MOSAIC: Generating Consistent, Privacy-Preserving Scenes from Multiple Depth Views in Multi-Room Environments

Zhixuan Liu[1] Haokun Zhu[1] Rui Chen[1] Jonathan Francis[1,2]
Soonmin Hwang[3] Ji Zhang[1] Jean Oh[1]
[1]Carnegie Mellon University [2]Bosch Center for AI [3]Hanyang University

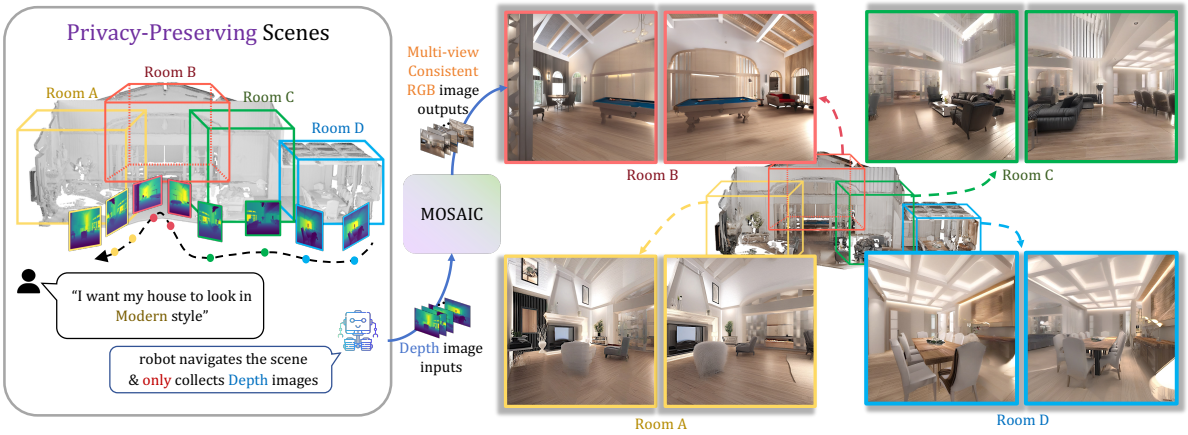

Figure 1. For privacy-preserving scenarios where RGB collection is restricted, MOSAIC generates consistent RGB images from depth data captured along robot paths, guided by text prompts. These outputs further enable 3D reconstruction of multi-room environments.

Abstract

*We introduce a diffusion-based approach for generating privacy-preserving digital twins of multi-room indoor environments from depth images only. Central to our approach is a novel **M**ulti-view **O**verlapped **S**cene **A**lignment with **I**mplicit **C**onsistency (MOSAIC) model that explicitly considers cross-view dependencies within the same scene in the probabilistic sense. MOSAIC operates through a multi-channel inference-time optimization that avoids error accumulation common in sequential or single-room constraints in panorama-based approaches. MOSAIC scales to complex scenes with zero extra training and provably reduces the variance during denoising process when more overlapping views are added, leading to improved generation quality. Experiments show that MOSAIC outperforms state-of-the-art baselines on image fidelity metrics in reconstructing complex multi-room environments. Resources and code are at* https://mosaic-cmubig.github.io.

1. Introduction

Autonomous scene reconstruction is a crucial capability in robotics and computer vision [1, 17, 24, 25], with applications spanning virtual reality, architectural design, and AI-driven simulation. However, existing multi-view 3D reconstruction methods rely heavily on capturing RGB images, which poses significant privacy risks in sensitive environments such as hospitals, factories, elder care facilities, and private residences. Capturing RGB data in these settings can inadvertently expose personal information [32, 47], limiting the practicality of generative AI solutions for real-world deployments where visual privacy must be preserved.

A promising solution for privacy-sensitive environments involves deploying mobile robots that collect only geometric data, such as depth images: this strategy preserves the scene's structural layout while preventing the capture of sensitive texture information. In order to create digital replicas of these scenes, we need algorithms that are capable of generating high-quality, multiview-consistent sequential images that align with ground truth geometry, while maintaining visual coherence across viewpoints.

