PIXWORLD: UNIFYING 3D SCENE GENERATION AND RECONSTRUCTION IN PIXEL SPACE

Anonymous authors
Paper under double-blind review



Figure 1: **PixWorld unifies 3D reconstruction and generation in pixel space. Left:** Multi-view reconstruction, generation conditioned on an arbitrary subset of clean views, and text-to-3D. PixWorld reconstructs the clean input views and generates the remaining target views. **Top right:** PixWorld-Streaming supports incremental 3D reconstruction and generation from streaming inputs. **Bottom right:** WorldScore and generation latency on one H100 using each method's native configuration. PixWorld achieves a WorldScore of 73.11 and a $497\times$ speedup over FantasyWorld.

ABSTRACT

3D reconstruction and generation typically follow separate paradigms: pixel-based regression for reconstruction and latent diffusion for generation. Recent methods seek to unify them in a shared latent space. However, autoencoder compression introduces information loss in both tasks, while latent-space diffusion objectives supervise encoded latents rather than the resulting 3D scenes. To address these limitations, we introduce PixWorld, a single model that unifies 3D reconstruction and generation under a pixel-space diffusion paradigm. By supervising diffusion on rendered images rather than encoded latents, PixWorld eliminates the autoencoder bottleneck and directly optimizes the 3D scene. Furthermore, we introduce PixWorld-Streaming through autoregressive diffusion distillation, extending this unified framework beyond a fixed view budget to support online reconstruction and generation over minute-long sequences. Experiments show that PixWorld outperforms latent-space generation methods and matches state-of-the-art reconstruction, while PixWorld-Streaming retains comparable quality under causal streaming inputs, demonstrating the effectiveness of a unified pixel-space approach.

1 INTRODUCTION

Building 3D scenes from visual observations is a long-standing goal in computer vision, with broad impact across gaming, robotics, embodied AI, and VR/AR (Ding et al., 2025; Kong et al., 2025; Ye et al., 2026b; Zhang et al., 2025). Two complementary directions have driven progress in this space:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 2: **Latent-space vs. pixel-space diffusion. Top:** Latent-space diffusion supervises encoded latents, which are decoded into a 3D scene. **Bottom:** PixWorld supervises rendered images in pixel space, directly optimizing the predicted scene through differentiable rendering.

reconstruction recovers 3D scenes from real-world captures, while generation synthesizes plausible scenes from limited or even imagined conditions (Zhang et al., 2025; Wang et al., 2026; Li et al., 2026). Together, they form the foundation for populating the digital worlds of tomorrow.

3D scene generation and reconstruction have long developed as two separate lines of research. Reconstruction is dominated by feed-forward methods that regress 3D representations directly from multi-view images (Hong et al., 2023; Charatan et al., 2024; Szymanowicz et al., 2024; Xu et al., 2025; Chen et al., 2024b). Generation has evolved from per-scene optimization with 2D priors via score distillation (Lin et al., 2023; Poole et al., 2022; Tang et al., 2023; Wang et al., 2023) to latent-space diffusion as the current mainstream (Yang et al., 2025b; Go et al., 2025b; Huang et al., 2026b; Li et al., 2025a), with recent extensions diffusing in the feature space of pretrained 3D foundation models or representation autoencoders (Gao et al., 2026; Sun et al., 2026; Jang et al., 2026). A recent work, Gen3R (Huang et al., 2026b), attempts to unify the two tasks by extending the latent-space generation pipeline to also handle reconstruction. However, this design introduces clear limitations: the diffusion objective is defined on intermediate latent features rather than the underlying 3D representation, preventing the 3D output from being directly optimized; in addition, both branches suffer from information loss introduced by the pretrained Variational Autoencoder (VAE) or Representation Autoencoder (RAE), which itself requires additional training.

To address these limitations, we reformulate unification under a pixel-space diffusion paradigm and introduce PixWorld, a single framework that jointly handles 3D reconstruction and generation (Figure 1). This design brings two key benefits. First, for both generation and reconstruction, eliminating the intermediate latent stage removes the information loss introduced by autoencoder compression and the need to train a separate VAE/RAE for 3D modeling, which is especially critical for reconstruction as a fidelity-bound task. Second, for generation, the diffusion objective directly supervises the 3D representation through differentiable rendering (see Fig. 2), aligning the training signal with the fidelity of the 3D scene rather than with targets in an intermediate latent space. Building on this pixel-space framework, we unify both tasks through a shared forward computation by partitioning multi-view inputs into clean and noisy subsets: clean views drive reconstruction, while noisy ones are generated conditioned on the clean ones, with both producing a pixel-aligned 3D Gaussian representation (Kerbl et al., 2023). When all input views are clean, the model performs reconstruction in a single forward pass; when noisy views are included, it jointly reconstructs the observed content and generates new content through iterative denoising.

To support online reconstruction and generation over minute-long sequences, we introduce PixWorld-Streaming through autoregressive diffusion distillation. PixWorld-Streaming incrementally incorporates incoming clean views into the reconstructed scene and generates unseen regions conditioned on past context. By processing inputs causally, it extends the unified pixel-space framework to growing scenes without requiring access to future views. Extensive experiments on RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), and WorldScore (Duan et al., 2025) demonstrate that PixWorld outperforms prior latent-space generation methods while matching state-of-the-art reconstruction methods. PixWorld-Streaming retains this quality in both streaming reconstruction and generation under causal inputs, demonstrating the benefits of unifying the two tasks in pixel space across offline and streaming settings.

Our main contributions are summarized as follows:

- We propose PixWorld, a unified pixel-space diffusion framework for 3D scene reconstruction and generation. Without a VAE or RAE, PixWorld avoids information loss from autoencoder compression for both tasks. Furthermore, its diffusion objective is defined on rendered images rather than encoded latents, enabling direct optimization of generated 3D scenes.
- We introduce PixWorld-Streaming via autoregressive diffusion distillation, enabling online reconstruction and generation over minute-long sequences conditioned on past context.
- Extensive experiments demonstrate strong reconstruction and generation performance of PixWorld, with PixWorld-Streaming preserving this quality under causal streaming inputs, validating pixel-space unification in both offline and streaming settings.

2 PRELIMINARY

Pixel-Space Diffusion. Latent diffusion encodes an image x into a compressed representation $z = E(x)$, performs denoising in latent space, and decodes the result as $\hat{x} = D(\hat{z})$. Pixel-space diffusion instead defines the noisy variable and denoising target directly in RGB space. Let $x \in \mathbb{R}^{H\times W\times 3}$ be a clean image, c its conditioning information, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ Gaussian noise. We use the linear forward process

$$x_\sigma = (1-\sigma)x + \sigma\epsilon, \qquad \hat{x} = f_\theta(x_\sigma, \sigma, c), \tag{1}$$

where $\sigma \in [0, 1]$ is the noise level: $\sigma = 0$ denotes a clean image and $\sigma = 1$ pure noise. Following JiT (Li & He, 2025), the network predicts the clean image, while its training objective is evaluated in velocity space. For $\sigma > 0$, the target velocity is $v = \epsilon - x$ and the predicted velocity is $\hat{v} = (x_\sigma - \hat{x})/\sigma$, giving

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x,\epsilon,\sigma}\left[\|\hat{v} - v\|_2^2\right] = \mathbb{E}_{x,\epsilon,\sigma}\left[\frac{\|\hat{x} - x\|_2^2}{\sigma^2}\right]. \tag{2}$$

This pixel-space parameterization is particularly attractive for our 3D setting. Latent diffusion defines its objective on encoded latents, leaving the diffusion process decoupled from both the rendered output and the underlying 3D representation it is meant to supervise. Pixel-space diffusion instead keeps the diffusion variable and the rendered output in the same RGB domain, so the diffusion objective can be supervised directly on rendered images, aligning optimization with 3D scene fidelity rather than with targets in an intermediate latent space.

Distribution Matching Distillation. Distribution matching distillation (DMD) (Yin et al., 2024b;a) transfers a multi-step diffusion teacher's distribution to a few-step generator. Given a generated image or multi-view image set $\hat{X}_\theta = H_\theta(z, c)$, we add noise as $X_\tau = (1-\tau)\hat{X}_\theta + \tau\epsilon$, with $0 < \tau < 1$. A frozen real-score model estimates the target score s_{real}, and a trainable fake-score model estimates the generator score s_{fake}. Matching their noisy distributions by reverse KL divergence gives

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_{z,c,\tau,\epsilon}\left[w(\tau)\big(s_{\mathrm{fake}}(X_\tau, \tau, c) - s_{\mathrm{real}}(X_\tau, \tau, c)\big)\frac{\partial \hat{X}_\theta}{\partial \theta}\right], \tag{3}$$

where both scores are evaluated under the same conditioning and $w(\tau)$ includes the forward-process factor $1 - \tau$. The DMD objective uses this score difference to guide the student toward the teacher's conditional distribution, transferring the teacher's generative capability to a model that requires only a few denoising steps.

3 UNIFIED PIXEL-SPACE 3D MODELING

Figure 3 illustrates reconstruction and generation under different view conditions; Figure 4 presents the framework. We first introduce PixWorld for unified 3D reconstruction and generation (Sec. 3.1). We then describe how DMD trains the 3DGS branch to learn the RGB branch's generative capability (Sec. 3.2). Finally, we present PixWorld-Streaming and its autoregressive diffusion distillation procedure (Sec. 3.3).

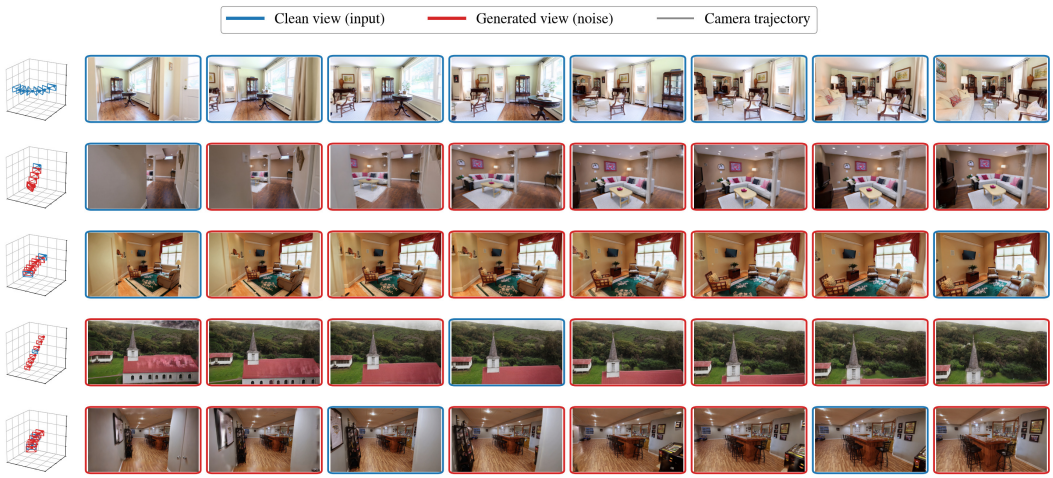


Figure 3: **Reconstruction and generation with different view conditions.** Rows: all-clean reconstruction, first-view generation, interpolation, bidirectional generation, and two-intermediate-view generation. Blue/red frustums mark observed/generated views. Each sequence contains 16 Gaussian-rendered views at 480×832; eight are shown.

3.1 PIXWORLD: UNIFIED RECONSTRUCTION AND GENERATION IN PIXEL SPACE

Clean and noisy input formulation. Given posed images $\mathcal{I} = \{\mathcal{I}^n\}_{n=1}^N$ with cameras $\mathcal{T} = \{\mathcal{T}^n\}_{n=1}^N$, we partition view indices into disjoint clean and noisy subsets, $\Omega_\mathrm{c} \cup \Omega_\mathrm{n} = \{1, \ldots, N\}$. The model jointly processes

$$\tilde{\mathcal{I}}_\sigma^n = \begin{cases} \mathcal{I}^n, & n \in \Omega_\mathrm{c}, \\ (1-\sigma)\mathcal{I}^n + \sigma\epsilon^n, & n \in \Omega_\mathrm{n}. \end{cases} \tag{4}$$

Clean views provide reconstruction evidence, while noisy views are generated conditioned on the available observations and optional text y. Setting $\Omega_\mathrm{n} = \varnothing$ gives pure reconstruction; an empty clean subset supports text-conditioned generation. All noisy views share σ in the non-streaming model, while clean observations remain noise-free.

Latent-to-pixel input alignment. Motivated by the success of latent-to-pixel adaptation in image generation (Chen et al., 2026c; Jiang et al., 2026), we adopt L2P initialization to inherit pretrained generative priors. Specifically, we initialize the shared transformer from Wan (Wan et al., 2025) and replace its latent input interface with multi-view pixel patchification. We compare L2P initialization with training from scratch in Table 8(c). Inspired by the patch embedder alignment in DC-VideoGen (Chen et al., 2025c), we first train the pixel patch embedder to match Wan's input embeddings for corresponding images using a squared-error loss:

$$\mathcal{L}_\mathrm{align} = \mathbb{E}_\mathcal{I}\left[\frac{1}{M}\left\|P_\mathrm{pix}(\mathcal{I}) - \mathrm{sg}[A(P_\mathrm{Wan}(E_\mathrm{Wan}(\mathcal{I})))]\right\|_F^2\right]. \tag{5}$$

Here M is the number of matched tokens, E_Wan and P_Wan are the frozen source encoder and patch embedder, A establishes the corresponding token grid, and sg stops gradients. After input alignment, joint scene training operates directly in pixel space.

Joint RGB and 3D Gaussian prediction. The aligned tokens are processed jointly by the shared DiT, conditioned on camera rays, noise level, and optional text. Pixel-space heads based on the lightweight convolutional design of DiP (Chen et al., 2025e) decode the shared context into RGB and 3DGS outputs, producing $(X_\mathrm{rgb}, \hat{\mathcal{D}}, \hat{\mathcal{G}}) = f_\theta(\tilde{\mathcal{I}}_\sigma, \sigma, \mathcal{T}, y)$. The Gaussian branch predicts per-pixel depth together with rotation, scale, opacity, and appearance. Gaussian centers are determined by the predicted depth: $\mu_p^n = \Pi^{-1}(p, \hat{d}_p^n, \mathcal{T}^n)$, where Π^{-1} unprojects a pixel into world coordinates. Aggregating these pixel-aligned Gaussians yields $\hat{\mathcal{G}}$; differentiable rendering gives $X_\mathrm{render}^n = \mathcal{R}(\hat{\mathcal{G}}, \mathcal{T}^n)$.

