OneWorld: Taming Scene Generation with 3D Unified Representation Autoencoder

Sensen Gao [1*], Zhaoqing Wang [2*], Qihang Cao [1], Dongdong Yu [2], Changhu Wang [2], Tongliang Liu [4,3†], Mingming Gong [5,3†], and Jiawang Bian [1†]

[1] Nanyang Technological University, Singapore
[2] AISphere, Beijing, China
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[4] University of Sydney, Sydney, Australia
[5] University of Melbourne, Melbourne, Australia

Abstract. Existing diffusion-based 3D scene generation methods primarily operate in 2D image/video latent spaces, which makes maintaining cross-view appearance and geometric consistency inherently challenging. To bridge this gap, we present OneWorld, a framework that performs diffusion directly within a coherent 3D representation space. Central to our approach is the 3D Unified Representation Autoencoder (3D-URAE); it leverages pretrained 3D foundation models and augments their geometry-centric nature by injecting appearance and distilling semantics into a unified 3D latent space. Furthermore, we introduce token-level Cross-View-Correspondence (CVC) consistency loss to explicitly enforce structural alignment across views, and propose Manifold-Drift Forcing (MDF) to mitigate train–inference exposure bias and shape a robust 3D manifold by mixing drifted and original representations. Comprehensive experiments demonstrate that OneWorld generates high-quality 3D scenes with superior cross-view consistency compared to state-of-the-art 2D-based methods.

Keywords: 3D Scene Generation · 3D Unified Representation Autoencoder · Cross-View Correspondence · Manifold-Drift Forcing

1 Introduction

3D scene generation maps conditioning inputs, such as text or images, to a 3D scene representation, producing either explicit outputs like point clouds or 3DGS [21], or implicit outputs like video. It has emerged as vital technology for gaming, robotics, and VR/AR by enabling the creation of photorealistic environments at scale [7, 23, 58, 61]. This capability provides essential training data for simulations and offers powerful new tools for creative content design. By jointly modeling scene geometry, appearance, and spatial structure, it serves as the backbone for building interactive and geometrically consistent digital worlds.

* Co-first authors; † Corresponding authors.

Fig. 1: (a) OneWorld generates 3DGS from a single view and renders novel views. (b) Architecture: FlashWorld [27] diffuses in compressed video latents; Gen3R [17] compresses 3D features to align a 3D foundation encoder with video latents but can only generate geometry and appearance separately, rather than jointly. OneWorld generates directly in a unified 3D representation space, without compression or separate generation. (c) Performance comparison on WorldScore [8] and DL3DV [30].

Early 3D scene generation methods often rely on pretrained 2D generative priors, either through optimization-based score distillation [28, 38, 46, 52] or by synthesizing multi-view images or videos followed by 3D reconstruction [11, 16, 32, 44, 55]. While these approaches can produce plausible results, they are limited by expensive per-scene optimization and inconsistent geometry due to weak explicit 3D modeling. More recently, several works move generation into multi-view or video latent spaces by freezing an image or video VAE (See Fig. 1 (b)) and learning a latent-to-3D decoder (often 3DGS [21]) while adapting the diffusion backbone [6, 12–14, 27, 56]. This improves efficiency, but 2D latent spaces provide limited explicit 3D coupling across views, which may weaken multi-view coherence. Meanwhile, transformer-based 3D feed-forward models [2, 24, 29, 49–51] have advanced rapidly. Models such as Dust3R [50], VGGT [49] and π^3 [51] achieve strong performance in a wide range of 3D downstream tasks. Inspired

by recent image generation approaches that operate directly in the representation space of vision foundation models (*e.g.*, DINOv2 [35]) rather than relying on VAEs [1,3,43,65], we explore whether it is possible to learn 3D scene generation directly within the representation space of a 3D foundation model.

The concurrent work Gen3R [17] compresses the 3D representation space to align with a VAE video latent space [48], limiting the representational capacity of the learned 3D representation. In addition, it generates geometry (point clouds) and appearance (video) separately (See Fig. 1 (b)), constraining a unified 3D representation with coherent geometry and high-fidelity appearance. Furthermore, its reliance on video-latent alignment constrains geometric modeling to small-baseline view variations, limiting generalization to large viewpoint changes.

To address these challenges, we present OneWorld, a framework that performs diffusion within a coherent 3D representation space, enabling sampling within a coherent world rather than per-view image or video latents. Central to our approach is the 3D Unified Representation Autoencoder (3D-URAE); it leverages pretrained 3D foundation models and augments their geometry-centric nature by injecting appearance and distilling semantics into a unified 3D latent space. The appearance-injection branch complements semantic tokens with appearance tokens, allowing unified 3D tokens to be decoded into a renderable 3DGS representation with consistent visual identity. The semantic-distillation branch transfers knowledge from a vision foundation model (VFM) to regularize the 3D tokens onto a compact, semantically meaningful manifold, reducing learning complexity and improving diffusion efficiency in 3D representation space.

Beyond the unified representation, we explicitly preserve cross-view structure during diffusion training by introducing a token-level Cross-View-Correspondence (CVC) consistency loss. Specifically, we enforce the target-view tokens to retain the correspondence patterns induced by the conditioning-view tokens, so that denoising preserves structural consistency rather than merely minimizing the mean error between 3D representations. Moreover, we identify sampling drift induced by train–inference exposure bias as a key factor that degrades 3D generation quality: at inference the model must denoise its own intermediate predictions, so small errors accumulate over steps, and this effect is amplified in 3D by coupled cross-view constraints. To mitigate this issue, we propose Manifold-Drift Forcing (MDF), which trains the 3DGS decoder with a mixture of diffusion-sampled (drifted, off-manifold) and original (on-manifold) unified representations during decoding, shaping a robust 3D representation manifold for stable diffusion sampling, and improving appearance consistency and cross-view coherence. Extensive experiments on RealEstate10K [66], DL3DV [30], and WorldScore [8] demonstrate that OneWorld generates high-quality 3D scenes with strong cross-view consistency, significantly outperforming prior 2D-based approaches.