Various approaches for scene-level multi-view image generation have been proposed, with autoregressive generation techniques [8, 10, 15, 44, 45] representing the current mainstream trend. These approaches leverage powerful inpainting models to iteratively render unseen parts of the scene from sequential camera perspectives. However, these approaches exhibit style drift, where images generated for

initial viewpoints often differ stylistically from those produced when revisiting the same location from different angles, creating visual inconsistencies. Moreover, warp-and-inpainting methods frequently encounter depth misalignment issues, leading to compounding errors over sequential generations [23, 36]. Unlike prior methods that perform diffusion synchronization tasks [2, 11, 18, 20, 22] in constrained settings (limited viewpoint variations), our task requires coherence between viewpoints based on arbitrary trajectory features, extensive perspective changes, and spatial discontinuities. This fundamental distinction makes existing methods ineffective for complex multi-room environments with widely-varying camera positions.

To overcome these limitations, we present **MOSAIC** (**M**ulti-view **O**verlapped **S**cene **A**lignment with **I**mplicit **C**onsistency), a training-free diffusion framework that converts privacy-preserving depth sequences into photorealistic, geometry-aligned RGB views across large, cluttered, multi-room environments. Previous work [37] enforces consistency by fine-tuning a modified architecture on trajectory-specific data, limiting scalability and generalization; MOSAIC instead performs a lightweight multi-view inference-time optimization that scales to unseen environments and enforces cross-view agreement at every denoising step, while preserving the intrinsic knowledge of a large pre-trained model [28]. A depth-weighted projection loss highlights the most reliable viewpoints, and a final pixel-space refinement converts latent-space coherence into precisely aligned images. As more overlapping views are added, MOSAIC reduces variance and tightens image–depth alignment, unlike autoregressive pipelines that compound errors, thus delivering stable reconstructions over arbitrarily long trajectories. Our experiments demonstrate superior performance over existing methods across multiple metrics, establishing a new state-of-the-art for scene level multi-view image generation with geometric priors.

In summary, (1) we introduce MOSAIC, a training-free diffusion pipeline that turns arbitrary depth sequences into photorealistic geometry-aligned RGB views, (2) we devise an inference-time sampler, with depth-weighted projection loss and late pixel refinement, that enforces strict cross-view coherence, (3) we theoretically prove that adding overlapping views monotonically reduces denoising variance, curbing error accumulation, and (4) we demonstrate state-of-the-art results in fidelity, prompt-following, and geometry alignment in cluttered, multi-room environments.

2. Related Work

Diffusion-guided scene-level multi-view generation. Recent advances in 2D diffusion-based generative models [9, 13, 28, 33, 35, 46] have catalyzed significant progress in scene-level multiview image generation, yet most pipelines rely on iterative warping and inpainting for view completion [8, 10, 15, 44, 45], where style drift accumulates across iterations and produces artifacts. A panorama-first strategy mitigates drift by re-projecting a single generated panorama [21, 34, 36, 40], but cannot cope with multi-room layouts or long, unconstrained trajectories. MVDiffusion [37] requires training an additional module for cross-view consistency, which limits the generalization to unseen environments and prompt following fidelity. SceneTex [7] assumes a pre-meshed scene and performs score-distillation texture optimization, requiring $\sim$20 hours per scene and often seeking to a single appearance mode; our training-free DDIM process finishes in few minutes while preserving generation diversity under single text prompt. Recent video-diffusion models provide temporal coherence [4, 14, 43, 48], but lack depth conditioning and break under large viewpoint changes. Our method, on the other hand, explicitly models cross-view consistency by optimizing MOSAIC objective and generalizes to arbitrary viewpoints without imposing layout constraints.

Diffusion Synchronization. Several works have explored synchronization techniques for diffusion models [2, 11, 18, 20, 22] to achieve consistency across multiple outputs. However, existing approaches primarily address specialized cases: panoramic view generation [2, 18, 20], orthogonal transformations [11], or object-level consistency with evenly distributed camera positions and well established UV map [18, 22]. Our work addresses a fundamentally different challenge: maintaining consistent image generation across scene-level camera trajectories with arbitrary viewpoints and significant perspective variations.