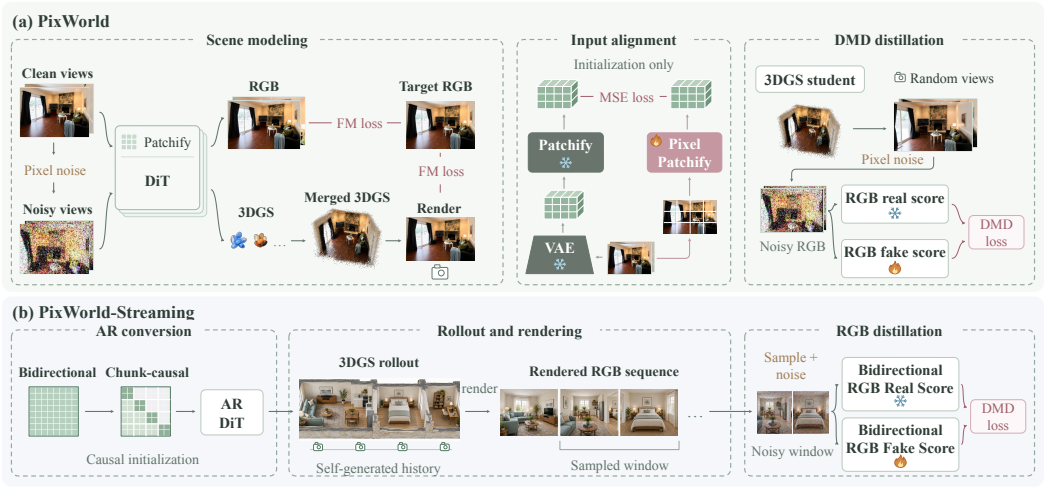


Figure 4: **Overview of PixWorld and PixWorld-Streaming.** (a) **PixWorld.** The model jointly processes clean and noisy views to predict RGB images and 3D Gaussians under pixel-space supervision. L2P input alignment adapts a pretrained latent diffusion model to pixel inputs. DMD trains the 3DGS branch to learn the RGB branch's generative capability through distribution matching on rendered images, enabling few-step 3D generation. (b) **PixWorld-Streaming.** Autoregressive diffusion distillation with a rolling window enables incremental reconstruction and generation conditioned on past context.

Shared RGB and rendering flow matching. Both outputs use a shared velocity-space objective over the clean and noisy input views. With $\sigma_\delta = \max(\sigma, \delta)$ for a small positive numerical floor δ, we define

$$\hat{v}^n_{\text{rgb}} = \frac{\tilde{\mathcal{I}}^n_\sigma - X^n_{\text{rgb}}}{\sigma_\delta}, \qquad \hat{v}^n_{\text{render}} = \frac{\tilde{\mathcal{I}}^n_\sigma - X^n_{\text{render}}}{\sigma_\delta}, \qquad v^n = \frac{\tilde{\mathcal{I}}^n_\sigma - \mathcal{I}^n}{\sigma_\delta}. \tag{6}$$

For $\Omega = \Omega_{\text{c}} \cup \Omega_{\text{n}}$, the joint objective is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\Big[\big\langle \|\hat{v}^n_{\text{rgb}} - v^n\|_2^2 \big\rangle_\Omega + \lambda_{\text{gs}}(1-\sigma)\,\big\langle \|\hat{v}^n_{\text{render}} - v^n\|_2^2 \big\rangle_\Omega\Big], \tag{7}$$

where $\langle\cdot\rangle_\Omega$ denotes a view average. We use $\lambda_{\text{gs}} = 1.0$. Gaussian rendering losses are typically larger than RGB losses and less stable at high noise. Weighting by $1-\sigma$ prioritizes RGB structure formation at high noise while retaining 3DGS supervision across $0 < \sigma < 1$ for RGB–3D consistency.

Depth, perceptual, and novel-view supervision. We obtain depth targets $d^\star$ from Depth Anything 3 (DA3) (Lin et al., 2025) and supervise valid pixels with a log-space loss, $\mathcal{L}_{\text{depth}} = \langle \rho(\log \hat{d} - \log d^\star)\rangle_{\text{valid}}$, where ρ is the Huber penalty and the predicted and reference depths use the same convention. We adopt PixelGen's low-noise gate (Ma et al., 2026), $g(\sigma) = \mathbf{1}[\sigma < 0.7]$. We retain its local LPIPS supervision (Zhang et al., 2018) and use frozen π^3 (Wang et al., 2025c) as the geometry-aware multi-view feature extractor Ψ. For each branch $b \in \{\text{rgb}, \text{render}\}$, the loss is

$$\mathcal{L}^b_{\text{perc}} = \mathbb{E}\big[g(\sigma)\big(\langle \text{LPIPS}(X^n_b, \mathcal{I}^n)\rangle_\Omega + \|\Psi(X_b) - \text{sg}[\Psi(\mathcal{I})]\|^2_{\text{mean}}\big)\big]. \tag{8}$$

Gradients pass through the frozen feature extractor to the prediction, while reference features are detached.

We additionally supervise held-out renderings with $\mathcal{L}_{\text{NVS}}$, a squared reconstruction error weighted by $\lambda_{\text{gs}}(1-\sigma)/\sigma_\delta^2$ to match the input-view 3DGS objective. The target images are excluded from the DiT inputs. We use PixelGen's LPIPS coefficient 0.1 (Ma et al., 2026) to weight our combined perceptual term, yielding the complete scene-training objective

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{NVS}} + 0.1\big(\mathcal{L}^{\text{rgb}}_{\text{perc}} + \mathcal{L}^{\text{render}}_{\text{perc}}\big). \tag{9}$$

3.2 PIXWORLD DISTILLATION: FEW-STEP 3D GENERATION FROM AN RGB TEACHER

To obtain a few-step 3D generator, we apply the DMD formulation in Sec. 2 to rendered images, following the cross-mode distillation principle of FlashWorld (Li et al., 2025a). The student G_θ uses RGB predictions for its first three denoising steps and predicts a Gaussian scene at the final step (Appendix B.4). Its renderings receive distribution-matching supervision. In Eq. 3, we therefore instantiate the image generator as

$$\hat{X}_\theta = \mathcal{R}(G_\theta(z, c), \mathcal{T}), \qquad \frac{\partial \hat{X}_\theta}{\partial \theta} = \frac{\partial \mathcal{R}}{\partial G_\theta} \frac{\partial G_\theta}{\partial \theta}. \tag{10}$$

FlashWorld performs cross-mode distillation on encoded renderings, $E(\mathcal{R}(G_\theta(z, c), \mathcal{T}))$, so the student gradient must pass through the VAE encoder as well as the renderer (Li et al., 2025a). Even with frozen encoder weights, backpropagation to the 3DGS student requires the encoder's input Jacobian. PixWorld instead matches distributions directly in RGB space, as in Eq. 10. This removes the VAE compression bottleneck from distillation and the additional encoder forward/backward path: RGB score differences supervise the rendered pixels directly and propagate through the renderer to the Gaussian generator.

The RGB real-score model $s^{\text{rgb}}_{\text{real}}$ is a frozen copy of the trained RGB branch, and the RGB fake-score model $s^{\text{rgb}}_{\text{fake}}$ is a trainable copy that tracks the student's rendered-image distribution. They process the same re-noised renderings $X_\tau = (1-\tau)\hat{X}_\theta + \tau\epsilon$ under identical camera, observation, and text conditions. Render cameras can include sampled novel viewpoints, enabling unpaired novel-view supervision through distribution matching. DMD trains a few-step 3DGS student by matching its rendered-image distribution to the multi-step RGB teacher's high-quality generative distribution. The score difference $s^{\text{rgb}}_{\text{fake}} - s^{\text{rgb}}_{\text{real}}$ guides student updates through the differentiable renderer. These updates alternate with fake-score denoising on detached renderings, while the RGB teacher remains frozen.

3.3 PIXWORLD-STREAMING: AUTOREGRESSIVE RECONSTRUCTION AND GENERATION

PixWorld-Streaming extends PixWorld to long sequences through autoregressive diffusion distillation. We obtain a causal few-step initialization following the first two stages of Causal Forcing++ (Zhao et al., 2026), then adopt the sliding-window generation mechanism of Rolling Forcing (Liu et al., 2025). A short active window jointly denoises chunks at progressively higher noise levels, with bidirectional attention inside the window and completed history as context. As the window advances, it emits the oldest clean chunk and appends a new noise chunk. For streaming 3DGS prediction, we follow ReCoSplat's incremental scene assembly design (Cheng et al., 2026): Gaussians from completed chunks are merged into a persistent scene in a shared world frame. The model incrementally reconstructs scenes from incoming observations and generates new content conditioned on past context.

Distribution matching on rendered rollouts. We follow Rolling Forcing's rollout-based training and apply distribution matching to RGB rendered from the student's Gaussian scenes. For a sampled clip $\mathcal{V}$, let $\mathcal{S}^k_\phi$ be the merged scene used to predict chunk k, including the available scene history. Its rendered sequence is

$$\hat{X}^{\mathcal{V}}_\phi = \{\mathcal{R}(\mathcal{S}^k_\phi, \mathcal{T}_k)\}_{k\in\mathcal{V}}, \qquad X^{\mathcal{V}}_\tau = (1-\tau)\hat{X}^{\mathcal{V}}_\phi + \tau\epsilon. \tag{11}$$

The scalar τ is the DMD perturbation level, separate from the staggered noise levels used for rolling generation. Following Rolling Forcing (Liu et al., 2025), we use the frozen bidirectional RGB teacher as the real-score model $s^{\text{bi,rgb}}_{\text{real}}$, while a trainable bidirectional RGB fake-score model $s^{\text{bi,rgb}}_{\text{fake}}$ tracks the student's rendered-image distribution. Both score the same perturbed clip under the same cameras and conditioning, giving

$$\nabla_\phi \mathcal{L}_{\text{stream}} = \mathbb{E}\left[w(\tau)\big(s^{\text{bi,rgb}}_{\text{fake}} - s^{\text{bi,rgb}}_{\text{real}}\big) \frac{\partial \hat{X}^{\mathcal{V}}_\phi}{\partial \phi} \right]. \tag{12}$$

The gradient passes through rendering into the 3DGS student. Fake-score denoising updates use detached student renders, while the real-score model remains frozen. This distribution matching transfers the RGB teacher's generative capability to a few-step streaming 3DGS student through its rendered images.

Table 1: **Multi-view reconstruction on RealEstate10K and DL3DV-10K.** Best/second-best scores are **bold**/underlined.

	RealEstate10K						DL3DV-10K					
	4-views			8-views			4-views			8-views		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MVSplat (Chen et al., 2024b)	22.58	0.762	0.264	21.64	0.719	0.301	17.11	0.501	0.410	15.75	0.432	0.491
DepthSplat (Xu et al., 2025)	25.16	0.832	0.194	27.77	0.872	0.154	20.38	**0.719**	0.320	19.26	**0.692**	0.360
AnySplat (Jiang et al., 2025)	20.07	0.731	0.286	20.52	0.752	0.262	20.11	0.671	0.318	20.02	0.664	0.327
YoNoSplat (Ye et al., 2026a)	25.86	0.841	0.143	28.35	0.889	0.107	22.89	0.710	0.228	21.92	0.678	0.262
PixWorld (Ours)	**26.15**	**0.845**	**0.136**	**28.74**	**0.895**	**0.103**	**23.26**	0.718	**0.221**	**22.44**	0.685	**0.249**

Table 2: **One-view generation on RealEstate10K and DL3DV-10K.** Best/second-best scores are **bold**/underlined.

	Novel View Synthesis			*Generation Quality*				*Camera Control*		
Method	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
RealEstate10K										
LVSM (Jin et al., 2025)	17.83	0.603	0.336	0.971	0.970	0.594	0.506	0.711	0.592	0.372
GF (Wu et al., 2025)	15.63	0.554	0.454	0.932	0.941	0.505	0.475	0.596	0.478	0.290
Gen3C (Ren et al., 2025)	17.26	0.624	0.391	0.951	0.957	0.561	0.524	0.648	0.514	0.334
FlashWorld (Li et al., 2025a)	16.51	0.626	0.404	0.959	0.961	0.616	0.550	0.843	0.759	0.546
Gen3R (Huang et al., 2026b)	17.59	0.631	0.382	0.974	0.971	0.552	0.536	0.633	0.434	0.148
PixWorld (Ours)	**19.21**	**0.716**	**0.312**	**0.980**	**0.980**	**0.622**	**0.560**	**0.875**	**0.800**	**0.622**
DL3DV-10K										
LVSM (Jin et al., 2025)	14.91	0.433	0.530	0.931	0.933	0.494	0.466	0.553	0.373	0.134
GF (Wu et al., 2025)	12.69	0.357	0.591	0.898	0.910	0.474	0.435	0.492	0.338	0.114
Gen3C (Ren et al., 2025)	15.58	0.514	0.479	0.927	0.933	0.532	0.496	0.552	0.377	0.128
FlashWorld (Li et al., 2025a)	15.42	0.473	0.461	0.942	0.950	0.619	0.559	0.769	0.674	0.420
Gen3R (Huang et al., 2026b)	15.76	0.503	0.496	0.945	0.942	0.547	0.531	0.593	0.399	0.118
PixWorld (Ours)	**16.63**	**0.541**	**0.433**	**0.951**	**0.957**	**0.627**	**0.569**	**0.810**	**0.718**	**0.500**

Table 3: **WorldScore-Static on the photorealistic subset.** Average summarizes seven metrics. Best and second-best scores are **bold** and underlined.

Method	Camera Control	Object Control	Content Alignment	3D Consistency	Photometric Consistency	Style Consistency	Subjective Quality	**Average**
AETHER (Zhu et al., 2025)	57.44	52.26	28.06	79.84	58.68	72.09	41.11	55.64
Uni3C (Cao et al., 2025)	62.94	45.83	47.40	78.59	85.48	88.32	57.00	66.51
Voyager (Huang et al., 2025b)	45.92	**57.69**	48.36	56.00	80.68	72.89	44.74	58.04
FantasyWorld (Dai et al., 2025)	57.05	34.46	38.45	83.31	86.11	**94.22**	57.40	64.43
FlashWorld (Li et al., 2025a)	83.86	51.07	**58.36**	84.93	87.84	78.62	55.26	71.42
PixWorld (ours)	**88.52**	50.27	57.15	**88.94**	**92.68**	74.39	**59.81**	**73.11**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We initialize PixWorld from Wan2.2-TI2V-5B (Wan et al., 2025) and train on RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), ACID (Liu et al., 2021), and 10K sequences selected from SpatialVID (Wang et al., 2025a), using 16 posed views at 480×832. Training comprises 2K steps of L2P input alignment, 50K steps of joint scene training, and 10K DMD2 updates (Yin et al., 2024a) to obtain a four-step generator. We evaluate reconstruction and conditional generation on 200 test scenes from each of RealEstate10K and DL3DV, using PSNR/SSIM/LPIPS for view fidelity, VBench (Huang et al., 2024; 2025d) for visual quality, and π^3-based pose AUC (Wang et al., 2025c;e) for camera control. WorldScore-Static (Duan et al., 2025) further evaluates scene generation. Training and evaluation details are provided in Appendix B.

