WorldAgents: Can Foundation Image Models be Agents for 3D World Models?

Ziya Erkoç, Angela Dai, and Matthias Nießner

Technical University of Munich, Germany
https://ziyaerkoc.com/worldagents

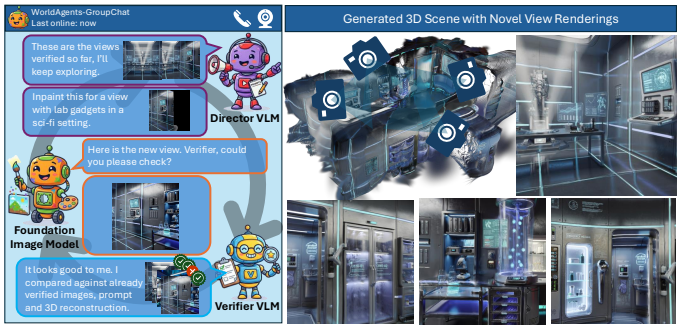

Fig. 1: WorldAgents employs 2D foundation models as agents in an iterative process to extract realistic, coherent 3D scenes from their learned distributions. The scene generation process is orchestrated by a VLM agent (director) that generates prompts for new views to synthesize, fulfilled by an image generation model that creates these views, and verified by a VLM agent that assesses both 2D and 3D consistency of the new synthesized image(s). The output is a sci-fi lab. scene reconstructed with 3DGS, and can be visualized from novel views for exploration.

Abstract. Given the remarkable ability of 2D foundation image models to generate high-fidelity outputs, we investigate a fundamental question: *do 2D foundation image models inherently possess 3D world model capabilities?* To answer this, we systematically evaluate multiple state-of-the-art image generation models and Vision-Language Models (VLMs) on the task of 3D world synthesis. To harness and benchmark their potential implicit 3D capability, we propose an agentic framing to facilitate 3D world generation. Our approach employs a multi-agent architecture: a VLM-based director that formulates prompts to guide image synthesis, a generator that synthesizes new image views, and a VLM-backed two-step verifier that evaluates and selectively curates generated frames from both 2D image and 3D reconstruction space. Crucially, we demonstrate that our agentic approach provides coherent and robust 3D reconstruction, producing output scenes that can be explored by rendering novel views. Through extensive experiments across various foundation models, we demonstrate that 2D models do indeed encapsulate a grasp of 3D worlds. By exploiting this understanding, our method successfully synthesizes expansive, realistic, and 3D-consistent worlds.

Keywords: 3D Scene Generation · Multi-agent Method · Text-to-scene Generation

1 Introduction

Recent rapid advances in 2D foundation models have revolutionized the field of computer vision. Text-to-image diffusion models demonstrate an unprecedented ability to generate high-fidelity, photorealistic images and exhibit deep semantic understanding of visual scenes [6, 16, 28, 36]. Trained on internet-scale datasets, these models encapsulate vast amounts of visual knowledge. While 2D generation has reached remarkable heights, the synthesis of immersive, 3D-consistent environments, often referred to as 3D world generation, remains a formidable challenge. Existing 3D generation methods [8, 10, 17, 32, 42, 44, 49] are frequently bottlenecked by the scarcity of diverse, high-quality 3D training data or the computational complexity of maintaining multi-view consistency through Score Distillation Sampling [29, 34, 45, 48].

Since 2D foundation models are trained on billions of 2D images, each of which represents 2D projections of our 3D spatial world, a compelling hypothesis emerges: these models may have implicitly learned the underlying spatial structures and physical rules of the environments they depict. This leads us to investigate a fundamental question: *do 2D foundation image models inherently possess 3D world model capabilities?* If these models in fact learn a robust prior of the 3D world, they could theoretically be leveraged to bypass the reliance on explicit 3D datasets, serving as powerful engines for 3D scene synthesis.

To answer this question, we systematically evaluate the implicit 3D spatial understanding of various state-of-the-art image generation models and VLMs. However, high-fidelity 3D reconstruction demands near pixel-perfect cross-view consistency, which single-pass prompting of a 2D model typically fails to guarantee. To harness and benchmark the potential implicit 3D capabilities of such 2D models, we propose a novel agentic method designed to orchestrate 2D foundation models for the task of consistent 3D world generation.

We thus cast 3D scene generation as a multi-agent process, comprising three specialized agents that work together to harness 2D image generation models to reconstruct coherent 3D worlds:

1. *VLM Director*, which acts as the high-level planner, dynamically formulating prompts to guide each new image generation and dictating the semantic evolution of the scene.
2. *Image Generator*, which employs a 2D image generation model that executes spatial navigation by sequentially inpainting to synthesize novel, geometrically aligned views. Since the image models do not have explicit control of the camera positions, we applied inpainting approach to guide image model to complete the scene by dictating what to paint.
3. *VLM 2-Stage Verifier*, which serves as the critical quality control mechanism. Unlike standard rigid pipelines, this verifier provides fine-grained evaluation

to selectively keep or discard generated frames. Crucially, it assesses consistency in two distinct stages: first from the 2D image space only, for semantic and structural coherence, and then from the 3D reconstruction space to guarantee strict geometric alignment.

We show that our method is able to effectively and efficiently elicit 3D scenes from 2D models. Our generator operates strictly via sampling from 2D image models without any 3D-aware weight updates. While we do employ 3D reconstruction to construct the 3D scenes and during the verification process, this 3D information never propagates back to the generative process and is instead resynthesized into updated text prompts and image conditions that guide 2D image sampling.