Diffusion Inference-Time Optimization. To circumvent the need for large-scale model fine-tuning, several approaches have explored test-time optimization with directional guidance. Diffusion-TTA [29] adapts discriminative models using pre-trained generative diffusion models, updating parameters through gradient backpropagation at inference time. Other methods [26, 27, 38] compute directional controls for single-instance generation (music, images) by calculating directional losses during diffusion and back-propagating to update denoising features. While these methods focus on intra-level control with easily quantifiable directions, our approach is the first to propose multi-channel test-time optimization for cross-view control using self-provided directional guidance, maintaining consistency across arbitrary viewpoints without requiring additional training.

3. Problem Definition

We seek to acquire digital replicas of real-world, multi-room indoor environments in a privacy-preserving manner. To safeguard privacy, we refrain from collecting sensitive real RGB data and capture only geometric structures instead. From these structures, we generate *synthetic*

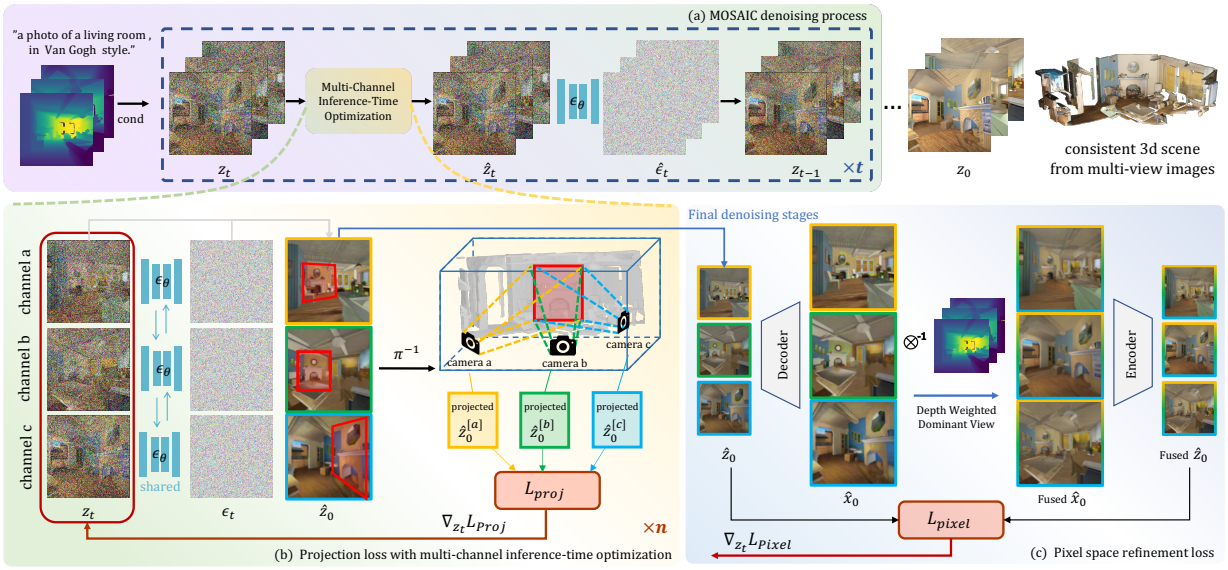

Figure 2. **MOSAIC overview.** (a) *Multi-channel denoising.* Each depth–text–conditioned view is assigned its own latent channel. A shared denoiser iteratively refines the latent set while a multi-channel inference-time optimizer keeps the channels synchronized. (b) *Projection loss.* At every step the predicted clean latents z_0 guided depth-weighted projection loss L_{proj} drives the channels toward a geometry-consistent solution. (c) *Pixel-space refinement.* During the final denoising stages, the pixel-level loss L_{pixel} fuses the views and enforces RGB consistency, yielding photorealistic, cross-view-aligned images that can be reconstructed into a coherent 3-D scene.

RGB data—producing complex, photorealistic scenes that closely align with reality while ensuring no actual environmental details are directly revealed. In this work, we assume that the geometric structures are collected as multiple depth images $\{d^{[1]}, \ldots, d^{[N]}\}$ by mobile robots deployed to real-world scenes. To fully cover multi-room environments, the mobile robot plans proper camera pose sequences to ensure overlapping views. Namely, for each $d^{[i]}$, there must exist some area that is also covered by at least one other view $d^{[j]}$ $(j \neq i)$. From $\{d^{[i]}\}_{i\in[N]}$, we are interested in generating corresponding RGB images $\{x^{[i]}\}_{i\in[N]}$ that are multi-view consistent: overlapping depth views must lead to consistent RGB outputs where they overlap in the 3D scene $\bar{x}$. Namely, for each pixel p in $\bar{x}$ that is covered by multiple depth views, i.e., $I_p := \{i \in [N] \mid p \in d^{[i]}\}$, the RGB output should agree:

$$x^{[i]}[p] = x^{[j]}[p] = \bar{x}[p], \forall i, j \in I_p, i \neq j \tag{1}$$

Then the generated RGB would form a complete 3D multi-room scene $\bar{x}$ once warped by the cameras poses $\{T^{[i]}\}_{i\in[N]}$ (e.g., $\bar{x} = \sum_{i\in[N]} T^{[i]} x^{[i]}$).

4. Method: MOSAIC

4.1. Preliminaries

Denoising Diffusion Implicit Models (DDIM) [35] train a generative model $p_\theta(z_0)$ to approximate a data distribution $q(z_0)$ given samples $z \in \mathcal{Z}$ from q. DDIM considers the following non-Markovian inference model:

$$q_\sigma\left(z_{1:T} \mid z_0\right) := q_\sigma\left(z_T \mid z_0\right) \prod_{t=2}^{T} q_\sigma\left(z_{t-1} \mid z_t, z_0\right) \tag{2}$$

where $\sigma \in \mathbb{R}^T_{\geq 0}$ is a real vector and $z_{1:T}$ are latent variables in $\mathcal{Z}$. The inference procedure $q_\sigma\left(z_{t-1} \mid z_t, z_0\right)$ is parameterized by a decreasing sequence $\alpha_{1:T} \in (0,1]^\top$. To approximate $q(z_0)$, DDIM learns a generative process

$$p_\theta\left(z_{0:T}\right) := p_\theta\left(z_T\right) \prod_{t=1}^{T} p_\theta^{(t)}\left(z_{t-1} \mid z_t\right). \tag{3}$$

Starting from a prior $p_\theta(z_T) = \mathcal{N}(0, I)$, z_t is sampled by $p_\theta^{(t)}(z_{t-1} \mid z_t) = q_\sigma\left(z_{t-1} \mid z_t, f_\theta^{(t)}(z_t)\right)$, where

$$f_\theta^{(t)}\left(z_t\right) := \left(z_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}\left(z_t\right)\right) / \sqrt{\alpha_t} \tag{4}$$

predicts z_0 with a noise prediction model $\epsilon_\theta^{(t)}$. Learning is performed by optimizing the standard variational objective

$$J_\sigma\left(\epsilon_\theta\right) := \mathbb{E}_{z_{0:T}\sim q_\sigma}\left[\log q_\sigma\left(z_{1:T} \mid z_0\right) - \log p_\theta\left(z_{0:T}\right)\right] \tag{5}$$

In this work, we generate RGB images from depth views d, via ControlNet [46]: $\epsilon_\theta^{(t)}(z_t, d)$. In addition, we sample latent RGB representations z from p_θ which can later be decoded into full images via a pre-trained VAE decoder [19].

In short, the generation process we consider is

$$z_0^{[i]} \sim p_\theta(z_0, d^{[i]}),\ x^{[i]} = g(z_0^{[i]}) \text{ for } i \in [N]. \tag{6}$$