4.2 EVALUATION OF UNIFIED RECONSTRUCTION AND GENERATION

Tables 1 and 2 evaluate reconstruction and one-view generation on both datasets. PixWorld achieves the highest PSNR and lowest LPIPS in all reconstruction settings, and leads across the one-view generation metrics. The two-view results in Table 14 further show strong perceptual fidelity and camera control. Figure 5 shows qualitative comparisons between PixWorld and existing methods along camera trajectories, while Figure 3 illustrates different reconstruction and generation conditions. On WorldScore-Static, PixWorld achieves an average score of 73.11 (Table 3). Four-step generation takes 1.29 s on one H100 under the native inference configuration; full timings are given in Appendix C.4.



Figure 5: **Single-view generation along camera trajectories.** PixWorld and each baseline use the same input image and camera path. Large panels show the input viewpoint; smaller panels show middle and final views. PixWorld results are four-step Gaussian renderings.



Figure 6: **RGB-to-3DGS distillation.** Gaussian renderings from the four-step distilled model (top) and the 50-step model (bottom), using the same input and camera trajectory. Distillation reduces ghosting and sharpens scene boundaries.

Table 4: **Streaming reconstruction on DL3DV (256 views).** †: predicted poses.

Method	PSNR↑	SSIM↑	LPIPS↓
YoNoSplat (offline)	21.549	0.749	<u>0.190</u>
OF^{3}GS†	15.677	0.403	0.547
ReCoSplat	<u>22.003</u>	<u>0.751</u>	0.202
PixWorld-Streaming	**22.172**	**0.760**	**0.186**

Table 5: **60-second streaming generation.** IQ: image quality; R-PSNR: revisit PSNR.

Method	CMC↓	IQ↑	R-PSNR↑
LingBot-World	<u>2.05</u>	**69.32**	<u>14.18</u>
HY-WorldPlay	2.83	52.18	12.85
SANA-WM	2.16	<u>68.74</u>	14.04
PixWorld-Streaming	**1.85**	63.39	**20.33**

4.3 Evaluation of Streaming Reconstruction and Generation

For streaming training, we collect 3,000 one-minute clips from YouTube, SEKAI Gaming, and SEKAI Real Walk, reserving 100 sequences across the three sources for held-out generation evaluation. Tables 4 and 5 evaluate reconstruction from 256 DL3DV views and generation on those 60-second test sequences, respectively. PixWorld-Streaming achieves the best reconstruction PSNR/SSIM/LPIPS, CMC, and revisit PSNR among the compared methods. Data and evaluation protocols are given in Appendix B.2 and B.6.2; speed and memory are reported in Appendix C.4.

4.4 Ablation Study

Tables 6 and 7 compare pixel space with SD-VAE (Rombach et al., 2022) and RAE (Zheng et al., 2025), using a 1.3B DiT trained from scratch for 100K steps at 256×256.

Table 6: **Pixel-space vs. latent-space modeling for 1-view generation.**

Representation	PSNR↑	LPIPS↓	IQ↑	AUC@5↑
SD-VAE	16.04	0.447	0.493	0.335
RAE	15.46	0.486	0.501	0.372
Pixel space	**17.95**	**0.353**	**0.603**	**0.538**

Table 7: **Pixel-space vs. latent-space 8-view reconstruction (RE10K).**

Representation	PSNR↑	SSIM↑	LPIPS↓
SD-VAE	27.47	0.865	0.172
RAE	26.89	0.853	0.188
Pixel space	**28.16**	**0.885**	**0.106**

Table 8: **Ablation study on RealEstate10K one-view generation.** (a) Losses and input alignment. (b) Pixel-space heads. (c) Initialization and distillation. (d) Fine-tuning scope ($n + n$: first/last n blocks). (e) Gaussian loss weight. Best/second-best: **bold**/underlined.

	Novel View Synthesis			Generation Quality				Camera Control		
Variant	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
(a) Component ablations: 20K steps										
Full model	**18.53**	**0.679**	**0.352**	**0.972**	**0.975**	**0.609**	**0.546**	**0.842**	**0.756**	**0.568**
w/o LPIPS	17.84	0.653	0.371	0.968	0.973	0.603	0.541	0.833	0.736	0.529
w/o geometry perception	18.15	0.668	0.364	0.971	**0.975**	0.604	0.544	0.840	0.741	0.543
w/o depth supervision	18.03	0.662	0.363	0.970	0.974	0.606	0.538	0.826	0.728	0.508
w/o L2P alignment	18.19	0.671	0.361	**0.972**	0.973	0.604	0.545	0.835	0.744	0.537
(b) Pixel-space head ablations (RGB and 3DGS): 20K steps										
JiT Head	18.06	0.655	0.382	0.962	0.965	0.594	0.527	0.823	0.735	0.564
DiP Head (default)	**18.53**	**0.679**	**0.352**	**0.972**	**0.975**	**0.609**	**0.546**	**0.842**	**0.756**	**0.568**
PiT Head	18.47	0.663	0.371	0.968	0.970	0.597	0.529	0.821	0.715	0.557
(c) Initialization and distillation										
Scratch (200K)	18.56	0.675	0.341	0.971	0.972	0.611	0.549	0.858	0.781	0.609
L2P (50K)	18.82	0.694	0.327	0.974	0.976	0.617	0.552	0.862	0.785	0.603
L2P + DMD	**19.21**	**0.716**	**0.312**	**0.980**	**0.980**	**0.622**	**0.560**	**0.875**	**0.800**	**0.622**
(d) Fine-tuning scope: 20K steps										
$5 + 5$ blocks (default)	**18.53**	**0.679**	**0.352**	**0.972**	**0.975**	**0.609**	**0.546**	**0.842**	**0.756**	**0.568**
$10 + 10$ blocks	17.93	0.661	0.368	0.965	0.968	0.601	0.539	0.826	0.751	0.517
Full fine-tuning (30)	18.02	0.665	0.362	0.971	0.967	0.583	0.526	0.816	0.703	0.498
(e) Gaussian loss weight: 20K steps										
$\lambda_{\text{gs}} = 0.5$	17.52	0.628	0.404	0.957	0.962	0.587	0.518	0.743	0.625	0.408
$\lambda_{\text{gs}} = 1.0$ (default)	**18.53**	**0.679**	**0.352**	**0.972**	**0.975**	**0.609**	**0.546**	**0.842**	**0.756**	**0.568**
$\lambda_{\text{gs}} = 2.0$	18.14	0.657	0.364	0.967	0.969	0.598	0.542	0.832	0.735	0.513

Pixel space achieves 17.95 dB generation PSNR and 0.106 eight-view reconstruction LPIPS. Under 20K-step fine-tuning, Table 8(a) ablates losses and L2P alignment: removing LPIPS increases LPIPS from 0.352 to 0.371, while removing depth reduces AUC@5 from 0.568 to 0.508. Following Z-Image-Pixel (Jiang et al., 2026), Table 8(b) compares JiT's linear (Li & He, 2025), DiP's convolutional (Chen et al., 2025e), and PiT's Transformer (Yu et al., 2025b) heads for RGB and 3DGS; DiP leads all ten metrics.

In Table 8(c), the same scratch model trains for 100K steps at 256×256, then continues for 100K at 480×832. Training from scratch takes 200K steps to approach L2P's performance after 20K scene-training steps. L2P (50K) further improves PSNR/LPIPS to 18.82 dB/0.327, showing faster convergence. Four-step DMD2 reaches 19.21 dB/0.312 and reduces NFE from 100 to 4 (Figure 6). Following L2P (Chen et al., 2026c), fine-tuning the first and last five blocks leads all ten 20K-step metrics in Table 8(d). In Table 8(e), $\lambda_{\text{gs}} = 1.0$ leads all ten metrics; settings appear in Appendix C.1.

5 Conclusion

We present PixWorld, a pixel-space diffusion framework that unifies 3D scene reconstruction and generation by jointly processing clean and noisy views. It eliminates the autoencoder bottleneck and directly supervises pixel-aligned 3D Gaussians through differentiable rendering. RGB-to-3DGS distillation transfers the RGB branch's generative capability to a four-step generator. We further introduce PixWorld-Streaming through autoregressive diffusion distillation, enabling online reconstruction and generation conditioned on past context over minute-long sequences. Experiments show stronger generation than latent-space methods and reconstruction quality comparable to state-of-the-art methods, with this quality retained under causal streaming inputs. These results highlight pixel-space diffusion as a promising paradigm for scalable, unified 3D scene modeling.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, to assist with translation, language editing, manuscript drafting and organization, related-work searches, implementation and reference checks, and schematic figure preparation. The authors developed the research ideas, model and training designs, and scientific conclusions, reviewed all AI-assisted materials, and take responsibility for the final paper and artifacts.

ETHICS STATEMENT

This work uses public research datasets, collected YouTube videos, and pretrained models for 3D scene reconstruction and generation. Potential misuse includes misleading depictions of real places and reconstruction of private environments without consent. Responsible use should respect privacy and applicable data and model licenses, identify generated content when it could be mistaken for real observations, and require additional validation and safeguards before safety-critical deployment.

REPRODUCIBILITY STATEMENT

The paper describes the PixWorld and PixWorld-Streaming architectures, training objectives, distillation procedures, and evaluation settings. It provides model configurations, optimization settings, loss definitions, streaming details, benchmark protocols, ablation settings, and runtime measurements. Experiments use public datasets, collected YouTube clips, and pretrained models; data collection and benchmark protocols are specified in the appendix.

REFERENCES

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models. *arXiv preprint arXiv:2510.18457*, 2025.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3C: Unifying precisely 3D-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. *arXiv preprint arXiv:2509.25162*, 2025a.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024a. URL https://arxiv.org/abs/2407.01392.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025b. URL https://arxiv.org/abs/2504.13074.

Hansheng Chen, Jan Ackermann, Minseo Kim, Gordon Wetzstein, and Leonidas Guibas. Asymmetric flow models. *arXiv preprint arXiv:2605.12964*, 2026a. URL https://arxiv.org/abs/2605.12964.

Junyu Chen, Wenkun He, Yuchao Gu, Yuyang Zhao, Jincheng Yu, Junsong Chen, Dongyun Zou, Yujun Lin, Zhekai Zhang, Muyang Li, Haocheng Xi, Ligeng Zhu, Enze Xie, Song Han, and Han Cai. DC-VideoGen: Efficient video generation with deep compression video autoencoder. *arXiv preprint arXiv:2509.25182*, 2025c. URL https://arxiv.org/abs/2509.25182.

Ruiyang Chen, Feiran Li, Chu Zhou, Zonglin Li, Zhanyu Ma, and Heng Guo. OF3GS: On-the-fly feed-forward 3D gaussian splatting from unposed images. *arXiv preprint arXiv:2606.03254*, 2026b. URL https://arxiv.org/abs/2606.03254.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025d.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024b.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Adv. Neural Inform. Process. Syst.*, 37:107064–107086, 2024c.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. DiP: Taming diffusion models in pixel space. *arXiv preprint arXiv:2511.18822*, 2025e.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Jiawei Chen, Zhuoqi Zeng, Wei Zhang, Chengjie Wang, Jian Yang, and Ying Tai. L2P: Unlocking latent potential for pixel generation. *arXiv preprint arXiv:2605.12013*, 2026c. URL https://arxiv.org/abs/2605.12013.

Freeman Cheng, Botao Ye, Xueting Li, Junqi You, Fangneng Zhan, and Ming-Hsuan Yang. ReCoSplat: Online feed-forward gaussian splatting via render-and-compare. *arXiv preprint arXiv:2603.09968*, 2026. URL https://arxiv.org/abs/2603.09968.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-Forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. URL https://arxiv.org/abs/2510.02283.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Adv. Neural Inform. Process. Syst.*, 36:39897–39914, 2023.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Sensen Gao, Zhaoqing Wang, Qihang Cao, Dongdong Yu, Changhu Wang, Tongliang Liu, Mingming Gong, and Jiawang Bian. Oneworld: Taming scene generation with 3d unified representation autoencoder. *arXiv preprint arXiv:2603.16099*, 2026.

Hyojun Go, Dominik Narnhofer, Goutam Bhat, Prune Truong, Federico Tombari, and Konrad Schindler. Text-to-3D by stitching a multi-view reconstruction network to a video generator. *arXiv preprint arXiv:2510.13454*, 2025a.

Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, and Changick Kim. Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21524–21536, 2025b.

Hyojun Go, Byeongjun Park, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Videorfsplat: Direct scene-level text-to-3d gaussian splatting generation with flexible pose and multi-view joint modeling. *arXiv preprint arXiv:2503.15855*, 2025c.

Jiaxin Guo, Tongfan Guan, Wenzhen Dong, Wenzhao Zheng, Wenting Wang, Yue Wang, Yeung Yam, and Yun-Hui Liu. SaLon3R: Structure-aware long-term feedforward 3D reconstruction from unposed images. *International Journal of Computer Vision*, 134:383, 2026. doi: 10.1007/s11263-026-02965-2. URL https://link.springer.com/article/10.1007/s11263-026-02965-2.

Junlin Hao, Peiheng Wang, Haoyang Wang, Xinggong Zhang, and Zongming Guo. Gaussvideodreamer: 3d scene generation with video diffusion and inconsistency-aware gaussian splatting. *arXiv preprint arXiv:2504.10001*, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18062–18071, 2025.

Guichen Huang, Ruoyu Wang, Xiangjun Gao, Che Sun, Yuwei Wu, Shenghua Gao, and Yunde Jia. LongSplat: Online generalizable 3D gaussian splatting from long sequence images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4994–5002, 2026a. doi: 10.1609/aaai.v40i6.42504. URL https://ojs.aaai.org/index.php/AAAI/article/view/42504.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers*, 2025a.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026b.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. *arXiv preprint arXiv:2506.04225*, 2025b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025c. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025d.

Allan Jabri, David J Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 14569–14589, 2023.

Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. *arXiv preprint arXiv:2603.22275*, 2026.

Dengyang Jiang, Ruoyi Du, Zhennan Chen, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Xiangpeng Yang, Huanqia Cai, Aiming Hao, Yuming Jiang, Peng Gao, Harry Yang, and Steven Hoi. An empirical study of training pixel-space text-to-image diffusion models. *arXiv preprint arXiv:2608.16887*, 2026. URL https://arxiv.org/abs/2608.16887.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *Int. Conf. Learn. Represent.*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, et al. 3d and 4d world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, Jia-Wang Bian, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR, 2026.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *Adv. Neural Inform. Process. Syst.*, 37:75125–75151, 2024.

Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025a.

Yang Li, Jinglu Wang, Lei Chu, Xiao Li, Shiu-hong Kao, Ying-Cong Chen, and Yan Lu. StreamGS: Online generalizable gaussian splatting reconstruction for unposed image streams. *arXiv preprint arXiv:2503.06235*, 2025b. URL https://arxiv.org/abs/2503.06235.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 300–309, 2023.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, 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.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14458–14467, 2021. URL https://infinite-nature.github.io/.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. URL https://arxiv.org/abs/2509.25161.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint arXiv:2511.19365*, 2025.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. PixelGen: Pixel diffusion beats latent diffusion with perceptual loss. *arXiv preprint arXiv:2602.02493*, 2026. URL https://arxiv.org/abs/2602.02493.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Pedro Quesado, Erkut Akdag, Yasaman Kashefbahrami, Willem Menu, and Egor Bondarev. LiveStre4m: Feed-forward live streaming of novel views from unposed multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026. URL https://arxiv.org/abs/2604.06740.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6121–6132, 2025.