We found that our agentic approach yields robust 3D reconstructions, allowing us to freely explore various generated environments by rendering arbitrary novel views. Through extensive experiments, we demonstrate that 2D foundation models do, in fact, encapsulate a profound grasp of 3D worlds. By exploiting this latent understanding through our carefully designed multi-agent orchestration, our method overcomes the limitations of independent 2D generation to successfully synthesize expansive, realistic, and strictly 3D-consistent worlds. In summary, our main contributions are as follows:

- We provide a comprehensive investigation into the implicit 3D world model capabilities of state-of-the-art 2D image generation models guided by VLMs.
- We introduce a multi-agent architecture comprising a VLM director, a view generator, and a two-step verifier, specifically designed to harness 2D models for consistent 3D synthesis.

2 Related Works

2.1 3D World and Scene Generation

There has been recent focus in building 3D worlds from text prompt or input views [2, 9, 19, 21, 40, 41, 50, 56, 57]. In particular, various methods have been proposed to leverage the powerful generative capacity of image models or video diffusion models, coupled with 3D-based control, typically through camera controlled conditioning. A line of work approaches the problem in the form of a panorama image generation [50, 58]. LayerPano3D [50] employs a layered image generation task. They fine-tune Flux [27] to generate panorama images from text input. Unseen regions of each layer are inpainted with the same fine-tuned model. Additionally, DreamScene360 [58] includes a text-to-image diffusion model to generate a panorama image with a self-refinement mechanism using a VLM. Additionally, one line of work also uses image diffusion models to create 3D scenes, using point-cloud conditioned inpainting [41]. To fully leverage existing image diffusion models, our approach operates entirely without additional pre-training. Furthermore, we introduce a multi-agent framework that actively guides the entire generation pipeline, rather than acting solely as a verifier. Crucially, our

verification process ensures consistency directly within the 3D reconstruction space, moving beyond standard 2D image-domain validation. Our generation process includes iterative inpainting that does not require generating panorama images.

One line of work approaching this problem with both image- and depth-inpainting [21, 55]. Both WonderWorld [55] and Text2Room [21] are using handcrafted prompts to synthesize new regions to create 3D scenes. In contrast, we do not use handcrafted prompts but rely on VLM-based agents to orchestrate the scene generation process to construct navigable 3D scenes. Additionally, we employ an iterative, image-based inpainting strategy for scene generation to demonstrate the high degree of 3D consistency achievable without relying on explicit depth inpainting. Another line of work in scene generation includes retrieval-based layout-generation methods [17, 30, 43, 44, 51, 52]. These methods rely on 3D layout data to be trained on which is orders-of-magnitude less than what image models have. Major direction in video-based scene generation is camera-controlled models [3, 4, 57]. Following advances in video synthesis, Stable Virtual Camera [57] demonstrated scene navigation and traversal through fine-tuning video diffusion models to provide camera-controlled multi-view generation. This creates compelling novel-view synthesis that can be used for further 3D reconstruction. WorldExplorer [40], took this approach further to generate large 3D scenes that can be 3D reconstructed and arbitrarily rendered from novel views. Unlike these approaches, we employ VLM-based agents to guide the process and verify the generated frames without requiring crafting of a trajectory generation process. Our approach does not use any fine-tuned camera-controlled model but relies on existing text- and image-to-image 2D foundation models.

2.2 2D Foundation Image Models

Past years have seen unprecedented advances in 2D image generation models [16, 22, 28, 33, 36, 46]. Recent models can be conditioned on both text and multiple-images at the same time to achieve text-conditioned editing capabilities. Various methods have thus been built on top of these models to explore the capabilities for other downstream tasks such as 3D reconstruction using SDS (Score Distillation Sampling) and personalization [18, 34, 35, 37, 38]. Such image foundation models show strong capabilities in those downstream tasks. Their powerful generative and perception capacity have also inspired our approach to leverage image foundation models to their full extent to understand if they would generate 3D-consistent views. Nano Banana [11, 46] and Flux.2 [28] are some of the most recent methods that can generate high-fidelity images within a few seconds. We aim to exploit the full power of those image synthesis models to generate traversable 3D scenes.

2.3 Agent-Driven Generation and VLM Evaluators

Recently, agent-based methods have achieved remarkable success across various domains [14, 17, 23, 43, 54]. These approaches leverage the robust visual and tex-

tual reasoning capabilities of Vision-Language Model (VLM) agents to tackle diverse tasks. Closest to our work is VIGA [54], which translates images into 3D scenes by generating corresponding Blender [12] code. Their experiments demonstrate that VLMs possess a deep semantic understanding of scenes and can effectively manipulate code representations for image-to-3D reconstruction. Inspired by the strong reasoning capabilities of VLMs and recent advancements in 2D foundation models, we introduce a method for frame-by-frame 3D world generation. Unlike VIGA, which relies on a proxy code representation for static 3D reconstruction, our method directly generates image frames, and our ultimate objective is the synthesis of interactive, navigable 3D worlds from text prompts.