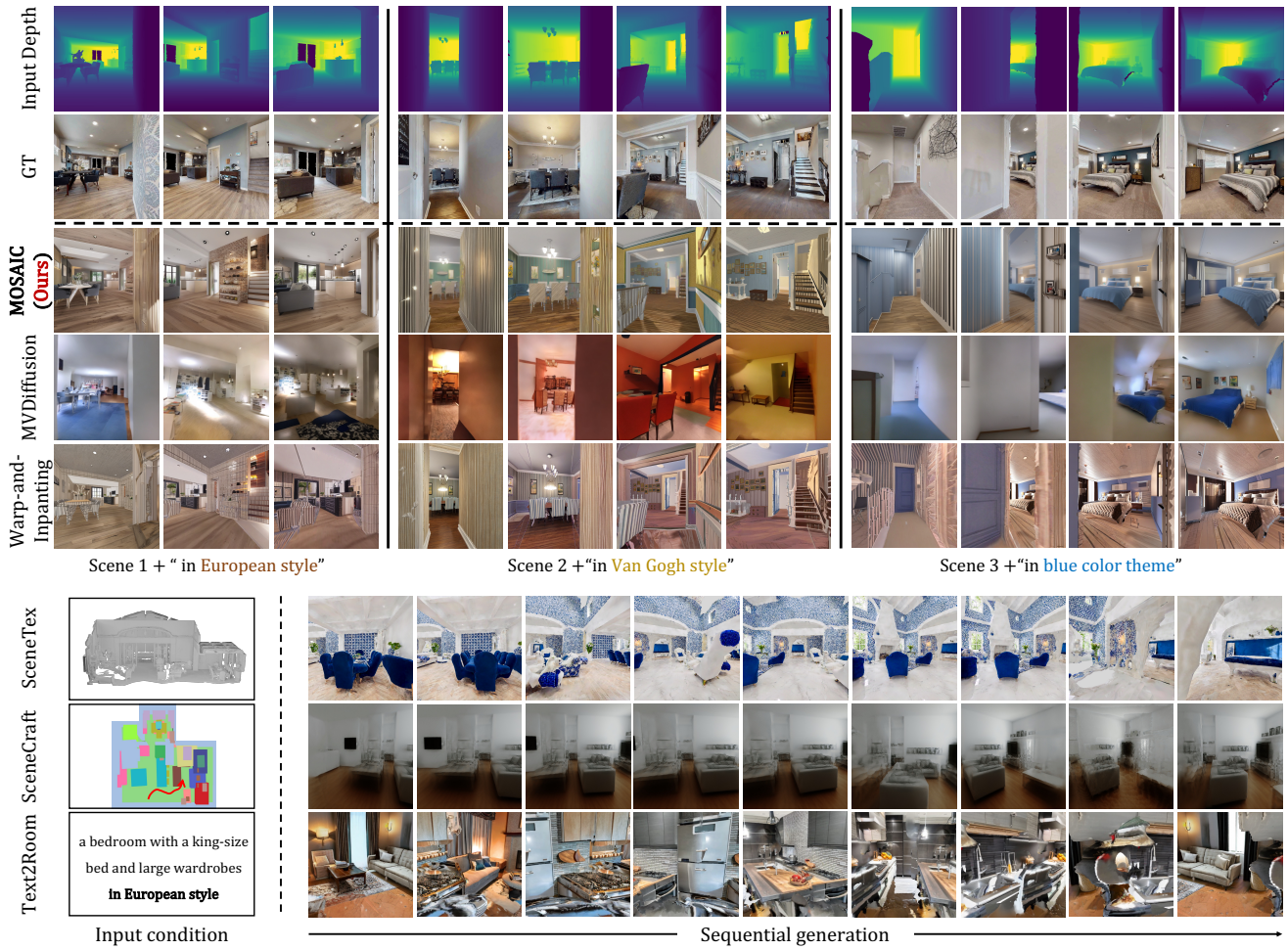

Figure 3. **Qualitative comparison with multi-view baselines.** For three indoor scenes (each conditioned on a style prompt) we show: input depth maps, ground-truth RGB, our MOSAIC result, and two baselines (MVDiffusion, Warp-and-Inpainting) sharing the identity input format. Below, we compare against baselines (SceneTex, SceneCraft, Text2Room) using their native inputs. MOSAIC maintains photorealism, cross-view consistency, and prompt fidelity, whereas competing methods exhibit blur, style drift, or geometric artifacts.

For clarity, we drop the depth dependency $d^{[i]}$ in notations in the rest of this paper. While the generation of a single image x_i can be readily solved via depth-conditioned diffusion models [46], the generation of complete scenes from multiple depth images remains unsolved.