Our main contributions are summarized as follows:

- We propose **OneWorld**, a diffusion-based 3D scene generation framework that operates in a 3D feature space built upon pretrained 3D foundation models and unified by our 3D Unified Representation Autoencoder (3D-

URAE), which injects appearance and distills semantics to jointly encode geometry, appearance, and semantics.
- We introduce a cross-view correspondence-preserving loss for conditional diffusion training in unified 3D space, explicitly enforcing structural correspondence between the target view and conditioning views to improve cross-view consistency.
- We identify sampling drift from train–inference mismatch in 3D generation and propose Manifold-Drift Forcing, which trains the 3DGS decoder on mixed diffusion-sampled and ground-truth latents to shape a robust 3D manifold for stable sampling, improving appearance consistency and cross-view coherence.

2 Related Work

2.1 3D Scene Generation

Diffusion-based Iterative 3D Scene Generation. Early 3D generation methods typically leverage pretrained 2D generative models [36, 37, 40, 62] to provide strong generative priors. One line of work employs Score Distillation Sampling (SDS) [28, 38, 46, 52] to optimize a 3D representation, such as 3DGS [21] or NeRF [34], by aligning rendered views with the distribution of a pretrained 2D diffusion model. Although these iterative generation approaches achieve notable progress, they often suffer from cross-view semantic inconsistency due to the lack of explicit multi-view constraints. Moreover, generating each individual 3D scene requires extensive iterative optimization, leading to substantial computational cost and limited scalability.

Multi-View Reconstruction-Based 3D Scene Generation. Another line of methods first synthesizes multi-view images or videos using pretrained 2D diffusion models, followed by 3D reconstruction through multi-view synthesis [4, 11, 16, 31, 32, 41, 44, 45, 55, 64] or incremental outpainting [5, 10, 42, 59, 60]. Compared to iterative optimization-based approaches, these methods significantly improve generation efficiency by avoiding per-scene optimization. However, despite this speed advantage, they still lack explicit 3D reasoning, as the underlying generative process is grounded in 2D RGB priors. As a result, they often produce inconsistent geometry and weak multi-view fidelity.

3D Scene Generation in 2D Latent Spaces. The difference between the 3D generation method introduced here and Multi-View Reconstruction-Based methods is that one uses a Diffusion model to generate multi-view images or videos, while this method generates multi-view latents [13, 26, 56] or video latents [6, 12, 14, 27]. The original image VAE or video VAE encoder is frozen, and a decoder for decoding 3D representations is trained, usually implemented as 3DGS. Early approaches, such as Prometheus [56] and SplatFlow [13], adapt SD-VAE [37] to reconstruct 3DGS representations from multi-view latents, and subsequently fine-tune the image generation model. Building upon this latent-to-3D reconstruction paradigm, later works including FlashWorld [27], VIST3A [12],

and FantasyWorld [6] replace SD-VAE [37] with Wan-VAE [48], enabling reconstruction from video latents and extending the framework to video generation models. Learning in the multi-view latent space offers improved computational efficiency and better compatibility with pretrained diffusion backbones compared to directly modeling multi-view image generation; however, since 2D image or video latent spaces do not explicitly enforce cross-view identity consistency, this formulation often results in geometric inconsistency.

2.2 Representation Autoencoder

A common line of work in image and video generation uses VAEs [9, 22, 47] to compress data into low-dimensional latent spaces, facilitating the training of latent diffusion models [40]. However, such compression inevitably results in information loss. Recently, some studies [1,3,43,65] have explored a different route: employing a frozen, pretrained representation encoder and training only the decoder to reconstruct images from high-dimensional semantic features. Known as Representation Autoencoders (RAEs), this approach demonstrates that training diffusion transformers [36] in such latent spaces achieves faster convergence and better performance than VAEs. Motivated by this shift, we design a 3D Unified RAE on top of a pretrained 3D foundation model so diffusion happens directly in a geometry-aware representation space.

3 Method

We present OneWorld, a diffusion framework that performs scene generation directly in a unified, geometry-aware 3D representation space (See Fig. 2). First, we build a 3D Unified Representation Autoencoder (3D-URAE) on top of a feed-forward 3D foundation model by injecting appearance cues and distilling semantic structure into geometry tokens, producing renderable and semantically organized 3D latents (Sec. 3.1). Next, we train a conditional diffusion model in this unified space and introduce a cross-view correspondence preservation regularizer to maintain consistent token-level correspondences between the target and conditioning views, improving cross-view structural coherence (Sec. 3.2). Finally, to address train–inference exposure bias that causes diffusion samples to drift off the 3D-URAE manifold, we propose manifold-drift forcing, which trains the downstream 3D decoder on interpolations of ground-truth and sampled latents to improve robustness and stabilize multi-view sampling (Sec. 3.3).

3.1 3D Unified Representation Autoencoder

Preliminary. We adopt a 3D foundation model for feed-forward scene reconstruction. In this work, the instantiated model is π^3 [51], denoted as $\mathcal{F}$, which infers multiple key 3D quantities of a scene from observed views together with

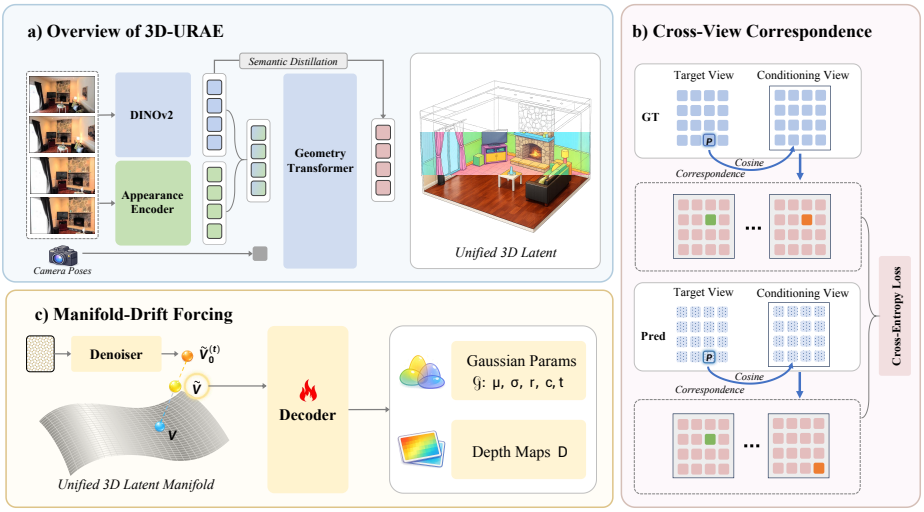


Fig. 2: Overview of the proposed OneWorld framework. (a) We construct a **unified 3D representation space** by introducing **appearance injection** and **semantic distillation**. **(b)** During DiT [36] training, we incorporate **cross-view correspondence**, preserving cross-view geometric token correspondences from the target view to the conditioned view. **(c) Manifold-drift forcing:** we augment the original 3D manifold by mixing ground-truth 3D features with sampled 3D features, enabling a more robust 3D decoder.