3 Method

Figure 2 presents an overview of our proposed method. We formulate 3D scene generation as a collaborative process orchestrated by three specialized agents: a Generator, a Verifier, and a Director. The Generator is a 2D image foundation model capable of text- and image-conditioned synthesis; we leverage it to inpaint specific regions based on scene captions generated by Director. To maintain global consistency, the Verifier evaluates each newly generated image against a history of previously accepted views. It concurrently maintains an intermediate 3D reconstruction to ensure the generated views form a geometrically coherent 3D space. The overall iterative process is guided by the Director, which analyzes the verified view history to propose descriptive prompts for novel viewpoints. Once the Director determines that the scene is comprehensively covered, the generation process terminates, and the accumulated views are utilized to reconstruct the final 3D Gaussian Splatting (3DGS) representation using AnySplat [24].

3.1 Problem Formulation

Given an input text description y_1, our aim is to generate a spatially coherent 3D scene $\mathcal{W}$ representative of y_1. Concretely, we produce a set of N posed images that collectively define a 3D world and can be reconstructed as 3D Gaussians [25] to enable navigation and exploration of the 3D world. We provide agents with a total of $\hat{R}$ number of tries to generate N images. We additionally include the text description for each frame in our world state. In total, our world state is composed of a series of verified 2D images, their camera poses and corresponding text prompt, which are acquired through agentic process employing 2D foundation models and VLMs. Formally, we represent this scene as:

$$\mathcal{W} = \{(I_i, P_i, y_i)\}_{i=1}^{N}$$

where each I_i is a high-fidelity image and P_i is its corresponding absolute camera pose in the global coordinate system. The initial frame generation step is a special case, as it does not involve Director agent. It is just a text-to-image generation task using y_1.

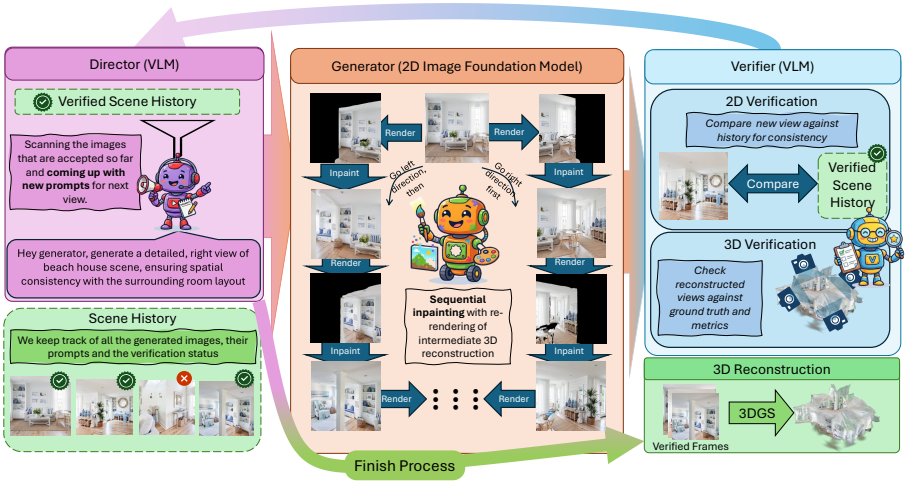

Fig. 2: Method overview. Our method employs a multi-agent approach comprising a Director, a Generator, and a Verifier to construct coherent 3D scenes using image diffusion models. The Director guides the overall process by formulating novel prompts. The Generator then leverages sequential inpainting to synthesize 3D-consistent views. Subsequently, the Verifier evaluates these generated views to ensure rigorous multi-view consistency. Finally, the verified frames are reconstructed into a 3D Gaussian Splatting (3DGS) representation.

We formulate 3D world generation as an iterative, agent-directed process. At each discrete time step t, a director agent $\mathcal{D}$ analyzes the current world state $\mathcal{W}_t = \{(I_i, P_i, y_i)\}_{i=1}^{t}$ to propose how to expand the region by generating text prompt y_{t+1}. When the overall generation has gone through $\frac{\hat{R}}{2}$ tries, it switches to exploring left. Next camera view is calculated as $P_{t+1} = P_t \cdot \Delta P_t$, where ΔP_t is a relative transformation of camera towards either right- or left-direction in a fixed amount, it also contains a random perturbation to create more diverse coverage for the next view.

Given the previous view I_t and the new camera pose P_{t+1}, a generator agent $\mathcal{G}$ relies on a 2D foundation model to synthesize a candidate view $\hat{I}_{t+1}$.

As 2D foundation models can be prone to structural hallucinations that violate multi-view geometry constraints, we introduce a strict 2-Stage Verifier $\mathcal{V}$. The Verifier acts as a binary gating function that evaluates the candidate $\hat{I}_{t+1}$ against the established world $\mathcal{W}_t$ across both 2D semantic space and 3D reconstruction space:

$$\mathcal{V}(\hat{I}_{t+1}, \mathcal{W}_t) = \begin{cases} 1, & \text{if } \hat{I}_{t+1} \text{ is semantically \& geometrically consistent with } \mathcal{W}_t \\ 0, & \text{otherwise} \end{cases}$$

The candidate view is appended to the global state (i.e., $\mathcal{W}_{t+1} = \mathcal{W}_t \cup \{(\hat{I}_{t+1}, P_{t+1}, y_{t+1})\}$) if and only if $\mathcal{V}(\hat{I}_{t+1}, \mathcal{W}_t) = 1$. If rejected, the candidate is discarded, and the generation step is re-sampled. By optimizing this discrete acceptance criterion, our approach guarantees that the final generated world $\mathcal{W}$

adheres to multi-view constraints while exploiting the superior visual fidelity of the underlying 2D foundation model.

