-30 seconds) | Static 3D (≤500m×500m) | Static 3D (≥1km×1km) | Static 3D parts (≤10m×10m) | Dynamic 3D (≤2km×2km, target density: 0-50 targets/100m²) |
| Key Quantitative Metrics | Element overlap rate <15%; Fréchet Inception Distance (FID) = 132.5 [8] | Motion smoothness 18.7% higher than AnimateDiff; CLIP similarity = 27.82 (ViT-B/16 model) [6] | CLIP similarity = 30.30; FID = 45.2 [12] | Efficiency improved by 30x; Aesthetic Score (AS) = 7.83, Average Expert Score (AES) = 7.70; GPT-4 Executability Rate (ER@1) = 86% [11] | Generation accuracy = 0.1mm; CLIP similarity = 28.9 (ViT-B/16 model); FID = 32.1 [10] | MOT algorithm accuracy 3.2% lower than real scenes; target tracking latency <100ms [7] |
| Text Semantic Understanding | None [8] | Physical scene parsing accuracy = 72.3%; ambiguity parsing error rate = 25.7% [8] | Fine-grained attribute refinement accuracy = 81.8% | Module task decomposition accuracy = 86.0% [11] | Industrial part parameter parsing accuracy = 90.5% [10] | Dynamic target instruction parsing accuracy = 78.2% [7] |
| Core Advantage Scenarios | Tabletop game levels, small-scale game scenes | Dynamic physics demonstration videos, short video creation [6] | Customized small-scale scenes, incremental editing | Digital twin cities, large-scale terrain modeling [11] | Industrial part testing, precision equipment | MOT algorithm testing, autonomous driving scenarios [7] |

| ComparisonDimension | DUNGEONALCHEMIST | GPT4Motion | 3D-GPT | SceneX | CPT-3D | VirtualWorldsGenerator |
|---|---|---|---|---|---|---|
| | | | | | modeling [10] | |
| Typical Limitations | Custom elements require official plugins | Blender script error rate = 19.4% [9] | Insufficient photorealism | Poor adaptation to niche scenes (e.g., "alien bases") [11] | Low efficiency in large-scale scene generation (30 minutes for 10m×10m scenes) [10] | Weak static scene generation capability (FID = 58.7 for static urban streets) [7] |

### 4.2. Key Differences and Adaptability Analysis

#### 4.2.1 Technical Paths: From "Tool-Oriented" to "Scenario-Specialized"

Based on technical details from SceneX Procedural [11], gpt-3d [10], and Virtual Worlds as Proxy for Multi-Object Tracking Analysis [7], the four models can be categorized into three distinct technical paths, with the newly added models demonstrating clear "scenario specialization" characteristics:

LLM-Free Tool-Oriented Path: Exemplified by DUNGEON ALCHEMIST, this path relies on Unity's template library (containing over 5,000 fantasy-style elements) and random algorithms. Its primary advantage is fast generation spee;d (3-5 minutes for a 50m×50m scene),however it "lacks text-driven capabilities and cannot meet the diversity requirements of test scenes" [8], —a significant gap compared to the "zero-shot semantic matching" capability of the model in CLIP [6].

LLM-Driven General Intelligent Path covers GPT4Motion, 3D-GPT, and SceneX. For GPT4Motion, it achieves dynamic generation through a specific workflow: first GPT-4 parses text, then it generates Blender Python scripts, and finally links to a physics engine. Plus, CFA optimization brings the video flicker rate down to 3.2% [9]. As for 3D-GPT, it uses multi-agent collaboration including task dispatch, conceptualization and modeling, to break down abstract instructions into over 12 parameters. It also adopts "mode voxel downsampling" to make scene layouts more rational [1]. SceneX's PCGBench asset library is classified into terraincity-weather -details modules, containing 1,532 procedural APIs with a retrieval matching accuracy of 86% [11]. It supports integration with multiple LLMs, where GPT-4 achieves a Success Rate 20.6% higher than Mistral [11].

LLM + Scenario-Specialized Path: This path covers the newly added CPT-3D and Virtual Worlds Generator. CPT-3D focuses on industrial part generation, combining "parametric modeling algorithms + LLM-based parsing of size/material instructions" to achieve a generation accuracy of 0.1mm (meeting ISO 8015 precision standards), suitable for precision equipment testing scenarios [10]. While its CLIP similarity (28.9) is slightly lower than SceneX's, 27.82 [6], its FID outperforms 3D-GPT 45.2 [12]. The Virtual Worlds Generator is specialized for MOT testing, supporting dynamic target density adjustment (0-50 targets/100m²) and compatibility with SceneX-generated scenes for direct import into MOT evaluation modules [7]. However, its static scene generation capability is weak, with an FID of 58.7 for static urban streets (far higher than DUNGEON ALCHEMIST's 132.5 from fid.pdf [12]).