Robbyant Team. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. URL https://arxiv.org/abs/2202.00512.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023.

Katja Schwarz, Denis Rozumny, Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. A recipe for generating 3d worlds from a single image. In *Int. Conf. Comput. Vis.*, pp. 3520–3530, 2025.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Changhao Song, Yuxuan Wang, Qibiao Li, Youcheng Cai, and Ligang Liu. StreamSplat: Streaming feed-forward 3D gaussian splatting. *arXiv preprint arXiv:2608.01659*, 2026. URL https://arxiv.org/abs/2608.01659.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025. URL https://arxiv.org/abs/2502.06764.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. URL https://arxiv.org/abs/2303.01469.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025. URL https://arxiv.org/abs/2512.14614.

Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. Vggt-world: Transforming vggt into an autoregressive geometry world model. *arXiv preprint arXiv:2603.12655*, 2026.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a. URL https://arxiv.org/abs/2509.09676.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025b.

Weijie Wang, Qihang Cao, Sensen Gao, Donny Y Chen, Haofei Xu, Wenjing Bian, Songyou Peng, Tat-Jen Cham, Chuanxia Zheng, Andreas Geiger, et al. Feed-forward 3d scene modeling: A problem-driven perspective. *arXiv preprint arXiv:2604.14025*, 2026.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c.

Yiming Wang, Lucy Chai, Xuan Luo, Michael Niemeyer, Manuel Lagunas, Stephen Lombardi, Siyu Tang, and Tiancheng Sun. Learning efficient fuse-and-refine for feed-forward 3D gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 38, 2025d. URL https://papers.neurips.cc/paper_files/paper/2025/hash/6e9d416f7ffeb2042729ddee8d3138a3-Abstract-Conference.html.

Zhaoqing Wang, Xiaobo Xia, Zhuolin Bie, Jinlin Liu, Dongdong Yu, Jia-Wang Bian, and Changhu Wang. Taming camera-controlled video generation with verifiable geometry reward. *arXiv preprint arXiv:2512.02870*, 2025e.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Adv. Neural Inform. Process. Syst.*, 36:8406–8441, 2023.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 21551–21561, 2024.

Zike Wu, Qi Yan, Xuanyu Yi, Lele Wang, and Renjie Liao. StreamSplat: Towards online dynamic 3D reconstruction from uncalibrated video streams. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c822d05dcc00695ed6b63c9e97bb3449-Abstract-Conference.html.

Chong Xia, Fangfu Liu, Yule Wang, Yize Pang, and Yueqi Duan. OnlineX: Unified online 3D reconstruction and understanding with active-to-stable state evolution. *arXiv preprint arXiv:2603.02134*, 2026. URL https://arxiv.org/abs/2603.02134.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a. URL https://arxiv.org/abs/2509.22622.

Yuanbo Yang, Jiahao Shao, Xinyang Li, Yujun Shen, Andreas Geiger, and Yiyi Liao. Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2857–2869, 2025b.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep Forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025. URL https://arxiv.org/abs/2512.05081.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *arXiv preprint arXiv:2412.07772*, 2024c. URL https://arxiv.org/abs/2412.07772.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6658–6667, 2024.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5916–5926, 2025a.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025b.

Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiao-Xiao Long, Hanxue Liang, Zexiang Xu, Hao Su, et al. Advances in feed-forward 3d reconstruction and view synthesis: A survey. *arXiv preprint arXiv:2507.14501*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pp. 3836–3847, 2023.

Renhe Zhang, Yuyang Tan, Jingyu Gong, Zhizhong Zhang, Lizhuang Ma, Yuan Xie, and Xin Tan. S2GS: Streaming semantic gaussian splatting for online scene understanding and reconstruction. *arXiv preprint arXiv:2603.14232*, 2026. URL https://arxiv.org/abs/2603.14232.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 586–595, 2018.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026. URL https://arxiv.org/abs/2605.15141.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4):1–12, 2018.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Int. Conf. Comput. Vis.*, pp. 8535–8546, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a. URL https://arxiv.org/abs/2605.15178.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026b. URL https://arxiv.org/abs/2602.02214.

APPENDIX

The appendix is organized into four parts:

- § A **Related Work**: 3D scene generation and reconstruction, pixel-space diffusion, and autoregressive diffusion distillation.
- § B **Implementation Details**: architecture, training, minute-long video collection, RGB and 3DGS sampling modes, distribution matching, streaming rollout, and evaluation protocols.
- § C **Additional Experimental Results**: ablation studies, detailed generation results, short-horizon streaming experiments, inference speed, and memory usage.
- § D **Additional Visualizations**: PixWorld and PixWorld-Streaming reconstruction and generation, with camera trajectories, RGB renderings, and depth maps.

A RELATED WORK

A.1 3D SCENE GENERATION

Diffusion-based Iterative 3D Scene Generation. Early 3D generation methods leverage pretrained 2D generative models (Rombach et al., 2022; Podell et al., 2023; Zhang et al., 2023; Peebles & Xie, 2023) as generative priors. A representative line employs score distillation, including Score Distillation Sampling (SDS) and its variants (Lin et al., 2023; Poole et al., 2022; Tang et al., 2023; Wang et al., 2023) to optimize a 3D representation such as 3DGS (Kerbl et al., 2023) or NeRF (Mildenhall et al., 2021) by aligning rendered views with a pretrained 2D diffusion model. View-independent diffusion priors can lead to cross-view inconsistencies. MVDream (Shi et al., 2023) introduces a multi-view diffusion prior, while ZeroNVS (Sargent et al., 2023) adapts view-conditioned diffusion and SDS anchoring to in-the-wild scenes. These approaches still require per-scene iterative optimization.

Multi-View Reconstruction-Based 3D Scene Generation. A second line uses diffusion-generated multi-view images or videos to reconstruct or refine a 3D representation (Chen et al., 2024c; Gao et al., 2024; Hao et al., 2025; Liu et al., 2024; 2023; Sun et al., 2024; Wu et al., 2024; Zhao et al., 2024). Related approaches progressively expand scenes through image outpainting and geometric fusion (Chung et al., 2023; Fridman et al., 2023; Schwarz et al., 2025; Yu et al., 2024; 2025a). Depending on the method, reconstruction uses feed-forward prediction or per-scene optimization, with multi-view conditioning and geometry-aware refinement used to improve consistency.

3D Scene Generation in Latent Spaces. More recent work moves diffusion into compressed latent spaces and learns a latent-to-3D decoder, typically targeting 3DGS. Several methods build on pretrained image or video latent diffusion models (Rombach et al., 2022; Wan et al., 2025) and learn mappings from generated latents to 3D Gaussians (Yang et al., 2025b; Go et al., 2025b; Li et al., 2024; 2025a; Dai et al., 2025; Go et al., 2025c). VIST3A (Go et al., 2025a) stitches a video generator to a pretrained multi-view reconstruction network and aligns the combined model through reward fine-tuning. Alongside work on Representation Autoencoders (RAEs) and foundation-model tokenizers (Zheng et al., 2025; Shi et al., 2025; Chen et al., 2025a; Bi et al., 2025), OneWorld (Gao et al., 2026) and GLD (Jang et al., 2026) perform generation in latent spaces derived from pretrained 3D foundation models. VGGT-World (Sun et al., 2026) instead forecasts the temporal evolution of frozen geometry features. Gen3R (Huang et al., 2026b) couples generation and reconstruction through aligned appearance and geometry latents.

A.2 3D SCENE RECONSTRUCTION

3D scene reconstruction aims to recover 3D representations from one or more captured views. Classical pipelines use structure-from-motion to estimate camera poses and sparse geometry, followed by multi-view stereo for dense reconstruction, while NeRF (Mildenhall et al., 2021) and 3D Gaussian Splatting (Kerbl et al., 2023) popularized per-scene optimization for high-fidelity novel view synthesis. To avoid expensive per-scene fitting, recent feed-forward methods learn to directly map single-view or multi-view inputs to a 3D representation in a single forward pass (Hong et al., 2023; Charatan et al., 2024; Szymanowicz et al., 2024; Chen et al., 2024b; Xu et al., 2025), with LRM (Hong et al.,

2023) regressing implicit triplanes and pixel-aligned Gaussian regressors (Charatan et al., 2024; Szymanowicz et al., 2024; Chen et al., 2024b) predicting per-pixel Gaussians.

Recent feed-forward methods extend 3DGS reconstruction to streaming observations without per-scene optimization. StreamGS (Li et al., 2025b) refines and aggregates per-frame Gaussians through cross-frame correspondences, and OF^{3}GS (Chen et al., 2026b) stabilizes unposed streaming reconstruction through decoupled intrinsic recovery and point refinement offsets. LongSplat (Huang et al., 2026a) encodes reconstruction history in a Gaussian-Image Representation, while SaLon3R (Guo et al., 2026) compresses Gaussians into saliency-aware anchors and refines them with a point transformer. ReCoSplat (Cheng et al., 2026) conditions scene updates on render-and-compare feedback and compresses its KV cache, whereas StreamSplat (Song et al., 2026) maintains a voxel-aligned cache for calibrated input streams. OnlineX (Xia et al., 2026) and S2GS (Zhang et al., 2026) additionally maintain language or instance-semantic fields alongside online Gaussian reconstruction. For dynamic scenes, SplatVoxel (Wang et al., 2025d) learns temporal Gaussian fusion, the dynamic StreamSplat of Wu et al. (2026) combines deformation with adaptive Gaussian fusion, and LiveStre4m (Quesado et al., 2026) couples Gaussian keyframe reconstruction with image-space interpolation for live novel-view video.

A.3 Pixel-Space Diffusion

Latent diffusion models (Rombach et al., 2022; Podell et al., 2023; Peebles & Xie, 2023) dominate large-scale image generation by operating in a compressed VAE latent space, introducing an indirection between the generative target and the pixel output. A complementary line of work performs diffusion directly in pixel space (Dhariwal & Nichol, 2021; Jabri et al., 2023; Hoogeboom et al., 2023; 2025; Wang et al., 2025b; Chen et al., 2025d; Yu et al., 2025b; Ma et al., 2025; Chen et al., 2025e; Li & He, 2025), with recent architectures improving generation quality and computational efficiency without a separate latent autoencoder.

Latent-to-pixel transfer reduces pixel-space training costs by reusing pretrained latent generators. L2P (Chen et al., 2026c) adapts pixel input/output layers on synthetic images while freezing pretrained intermediate layers. AsymFlow (Chen et al., 2026a) transfers latent priors through low-rank subspace alignment and asymmetric flow parameterization. Jiang et al. (2026) establish a scalable post-training recipe with Z-Image-Pixel, demonstrating the benefits of latent pretraining followed by pixel-space adaptation.

A.4 Autoregressive Diffusion Distillation

Diffusion Forcing (Chen et al., 2024a) enables flexible sequential generation through per-token noise levels, with extensions to video history conditioning and long-video synthesis (Song et al., 2025; Chen et al., 2025b). Few-step generation builds on progressive distillation (Salimans & Ho, 2022), consistency models (Song et al., 2023), and distribution matching distillation (Yin et al., 2024b;a). CausVid (Yin et al., 2024c) distills a bidirectional video model into a few-step causal generator, while Self Forcing (Huang et al., 2025c) reduces exposure bias by training on self-generated histories. Rolling Forcing (Liu et al., 2025), Self-Forcing++ (Cui et al., 2025), and LongLive (Yang et al., 2025a) further improve long-horizon generation through rolling denoising or training on extended rollouts; Deep Forcing (Yi et al., 2025) instead improves rollout stability through training-free cache management. Causal Forcing (Zhu et al., 2026b) addresses the teacher–student causality mismatch during initialization, and Causal Forcing++ (Zhao et al., 2026) makes this stage more efficient with causal consistency distillation.

B Implementation Details

B.1 Architecture and Training

Architecture. The Wan-based PixWorld backbone uses 30 transformer blocks, a hidden width of 3072, and 24 attention heads. Pixel patchification uses spatial patches of 32×32, with one view per token frame. Each input view contains 13 channels: noisy RGB (3), camera rays (6), conditioning RGB (3), and an observation mask (1). Shape-compatible transformer and conditioning weights are initialized from Wan2.2-TI2V-5B; the input interface is adapted through the token alignment in Sec. 3.1. We adopt the lightweight convolutional design of DiP (Chen et al., 2025e) for the pixel-space heads that predict RGB and 3DGS from the shared DiT context. The Gaussian branch predicts the depth-derived centers and remaining attributes.

Table 9: **PixWorld backbone and pixel interfaces.**

Component	Configuration
Shared DiT	30 blocks, width 3072, 24 attention heads
Pixel patchification	$(1, 32, 32)$ over view, height, width
Per-view input	RGB 3 + camera rays 6 + conditioning RGB 3 + mask 1
RGB output	DiP-based pixel-space head with shared DiT context
Gaussian output	DiP-based pixel-space head with shared DiT context
Gaussian centers	Depth-based unprojection using view cameras

From-scratch ablation model. For the representation study and the from-scratch baseline, we use a compact DiT with 30 transformer layers, hidden width 1536, and approximately 1.3B parameters. Its pixel-space configuration uses patch size $(1, 16, 16)$ over view, height, and width, with DiP-based heads for both RGB and 3DGS prediction. All parameters of this model are trained from random initialization. We choose this size to make the ablations computationally affordable; its parameter count is comparable in scale to the trainable subset of our Wan2.2-TI2V-5B adaptation, which updates the first and last five blocks and the pixel interfaces. The transferred model retains the larger pretrained backbone specified in Table 9.

Data and optimization settings. We train on RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), ACID (Liu et al., 2021), and a 10K-sequence subset of SpatialVID (Wang et al., 2025a). The training pipeline samples clean, noisy, and held-out target views from posed multi-view sequences. Training uses 32 NVIDIA H100 GPUs with a global batch size of 64 scenes; each scene contains 16 input views at 480×832. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01 for the scene backbone and 10^{-6} for the Gaussian branch, and gradient clipping at norm 1.

Training schedule. We initialize the shared backbone from Wan2.2-TI2V-5B (Wan et al., 2025) and adapt it to 16 posed views at 480×832. L2P input alignment trains the pixel interface for 2K steps before 50K steps of joint scene training (Sec. 3.1). The base-training configuration uses AdamW (Loshchilov & Hutter, 2017) with learning rate 5×10^{-5}, 2K warmup steps, cosine decay to 5×10^{-6}, and EMA decay 0.999 over 50K steps. The first and last five transformer blocks and the pixel interfaces are trainable, while intermediate backbone blocks remain frozen. We use the joint RGB/rendering velocity objective, held-out NVS supervision, depth supervision, and noise-gated 2D/3D perceptual losses defined in Eq. 9. Few-step DMD2 distillation (Yin et al., 2024a) uses four sampling steps and 10K alternating optimizer updates, with four fake-score updates per generator update (8K fake-score and 2K generator updates); the generator and fake-score learning rates are 10^{-6} and 5×10^{-7}. PixWorld-Streaming trains with global batch size 64 in three stages: 20K steps of causal adaptation, 5K steps of few-step causal distillation, and 3K steps of Rolling Forcing post-training on Gaussian rollouts (Sec. 3.3).