their camera parameters. Given N input images $\mathcal{I} \in \mathbb{R}^{N\times H\times W\times 3}$, π^3 first leverages a vision foundation model (VFM) (*e.g.*, DINOv2 [35]) as an image patchifier or tokenizer, denoted as $\mathcal{E}_{\text{patch}}$, to obtain per-view visual tokens:

$$\mathcal{E}_{\text{patch}} : \mathcal{I} \to \mathcal{Z} \in \mathbb{R}^{N\times h\times w\times C}, \tag{1}$$

where $h\times w$ is the token grid resolution and C is the token channel dimension. To enable camera-controllable generation and to directly use the dataset-provided ground-truth camera coordinate system without additional conversions, π^3 takes the per-view camera parameters $\mathcal{T} \in \mathbb{R}^{N\times C_T}$ (*e.g.*, intrinsics and extrinsics in the dataset convention) as explicit inputs. The geometry encoder $\mathcal{E}_{\mathcal{V}}$ then jointly encodes the patchified tokens and camera parameters into geometry tokens:

$$\mathcal{E}_{\mathcal{V}} : (\mathcal{Z}, \mathcal{T}) \to \mathcal{V} \in \mathbb{R}^{N\times h_v\times w_v\times C_v}, \tag{2}$$

where $h_v \times w_v$ is the spatial resolution of geometry tokens and C_v is the corresponding feature dimension. Finally, the geometry tokens are decoded by prediction heads $\mathcal{D}_{\mathcal{V}}$ into a 3D Gaussian Splatting representation and depth maps:

$$\mathcal{D}_{\mathcal{V}} : \mathcal{V} \to (\mathcal{G}, \mathcal{D}), \qquad \mathcal{G} \in \mathbb{R}^{N\times H\times W\times C_{\text{GS}}},\ \mathcal{D} \in \mathbb{R}^{N\times H\times W\times 1}, \tag{3}$$

where C_{GS} denotes the dimensionality of the 3DGS parameterization used in our implementation.

Table 1: Ablation analysis of **appearance injection** and **semantic distillation**. **(a)** We analyze the effect of **appearance injection** from a reconstruction perspective. **(b)** We examine the benefit of **semantic distillation** for training the generative model.

(a) Ablation on Appearance Injection Branch.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o App. Inject	21.14	0.669	0.293
3D-URAE	28.19	0.932	0.102

(b) Ablation on Semantic Distillation Branch.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Sem. Distill	17.45	0.644	0.352
OneWorld	21.57	0.735	0.231

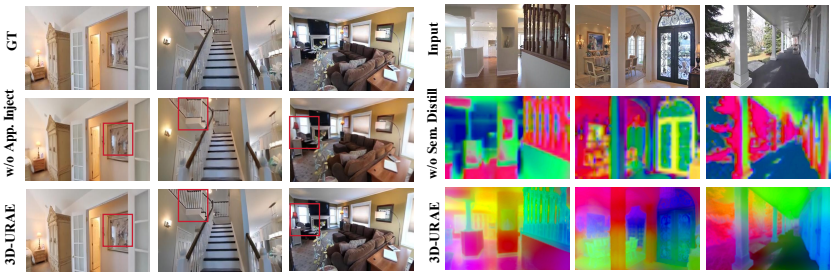
Appearance Injection Branch. The 3D foundation model $\mathcal{F}$ typically adopts a vision foundation model $\mathcal{E}_{\text{patch}}$ (*e.g.*, DINOv2) to patchify $\mathcal{I}$ into semantic tokens $\mathcal{Z}$, which may discard fine-grained appearance details due to heavy semantic abstraction. More detailed results are shown in Tab. 1a and Fig. 3 (a). To compensate for this loss, we introduce an appearance injection branch that extracts appearance-preserving tokens from the original images and augments the conditioning of the geometry encoder. Specifically, we implement a lightweight convolutional encoder $\mathcal{E}_{\text{app}}$ to map $\mathcal{I}$ into appearance tokens $\mathcal{Z}_{\text{app}} \in \mathbb{R}^{N\times h\times w\times C}$, aligned with $\mathcal{Z}$ in both resolution and channel dimension. We then concatenate $\mathcal{Z}$ and $\mathcal{Z}_{\text{app}}$ along the channel axis and feed the augmented tokens with camera parameters $\mathcal{T}$ into $\mathcal{E}_{\mathcal{V}}$, yielding:

$$\mathcal{E}_{\mathcal{V}}:\ \big([\mathcal{Z}\,\|\,\mathcal{Z}_{\text{app}}],\,\mathcal{T}\big)\ \rightarrow\ \mathcal{V}\in\mathbb{R}^{N\times h_v\times w_v\times C_v}. \tag{4}$$

Semantic Distillation Branch. Although π^3 (and similar feed-forward reconstructors such as VGGT [49]) uses VFM semantic tokens (*e.g.*, DINOv2) as inputs, the pure 3D reconstruction supervision often yields geometry tokens $\mathcal{V}$ that are geometry-dominant and weak in semantic structure (See Fig. 3 (b)). To obtain a more compact and semantically organized 3D latent manifold that is easier for diffusion to model (More detailed results are shown in Tab. 1b), we introduce a semantic distillation branch to inject semantics into $\mathcal{V}$. Specifically, we distill from the original VFM tokens $\mathcal{Z} = \mathcal{E}_{\text{patch}}(\mathcal{I})$ and align them with geometry tokens $\mathcal{V} \in \mathbb{R}^{N\times h_v\times w_v\times C_v}$. A lightweight adapter $\mathcal{A}_{\text{sem}}$ maps $\mathcal{Z}$ into the geometry-token space, *i.e.*, $\mathcal{Z}_{\text{sem}} = \mathcal{A}_{\text{sem}}(\mathcal{Z}) \in \mathbb{R}^{N\times h_v\times w_v\times C_v}$. Following VA-VAE [57], we employ a marginal cosine similarity loss and a marginal distance matrix similarity loss; both are computed per view and then averaged over the N views,