4.2. MOSAIC Formulation

Intuitively, the sampling procedure in Eq. (6) would generate inconsistent RGB views because the generative process p_θ is trained with independent samples $z_0 \sim q(z_0)$. In our case, however, the samples $z_0^{[1:N]}$ are indeed *dependent* since they are taken from the same complete scene $\bar{z}_0$. To fundamentally address the consistency issue, our key insight is to *explicitly model such dependency by incorporating extra projection-based conditionals* $q(z_0^{[i]} \mid \bar{z}_0)$ in the inference procedures. In practice, given the scene in the latent representation $\bar{z}_0$, $z_0^{[i]}$ can be collected by sampling a camera pose with projection function $\pi^{[i]}$ and setting $z_0^{[i]} = \pi^{[i]}(\bar{z}_0)$; subsequent latents $z_{1:T}^{[i]}$ can be sampled as usual. The updated inference procedure for N depth views from the same scene $\bar{z}_0$ can be rewritten from Eq. (2) as:

$$q_\sigma(z_{0:T}^{[1:N]} \mid \bar{z}_0) = \prod_{i\in[N]} q(z_0^{[i]} \mid \bar{z}_0) q_\sigma(z_{1:T}^{[i]} \mid z_0^{[i]}). \quad (7)$$

The corresponding generative process can be extended to

$$p_{\theta,\phi}(z_{0:T}^{[1:N]}, \bar{z}_0) = p_\phi(\bar{z}_0 \mid z_0^{[1:N]}) \prod_{i\in[N]} p_\theta(z_{0:T}^{[i]}), \quad (8)$$

where ϕ parameterizes the reverse process of image projection $q(z_0^{[i]} \mid \bar{z}_0)$. We name the model $p_{\theta,\phi}$ as **M**ulti-view **O**verlapped **S**cene **A**lignment with **I**mplicit **C**onsistency (MOSAIC), which can be learned by minimizing:

$$J_{\sigma,\text{MOSAIC}}(\epsilon_\theta, \phi) := \mathbb{E}_{(z_{0:T}^{[1:N]}, \bar{z}_0)\sim q_\sigma} \left[\log q_\sigma \left(z_{0:T}^{[1:N]} \mid \bar{z}_0 \right) - \log p_{\theta,\phi} \left(z_{0:T}^{[1,N]}, \bar{z}_0 \right) \right]. \quad (9)$$

4.3. Sampling from MOSAIC with Multi-Channel Inference-Time Optimization

While MOSAIC captures the dependency between depth views from the same scene, training directly using Eq. (9) is infeasible for two reasons: (a) optimizing the expectation in Eq. (9) via stochastic batches can be computationally intractable, since sufficient combinations of N views are required even from a single scene. Meanwhile, N should be sufficiently large to effectively cover multi-room environments; and (b) $p_{\theta,\phi}$ trained with some fixed N cannot easily generalize to various depth view numbers which further limits the applicability. Hence, we desire a tractable strategy to sample from MOSAIC that is also compatible with an arbitrary number of depth views N. Factorizing q_σ using Eq. (7) and $p_{\theta,\phi}$ using Eq. (8), we obtain:

$$\begin{aligned} &J_{\sigma,\text{MOSAIC}}\left(\epsilon_\theta, \phi\right) \\ &\equiv \sum_{i\in[N]} \underbrace{\mathbb{E}_{q_\sigma}\left[\log q_\sigma\left(z_{1:T}^{[i]} \mid z_0^{[i]}\right) - \log p_\theta\left(z_{0:T}^{[i]}\right)\right]}_{J_\sigma^{[i]}(\epsilon_\theta) \, Eq.\ (5)} \\ &\quad - \mathbb{E}_{q_\sigma}\left[\log p_\phi(\bar{z}_0 \mid z_0^{[1:N]})\right]. \end{aligned} \tag{10}$$

Note that $p_\phi(\bar{z}_0 \mid z_0^{[1:N]})$ essentially approximates the posterior of the projection operation in forward process and evaluates how well the original scene $\bar{z}_0$ is reconstructed from separate views with some parameterization ϕ. Hence, we can reasonably set p_ϕ to have an inverse exponential dependency on the total re-projection error:

$$p_\phi(\bar{z}_0 \mid z_0^{[1:N]}) \propto \exp(-\textstyle\sum_i \|\bar{z}_0 - f_\phi^{[i]}(z_0^{[i]})\|) \tag{11}$$

where $f_\phi^{[i]}$ projects individual views back to the original and error is considered within the back projected region. Hence the second term in Eq. (10) can be written as $\mathbb{E}_{q_\sigma}\left[\sum_{i\in[N]} \|\bar{z}_0 - f_\phi^{[i]}(z_0[i])\|\right]$. Since $\bar{z}_0$ is in general unavailable, we can instead minimize the total projected cross-view error:

$$J_{\text{Proj}}(\phi) = \mathbb{E}_{q_\sigma}\left[\textstyle\sum_{i,j} \|f_\phi^{[i]}(z_0^{[i]}) - f_\phi^{[j]}(z_0^{[j]})\|\right] \geq 0 \tag{12}$$