The process ends when we reach $\hat{R}$ maximum number of tries, N images, or the director agent concludes that all of the scene is observed and gives stop signal.

3.2 Director Agent

The Director agent, denoted as $\mathcal{D}$, serves as the semantic orchestrator of the 3D world synthesis process. To prevent the semantic drift and unconstrained wandering typical of autoregressive video generation, $\mathcal{D}$ dynamically computes the next logical viewpoint based on the exploration history.

At each time step t, the Director observes the current state of the generated world $\mathcal{W}_t$ alongside the overarching global text prompt y, which is equivalent to y_1. It is parameterized by a Vision-Language Model (VLM) that acts as a policy, mapping this environmental context to a view-specific text prompt y_{t+1} by checking out the world state and the previous prompts: $y_{t+1} = \mathcal{D}(\mathcal{W}_t)$.

Concurrently, y_{t+1} explicitly defines the expected visual content from the new perspective, providing strict semantic conditioning for the Generator. It includes textual description of where to investigate and what to be included in that part of the scene when the camera pose changes. The y_1 is already provided as an input, therefore, the Director agent is not involved in the generation of first frame.

By prompting the VLM to iteratively predict y_{t+1}, our framework functions as an autonomous, context-aware semantic operator, ensuring that the exploration trajectory creates meaningful scenes strictly aligned with the global semantic prior y. For instance, our director agent suggests following prompts in sci-fi scene for one iteration *"expand further right, seamlessly continuing the sleek metallic wall panels ... wrapping blue and cyan neon strips ... a large, translucent cylindrical containment unit with softly pulsing blue lights ... embed a recessed digital control panel"*. It provides comprehensive and semantically-rich prompts for the next view by also keeping the overall context in the sci-fi scene.

Our trajectory procedure starts from the first frame and first goes in the right direction and then the left direction. We prompt Director about which direction we are heading now. After that, we apply fixed rotation of ϕ degrees around the up-axis to form R_{fixed}. To increase coverage diversity we apply random transformation, T_{random} on top of that.

$$\underbrace{P_{t+1}}_{4\times4} = \underbrace{P_t}_{4\times4} \cdot \underbrace{T_{\text{random}}}_{4\times4} \cdot \underbrace{R_{\text{fixed}}}_{4\times4}$$

We calculate P_{t+1} using that formula, where $\cdot$ means matrix multiplication, and $\Delta P_t = T_{\text{random}} \cdot R_{\text{fixed}}$. If the process has gone through $\frac{\hat{R}}{2}$ tries, the director switches the process to exploring left of the initial frame.

3.3 Generator Agent

The Generator agent, $\mathcal{G}$, is tasked with synthesizing a high-fidelity candidate view $\hat{I}_{t+1}$ that adheres to the semantic conditioning y_{t+1} provided by the Director and geometric transformation P_{t+1}. To embed 3D structure and camera awareness into the 2D generation process, we reinterpret the 2D generative model as sequential inpainting. Each new image is conditioned on re-rendered views based on the reconstruction from previously generated views, ensuring geometric consistency across the scene.

In order to generate a new image, we first collect the $\mathcal{W}_t$ to reconstruct a 3DGS scene. We then re-render the scene from a new view to provide it as an input to our image diffusion model. Specifically, we utilize AnySplat [24] to lift $\mathcal{W}_t$ into a global set of 3D Gaussians, denoted as Θ_t, $\Theta_t = \mathcal{F}_{\text{AnySplat}}(\mathcal{W}_t)$.

To continue exploring the environment and synthesize the subsequent novel view, we compute a target camera pose P_{t+1}, which includes fixed rotation towards either left- or right-direction. Rather than relying on a strictly deterministic trajectory, we introduce a stochastic exploration mechanism by applying a randomly perturbed transformation so that the generator can get a more diverse coverage of the scene. Finally, we employ the Gaussian rasterizer $\mathcal{R}$ to render the reconstructed scene from the novel viewpoint P_{t+1}. This yields the rendered image I_{t+1}^{warp}, $I_{t+1}^{\text{warp}} = \mathcal{R}(\Theta_t, P_{t+1})$.

Due to camera translation and rotation, I_{t+1}^{warp} inevitably contains missing regions caused by disocclusions and new camera field of view. We leverage a pre-trained 2D foundation model, $\mathcal{G}_{\text{inpaint}}$, to complete the missing visual information. The model is conditioned on the known warped pixels, and the localized text prompt y_{t+1} from director, $\hat{I}_{t+1} = \mathcal{G}_{\text{inpaint}}(I_{t+1}^{\text{warp}}, y_{t+1})$.

By grounding the generation process in explicit 3D reprojection before applying the 2D generative prior, the Generator ensures that the overlapping regions between I_t and $\hat{I}_{t+1}$ remain geometrically rigidly aligned, while the foundation model is constrained purely to filling in the structurally logical, disoccluded regions.

3.4 2D & 3D Verifier Agents

Because 2D foundation models are prone to structural hallucinations and perspective distortions, we introduce a rigorous 2-Stage Verifier agent, $\mathcal{V}$, to act as a definitive gating mechanism for which images should compose the 3D world. The Verifier ensures that the candidate view $\hat{I}_{t+1}$ is both semantically aligned with the Director's intent and strictly geometrically consistent with the established 3D world $\mathcal{W}_t$. The verification is decomposed into a 2D semantic check and a 3D reconstruction-space check.