$$\mathcal{L}_{\text{mcos}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h_v w_v}\sum_{i=1}^{h_v}\sum_{j=1}^{w_v}\text{ReLU}\left(1-m_1-\frac{\mathbf{z}_{n,i,j}\cdot\mathbf{v}_{n,i,j}}{\|\mathbf{z}_{n,i,j}\|\,\|\mathbf{v}_{n,i,j}\|}\right), \tag{5}$$

where $\mathbf{z}_{n,i,j}$ and $\mathbf{v}_{n,i,j}$ denote the feature vectors from $\mathcal{Z}_{\text{sem}}$ and $\mathcal{V}$ at location (i,j) for the n-th view, respectively. Moreover, letting $N_p = h_v w_v$ and flattening the spatial grid into indices $p, q \in \{1,\dots,N_p\}$, we define the marginal distance



(a) Effect of Appearance Injection. (b) Effect of Semantic Distillation.

Fig. 3: Visualization of the effects of appearance injection and semantic distillation. We visualize the reconstruction results with and without **appearance injection** and present feature visualizations with and without **semantic distillation**.

matrix similarity loss (also per view, then averaged) as

$$\mathcal{L}_{\text{mdms}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{N_p^2}\sum_{p,q}\text{ReLU}\left(\left|\frac{\mathbf{z}_{n,p}\cdot\mathbf{z}_{n,q}}{\|\mathbf{z}_{n,p}\|\,\|\mathbf{z}_{n,q}\|}-\frac{\mathbf{v}_{n,p}\cdot\mathbf{v}_{n,q}}{\|\mathbf{v}_{n,p}\|\,\|\mathbf{v}_{n,q}\|}\right|-m_2\right). \quad (6)$$

Finally, we combine the two terms as $\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{mcos}} + \lambda_{\text{mdms}}\mathcal{L}_{\text{mdms}}$.

Training Objective for 3D-URAE. With the appearance injection branch and the semantic distillation branch, the 3D foundation model is transformed into 3D-URAE, and the resulting tokens $\mathcal{V}$ form a unified 3D representation that encodes geometry together with appearance and semantics. We train 3D-URAE with a differentiable 3DGS rendering loss and the semantic distillation loss; for each scene we predict $(\mathcal{G}, \mathcal{D})$ and render N_{novel} novel views using their camera parameters, where $\mathcal{D}$ is projected to modulate the opacity of $\mathcal{G}$,

$$\mathcal{L}_{\text{render}} = \frac{1}{N_{\text{novel}}}\sum_{n=1}^{N_{\text{novel}}}\Big(\|\hat{\mathcal{I}}_n - \mathcal{I}_n^{\text{gt}}\|_2^2 + \lambda_{\text{lpips}}\,\text{LPIPS}(\hat{\mathcal{I}}_n, \mathcal{I}_n^{\text{gt}})\Big). \quad (7)$$

We optimize the overall objective $\mathcal{L}_{\text{URAE}} = \mathcal{L}_{\text{render}} + \lambda_{\text{sem}}\,\mathcal{L}_{\text{sem}}$.

3.2 Cross-view Correspondence

We train a conditional diffusion model on the unified 3D tokens produced by 3D-URAE. For each scene, we sample a target view token grid $\mathcal{V}^{(\text{tgt})} \in \mathbb{R}^{h_v \times w_v \times C_v}$ and flatten it as $\mathbf{x}_0 \in \mathbb{R}^{N_p \times C_v}$ with $N_p = h_v w_v$. During training, we perturb $\mathbf{x}_0$ with a standard forward process. Since the unified 3D token space is high-dimensional, we adopt an $\mathbf{x}_0$-prediction parameterization (as observed effective in JiT [25]), while computing the training objective in the equivalent v-space via a deterministic conversion for numerical stability and compatibility with standard diffusion formulations. The denoiser is conditioned on a single clean conditioning view, its camera parameters, the target camera parameters, and an optional text prompt embedding. Importantly, the conditioning-view tokens

are encoded by 3D-URAE independently from the multi-view encoding used to obtain the target token grid $\mathbf{x}_0$, ensuring consistent inference when only the conditioning view is available,

$$\hat{\mathbf{x}}_0 = \mathcal{D}_\theta\Big(\mathbf{x}_t,\, t,\, \mathbf{c}_0^{(\text{cond})},\, \mathcal{T}^{(\text{cond})},\, \mathcal{T}^{(\text{tgt})},\, \mathbf{e}_{\text{text}}\Big), \hat{\mathbf{v}}_\theta = \frac{\alpha_t \mathbf{x}_t - \hat{\mathbf{x}}_0}{\sigma_t}, \hat{\mathbf{x}}_0 = \alpha_t \mathbf{x}_t - \sigma_t \hat{\mathbf{v}}_\theta, \tag{8}$$

where $\mathbf{c}_0^{(\text{cond})}$ is the flattened clean token grid of the conditioning view and $\mathcal{T}$ denotes the corresponding camera parameters. We optimize the standard v-prediction objective:

$$\mathcal{L}_v = \mathbb{E}\big[\|\hat{\mathbf{v}}_\theta - \mathbf{v}\|_2^2\big]\,. \tag{9}$$

Cross-view Correspondence Preservation Loss. While $\mathcal{L}_v$ enforces proximity in token space, it does not explicitly preserve cross-view structural correspondence. We therefore introduce a correspondence-preserving regularizer that aligns the nearest-neighbor matching pattern between the target view and the conditioning view. Concretely, let $\mathbf{c}_0^{(\text{cond})} \in \mathbb{R}^{N_p \times C_v}$ denote the flattened clean tokens of the conditioning view. For each target view token p, we compute cosine similarities to all conditioning locations, select the most similar index $q_p^\star$, and keep it only if the confidence exceeds a threshold $\tau = 0.9$. We then apply a cross-entropy loss on the correspondence distribution induced by the predicted clean tokens $\hat{\mathbf{x}}_0$,

$$q_p^\star = \arg\max_q \ \cos(\mathbf{x}_{0,p}, \mathbf{c}_{0,q}^{(\text{cond})}), \qquad \mathbb{1}_p = \mathbb{1}\left(\max_q \cos(\mathbf{x}_{0,p}, \mathbf{c}_{0,q}^{(\text{cond})}) \geq \tau\right), \tag{10}$$

$$\mathcal{L}_{\text{cvc}} = \frac{1}{\sum_p \mathbb{1}_p} \sum_{p=1}^{N_p} \mathbb{1}_p \cdot \left(-\log \frac{\exp\big(\cos(\hat{\mathbf{x}}_{0,p}, \mathbf{c}_{0,q_p^\star}^{(\text{cond})})/T\big)}{\sum_{q=1}^{N_p} \exp\big(\cos(\hat{\mathbf{x}}_{0,p}, \mathbf{c}_{0,q}^{(\text{cond})})/T\big)} \right), \tag{11}$$

where T is a temperature hyperparameter. Finally, we combine the two terms:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_v + \lambda_{\text{cvc}} \mathcal{L}_{\text{cvc}}. \tag{12}$$