Hence, Eq. (10) is equivalent to

$$J_{\sigma,\text{MOSAIC}}\left(\epsilon_\theta, \phi\right) \equiv \textstyle\sum_{i\in[N]} J_\sigma^{[i]}(\epsilon_\theta) + J_{\text{Proj}}(\phi). \tag{13}$$

Although it is impractical to directly train $p_{\theta,\phi}$ using Eq. (13), it is possible to approximate the samples from $p_{\theta,\phi}$ by fine-tuning the output of pre-trained models p_θ, e.g., Latent Diffusion Model (LDM) [33]. We observe that for MOSAIC, $p_{\theta,\phi}$ fits trajectories $z_{1:T}^{[1]}, \ldots, z_{1:T}^{[N]}$ that are mutually dependent at each step $t = 1, \ldots, T$ through $z_0^{1:N}$, while such dependency is missing from LDM p_θ. Furthermore, this dependency is exactly captured by $J_{\text{Proj}}(\phi)$. If $p_{\theta,\phi}^*$ optimally solves $J_{\sigma,\text{MOSAIC}}$ and $z_0^{[1:N]} \sim p_{\theta,\phi}^*$, there must exist some $\bar{z}_0$ and corresponding projections $\pi^{[1:N]}$ such that $z_0^{[i]} = \pi^{[i]}(\bar{z}_0),\ \forall i \in [N]$. Then, $J_{\text{Proj}}(\phi)$ must have been minimized to 0 with $f_{\phi^*}^{[i]} = (\pi^{[i]})^{-1},\ \forall i$. In other words, samples from an ideal MOSAIC model should have zero projected cross-view error in expectation as long as $f_\phi^{[i]}$ exactly matches the inverse projection $(\pi^{[i]})^{-1}$. While it is generally hard to fit an inverse projection, it is unnecessary in our case because the ground truth projections $\pi^{[1:N]}$ are directly available when collecting depth conditions $d^{[1:N]}$[1].

Hence, we can approximately sample from MOSAIC by fine-tuning LDM (i.e., p_θ) output at each denoising step $t-1$ by solving the following *inference-time optimization* through gradient decent:

$$\min_{z_{t-1}^{[1:N]}} L_{\text{Proj}}(z_{t-1}^{[1:N]}) \ \textbf{s.t.}\ z_{t-1,\text{init}}^{[i]} \sim p_\theta^{(t)}(z_{t-1}^{[i]} \mid z_t^{[i]}), \tag{14}$$

where the empirical projection loss is given by

$$L_{\text{Proj}}(z_t^{[1:N]}) = \sum_{i,j} \|(\pi^{[i]})^{-1}(\hat{z}_0^{[i]}) - (\pi^{[j]})^{-1}(\hat{z}_0^{[j]})\|. \tag{15}$$

$\hat{z}_0^{[i]}$ is the estimated $z_0^{[i]}$ from $z_t^{[i]}$, i.e., $\hat{z}_0^{[i]} = f_\theta^{(t)}(z_t^{[i]})$.

Depth Weighted Projection Loss. When calculating L_{Proj}, we observe that depth information provides a natural weighting mechanism for view contributions. For points visible from multiple views, the view with smaller depth value (i.e., closest to the camera) likely provides more accurate RGB information due to higher sampling density and reduced occlusion. To incorporate this insight, we modify our projection loss to weight view contributions based on their relative depth values:

$$L_{\text{Proj}}^{\text{depth}}(z_t^{[1:N]}) = \sum_{i,j} w_{i,j} \cdot \|(\pi^{[i]})^{-1}(\hat{z}_0^{[i]}) - (\pi^{[j]})^{-1}(\hat{z}_0^{[j]})\| \tag{16}$$

where $w_{i,j}$ weights the importance of consistency between views i and j based on their relative depth values, weights are calculated per pixel. To prioritize views with smaller depth values, we employ weighting scheme:

$$w_{i,j} = \frac{\exp(-\alpha \cdot (\pi^{[i]})^{-1} d^{[i]})}{\exp(-\alpha \cdot (\pi^{[i]})^{-1} d^{[i]}) + \exp(-\alpha \cdot (\pi^{[j]})^{-1} d^{[j]})}, \tag{17}$$

where α is a hyperparameter controlling the selectivity of the weighting. With this formulation, when aggregating across all views, the final RGB values for overlapping regions will naturally favor views with minimal depth values.