Optimization objectives. The transferred model first aligns its pixel input tokens to frozen Wan token targets, then optimizes the joint scene objective in Eq. 9. Both direct RGB predictions and Gaussian renderings are supervised in velocity space. The Gaussian term carries the multiplier $\lambda_{\text{gs}}(1 - \sigma)$, with $\lambda_{\text{gs}} = 1.0$ by default. Following PixelGen (Ma et al., 2026), we fix the perceptual

coefficient to its LPIPS weight 0.1 and adopt its low-noise cutoff, applying both to our combined 2D/3D perceptual term. The term is active only for $\sigma < 0.7$; this cutoff follows from PixelGen's timestep threshold $t = 0.3$ under $t = 1 - \sigma$. Depth supervision is evaluated separately on valid labels. Few-step distillation alternates Gaussian-student updates with RGB fake-score updates, keeping the RGB real-score model frozen. Streaming distillation uses Rolling Forcing with staggered-noise windows, bidirectional attention among active chunks, and the student's own completed history as described in Appendix B.5.

From-scratch training schedule. The from-scratch pixel-space model follows a two-stage schedule: the first 100K steps use 256×256, and the next 100K steps continue training the same model at 480×832 from its low-resolution checkpoint, for 200K steps in total. The representation comparisons in Tables 6 and 7 evaluate the pixel-space, SD-VAE, and RAE models after 100K steps at 256×256. The performance comparison in Table 8(c) uses the same pixel-space model after completing both stages.

B.2 Minute-Long Video Collection

For minute-long streaming experiments, we assemble 3,000 one-minute clips from three sources (Table 10). Following the scene-video collection approach of RealEstate10K (Zhou et al., 2018), we collect 1,000 continuous YouTube clips without shot cuts, plus 1,000 clips each from SEKAI Gaming and SEKAI Real Walk. We estimate camera poses and intrinsics for these clips with ViPE (Huang et al., 2025a). We reserve 100 sequences across these three sources as a held-out test set for 60-second streaming generation; the remaining 2,900 are used for training.

Table 10: **Minute-long video collection for streaming experiments.**

Source	Clips	Duration per clip
YouTube (continuous, no shot cuts)	1,000	1 min
SEKAI Gaming	1,000	1 min
SEKAI Real Walk	1,000	1 min
Total	3,000	1 min

B.3 Sampling with RGB and 3DGS Modes

PixWorld supports both direct RGB and 3DGS generation modes through its shared backbone. We denote the normalized sampling timestep by σ, consistent with the noise level in Sec. 3.1; sampling proceeds from $\sigma = 1$ (noise) toward $\sigma = 0$ (clean). At high noise, Gaussian prediction and rendering are unstable, so we use direct RGB predictions to establish a coarse scene layout. We then switch to Gaussian renderings at lower noise to refine the scene with explicit 3D consistency. For each generated view, the clean-image estimate used by the sampler is

$$\hat{X}(\sigma) = \begin{cases} X_{\mathrm{rgb}}, & \sigma > \sigma_{\mathrm{sw}}, \\ X_{\mathrm{render}} = \mathcal{R}(\hat{\mathcal{G}}, \mathcal{T}), & 0 < \sigma \le \sigma_{\mathrm{sw}}. \end{cases} \tag{13}$$

The selected prediction defines the corresponding velocity as in Sec. 3.1. Conditioning views remain clean throughout sampling, and the final output is the predicted Gaussian scene. We set $\sigma_{\mathrm{sw}} = 0.5$ by default and specify the threshold ablation in Appendix C.1.

B.4 Rendered-Image Distribution Matching

Three networks have distinct roles. The few-step 3DGS student generates a Gaussian scene and its rendered images. The frozen RGB real-score model estimates the target image distribution; the trainable RGB fake-score model estimates the current student-render distribution, following FlashWorld (Li et al., 2025a). Both score models receive the same re-noised rendered views and the same camera, observation, and text conditions. Observed conditioning views remain fixed and are excluded from the generative score loss.

Four-step student schedule. For non-streaming DMD2 training and inference, we follow FlashWorld's timestep schedule (Li et al., 2025a), normalized to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1.0, 0.9, 0.75, 0.5)$. With the switching rule in Eq. 13, our student uses RGB mode for the first three steps and 3DGS mode for the final step. The last step predicts the Gaussian scene whose renderings receive the distribution-matching supervision. The student uses no classifier-free guidance, giving four network function evaluations. These generator timesteps are distinct from the perturbation level τ used to query the real and fake score models below. Training alternates four fake-score updates with one generator update, totaling 8K and 2K updates, respectively.

FlashWorld applies DMD to VAE-encoded renderings, so updating its Gaussian student requires backpropagation through both the renderer and the frozen VAE encoder's input Jacobian (Li et al., 2025a). Our RGB-space matching sends the score-difference gradient directly through the renderer to the Gaussian student, avoiding that encoder pass.

Let $\mu_j(X_\tau, \tau, c)$ be the clean-image prediction of score model $j \in \{\text{real}, \text{fake}\}$. Under the linear forward process, its distribution score estimate is

$$s_j(X_\tau, \tau, c) = \frac{(1-\tau)\mu_j(X_\tau, \tau, c) - X_\tau}{\tau^2}, \qquad 0 < \tau < 1. \tag{14}$$

Here s_j abbreviates s_j^{rgb} in non-streaming distillation and $s_j^{\text{bi,rgb}}$ in streaming distillation. This converts both RGB predictions to scores under the same noise convention. For a reverse-KL distribution-matching objective on noisy samples, the student gradient contains $(1-\tau)(s_{\text{fake}} - s_{\text{real}})$; Eq. 3 absorbs this forward-process factor into $w(\tau)$. During a student update, score evaluations are treated as a fixed gradient direction. One equivalent surrogate is

$$\mathcal{L}_{\text{sur}} = \frac{1}{2} \left\| \hat{X}_\theta - \text{sg}\Big[\hat{X}_\theta - w(\tau)(s_{\text{fake}} - s_{\text{real}}) \Big] \right\|_2^2. \tag{15}$$

Its gradient follows the score difference through the rendered image into the Gaussian student. During a fake-score update, the student renderings are detached and treated as clean denoising targets. The fake-score model minimizes velocity-space error on their perturbed versions, while the real-score model remains frozen. Alternating these updates lets the fake score follow the changing student distribution.

B.5 Streaming Rollout and Scene Assembly

Each chunk contains four temporally ordered views ($K = 4$). Reconstruction ingests clean posed views, while generation predicts new views from available observations and completed history. Following Rolling Forcing (Liu et al., 2025), active chunks at staggered noise levels attend within a sliding window; advancing the window completes its oldest chunk and appends a noisy one. Causal adaptation and few-step initialization follow Causal Forcing++ (Zhao et al., 2026).

Following ReCoSplat (Cheng et al., 2026), completed chunks contribute Gaussians to a persistent scene in a shared world frame. Direct RGB predictions provide visual attention history, and Gaussian renderings provide scene supervision. Training scores rendered clips from the student's own rollout with the real and fake RGB score models described in Appendix B.4. The chunk experiments appear in Sec. C.3.

B.6 Evaluation Protocols and Metrics

B.6.1 Offline Evaluation

We sample 200 test scenes with a large camera-pose range from each of RealEstate10K and DL3DV-10K. Reconstruction uses four or eight posed observations and disjoint target views. First-Frame generation conditions the forward trajectory on its first image; Bidirectional generation uses a middle image and generates toward both ends. Interpolation uses endpoint anchors, whereas Extrapolation uses anchors at one end of the trajectory. One-view and two-view results each average 100 cases per configuration. All generation methods receive camera conditions; all except LVSM also receive text. Figure 3 illustrates view selection. In Figure 5, PixWorld uses 16 views at 480×832 and the same input image and camera path as each baseline.

Evaluation metrics. For PSNR, SSIM, and LPIPS, we run each method at its native training resolution (AnySplat at 336×448), then center-crop its rendering and ground truth to 1:1 and resize both to 256×256 with identical interpolation. We assess generation quality with VBench's I2V Subject, I2V Background, Image Quality, and Aesthetic Quality metrics (Huang et al., 2024; 2025d), and camera control with π^3-estimated pose AUC at $30°$, $15°$, and $5°$ (Wang et al., 2025c;e). VBench, pose, and WorldScore retain their prescribed processing.

WorldScore. Table 3 uses the WorldScore-Static photorealistic subset (Duan et al., 2025). Camera and Object Control assess controllability; Content Alignment assesses prompt adherence; 3D, Photometric, and Style Consistency assess stability; Subjective Quality measures perceptual quality. The average combines all seven. Baseline scores for AETHER, Uni3C, Voyager, and FantasyWorld follow FantasyWorld (Dai et al., 2025); we evaluate FlashWorld and PixWorld on the same subset.

B.6.2 Streaming Evaluation

For Table 4, we use ReCoSplat's 140-scene DL3DV test protocol and baseline scores (Cheng et al., 2026): 256 context views from up to the first 360 frames and eight disjoint targets. Its fine-tuned YoNoSplat checkpoint and ReCoSplat use provided poses and intrinsics; calibrated OF^{3}GS uses ground-truth intrinsics and predicted poses (†).

Streaming generation uses a held-out test set of 100 one-minute sequences from the three collection sources in Sec. B.2, under our 60-second camera-sequence protocol with ViPE-estimated camera poses and intrinsics (Huang et al., 2025a). Table 5 compares LingBot-World (Robbyant Team, 2026), HY-WorldPlay (Sun et al., 2025), SANA-WM (Zhu et al., 2026a), and PixWorld-Streaming on this test set. CMC measures camera-motion consistency, IQ is VBench's imaging-quality score (Huang et al., 2024), and revisit PSNR measures appearance consistency at near-return viewpoints.

C Additional Experimental Results

We provide complete ablation and conditional-generation results, followed by chunk-based streaming experiments and measurements of inference speed and memory.

C.1 Ablation Studies

Tables 11 and 12 report all ten generation metrics and both reconstruction view counts, respectively. Every representation uses the compact from-scratch DiT described in Appendix B.1; these comparisons evaluate the checkpoints after 100K steps at 256×256.

Table 11: **Pixel-space vs. latent-space 1-view generation: complete metrics.** Best/second-best scores are **bold**/underlined.

	Novel View Synthesis			Generation Quality				Camera Control		
Representation	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
SD-VAE (Rombach et al., 2022)	<u>16.04</u>	<u>0.581</u>	<u>0.447</u>	0.923	<u>0.930</u>	0.493	0.470	0.629	0.502	0.335
RAE (Zheng et al., 2025)	15.46	0.548	0.486	<u>0.924</u>	0.929	<u>0.501</u>	<u>0.492</u>	<u>0.675</u>	<u>0.534</u>	<u>0.372</u>
Pixel space	**17.95**	**0.647**	**0.353**	**0.976**	**0.972**	**0.603**	**0.547**	**0.825**	**0.736**	**0.538**

Table 12: **Pixel-space vs. latent-space reconstruction on RealEstate10K.** Results use 4 or 8 input views. Best/second-best scores are **bold**/underlined.

	4 views			8 views		
Representation	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SD-VAE (Rombach et al., 2022)	<u>24.93</u>	<u>0.831</u>	<u>0.192</u>	<u>27.47</u>	<u>0.865</u>	<u>0.172</u>
RAE (Zheng et al., 2025)	24.06	0.803	0.221	26.89	0.853	0.188
Pixel space	**25.82**	**0.847**	**0.141**	**28.16**	**0.885**	**0.106**

The component study uses matched 20K-step fine-tuning, data, and sampler settings. Removing LPIPS or geometry perception disables that loss on both direct RGB predictions and Gaussian

renderings. Removing depth supervision retains the depth prediction needed for Gaussian placement. Removing L2P skips input alignment while retaining the transferred backbone. All component variants are evaluated before distillation with a 50-step sampler and CFG scale 5. The pixel-space head study uses the same 20K-step protocol, varying the output-head architecture for both RGB and 3DGS prediction.

Initialization and distillation. For Table 8(c), we continue training the same approximately 1.3B-parameter pixel-space model after its first 100K steps at 256×256, using 480×832 for the next 100K steps. The resulting Scratch (200K) checkpoint thus completes both stages of a single training run. The from-scratch model requires 200K training steps to approach the performance of L2P after only 20K scene-training steps (Table 8(a)). L2P uses 2K alignment steps before scene training; after 50K scene-training steps at 480×832, it outperforms Scratch on nine of ten metrics. These results demonstrate slower convergence from scratch and motivate transferring pretrained generative priors.

Starting from the trained L2P model, DMD2 adds 8K fake-score and 2K generator updates. In this comparison, the teacher uses 50 sampling steps with CFG scale 5 (NFE 100), whereas the student uses four steps without CFG (NFE 4).

Fine-tuning scope. Following L2P (Chen et al., 2026c), we adapt the pretrained backbone by fine-tuning its first and last blocks while freezing the intermediate blocks. Table 8(d) compares the extent of these updates. For our 30-block backbone, $5+5$ updates blocks 1–5 and 26–30 (10 trainable blocks) and freezes blocks 6–25; $10 + 10$ updates blocks 1–10 and 21–30 (20 trainable blocks) and freezes blocks 11–20. Full fine-tuning updates all 30 blocks. The pixel input projection and both RGB and 3DGS heads are fully trainable in every variant.

All variants start from the same checkpoint after 2K-step input alignment and use 20K scene-training steps with the same data, losses, DiP heads, resolution, view count, global batch size, optimizer schedule, and pre-distillation evaluation protocol. The $5 + 5$ setting performs best across all ten metrics, achieving 18.53 dB PSNR, 0.352 LPIPS, and 0.568 AUC@5. Expanding to $10 + 10$ gives 17.93 dB, 0.368, and 0.517, respectively; full fine-tuning gives 18.02 dB, 0.362, and 0.498. These results support our default choice of updating the first and last five blocks under this training budget. The default row reuses the result from Table 8(a).

Gaussian loss weight. Table 8(e) compares $\lambda_{\text{gs}} \in \{0.5, 1.0, 2.0\}$ under the same 20K-step protocol, keeping the $1 - \sigma$ noise weight and all other settings fixed. The default $\lambda_{\text{gs}} = 1.0$ uses the Full model results in Table 8(a).