3.3 Manifold-Drift Forcing

Although the correspondence-preserving diffusion training improves cross-view structural consistency during training, inference still suffers from a train–inference exposure bias: the denoiser is trained to predict from ground-truth noised latents following the forward process, while at inference it conditions on its own intermediate samples. This sampling drift gradually pushes the sampled latents away from the 3D-URAE manifold, and the discrepancy can be amplified in multi-view generation because cross-view constraints couple all views through shared 3D structure (a detailed analysis is provided in the appendix).

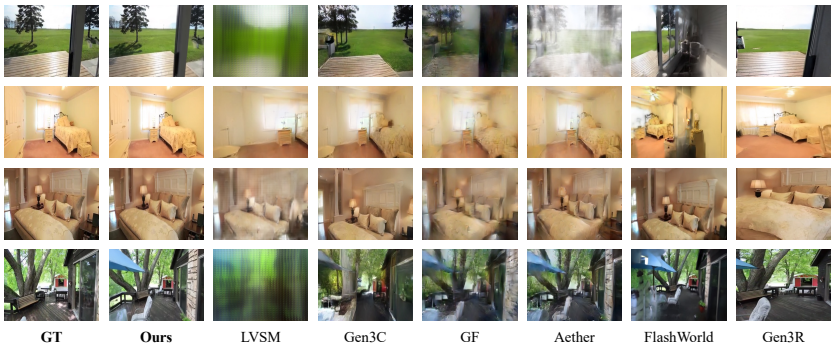


Fig. 4: Qualitative visual results on one-view-based novel view generation.

To mitigate this issue, we propose a manifold-drift forcing strategy that explicitly trains the downstream 3D decoder to be robust to off-manifold latents produced during sampling. Concretely, for each scene we run the diffusion model to obtain intermediate predictions within a step interval $t \sim \mathcal{U}([T_1, T_2])$ and take the corresponding predicted clean latent $\hat{\mathcal{V}}_0^{(t)}$ (*i.e.*, reshaped from $\hat{\mathbf{x}}_0$). Meanwhile, we have the ground-truth unified tokens $\mathcal{V}$ from 3D-URAE. We then construct a drifted training latent by mixing the predicted latent with the ground truth using a ratio $\alpha \in [0, 1]$:

$$\tilde{\mathcal{V}} = \alpha\, \hat{\mathcal{V}}_0^{(t)} + (1 - \alpha)\, \mathcal{V}, \qquad t \sim \mathcal{U}([T_1, T_2]),\ \alpha \sim \mathcal{U}([0, 1]). \tag{13}$$

We feed $\tilde{\mathcal{V}}$ into the 3D decoder (*i.e.*, the prediction heads) to obtain $(\tilde{\mathcal{G}}, \tilde{\mathcal{D}})$ and optimize the same differentiable 3DGS rendering objective as in Sec. 3.1. Different from the 3D-URAE training, since most views involved in diffusion sampling are already "noisy" (including both sampled views and conditioning views), we do not sample additional target views; instead, we render and supervise all available views (the noise views and the conditioning views) under their camera parameters. The objective is

$$\tilde{\mathcal{L}}_{\text{render}} = \frac{1}{N} \sum_{n=1}^{N} \Big(\|\tilde{\mathcal{I}}_n - \mathcal{I}_n^{\text{gt}}\|_2^2 + \lambda_{\text{lpips}}\, \text{LPIPS}(\tilde{\mathcal{I}}_n, \mathcal{I}_n^{\text{gt}}) \Big), \tag{14}$$

where $\tilde{\mathcal{I}}_n$ is rendered from $(\tilde{\mathcal{G}}, \tilde{\mathcal{D}})$ using the n-th view camera parameters in the dataset convention. By training the decoder on the interpolated latent $\tilde{\mathcal{V}}$ across a range of sampling steps, the decoder learns to tolerate diffusion-induced manifold drift and yields more stable 3D reconstruction quality at inference.

4 Experiment

4.1 Training Details

Datasets. We train on two large-scale calibrated multi-view datasets: RealEstate10K (Re10K) [66] and DL3DV-10K [30]. Together they contain ∼70K multi-view-

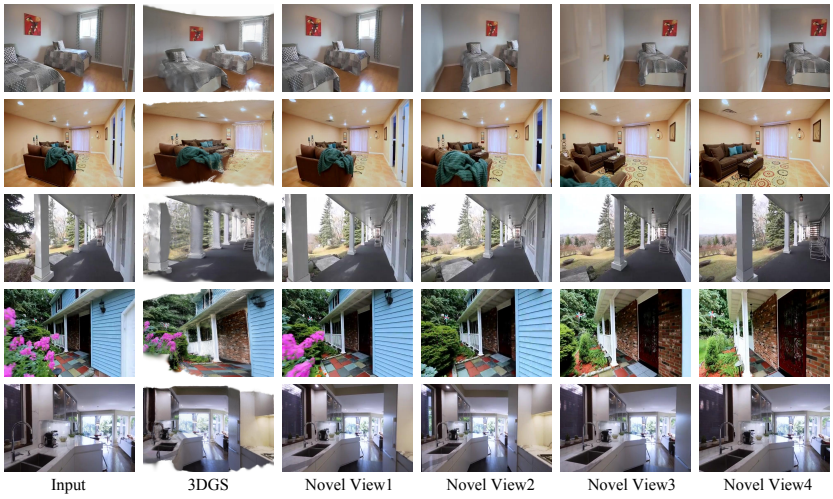


Fig. 5: Visualization of 3DGS and rendered novel views.

consistent scenes covering diverse real-world environments. We use the official train/test split for Re10K. For DL3DV-10K, we randomly split scenes into 90% for training and 10% for testing. Scene-level text prompts are generated with a multimodal large language model [15].