Image-Space Verification First, we employ a Vision-Language Model (VLM) to assess the semantic coherence and visual quality of the candidate image. The VLM, denoted as $\mathcal{V}_{\text{2D}}$, takes the candidate view $\hat{I}_{t+1}$, the world state $\mathcal{W}_t$, and the director's prompt y_{t+1} to detect obvious visual artifacts, domain shifts, or

prompt misalignment. The output of the verifier is a binary decision: $v_{\text{2D}} = \mathcal{V}_{\text{2D}}(\hat{I}_{t+1}, \mathcal{W}_t, y_{t+1}) \in \{0, 1\}$.

3D Reconstruction-Space Verification Even if a candidate frame is semantically plausible in 2D, it may harbor subtle geometric distortions that violate multi-view consistency. To enforce strict global 3D consistency, we assess how the introduction of the candidate view $\hat{I}_{t+1}$ impacts the overall integrity of the 3D reconstruction of the scene. We define a provisional global state $\mathcal{W}'_{t+1} = \mathcal{W}_t \cup \{(\hat{I}_{t+1}, P_{t+1}, y_{t+1})\}$, representing all verified frames up to step t plus the new candidate. We lift this provisional set into a unified 3D representation using AnySplat [24], yielding a candidate 3DGS model: $\Theta'_{t+1} = \mathcal{F}_{\text{AnySplat}}(\mathcal{W}'_{t+1})$.

To quantify the global structural integrity, we render the scene from all historical camera poses $P_i \in \{P_1, \ldots, P_{t+1}\}$ to obtain a set of reconstructed views $\{I_1^{\text{render}}, \ldots, I_{t+1}^{\text{render}}\}$. We then compute standard novel view synthesis metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics are computed between every input frame I_i (where $I_{t+1} = \hat{I}_{t+1}$) and its corresponding rendered view I_i^{render}:

$$s_{\text{metrics}}^{(i)} = \left\{\text{PSNR}(I_i, I_i^{\text{render}}), \text{SSIM}(I_i, I_i^{\text{render}}), \text{LPIPS}(I_i, I_i^{\text{render}})\right\}$$

An inconsistent candidate view will force the AnySplat model to compromise its geometry, resulting in a noticeable degradation in the reconstruction quality of the previously verified frames. We aggregate these metrics into a global quality profile $S_{\text{global}} = \{s_{\text{metrics}}^{(i)}\}_{i=1}^{t+1}$.

Rather than relying on rigid, empirically defined thresholds, we pass these quantitative scores S_{global} alongside the visual image pairs $(I_i, I_i^{\text{render}})_{i=1}^{t+1}$ to our dedicated verifier VLM, $\mathcal{V}_{\text{3D}}$. This multimodal agent holistically evaluates the global geometric stability to determine if the candidate maintains the structural integrity of the 3D world:

$$v_{\text{3D}} = \mathcal{V}_{\text{3D}}(\mathcal{W}_t, \{I_i, I_i^{\text{render}}\}_{i=1}^{t+1}, S_{\text{global}}) \in \{0, 1\}$$

Final Decision The final acceptance decision is the logical conjunction of both verification steps: $\mathcal{V}(\hat{I}_{t+1}, \mathcal{W}_t) = v_{\text{2D}} \wedge v_{\text{3D}}$. If $\mathcal{V}(\hat{I}_{t+1}, \mathcal{W}_t) = 1$, the candidate is accepted into the global state, updating the verified history. If rejected, the frame is discarded, and the Generator is prompted to re-sample a new candidate. This closed-loop evaluation guarantees that the final synthesized world strictly adheres to multi-view geometry without sacrificing the high visual fidelity of the 2D foundation model. If the generator cannot generate a successfully verified image after $\hat{r}$ tries, both the director output y_{t+1} and P_{t+1} are recalculated to give the model another opportunity to explore.

4 Experiments

We evaluate our method for 3D world generation, analyzing various 2D image generative models [28, 46] and VLMs [1, 5], as well as comparing with state-of-

the-art 3D generative methods Stable Virtual Camera (SEVA) [57], WonderWorld [55], Text2Room [21], and WorldExplorer [40]. We used CLIP Score [20], Inception Score [39], CLIP-IQA [47] and Reprojection Error [15] metrics to numerically evaluate all methods, with novel view renderings. We have used various 2D foundation models and VLMs for our agents. For the former, we have Flux.2 [Klein] 9B, Flux.2 [Pro] and Nano Banana v1 (Gemini 2.5 Flash). We run "Klein" version locally but use API access for "Pro" version. For VLMs, we use OpenAI API for GPT4.1 access but use Qwen3-VL 8B locally. Finally, for Nano Banana we used the official API. We use RTX A6000 GPU for the local deployments. In our experiments, we pick N as 14, $\hat{R}$ as 28 and $\hat{r}$ as 2. It takes approximately 25 minutes to generate a scene with Flux.2 [Pro] and GPT 4.1. We present results from our method in Figure 3.