Sampling-mode switching timestep. Table 13 compares $\sigma_{\text{sw}} \in \{0.7, 0.6, 0.5, 0.4\}$ for RealEstate10K one-view generation using the same 20K-step checkpoint, 50-step sampler, CFG scale 5, and evaluation protocol as the component study. Only the sampling threshold changes: larger values switch to 3DGS earlier, while smaller values retain RGB mode longer. The default $\sigma_{\text{sw}} = 0.5$ performs best across all five metrics, achieving 18.53 dB PSNR, 0.352 LPIPS, and 0.568 AUC@5. Both earlier switching (0.7 or 0.6) and later switching (0.4) reduce performance, supporting a balance between establishing the scene layout in RGB mode and refining it in 3DGS mode. The default row matches the Full model results in Table 8(a).

Table 13: **Sampling-mode switching timestep on RealEstate10K one-view generation.** Results use the 20K-step model before distillation. Best/second-best: **bold**/underlined.

σ_{sw}	PSNR↑	SSIM↑	LPIPS↓	I.Q.↑	AUC@5↑
0.7	17.53	0.641	0.375	0.572	0.476
0.6	18.16	<u>0.669</u>	<u>0.358</u>	0.595	<u>0.549</u>
0.5 (default)	**18.53**	**0.679**	**0.352**	**0.609**	**0.568**
0.4	<u>18.27</u>	0.661	0.383	<u>0.602</u>	0.537

C.2 One-View and Two-View Generation

Tables 15 and 16 report the per-configuration results. Each entry in the main-paper one-view table is the arithmetic mean of the corresponding First-Frame and Bidirectional results; each entry in the

two-view summary table is the arithmetic mean of the corresponding Interpolation and Extrapolation results, computed separately for each dataset and metric.

Table 14: **Two-view generation on RealEstate10K and DL3DV-10K.** Best/second-best scores are **bold**/underlined.

	Novel View Synthesis			*Generation Quality*				*Camera Control*		
Method	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
RealEstate10K										
LVSM (Jin et al., 2025)	23.61	**0.819**	0.215	0.971	0.964	0.607	0.516	0.861	0.789	0.611
GF (Wu et al., 2025)	18.27	0.647	0.353	0.925	0.939	0.507	0.464	0.630	0.473	0.224
Gen3C (Ren et al., 2025)	20.13	0.715	0.300	0.948	0.947	0.567	0.518	0.698	0.539	0.255
FlashWorld (Li et al., 2025a)	21.48	0.771	0.257	0.964	0.962	0.619	0.547	0.877	0.812	0.637
Gen3R (Huang et al., 2026b)	21.33	0.724	0.284	0.970	**0.972**	0.550	0.541	0.728	0.576	0.258
PixWorld (Ours)	**23.68**	0.813	**0.208**	**0.972**	0.971	**0.623**	**0.561**	**0.884**	**0.813**	**0.657**
DL3DV-10K										
LVSM (Jin et al., 2025)	19.18	0.589	0.343	0.916	0.917	0.533	0.502	0.741	0.609	0.375
GF (Wu et al., 2025)	15.39	0.459	0.470	0.897	0.912	0.479	0.445	0.563	0.379	0.147
Gen3C (Ren et al., 2025)	17.62	0.543	0.412	0.928	0.934	0.537	0.502	0.628	0.434	0.176
FlashWorld (Li et al., 2025a)	18.27	0.562	0.360	0.938	0.949	0.600	0.558	0.803	0.714	0.515
Gen3R (Huang et al., 2026b)	18.05	0.558	0.392	0.942	0.944	0.535	0.530	0.726	0.560	0.246
PixWorld (Ours)	**19.64**	**0.609**	**0.320**	**0.952**	**0.959**	**0.607**	**0.563**	**0.817**	**0.733**	**0.522**

Table 15: **One-view generation by conditioning configuration.** Best/second-best scores are **bold**/underlined.

	Novel View Synthesis			*Generation Quality*				*Camera Control*		
Method	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
RealEstate10K – 1-view First Frame										
LVSM (Jin et al., 2025)	17.95	0.604	0.335	0.968	0.971	0.584	0.505	0.695	0.576	0.359
GF (Wu et al., 2025)	15.92	0.575	0.436	0.942	0.947	0.514	0.487	0.609	0.488	0.301
Gen3C (Ren et al., 2025)	17.12	0.622	0.394	0.959	0.961	0.569	0.530	0.664	0.521	0.338
FlashWorld (Li et al., 2025a)	16.26	0.613	0.417	0.951	0.954	**0.617**	0.544	0.849	0.766	0.553
Gen3R (Huang et al., 2026b)	17.43	0.628	0.383	0.973	0.970	0.547	0.531	0.653	0.443	0.145
PixWorld (Ours)	**19.17**	**0.689**	**0.313**	**0.977**	**0.980**	0.605	**0.552**	**0.868**	**0.785**	**0.608**
RealEstate10K – 1-view Bidirectional										
LVSM (Jin et al., 2025)	17.70	0.601	0.337	0.974	0.969	0.603	0.506	0.726	0.608	0.385
GF (Wu et al., 2025)	15.34	0.532	0.471	0.921	0.935	0.495	0.462	0.583	0.467	0.279
Gen3C (Ren et al., 2025)	17.39	0.626	0.388	0.943	0.952	0.553	0.517	0.632	0.507	0.329
FlashWorld (Li et al., 2025a)	16.75	0.639	0.390	0.966	0.967	0.614	0.555	0.837	0.751	0.539
Gen3R (Huang et al., 2026b)	17.74	0.633	0.381	0.975	0.972	0.557	0.541	0.612	0.424	0.150
PixWorld (Ours)	**19.24**	**0.742**	**0.311**	**0.983**	**0.979**	**0.638**	**0.567**	**0.882**	**0.814**	**0.635**
DL3DV-10K – 1-view First Frame										
LVSM (Jin et al., 2025)	14.85	0.434	0.533	0.927	0.932	0.491	0.463	0.537	0.354	0.127
GF (Wu et al., 2025)	12.50	0.352	0.598	0.894	0.907	0.468	0.431	0.482	0.329	0.109
Gen3C (Ren et al., 2025)	15.72	0.516	0.482	0.925	0.928	0.526	0.491	0.556	0.371	0.121
FlashWorld (Li et al., 2025a)	15.31	0.468	0.463	0.939	0.950	0.616	0.558	0.765	0.667	0.412
Gen3R (Huang et al., 2026b)	15.57	0.499	0.504	0.941	0.940	0.543	0.529	0.614	0.408	0.102
PixWorld (Ours)	**16.68**	**0.553**	**0.418**	**0.948**	**0.955**	**0.622**	**0.568**	**0.815**	**0.735**	**0.483**
DL3DV-10K – 1-view Bidirectional										
LVSM (Jin et al., 2025)	14.96	0.432	0.526	0.934	0.933	0.497	0.468	0.568	0.391	0.141
GF (Wu et al., 2025)	12.88	0.361	0.584	0.902	0.913	0.479	0.439	0.501	0.346	0.118
Gen3C (Ren et al., 2025)	15.44	0.512	0.476	0.929	0.937	0.537	0.501	0.548	0.382	0.134
FlashWorld (Li et al., 2025a)	15.53	0.478	0.458	0.944	0.950	0.621	0.559	0.773	0.681	0.428
Gen3R (Huang et al., 2026b)	15.94	0.507	0.487	0.948	0.944	0.551	0.532	0.571	0.389	0.133
PixWorld (Ours)	**16.57**	**0.529**	**0.447**	**0.954**	**0.958**	**0.631**	**0.570**	**0.804**	**0.701**	**0.517**

C.3 Chunk-Based Streaming Experiments

We use four views per chunk for streaming reconstruction and generation (Sec. B.5). Reconstruction sequentially incorporates clean observations and accumulates their predicted Gaussians; generation conditions each new chunk on the available history. The minute-long video collection is described in Sec. B.2.

Table 16: **Two-view generation by conditioning configuration.** Best/second-best scores are **bold**/underlined.

	Novel View Synthesis			*Generation Quality*				*Camera Control*		
Method	PSNR↑	SSIM↑	LPIPS↓	I2V Subj.↑	I2V BG↑	I.Q.↑	Aes.Q.↑	AUC@30↑	AUC@15↑	AUC@5↑
RealEstate10K – 2-view Interpolation										
LVSM (Jin et al., 2025)	**24.26**	**0.832**	**0.200**	0.969	0.963	0.613	0.518	0.871	0.806	0.638
GF (Wu et al., 2025)	18.08	0.621	0.373	0.936	0.951	0.525	0.476	0.642	0.487	0.236
Gen3C (Ren et al., 2025)	20.17	0.726	0.298	0.967	0.949	0.581	0.527	0.714	0.553	0.261
FlashWorld (Li et al., 2025a)	21.66	0.776	0.252	0.958	0.959	0.619	0.546	0.875	0.810	0.634
Gen3R (Huang et al., 2026b)	22.42	0.752	0.260	0.964	0.973	0.544	0.539	0.779	0.631	0.295
PixWorld (Ours)	23.78	0.819	0.203	**0.972**	**0.975**	**0.621**	**0.560**	**0.882**	**0.811**	**0.653**
RealEstate10K – 2-view Extrapolation										
LVSM (Jin et al., 2025)	22.95	**0.806**	0.230	0.972	0.965	0.601	0.514	0.851	0.771	0.583
GF (Wu et al., 2025)	18.46	0.672	0.332	0.914	0.927	0.489	0.451	0.618	0.458	0.211
Gen3C (Ren et al., 2025)	20.08	0.703	0.302	0.928	0.945	0.552	0.509	0.682	0.524	0.249
FlashWorld (Li et al., 2025a)	21.30	0.765	0.261	0.970	0.965	0.618	0.548	0.878	0.813	0.640
Gen3R (Huang et al., 2026b)	20.24	0.695	0.307	**0.976**	**0.970**	0.556	0.542	0.677	0.521	0.220
PixWorld (Ours)	**23.57**	**0.806**	**0.213**	0.972	0.966	**0.625**	**0.561**	**0.885**	**0.814**	**0.661**
DL3DV-10K – 2-view Interpolation										
LVSM (Jin et al., 2025)	19.37	0.594	0.328	0.918	0.917	0.538	0.511	0.758	0.630	0.403
GF (Wu et al., 2025)	15.16	0.452	0.480	0.903	0.918	0.487	0.452	0.573	0.395	0.156
Gen3C (Ren et al., 2025)	17.81	0.546	0.406	0.931	0.938	0.541	0.508	0.641	0.446	0.183
FlashWorld (Li et al., 2025a)	18.16	0.559	0.363	0.937	0.951	0.602	0.564	0.812	0.724	0.528
Gen3R (Huang et al., 2026b)	18.21	0.562	0.391	0.940	0.946	0.531	0.536	0.745	0.582	0.260
PixWorld (Ours)	**19.53**	**0.604**	**0.320**	**0.959**	**0.962**	**0.609**	**0.570**	**0.815**	**0.732**	**0.531**
DL3DV-10K – 2-view Extrapolation										
LVSM (Jin et al., 2025)	18.98	0.584	0.358	0.913	0.916	0.528	0.493	0.723	0.587	0.346
GF (Wu et al., 2025)	15.61	0.466	0.459	0.891	0.906	0.471	0.438	0.553	0.362	0.137
Gen3C (Ren et al., 2025)	17.42	0.539	0.418	0.924	0.930	0.532	0.496	0.614	0.421	0.169
FlashWorld (Li et al., 2025a)	18.37	0.565	0.356	0.938	0.946	0.598	0.552	0.793	0.703	0.501
Gen3R (Huang et al., 2026b)	17.88	0.554	0.392	0.943	0.941	0.539	0.523	0.706	0.538	0.231
PixWorld (Ours)	**19.74**	**0.613**	**0.319**	**0.945**	**0.956**	**0.604**	**0.556**	**0.819**	**0.734**	**0.513**

C.3.1 Short-Horizon Reconstruction and Generation

Table 17 compares PixWorld and PixWorld-Streaming on RealEstate10K and DL3DV-10K at 480×832, using four views per streaming chunk. Reconstruction incorporates eight clean observations in two chunks and evaluates the assembled scene at held-out cameras. First-Frame generation covers 16 views: one observed image and 15 generated views, forming four chunks including the conditioning image. Both generators use four unguided denoising steps with matched observations and target cameras. Offline reference scores come from Tables 1 and 15.

Table 17: **Short-horizon reconstruction and generation.** Each streaming chunk contains four views. Generation uses the First-Frame setting. Best/second-best scores are **bold**/underlined.

	8-view reconstruction 8 clean views; 2 chunks			**16-view generation** 1 clean + 15 generated views; 4 chunks			
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	LPIPS↓	IQ↑	AUC@5↑
RealEstate10K							
PixWorld	**28.74**	**0.895**	**0.103**	**19.17**	**0.313**	**0.605**	0.608
PixWorld-Streaming	28.66	0.892	0.107	19.04	0.331	0.604	**0.614**
DL3DV-10K							
PixWorld	22.44	0.685	0.249	16.68	0.418	**0.622**	0.483
PixWorld-Streaming	**22.57**	**0.692**	**0.246**	**16.94**	**0.395**	0.619	**0.512**

PixWorld-Streaming retains comparable short-horizon quality under causal inputs. On RealEstate10K, reconstruction PSNR changes from 28.74 to 28.66 dB, while generation PSNR changes from 19.17 to 19.04 dB. On DL3DV-10K, streaming improves reconstruction PSNR from 22.44 to 22.57 dB and generation PSNR from 16.68 to 16.94 dB, with lower LPIPS in both tasks. Camera-control AUC@5 increases on both datasets.

C.4 Inference Speed and Memory

C.4.1 Offline Generation and Output Decoding

Table 18 uses the same input image and native model configurations. Total latency covers the warmed generation request, including decoding; model loading and file export are outside the timing interval. With temporal compression by four, $N_{\text{latent}} = 1 + (N-1)/4$, versus 24 and 16 uncompressed diffusion frames for FlashWorld and PixWorld. PixWorld's BF16 DiP head takes 0.049 s per invocation. † Voyager decodes 13 latent frames into 49 RGBD frames, removing four reference frames to return 45 RGB views.

Table 18: **Generation and decoder latency on one H100 80GB.** Batch size one; native configurations. Decoder latency is per invocation after warmup. Lowest/second-lowest latencies: **bold**/underlined.

Method	Views	Latent frames	Resolution	Steps	Total (s)↓	Decoder	Latency (s)↓
Aether (Zhu et al., 2025)	41	11	480×720	50	89.73	RGB VAE	2.782
Uni3C (Cao et al., 2025)	81	21	480×768	40	393.46	RGB VAE	2.440
Voyager (Huang et al., 2025b)	45	13	512×768	50	570.62	RGBD VAE†	8.530
FantasyWorld (Dai et al., 2025)	81	21	480×704	50	640.99	RGB VAE	5.494
FlashWorld (FP8) (Li et al., 2025a)	24	24	480×704	4	<u>3.21</u>	3DGS decoder	<u>0.255</u>
PixWorld (4-step)	16	16	480×832	4	**1.29**	DiP head	**0.049**

C.4.2 Online Reconstruction Memory

Table 19 compares joint offline and four-view online reconstruction at 480×832, batch size one, on one H100 80GB using the same checkpoint. Online processing keeps recent visual context and all completed Gaussians on the GPU; peak allocated memory includes the scene and, for the contiguous variant, its temporary concatenation copy. Clean-view reconstruction takes one forward pass.