3D-URAE. We fine-tune the 3D foundation reconstructor π^3 to obtain our 3D Unified Representation Autoencoder (3D-URAE). Training uses a 1:1 mixture of Re10K and DL3DV-10K. For each scene, we sample N=8 views and supervise N_{novel}=4 novel views; all images are resized to 224×448. We optimize a weighted sum of a differentiable 3DGS rendering loss (with λ_{lpips}=0.05) and a semantic distillation objective (with λ_{sem}=0.1, λ_{mdms}=1.0 and margins m_1=m_2=0.05). We initialize from the π^3 checkpoint and fine-tune with AdamW [33], using a linearly decayed learning rate from 2×10^{-4} to 2×10^{-5}. The per-GPU batch size is 2 (global batch size 64). All experiments are run on 32 NVIDIA A100-SXM4-80G GPUs for 30K steps. Additional implementation details and hyperparameter analyses are provided in the Appendix due to space constraints.

Diffusion. We train a conditional DiT [36] denoiser in the unified 3D representation space from 3D-URAE (Sec. 3.2). The model predicts $\hat{\mathbf{x}}_0$ and is optimized with the v-prediction loss, $\mathcal{L}_v = \mathbb{E}\big[\|\hat{\mathbf{v}}_\theta - \mathbf{v}\|_2^2\big]$. We additionally enforce cross-view correspondence preservation on $\hat{\mathbf{x}}_0$, $\mathcal{L}_{\text{diff}} = \mathcal{L}_v + \lambda_{\text{cvc}}\mathcal{L}_{\text{cvc}}$, with τ=0.9 and λ_{cvc}=0.2. We initialize DiT from Wan-2.1-T2V-1.3B and condition on 3D-URAE latents from a single view (with camera parameters). To focus on non-text-conditioned generation, we apply classifier-free text dropping with rate 0.5. Training uses the Re10K and DL3DV-10K mixture with batch size 256 (8 per GPU), linear LR decay $1 \times 10^{-4} \to 1 \times 10^{-5}$, EMA decay 0.9995, and runs for 100K steps. Due to space constraints, additional implementation details and hyperparameter analyses are deferred to the Appendix.

Manifold-drift forcing. In the final stage (Sec. 3.3), we train the 3D decoder to be robust to off-manifold latents from diffusion sampling. We sample intermediate timesteps $t \sim \mathcal{U}([T_1, T_2])$ with T_1=10 and T_2=20, obtain the predicted clean latent $\hat{\mathcal{V}}_0^{(t)}$, and form a drifted latent $\tilde{\mathcal{V}} = \alpha\, \hat{\mathcal{V}}_0^{(t)} + (1-\alpha)\, \mathcal{V}$, where $\alpha \sim \mathcal{U}([0,1])$. We feed $\tilde{\mathcal{V}}$ into the 3D prediction heads to obtain $(\tilde{\mathcal{G}}, \tilde{\mathcal{D}})$ and optimize the same differentiable 3DGS rendering loss as in Sec. 3.1, supervising all views used for conditioning and sampling. We freeze the 3D-URAE encoder $\mathcal{E}_\mathcal{V}$ and update only the decoder heads $\mathcal{D}_\mathcal{V}$, generating drifted latents with the DiT from the previous stage. We train with a learning rate of 2×10^{-5}, batch size 256 (8 per GPU), for 10K steps.

4.2 Evaluation Protocols

We evaluate under two protocols: ground-truth-based novel view synthesis (NVS) and the reference-free WorldScore benchmark [8]. In NVS, generation is conditioned only on a single input image together with its camera parameters. We randomly sample 500 scenes from Re10K and 500 scenes from DL3DV-10K, synthesize target views at calibrated camera poses, and compare them with the corresponding ground-truth images. We report PSNR, SSIM [53], and LPIPS [63] between the generated and ground-truth images, and additionally report the VBench Score [18, 19] to assess generative capability under image conditioning, focusing on I2V Subject (I2V Subj.), I2V Background (I2V BG), and Imaging Quality (I.Q.). For fair comparison, each method generates at its own native resolution (*e.g.*, 704×480 for FlashWorld [27], 560×560 for Gen3R [17]), and all outputs are resized to 256×256 before computing resolution-sensitive metrics (*e.g.*, PSNR and SSIM).

For WorldScore, each test case provides a single reference image together with a text prompt and camera trajectory, and quality is measured by the standard WorldScore protocol and scoring metrics without paired ground-truth novel views. We follow the benchmark's photorealistic indoor split (500 scenes) for evaluation. However, the benchmark's outdoor distribution differs substantially from our training data distribution and our training set is relatively small; therefore, for outdoor evaluation we construct a WorldScore-style set of 500 scenes from DL3DV-10K by treating a single-view image and its text prompt from each outdoor scene as the reference, while matching the camera trajectories from WorldScore to drive generation. We then evaluate using the standard WorldScore protocol and metrics.

We compare against recent baselines spanning view synthesis and geometry-aware world generation: LVSM [20], a large transformer-based NVS model with minimal explicit 3D inductive bias; Gen3C [39], which improves camera controllability and temporal consistency via an explicit 3D cache; GF (Geometry Forcing) [54], which encourages diffusion models to internalize 3D structure for better geometric consistency; Aether [67], a geometry-aware unified world modeling framework for joint reconstruction and generation; FlashWorld [27], which directly generates 3D Gaussian scene representations for fast rendering-based

Table 2: Experimental results on 1-view-based novel view generation. We report the performance of different methods on the RealEstate10K and DL3DV-10K datasets. Best in **bold**; second best underlined.