4.1 Comparisons with State of the Art

We evaluate our method against state-of-the-art text-to-3D scene generation baselines: the image diffusion-based WonderWorld [55] & Text2Room [21] and the video diffusion-based SEVA [57] & WorldExplorer [40]. While baselines successfully generate 3D environments from text prompts, our approach demonstrates significant improvements in overall rendering fidelity and scene complexity. As illustrated in Figure 5, our method generates a sci-fi room characterized by rich geometric details and high object density. In contrast, the baselines produce sparser scenes that lack structural realism. Furthermore, in the kitchen scenario, the baseline methods exhibit severe structural artifacts and noticeable blurring around object boundaries. These qualitative observations are supported by our quantitative evaluation (Table 1). Our method achieves better performance across standard metrics, confirming its enhanced prompt alignment and overall scene quality. We additionally evaluated 3D consistency by calculating reprojection error using the method in WorldScore [15]. The results are presented in Table 2. Our method performs better than the baselines in terms of 3D consistency.

4.2 Image Model and VLM Analysis

We evaluated various combinations of open- and closed-source 2D image foundation models and Vision-Language Models (VLMs). Generally, all evaluated models yield plausible novel views and coherent subsequent 3D reconstructions. However, generation quality varies depending on the capacity of the underlying models. For instance, as illustrated in Figure 4, Flux.2 [Klein] occasionally produces geometrically inconsistent intersecting objects and fails to complete certain views. Similarly, Nano Banana 1 demonstrates lower efficacy in inpainting tasks, sometimes leaving target regions unpainted. We additionally applied BAGEL [13] model as the generator. It did not perform as well as other evaluated combinations. Furthermore, while Qwen3 exhibits strong general capabilities, it occasionally issues less accurate directives and verifications, leading to performance degradation. These qualitative observations are consistent with the

quantitative results reported in Table 1. Empirically, the combination of Flux.2 [Pro] and GPT-4.1 achieves the most favourable performance.

Fig. 3: Visual results from WorldAgents. Our method can generate diverse scenes that are populated with various objects in a clean and coherent way by following the text prompt.

4.3 Ablations

To evaluate the effectiveness of individual components, we conduct a series of ablation studies. In our initial baseline, we isolate the image generator. It has access to all of the verified frames and the text prompt. It tries to generate a new frame by also being consistent with the previous frames. Without the verifier module, the method accepts all synthesized frames unconditionally. To proxy the director agent's functionality, we augment the input text with a stochastic camera prompt specifying relative viewpoint shifts (left, right, or backward) with respect to the preceding frame. In the subsequent setting, we integrate the verifier module into this baseline to actively curate and filter the generated frames. Then, we append our director agents as well as the inpainting approach. Figure 6 shows

Fig. 4: Qualitative comparison of different image models and VLMs. The image models that we experimented with all showed satisfactory 3D scene generation results. However, there are subtle differences between them aligning with the complexity of the individual model.

Table 1: Quantitative comparison of 3D world generation methods. We evaluate prompt & semantic alignment and image quality using CLIP Score (CS) and Inception Score (IS), and CLIP Image Quality Assessment (IQA). Our method shows improvement over baselines. We also show multiple image models and VLMs to show our approach is applicable to multiple models and they have varying performances.

Method	**CS ↑**	**IS ↑**	**IQA ↑**
SEVA [57]	25.02	2.06	0.61
WonderWorld [55]	26.67	2.48	0.67
Text2Room [21]	22.27	**2.79**	0.27
WorldExplorer [40]	24.49	2.12	0.58
Ours (Flux.2 [Pro] + GPT 4.1)	**26.79**	2.26	**0.89**
Ours (Flux.2 [Pro] + Qwen3-VL 8B)	24.49	2.23	0.75
Ours (Flux.2 [Klein] 9B + GPT 4.1)	26.47	2.03	0.84
Ours (Flux.2 [Klein] 9B + Qwen3-VL 8B)	24.85	2.32	0.75
Ours (Nano Banana 1 + GPT 4.1)	25.89	2.14	0.70
Ours (BAGEL [13] + GPT 4.1)	18.97	2.41	0.55

the novel view renderings from each ablated method for an urban loft scene. Table 3 contains numerical comparison of the ablation of our components.

What is the impact of Verifier Agent? The verifier agent mitigates the risk of incorporating geometrically inconsistent views that would otherwise degrade the 3D reconstruction. Because image diffusion models can occasionally hallucinate artifacts or generate structurally incompatible layouts, injecting these anomalies into the pipeline causes irreversible corruption to the global geometry. To prevent this, the verifier autonomously filters erroneous generations. As demonstrated in Table 3, the inclusion of this module effectively removes catastrophic failures, leading to an improvement in reconstruction quality. As shown in Figure 6, it reduces the significant blur since it avoids inconsistent frames.

How much Director agent contributes to the overall process? Naively prompting the image diffusion model for iterative frame generation relies on static scene

Fig. 5: Qualitative baseline comparison. Our method generates visually appealing results with multiple objects placed in the scene nicely, without having artifacts or objectless regions unlike baselines.