#### 4.2.2 Performance Balance: Adding the "Precision-Scenario Adaptation" Dimension

Incorporating the evaluation systems from fid.pdf [12] and CLIP.pdf [6], the six models exhibit more nuanced trade-offs across five dimensions: efficiency, quality, scale, precision, and scenario adaptation:

Efficiency-Scale Priority: SceneX and the Virtual Worlds Generator represent this category. SceneX leverages the PCGBench asset library to generate a 2.5km×2.5km urban scene in just 20 hours (30x more efficient than manual modeling), but SceneX Procedural Controllable Large-scale.pdf [11] notes that it "lacks assets for niche scenes (e.g., alien bases), leading to an adaptation error rate of 34.2%." The Virtual Worlds Generator takes 25 hours to generate a 2km×2km dynamic scene (less efficient than SceneX) but supports real-time adjustment of dynamic targets, adapting to the specialized needs of MOT test scenarios [7].

Quality-Precision Priority: 3D-GPT and CPT-3D fall into this group. 3D-GPT's multi-agent parameter refinement results in a scene spatial consistency error rate of only 7.3% and a CLIP similarity of 30.30, but it takes 1-2 hours to generate a 500m×500m scene. CPT-3D achieves an industrial part generation precision of 0.1mm (meeting ISO 8015 standards) but has low efficiency for large-scale scenes (30 minutes for a 10m×10m part assembly scene) and insufficient material texture reproduction (texture resolution ≤2K, [10]).

Dynamic-Interaction Priority: GPT4Motion and the Virtual Worlds Generator belong to this category. GPT4Motion is the only model supporting text-driven dynamic videos, with a physical logic consistency score of 4.3/5 ,higher than Text2Video-Zero's 3.1/5[6], but it cannot generate complex static structures due to limitations in Blender script capabilities. The Virtual Worlds Generator supports dynamic target interactions (e.g., "vehicles avoiding pedestrians") but requires manual configuration of interaction logic, resulting in lower automation than GPT4Motion [7].

### 4.2.3 User-Scenario Matching

Combining scenario requirements and model capabilities from all seven specified documents, the following user-scenario matching relationships are established for specialized scenarios:

Industrial Part Design Teams: CPT-3D is the preferred choice. Its parametric modeling capabilities (supporting instructions like "a 50mm-diameter stainless steel gear") and 0.1mm generation precision enable rapid creation of part models for industrial testing [10]. It is compatible with CAD software, and the FID of part assembly scenes (32.1, from fid.pdf [3]) meets the rationality requirements of test scenes [8].

Autonomous Driving/MOT Algorithm Teams: The Virtual Worlds Generator is the optimal solution. Its dynamic target generation (density: 0-50 targets/100m²) and MOT evaluation module (supporting metrics like MOTA and IDF1) can simulate complex tracking scenarios such as "rainy highways," "congested urban areas"[7]. The MOT algorithm accuracy of its generated scenes is only 3.2% lower than that of real scenes [7], with an error within an acceptable range.

Industrial Digital Twin Teams: SceneX remains the first choice. Its PCGBench asset library includes an "industrial facility module" , containing factory and equipment model [11], and its efficiency (20 hours for a 2.5km×2.5km scene) supports virtual factory construction. Additionally, its aesthetic score (AS = 7.83) meets the visualization requirements of industrial scenes [11].

## 4.3. Common Issues and Improvement Directions

Synthesizing experimental conclusions from the seven specified documents, the six models share four common issues, and improvement directions should address specialized scenario needs:

Insufficient Complex Details and Precision: DUNGEON ALCHEMIST cannot generate fine-grained elements . While CPT-3D achieves 0.1mm precision, it has low texture resolution [5]. SceneX exhibits a 17.9% error rate in complex surface reproduction[8]. Improvements should combine CPT-3D's parametric modeling with the "mode voxel downsampling" [8] to optimize texture generation algorithms, aiming to increase part texture resolution to 4K and reduce surface error rate to below 10%.

Inadequate Deep Text Semantic Parsing: Except for 3D-GPT, all other models achieve a parsing accuracy of less than 76.5% for "spatially nested relationships"[8]. The Virtual Worlds Generator has a 31.8% error rate in parsing "dynamic target interaction instructions" [7]. Improvements should integrate the "text-image contrastive learning" framework[6] and fine-tune LLMs to incorporate

dynamic scene semantic knowledge[7], with the goal of increasing nested relationship parsing accuracy to over 90%.