Method	RealEstate10K						DL3DV-10K					
	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑
LVSM [20]	18.54	0.694	0.336	0.993	0.991	0.487	15.18	0.516	0.482	0.962	0.966	0.421
Gen3C [39]	19.88	0.697	0.271	0.991	0.990	0.514	15.79	0.531	0.497	0.941	0.953	0.414
GF [54]	15.97	0.518	0.424	0.986	0.976	0.553	11.62	0.322	0.621	0.932	0.929	0.533
Aether [67]	16.13	0.611	0.413	0.991	0.989	0.536	13.37	0.492	0.566	0.957	0.963	0.451
FlashWorld [27]	20.18	0.724	0.256	0.993	0.994	0.584	16.02	0.566	0.451	0.964	0.969	0.538
Gen3R [17]	20.09	0.714	0.269	**0.994**	0.994	0.593	15.94	0.557	0.468	0.964	0.970	0.543
OneWorld (Ours)	**21.57**	**0.735**	**0.231**	0.993	**0.995**	**0.604**	**17.19**	**0.589**	**0.418**	**0.966**	**0.973**	**0.556**

synthesis; and Gen3R [17], which combines reconstruction and diffusion priors to generate both appearance and geometry.

4.3 3D Scene Generation

1-view NVS on calibrated benchmarks.

We first evaluate 1-view, camera-conditioned novel view generation under the GT-based NVS protocol on RealEstate10K and DL3DV-10K (Tab. 2 for quantitative results; Fig. 4 and Fig. 5 for qualitative visualizations). OneWorld achieves the best overall results on both benchmarks. On RealEstate10K, it reaches 21.57 PSNR and 0.735 SSIM with the lowest LPIPS of 0.231, and it also attains the highest image quality score (I.Q. 0.604). Identity consistency is near-saturated, with I2V Subj. 0.993 and I2V BG 0.995. On DL3DV-10K, OneWorld again ranks first, with 17.19 PSNR, 0.589 SSIM, and the lowest LPIPS of 0.418, while improving I2V Subj./BG to 0.966/0.973 and achieving the best I.Q. of 0.556. These results show that OneWorld improves fidelity and perceptual quality while maintaining strong subject and background consistency across diverse real-world scenes. Additionally, we visualize and compare methods capable of generating 3D representations (*e.g.*, point clouds and 3DGS), as shown in Fig. 6. The results indicate that our method surpasses current SOTA baselines in producing coherent 3D scenes.

Fig. 6: Comparison of 3D scenes generated by different methods: Gen3R [17] uses point clouds, while FlashWorld [27] and OneWorld use 3DGS.

WorldScore-style reference-free evaluation. We further assess single-image world generation with the reference-free WorldScore protocol (Tab. 3). On WorldScore-Indoor, OneWorld achieves the best 3D Consistency at 84.98

Table 3: WorldScore-style reference-free evaluation results. We report results on WorldScore-Indoor and on DL3DV as an outdoor benchmark. Best in **bold**; second best underlined.

Method	WorldScore-Indoor						DL3DV					
	3D Consist.	Photo. Consist.	Obj. Cont.	Cont. Align.	Style Consist.	Subj. Quality	3D Consist.	Photo. Consist.	Obj. Cont.	Cont. Align.	Style Consist.	Subj. Quality
LVSM [20]	74.32	63.94	47.21	23.79	66.02	36.38	64.37	55.58	40.21	18.88	57.65	29.59
Gen3C [39]	74.11	78.74	**51.36**	27.12	71.18	40.05	69.41	69.96	44.18	21.92	62.23	33.07
GF [54]	78.05	69.21	46.95	24.62	68.71	38.66	67.12	58.44	41.67	20.35	59.92	31.74
Aether [67]	77.82	66.66	49.28	25.04	69.06	38.17	67.47	57.98	42.02	19.83	60.37	31.11
FlashWorld [27]	83.57	80.19	49.61	**49.27**	75.32	**54.09**	76.74	72.76	43.25	**42.32**	69.08	**46.62**
Gen3R [17]	80.19	78.64	45.80	45.61	**79.10**	53.76	76.69	71.30	**44.75**	40.38	**72.13**	43.84
OneWorld (Ours)	**84.98**	**81.67**	48.92	46.88	76.74	51.73	**78.21**	**74.09**	42.58	40.12	70.62	45.98

Fig. 7: Qualitative visualization of the ablation study under the 1-view-based NVS setting. We compare the full model with variants without Cross-View Correspondence and without Manifold-Drift Forcing.

and the best Photometric Consistency at 81.67, indicating stronger multi-view coherence and more stable appearance along the trajectory. It also ranks second on both Content Alignment (46.88) and Style Consistency (76.74), trailing only FlashWorld and Gen3R respectively. For Object Control and Subjective Quality, OneWorld remains competitive with scores of 48.92 and 51.73, although other methods achieve the top scores on these axes. On the DL3DV outdoor benchmark, OneWorld again ranks first in 3D Consistency, reaching 78.21, together with the best Photometric Consistency of 74.09. It further attains the second-best Style Consistency of 70.62 and the second-best Subjective Quality of 45.98, while staying close to the leading baselines on Object Control (42.58) and Content Alignment (40.12). Overall, the reference-free results show that OneWorld improves long-range 3D and appearance consistency across both indoor and outdoor settings, while maintaining competitive controllability and perceptual quality.

Table 4: Ablation on generation components evaluated on RealEstate10K under the 1-view NVS setting.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑
OneWorld (full)	21.57	0.735	0.231	0.993	0.995	0.604
w/o CVC	19.10	0.682	0.284	0.989	0.990	0.566
w/o MDF	20.59	0.714	0.256	0.992	0.994	0.589

4.4 Ablation Study

As shown in Tab. 4 and Fig. 7, both components contribute substantially to view-consistent generation quality. Removing the correspondence regularizer (w/o CVC) leads to a clear degradation across all appearance metrics, with PSNR dropping from 21.57 to 19.10, SSIM decreasing from 0.735 to 0.682, and LPIPS increasing from 0.231 to 0.284. This verifies that explicitly enforcing cross-view correspondence during diffusion training is crucial for preserving multi-view alignment in the unified 3D representation space, and helps suppress view-dependent artifacts that otherwise accumulate when synthesizing novel viewpoints from a single observation.

We also observe a consistent, though milder, degradation when removing manifold-drift forcing (w/o MDF). Without MDF, the reconstruction scores drop from 21.57 to 20.59 and from 0.735 to 0.714, while the perceptual distance increases from 0.231 to 0.256. This suggests that the 3D decoding heads become more sensitive to the shifted latents encountered along the sampling trajectory. Overall, the full model achieves the best fidelity and perceptual quality. The results imply that CVC mainly improves cross-view structural consistency during generation, whereas MDF further stabilizes decoding under the inference-time shift, leading to more reliable renderings in the 1-view NVS setting.