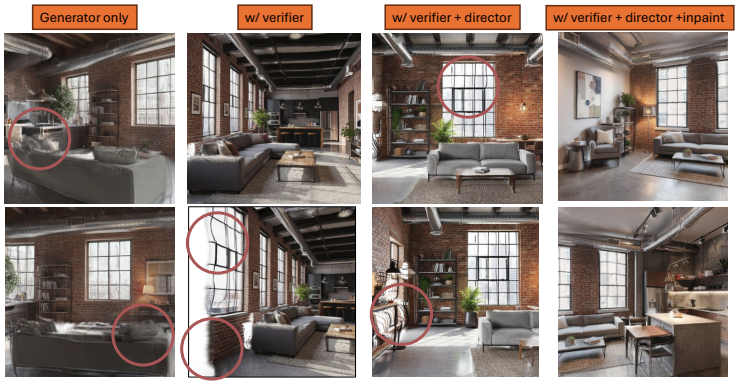

Fig. 6: Qualitative comparison of ablations. Relying solely on the generator yields blurry results. The addition of the Verifier reduces blur and improves consistency; however, the scene remains incomplete with visible artifacts in the windows. While the Director helps with completion of the scene, window misalignments persist. Integrating all components, successfully resolves these issues to synthesize a coherent scene.

Table 2: Quantitative comparison on 3D reprojection metric against baselines. We calculate the reprojection error using WorldScore [15]. Our method shows improved 3D consistency.

Method	Reproj. Error ↓
SEVA [57]	0.61
WonderWorld [55]	0.46
Text2Room [21]	0.45
WorldExplorer [40]	0.39
Ours	**0.36**

descriptions coupled with random camera poses (e.g., "left", "right"). However, this approach often leads to semantic redundancy and object duplication; for instance, continuously conditioning the generator on same prompt biases the model to hallucinate duplicate objects in the views rather than exploring new contextual elements that limits diversity of objects in the scene. To address this, we introduce the director agent, which leverages a Vision-Language Model (VLM) prior to dynamically generate context-aware, explorative prompts conditioned on previously verified views. Table 3 validates that VLM-based prompting strategy significantly enhances the completeness, structural diversity and the overall fidelity of the generated 3D environments.

Does sequential inpainting help with increasing generation quality? Relying solely on unconstrained image generation does not strictly enforce multi-view consistency, even when the model is conditioned on previous frames and precise textual prompts. To explicitly anchor novel views to the existing geometry, our method utilizes a 3D-aware inpainting step. Specifically, we render an intermediate 3D representation from a novel, unexplored camera pose and instruct the diffusion

Table 3: Quantitative ablation results. We conducted the study with Flux.2 [Pro] + GPT 4.1 setting. We evaluate CLIP Score (CS), Inception Score (IS), and CLIP-IQA (IQA). Our individual components contribute to the overall rendering quality and prompt alignment performance.

Generator	Verifier	Director	Inpaint	CS ↑	IS ↑	IQA ↑
✓				19.07	2.23	0.60
✓	✓			20.24	2.43	0.62
✓	✓	✓		21.80	**2.94**	0.69
✓	✓	✓	✓	**26.79**	2.26	**0.89**

model to inpaint the unobserved regions, guided by the director agent's localized prompt. As shown in Table 3, omitting this inpainting formulation results in degraded prompt alignment and scene coherence, highlighting its importance for maintaining rigorous 3D consistency throughout the generation process. Figure 6 also shows that we eventually get complete scenes.

Limitations In this work, we employ VLMs as high-level directors and evaluators to distill 3D-consistent, navigable environments from 2D image diffusion models. A highly promising direction for future research is extending this VLM-guided method to existing video diffusion models [7, 26, 31, 53]. While modern video priors naturally exhibit temporal coherence over short sequences, they frequently accumulate geometric drift and multi-view inconsistencies over longer, exploratory spatial trajectories. Integrating VLM-based agents into a video generation method could effectively regularize these models, thereby mitigating long-horizon degradation and enabling the synthesis of even larger 3D worlds.

5 Conclusion

In this paper, we investigated the fundamental question of whether 2D foundation image models inherently possess 3D world model capabilities. To explore this, we introduced a novel multi-agent framework comprising a VLM-based director, an image generator, and a two-step verifier that operates across both 2D and 3D spaces. Our extensive evaluations demonstrate that by carefully framing the generation process, we can successfully extract implicit 3D knowledge from these 2D models. Our agentic approach yields expansive, robust, and 3D-consistent scene reconstructions that support novel view rendering. Our findings confirm that 2D foundation models do encapsulate a grasp of 3D worlds. Future work can explore extending this multi-agent framework to video diffusion models or dynamic 4D scene generation for interactive world synthesis.

Acknowledgments This work was supported by Chaos Software, as well as the ERC Consolidator Grant Gen3D (101171131) of Matthias Nießner, the ERC Starting Grant SpatialSem (101076253) of Angela Dai and the Georg Nemetschek Institute (GNI) NeRF2BIM grant. We would like to thank Marc Benedí for helpful discussions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bahmani, S., Shen, T., Ren, J., Huang, J., Jiang, Y., Turki, H., Tagliasacchi, A., Lindell, D.B., Gojcic, Z., Fidler, S., Ling, H., Gao, J., Ren, X.: Lyra: Generative 3d scene reconstruction via self-distillation with video diffusion models. In: International Conference on Learning Representations (ICLR) (2026)
3. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers (2025), https://arxiv.org/abs/2411.18673
4. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Vd3d: Taming large video diffusion transformers for 3d camera control (2025), https://arxiv.org/abs/2407.12781
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
7. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), https://arxiv.org/abs/2311.15127
8. Bokhovkin, A., Meng, Q., Tulsiani, S., Dai, A.: Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation (2024), https://arxiv.org/abs/2412.01801
9. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis (2025), https://arxiv.org/abs/2503.13265
10. Chen, Y., He, T., Huang, D., Ye, W., Chen, S., Tang, J., Chen, X., Cai, Z., Yang, L., Yu, G., Lin, G., Zhang, C.: Meshanything: Artist-created mesh generation with autoregressive transformers (2024), https://arxiv.org/abs/2406.10163
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org, accessed: 30 June 2026
13. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining (2025), https://arxiv.org/abs/2505.14683
14. Deng, Z., Shi, Y., Ji, K., Xu, L., Huang, S., Wang, J.: Human-object interaction via automatically designed vlm-guided motion policy (2026), https://arxiv.org/abs/2503.18349
15. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation (2025), https://arxiv.org/abs/2504.00983