Table 19: **Reconstruction memory on one H100 80GB.** Views: 480×832; all Gaussians retained. OOM: allocation failure. Lowest/second-lowest peak memory per row: **bold**/underlined.

Input views	Offline peak (GiB)↓	Online peak (GiB)↓	Online + contiguous scene peak (GiB)↓	Retained scene (GiB)
16	<u>27.06</u>	**16.21**	**16.21**	1.81
32	44.47	**18.56**	<u>19.56</u>	3.62
64	70.92	**22.44**	<u>27.06</u>	7.24
128	OOM	**30.21**	<u>42.07</u>	14.47
256	OOM	**44.96**	<u>71.28</u>	28.95

Offline fits 64 views but runs out of memory at 128. Online reaches 256 views at 44.96 GiB (71.28 GiB with concatenation), including 28.95 GiB of retained Gaussians. At 64 views, online peak memory is $3.16\times$ lower.

C.4.3 Streaming Chunk Rollout Throughput

We benchmark four-step chunk rollouts of PixWorld-Streaming on one H100 80GB (480×832, batch size one, BF16), following Rolling Forcing's rolling-window schedule (Liu et al., 2025). We time denoising and the DiP image head. After warmup, three 48-chunk rollouts each cover 192 views over a nominal 60-second trajectory (four views per chunk, up to four active chunks). Equivalent FPS at a 16-FPS trajectory rate is $16 \times 60/T$, where T is the rollout time in seconds.

Table 20: **Streaming chunk rollout latency on one H100 80GB.**

Execution	Chunks	Steps/CFG	Window (s)	Rollout (s)	Equivalent FPS↑ @ 16-FPS trajectory
PixWorld-Streaming	48	4/1	0.509	24.41	39.33

Table 20 gives 24.41 s (39.33 equivalent FPS at a 16-FPS trajectory rate), 1.18 s to the first chunk, and 55.01 GiB peak memory.

Table 21: **Streaming chunk rollout latency by output module.** Times are in seconds; VAE rows estimate replacement at fixed backbone cost. Best/second-best output time, total time, and throughput: **bold**/underlined.

Output path	DiT/KV	Other	Output↓	Total↓	Equivalent FPS↑ @ 16-FPS trajectory
PixWorld-Streaming	21.99	0.05	**2.36**	**24.41**	**39.33**
BF16 VAE (estimate)	21.99	0.05	<u>5.48</u>	<u>27.53</u>	<u>34.88</u>
FP32 VAE (estimate)	21.99	0.05	10.70	32.74	29.32

Table 21 compares profiled output costs with VAE replacement estimates.

D Additional Visualizations

D.1 Reconstruction and Generation

Figures 7, 8, and 9 show PixWorld under different input-view selections, single-view generation, and text-to-3D generation, respectively. RGB images and depth maps are rendered from the predicted 3D Gaussians.

D.2 Comparison with Further Baselines

Figure 10 extends the qualitative comparison in Figure 5 to four methods it does not cover.

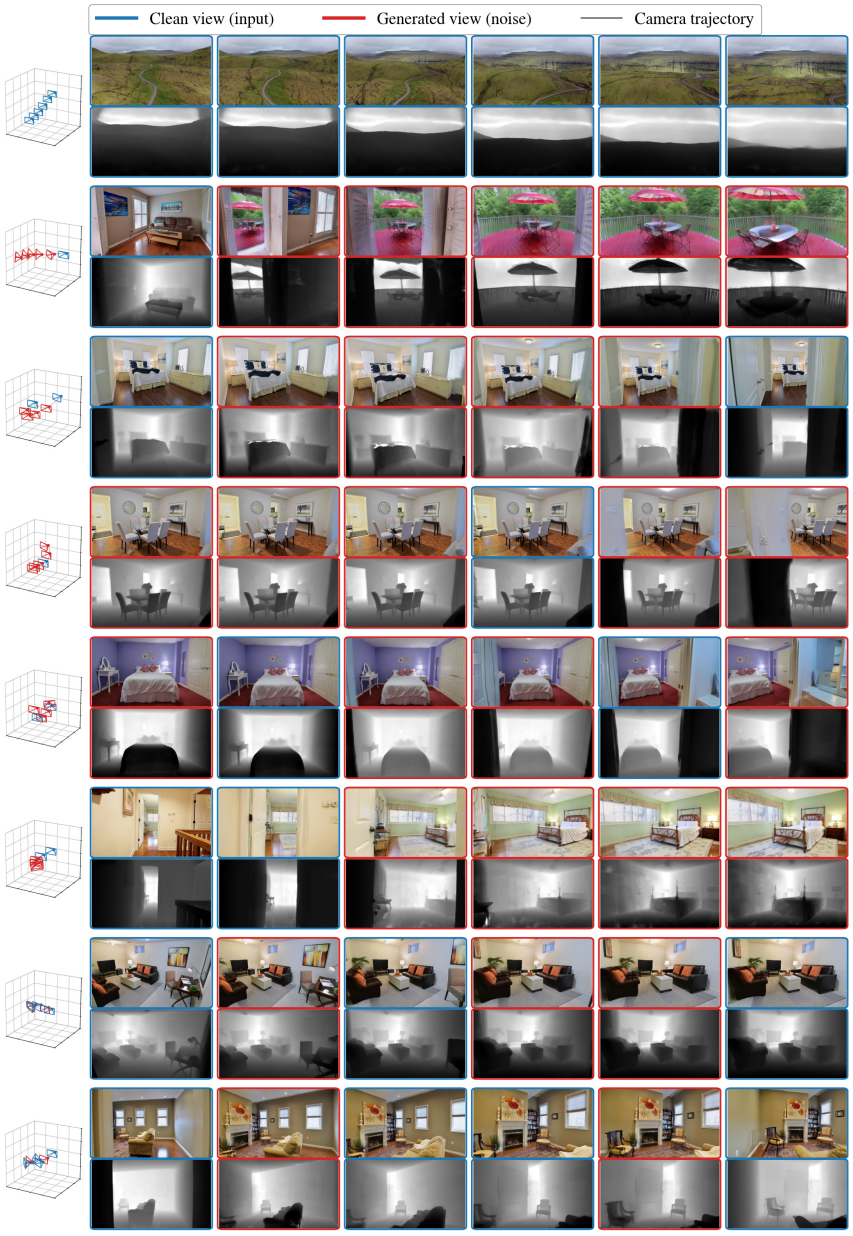


Figure 7: **Reconstruction and generation with PixWorld.** Camera trajectories, RGB renderings, and depth maps under different input-view selections.

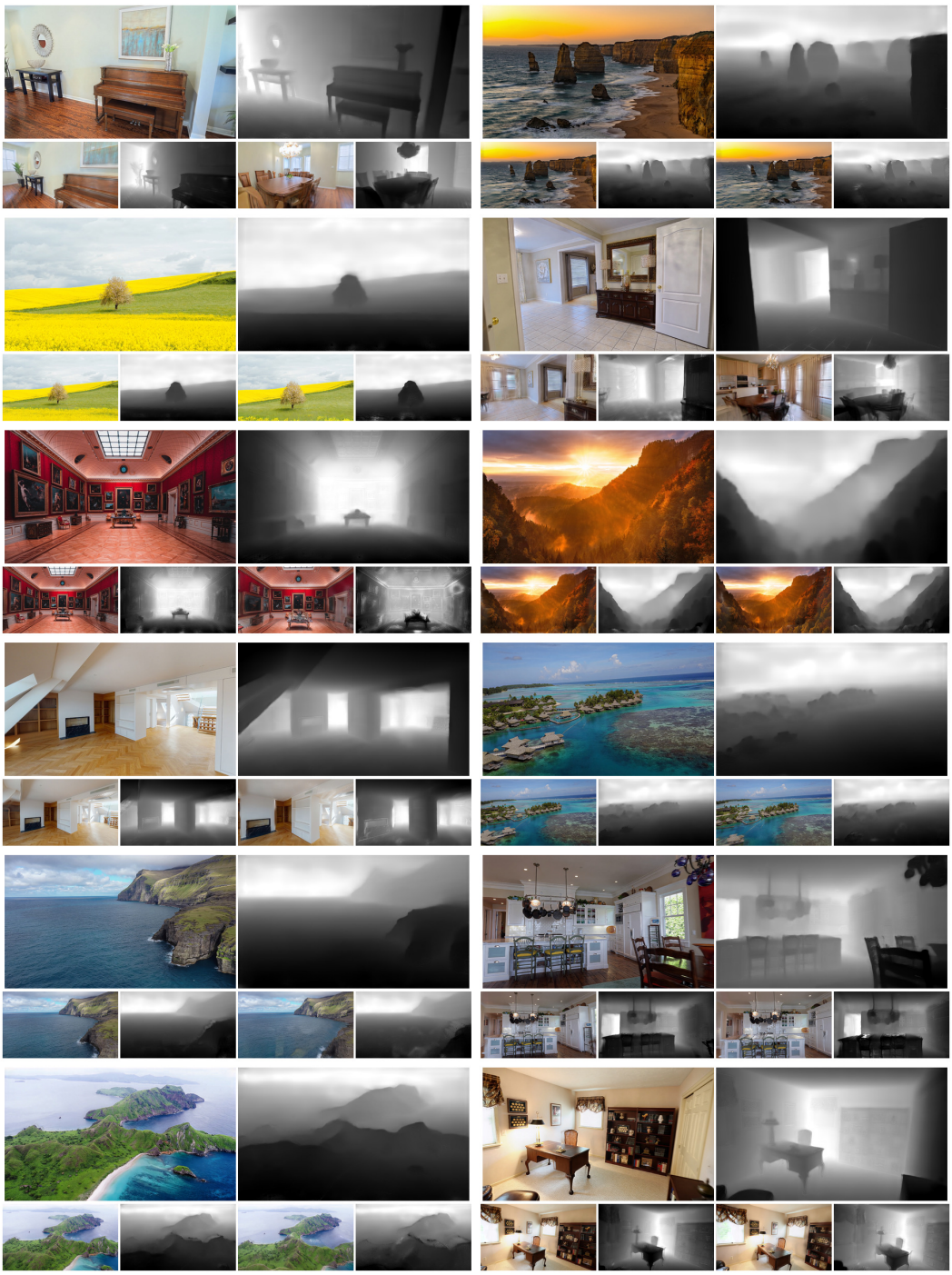


Figure 8: **Single-view generation with PixWorld.** RGB renderings and corresponding depth maps.

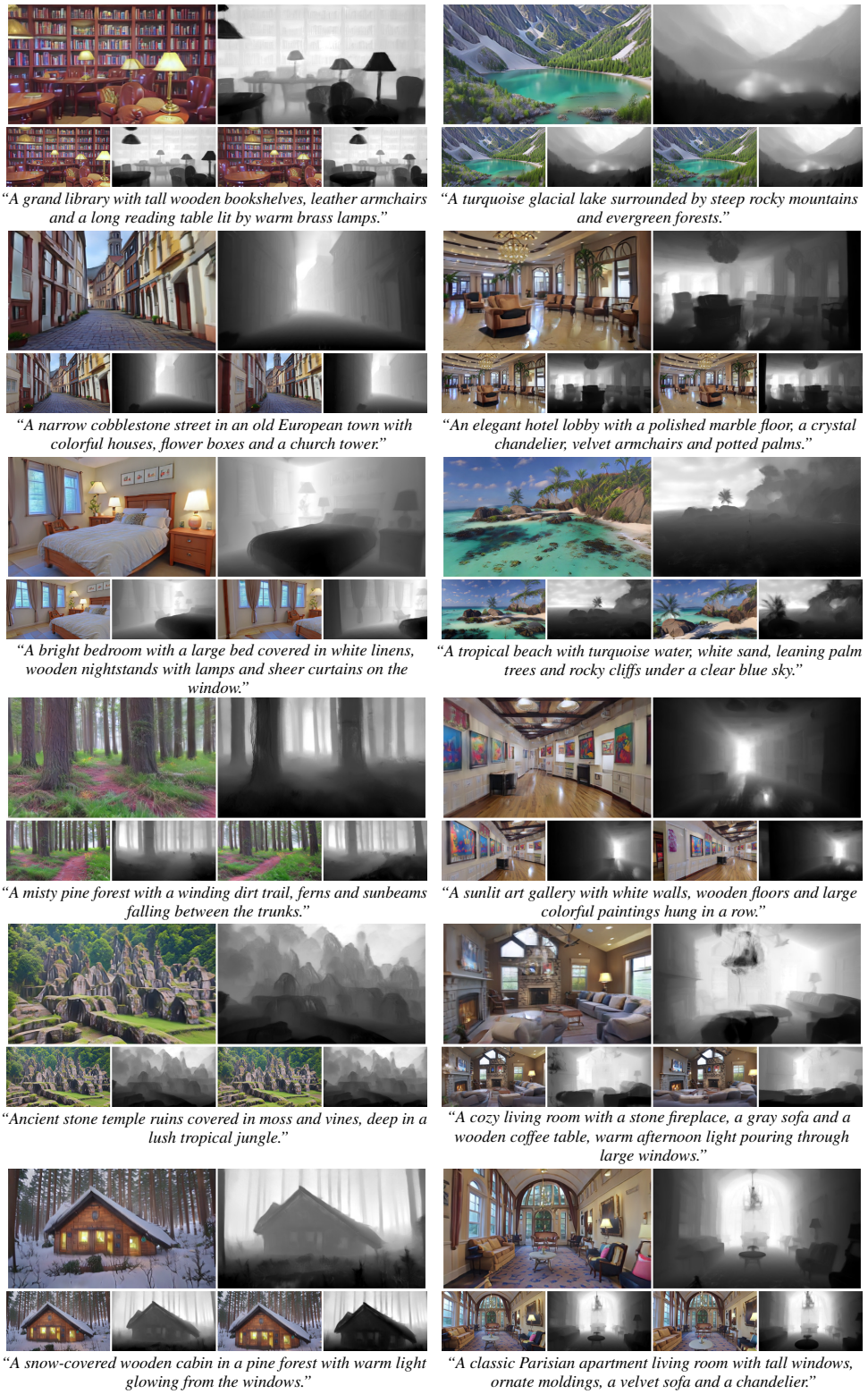


Figure 9: **Text-to-3D generation with PixWorld.** RGB renderings and depth maps; prompts are shown below each scene.