5 Conclusion

We present OneWorld, a diffusion framework that generates 3D scenes directly in the representation space of a pretrained 3D foundation model, overcoming the structural inconsistency and inefficiency of per-view 2D latent pipelines. By unifying geometry, appearance, and semantics in a coherent 3D representation space and explicitly enforcing cross-view structure while stabilizing sampling, OneWorld enables consistent and high-fidelity 3D scene generation. Our findings in this paper suggest that generative modeling in 3D foundation representation space marks a promising paradigm toward scalable and unified 3D world generation.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Research Foundation Fellowship in Artificial Intelligence (NRF Award NRFFIAI1-2024-0010) and the NTU FA-SUG.

References

1. Bi, T., Zhang, X., Lu, Y., Zheng, N.: Vision foundation models can be good tokenizers for latent diffusion models. arXiv preprint arXiv:2510.18457 (2025)
2. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1050–1060 (2025)
3. Chen, B., Bi, S., Tan, H., Zhang, H., Zhang, T., Li, Z., Xiong, Y., Zhang, J., Zhang, K.: Aligning visual foundation encoders to tokenizers for diffusion models. arXiv preprint arXiv:2509.25162 (2025)
4. Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.J., Cai, J.: Mvsplat360: Feed-forward 360 scene synthesis from sparse views. Adv. Neural Inform. Process. Syst. **37**, 107064–107086 (2024)
5. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
6. Dai, Y., Jiang, F., Wang, C., Xu, M., Qi, Y.: Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. arXiv preprint arXiv:2509.21657 (2025)
7. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys **58**(3), 1–38 (2025)
8. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. arXiv preprint arXiv:2504.00983 (2025)
9. 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: Int. Conf. Mach. Learn. (2024)
10. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. Adv. Neural Inform. Process. Syst. **36**, 39897–39914 (2023)
11. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
12. Go, H., Narnhofer, D., Bhat, G., Truong, P., Tombari, F., Schindler, K.: Vist3a: Text-to-3d by stitching a multi-view reconstruction network to a video generator. arXiv preprint arXiv:2510.13454 (2025)
13. Go, H., Park, B., Jang, J., Kim, J.Y., Kwon, S., Kim, C.: Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21524–21536 (2025)
14. Go, H., Park, B., Nam, H., Kim, B.H., Chung, H., Kim, C.: Videorfsplat: Direct scene-level text-to-3d gaussian splatting generation with flexible pose and multi-view joint modeling. arXiv preprint arXiv:2503.15855 (2025)
15. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
16. Hao, J., Wang, P., Wang, H., Zhang, X., Guo, Z.: Gaussvideodreamer: 3d scene generation with video diffusion and inconsistency-aware gaussian splatting. arXiv preprint arXiv:2504.10001 (2025)
17. Huang, J., Yang, Y., Yang, B., Ma, L., Ma, Y., Liao, Y.: Gen3r: 3d scene generation meets feed-forward reconstruction. arXiv preprint arXiv:2601.04090 (2026)
18. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21807–21818 (2024)

19. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
20. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: Int. Conf. Learn. Represent.
21. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., et al.: 3d and 4d world modeling: A survey. arXiv preprint arXiv:2509.07996 (2025)
24. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: Eur. Conf. Comput. Vis. pp. 71–91. Springer (2024)
25. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
26. Li, X., Lai, Z., Xu, L., Qu, Y., Cao, L., Zhang, S., Dai, B., Ji, R.: Director3d: Real-world camera trajectory and 3d scene generation from text. Adv. Neural Inform. Process. Syst. **37**, 75125–75151 (2024)
27. Li, X., Wang, T., Gu, Z., Zhang, S., Guo, C., Cao, L.: Flashworld: High-quality 3d scene generation within seconds. arXiv preprint arXiv:2510.13678 (2025)
28. 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. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 300–309 (2023)
29. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
30. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22160–22169 (2024)
31. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: Reconx: Reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024)
32. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024), `https://openreview.net/forum?id=a68SUt6zFt`, featured Certification
36. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Int. Conf. Comput. Vis. pp. 4195–4205 (2023)

37. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
39. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6121–6132 (2025)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
41. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single real image. arXiv preprint arXiv:2310.17994 (2023)
42. Schwarz, K., Rozumny, D., Bulò, S.R., Porzi, L., Kontschieder, P.: A recipe for generating 3d worlds from a single image. In: Int. Conf. Comput. Vis. pp. 3520–3530 (2025)
43. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025)
44. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
45. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024)
46. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
47. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Adv. Neural Inform. Process. Syst. **30** (2017)
48. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
49. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
50. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20697–20709 (2024)
51. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
52. 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. Adv. Neural Inform. Process. Syst. **36**, 8406–8441 (2023)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
54. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)

55. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21551–21561 (2024)
56. Yang, Y., Shao, J., Li, X., Shen, Y., Geiger, A., Liao, Y.: Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2857–2869 (2025)
57. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
58. Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y.L., Zhu, C., Xiang, J., et al.: World action models are zero-shot policies. arXiv preprint arXiv:2602.15922 (2026)
59. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5916–5926 (2025)
60. Yu, H.X., Duan, H., Hur, J., Sargent, K., Rubinstein, M., Freeman, W.T., Cole, F., Sun, D., Snavely, N., Wu, J., et al.: Wonderjourney: Going from anywhere to everywhere. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6658–6667 (2024)
61. Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.X., Liang, H., Xu, Z., Su, H., et al.: Advances in feed-forward 3d reconstruction and view synthesis: A survey. arXiv preprint arXiv:2507.14501 (2025)
62. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Int. Conf. Comput. Vis. pp. 3836–3847 (2023)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
64. Zhao, Y., Lin, C.C., Lin, K., Yan, Z., Li, L., Yang, Z., Wang, J., Lee, G.H., Wang, L.: Genxd: Generating any 3d and 4d scenes. arXiv preprint arXiv:2411.02319 (2024)
65. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
66. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: learning view synthesis using multiplane images. ACM Trans. Graph. **37**(4), 1–12 (2018)
67. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: Int. Conf. Comput. Vis. pp. 8535–8546 (2025)