Conflict Between Lightweight Deployment and Specialized Adaptation: Models like GPT4Motion and SceneX rely on heavyweight tools. The Virtual Worlds Generator needs 16GB VRAM for local deployment [7]. Improvements can reference DUNGEON ALCHEMIST's "template + traditional algorithm" approach, combined with lightweight LLMs [11], and tailor functions for specialized scenarios, retaining only the dynamic target generation module for MOT scenarios. The target is to reduce VRAM requirements for deployment to below 8GB.

Single Evaluation System: Most models only rely on metrics like FID and CLIP[12][6] and lack specialized scenario metrics [7]. Improvements should establish a dual evaluation system of "general metrics + specialized metrics"—for example, adding "dimensional precision error" for industrial scenes and "target tracking latency" for MOT scenarios. Referencing the "Two Time-Scale Update Rule (TTUR)" [3], the system should align evaluations with scenario requirements.

## 5. Conclusion

This study analyzes large model-driven 3D scene generation models, focusing on personal deployment and usability—critical for individual creators and small teams. Technically, three distinct paths have emerged, differing significantly in deployment requirements and practical capabilities.

The LLM-free tool path, exemplified by DUNGEON ALCHEMIST, offers lightweight advantages: sub-2GB installation, under 4GB memory usage, and 50m×50m fantasy static scenes generated in 3–5 minutes with <15% element overlap—serving basic basic needs for tabletop games or small game levels. Limitations include lack of text-driven functionality and restricted fantasy-style support, limiting adaptability

In contrast, LLM-driven general paths and LLM+specialized paths excel in semantic understanding and adaptability, handling dynamic scenes and precision industrial modeling. However, they face deployment challenges: reliance on heavy tools (Blender needs ≥8GB RAM; GPT-4 requires ≥24GB VRAM for local use) results in <30% personal deployment success. Virtual Worlds Generator demands 16GB VRAM, remaining inaccessible to most users—conflicting with lowering entry barriers.

Future development should prioritize optimizing personal deployment and usability through three directions:

1) Lightweight architectures inspired by DUNGEON ALCHEMIST's templates, integrating small LLMs for basic text parsing, with <500MB asset libraries and <8GB VRAM requirements for mid-range PCs.

2) User-friendly interfaces combining drag-and-drop and text shortcuts to prevent "free material stacking" while maintaining <15% element overlap.

3) Expanded templates beyond fantasy styles to include minimalist living rooms, 2D game levels, and short video backgrounds—balancing performance and practicality for skilled and non-professional creators alike.

## References

[1] Li X Y, Zhang Q, Kang D et al. Advances in 3D Generation A Survey. arXiv:2401.17807v1 [cs.CV] 31 Jan 2024.

[2] Liu, D Z. Liu Y, Huang W C, and Hu W. A Survey on Text-guided 3D Visual Grounding.arXiv:2406.05785v2 [cs.CV] 22 Jul 2024.

[3] Wen B C, Xie H Z, Chen Z X, Hong F Z, Liu Z W. 3D Scene Generation A Survey.arXiv:2505.05474v1 [cs.CV] 8 May 2025.

[4] Li H R, Tian Y R, Lan K, Liao Y. DreamScene 3D Gaussian-based Text-to-3D. arXiv:2507.13985v2 [cs.CV] 29 Jul 2025.

[5] Li H R, Shi H L, Zhang W L et al. DreamScene 3D Gaussian-based End-to-end Text-to-3D Scene Generation. arXiv:2507.13985. 2025.

[6] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020, 2021.

[7] Gaidon A, Wang Q, Cabon Y, Vig E. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. Dynamic Scene Generation and Evaluation. 2024.

[8] Xing T, Wu Y, Zhao Q C, et al. Digital Testing Scenario Generative Methods for Intelligent Algorithms Based on Large Language Models. Journal of Command and Control, 2025, 11(2): 239-247.

[9] Lv J X, Huang Y, Yan M F et al. GPT4Motion in Blender: Text-driven Dynamic Scene Generation with Physical Simulation. arXiv:2311.12631. 2024.

[10] Sun C Y, Han J L, Deng W J, Wang X L, Qin Z S, Stephen G. 3D-GPT: PROCEDURAL 3D MODELING WITH LARGE LANGUAGE MODELS. arXiv:2310.12945v2 [cs.CV] 29 May 2024.

[11] Zhou M, Hou J, Luo C, et al. SceneX: Procedural Controllable Large-scale Scene Generation via Large-language Models. arXiv preprint arXiv:2403.15698, 23 Mar 2024.

[12] Heusel M, Ramsauer H, Unterthiner T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium//31st Conference on Neural Information Processing Systems. 2017.