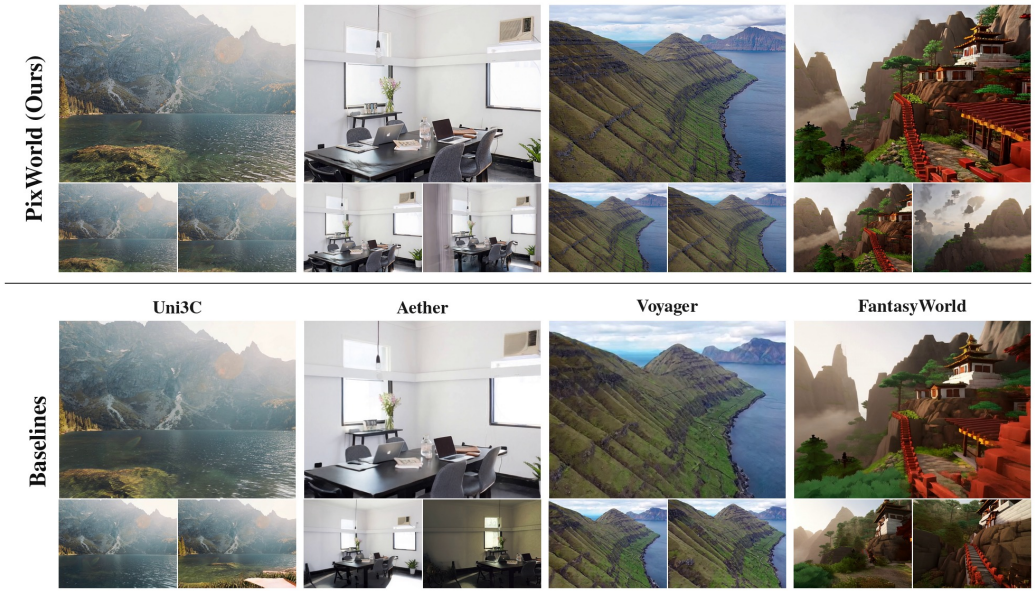
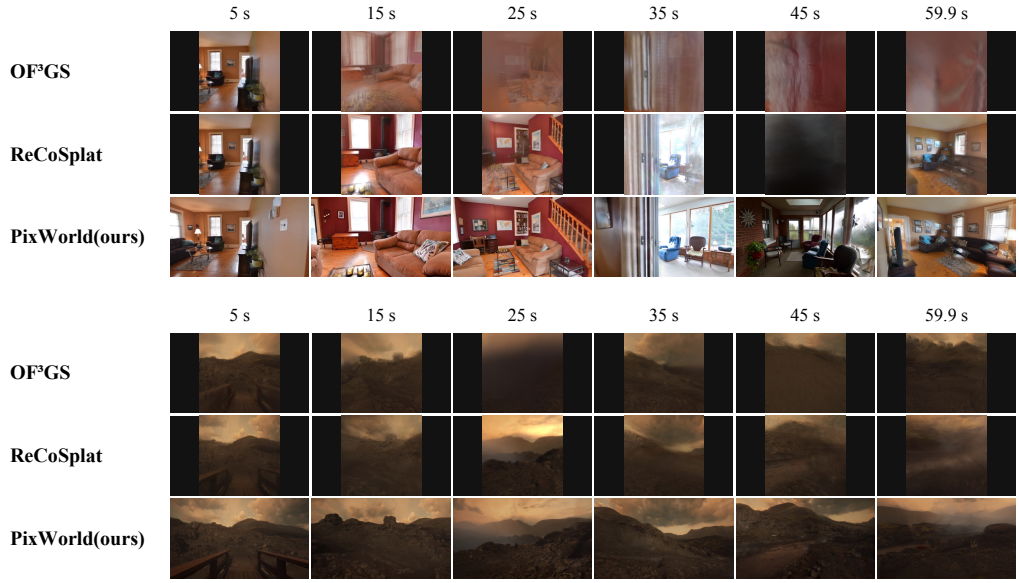


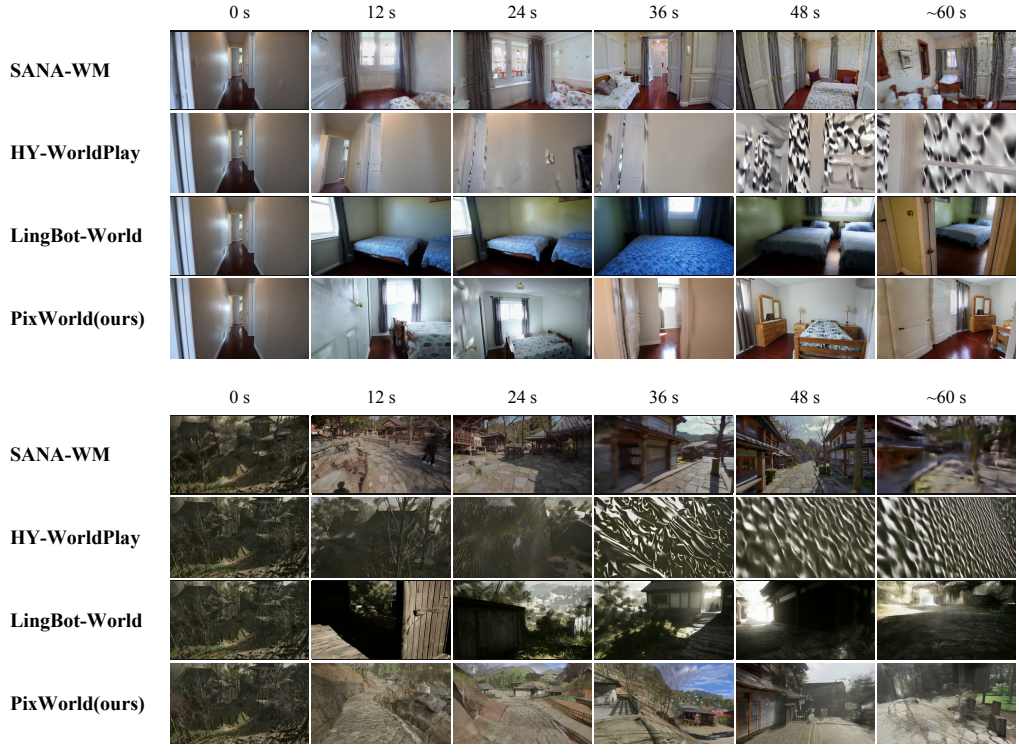
Figure 10: **Single-view generation compared with additional baselines.** The evaluation protocol and panel layout follow Figure 5. PixWorld results are four-step Gaussian renderings.

D.3 Streaming Comparisons with Other Methods

Figure 11 compares PixWorld-Streaming with other methods on streaming reconstruction and generation along indoor and outdoor camera trajectories. The sampled views show scene structure and visual consistency over time.



(a) Streaming reconstruction.



(b) Streaming generation.

Figure 11: **Streaming reconstruction and generation compared with other methods.** PixWorld-Streaming is compared with OF^{3}GS and ReCoSplat for reconstruction (a), and with SANA-WM, HY-WorldPlay, and LingBot-World for generation (b). Each panel shows indoor (top) and outdoor (bottom) scenes, with columns ordered by timestamp along the camera trajectories.

D.4 Streaming Reconstruction and Generation

We present additional indoor and outdoor examples of PixWorld-Streaming. Streaming reconstruction progressively integrates incoming observations, while streaming generation extends a scene along the camera trajectory. Figures 12 and 13 show scene overviews, RGB images, and depth maps, all rendered from the predicted 3D Gaussians.

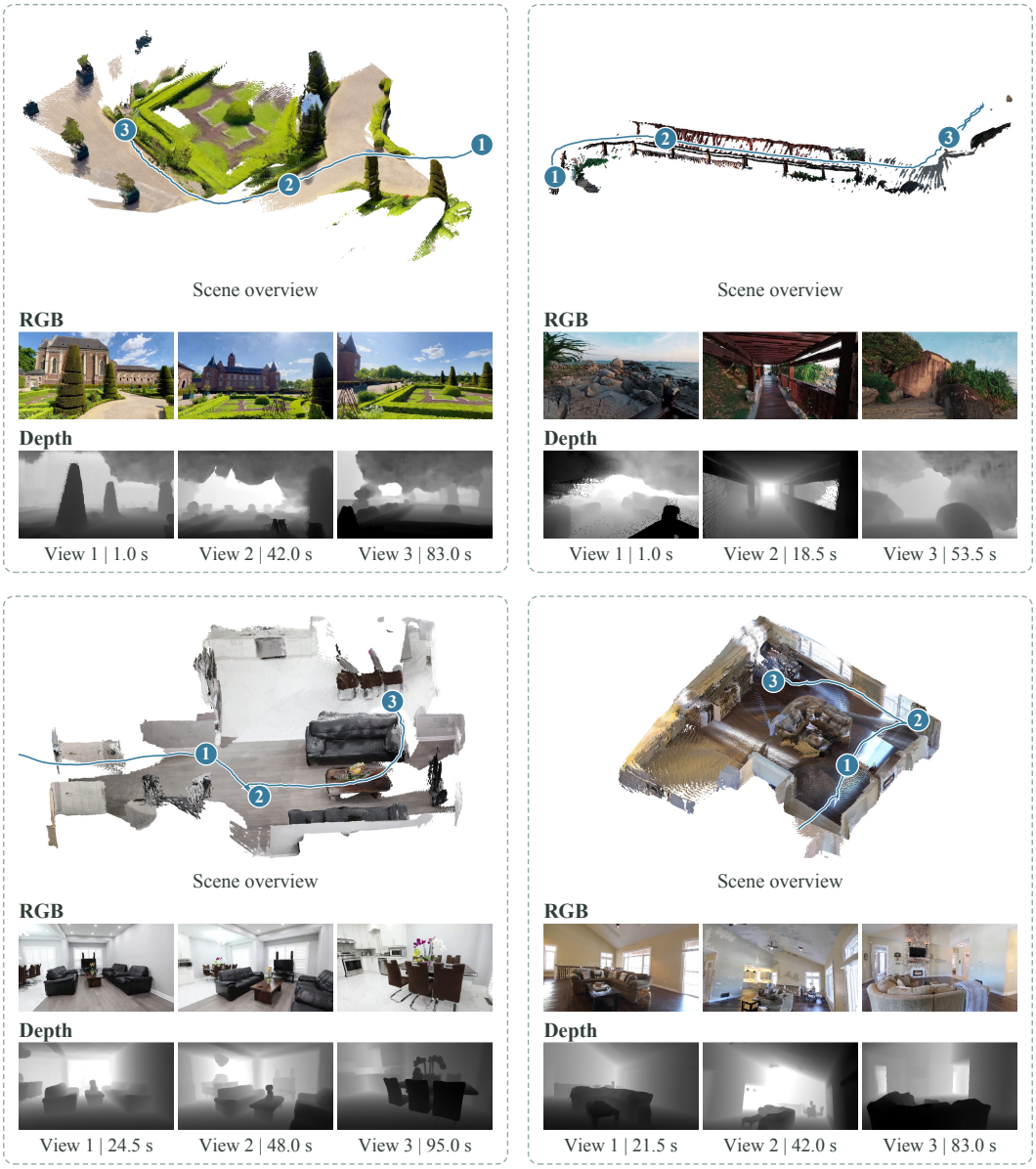


Figure 12: **Streaming reconstruction with PixWorld-Streaming.** Scene overviews and RGB/depth views rendered from predicted 3D Gaussians.

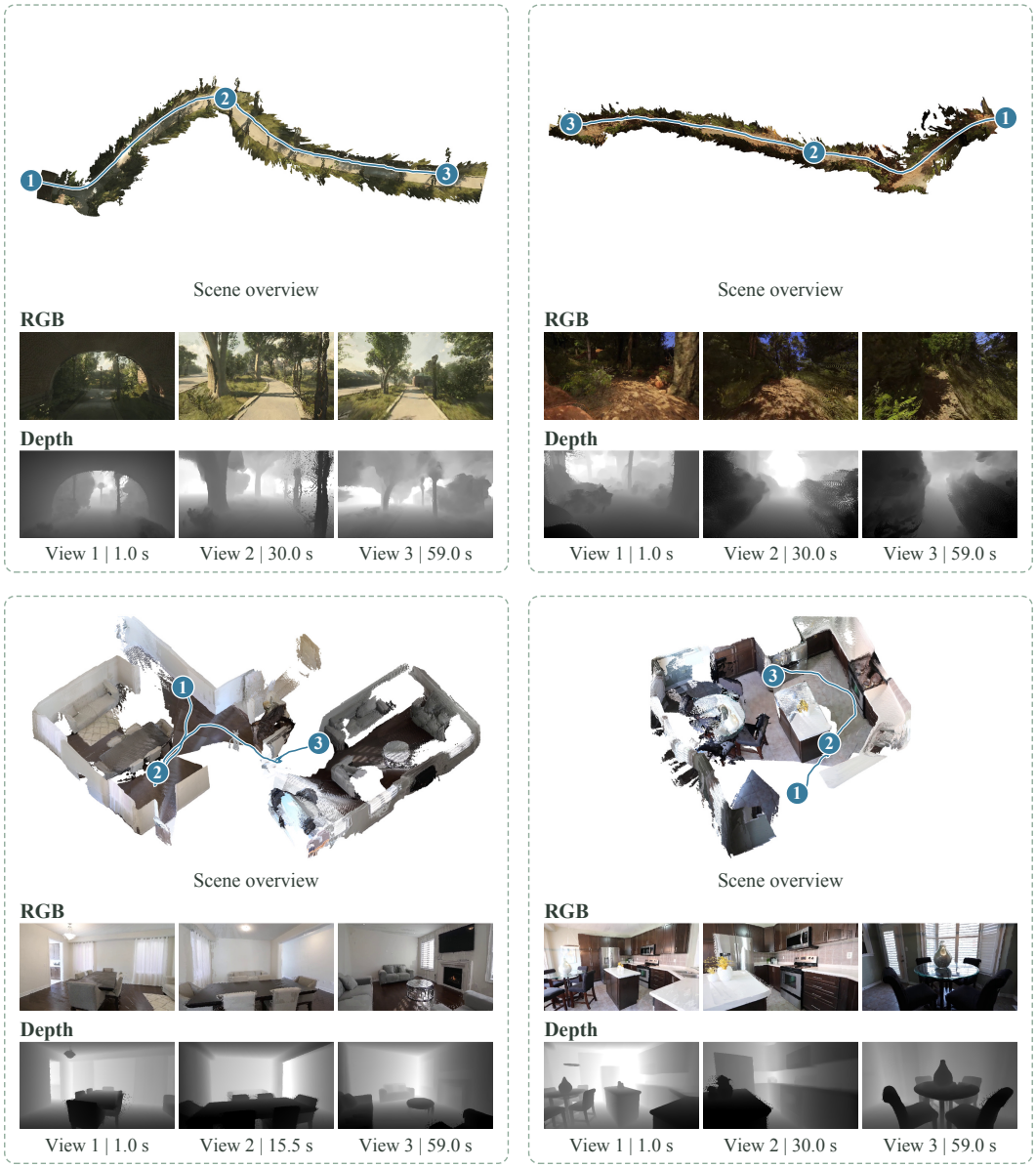


Figure 13: **Streaming generation with PixWorld-Streaming.** Scene overviews and RGB/depth views rendered from predicted 3D Gaussians.