16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
17. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems **36**, 18225–18250 (2023)
18. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
19. Garcin, S., Walker, T., McDonagh, S., Pearce, T., Bilen, H., He, T., Wang, K., Bian, J.: Beyond pixel histories: World models with persistent 3d state (2026), https://arxiv.org/abs/2603.03482
20. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022), https://arxiv.org/abs/2104.08718
21. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023)
22. Hu, D., Chen, J., Huang, X., Coskun, H., Sahni, A., Gupta, A., Goyal, A., Lahiri, D., Singh, R., Idelbayev, Y., Cao, J., Li, Y., Cheng, K.T., Chan, S.H., Gong, M., Tulyakov, S., Kag, A., Xu, Y., Ren, J.: Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. arXiv:2412.09619 [cs.CV] (2024)
23. Jain, S., Gupta, K., Gupta, K., Chandraker, M.: Nerfify: A multi-agent framework for turning nerf papers into code (2026), https://arxiv.org/abs/2603.00805
24. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
26. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025), https://arxiv.org/abs/2412.03603
27. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
28. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
29. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation (2023), https://arxiv.org/abs/2211.10440
30. Lin, C., Mu, Y.: Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior (2024), https://arxiv.org/abs/2402.04717

31. Menapace, W., Siarohin, A., Skorokhodov, I., Deyneka, E., Chen, T.S., Kag, A., Fang, Y., Stoliar, A., Ricci, E., Ren, J., et al.: Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7038–7048 (2024)
32. Meng, Q., Li, L., Nießner, M., Dai, A.: Lt3sd: Latent trees for 3d scene diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 650–660 (2025)
33. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
34. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion (2022), https://arxiv.org/abs/2209.14988
35. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Mildenhall, B., Ruiz, N., Zada, S., Aberman, K., Rubenstein, M., Barron, J., Li, Y., Jampani, V.: Dreambooth3d: Subject-driven text-to-3d generation. ICCV (2023)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
38. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models (2024), https://arxiv.org/abs/2307.06949
39. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf
40. Schneider, M.A., Höllein, L., Nießner, M.: Worldexplorer: Towards generating fully navigable 3d scenes. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
41. Schwarz, K., Rozumny, D., Bulò, S.R., Porzi, L., Kontschieder, P.: A recipe for generating 3d worlds from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3520–3530 (2025)
42. Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers (2023), https://arxiv.org/abs/2311.15475
43. Sun, F.Y., Liu, W., Gu, S., Lim, D., Bhat, G., Tombari, F., Li, M., Haber, N., Wu, J.: Layoutvlm: Differentiable optimization of 3d layout via vision-language models (2025), https://arxiv.org/abs/2412.02193
44. Tang, J., Nie, Y., Markhasin, L., Dai, A., Thies, J., Nießner, M.: Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20507–20518 (2024)
45. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation (2024), https://arxiv.org/abs/2309.16653

46. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
47. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring clip for assessing the look and feel of images (2022), https://arxiv.org/abs/2207.12396
48. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation (2023), https://arxiv.org/abs/2305.16213
49. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation (2025), https://arxiv.org/abs/2412.01506
50. Yang, S., Tan, J., Zhang, M., Wu, T., Li, Y., Wetzstein, G., Liu, Z., Lin, D.: Layerpano3d: Layered 3d panorama for hyper-immersive scene generation (2025), https://arxiv.org/abs/2408.13252
51. Yang, Y., Jia, B., Zhi, P., Huang, S.: Physcene: Physically interactable 3d scene synthesis for embodied ai (2024), https://arxiv.org/abs/2404.09465
52. Yang, Y., Sun, F.Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., Wu, J., Haber, N., Krishna, R., Liu, L., Callison-Burch, C., Yatskar, M., Kembhavi, A., Clark, C.: Holodeck: Language guided generation of 3d embodied ai environments (2024), https://arxiv.org/abs/2312.09067
53. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
54. Yin, S., Ge, J., Wang, Z.Z., Li, X., Black, M.J., Darrell, T., Kanazawa, A., Feng, H.: Vision-as-inverse-graphics agent via interleaved multimodal reasoning. arXiv preprint arXiv:2601.11109 (2026)
55. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5916–5926 (2025)
56. Zhang, Y., Cao, C., Wang, T., Zuo, X., Wu, J., Zhu, J., Guo, C.: Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories (2026), https://arxiv.org/abs/2603.02049
57. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12405–12414 (2025)
58. Zhou, S., Fan, Z., Xu, D., Chang, H., Chari, P., Bharadwaj, T., You, S., Wang, Z., Kadambi, A.: Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting (2024), https://arxiv.org/abs/2404.06903