4.4. Pixel Space Refinement Loss

While our projection loss effectively ensures consistency in latent space, the non-linear nature of the VAE [19] decoder

[1]With mobile robots, $\pi^{[i]}$ can be computed from the camera/robot poses in the world coordinate when taking each depth view $d^{[i]}$.

means that consistency in latent space does not necessarily translate to pixel-space consistency. This discrepancy is particularly pronounced when projection transformations are far from orthogonal. We address this issue through a pixel space refinement process during the final denoising stages. For each view i, we decode its predicted latent into pixel space: $\hat{x}^{[i]} = g(\hat{z}_0^{[i]})$. It is infeasible to calculate L_{Proj} in pixel space, because back-propagation through multi-channel VAE decoder poses computation difficulties. We propose to warp these decoded images between views and compute representations for each view point by depth-weighted fusion of the overlapped views:

$$w_{ij} = \frac{\exp(-\alpha \cdot \pi_{ij} d^{[j]})}{\sum_{k=1}^{N} \exp(-\alpha \cdot \pi_{ik} d^{[k]})}, \; x^{[i]*} = \sum_{j=1}^{N} w_{ij} \cdot \hat{x}^{[j]} \tag{18}$$

where α controls the selectivity of weighting and π_{ij} is the image warping from j view to i view. These optimal pixel-space representations are re-encoded to latent space: $z_0^{[i]*} = \beta \cdot f(x^{[i]*})$, where f is the mapping from pixel to latent. The Pixel Space Refinement Loss is defined as:

$$L_{Pixel} = \sum_{i=1}^{N} \|\hat{z}_0^{[i]} - z_0^{[i]*}\| \tag{19}$$

This directly addresses the latent-to-pixel mapping inconsistencies by forcing latent representations to align with those derived from optimally blended pixel-space images. We utilize the L_{Pixel} at the final denoising stages.

4.5. Properties of MOSAIC

Training-free inference time scale-up. The sampling scheme in Eq. (14) essentially consists of two parts: (a) encouraging multi-view consistency via L_{Proj} which is defined for an arbitrary number of views N; and (b) progressing the reverse process by invoking LDM independently on each view. Hence, unlike full training using Eq. (9), sampling from MOSAIC is agnostic of N, allowing our generation process to easily scale and adapt to larger scenes with more views during inference time, with no extra training.

Variance reduction for pre-trained LDMs. Since LDMs are normally pre-trained in scale, we can roughly view them to have zero prediction error in expectation. However, as will be shown later, LDMs can generate results with varying quality (e.g., in terms of depth preservation) due to the stochasticity of the denoising process, which can cause errors to accumulate until the whole process fails eventually in the warp-inpainting approach. To this regard, one important advantage of MOSAIC is its ability to *stabilize the denoising process given more overlapping views*. This can be seen by analyzing the expected variance of a scene $\bar{z}_0$ given a varying number of views $z_0^{[1:N]}$ that overlap. Consider a partition $z_0^{[1:N]} = \{z_0^{[I_1]}, z_0^{[I_2]}\}$ where $I_1 \cup I_2 = \{1, \ldots, N\}$,

$$\begin{aligned}\Sigma(\bar{z}_0 \mid z_0^{[I_1]}) &= \mathbb{E}[\Sigma(\bar{z}_0 \mid z_0^{[I_1]}, z_0^{[I_2]})] + \underbrace{\Sigma(\mathbb{E}[\bar{z}_0 \mid z_0^{[I_1]}, z_0^{[I_2]}])}_{\Sigma \text{ explained by } z_0^{[I_2]}} \\ &\succeq \mathbb{E}[\Sigma(\bar{z}_0 \mid z_0^{[I_1]}, z_0^{[I_2]})]. \end{aligned} \tag{20}$$

Taking expectation over $z_0^{[I_1]}$, we have

$$\mathbb{E}[\Sigma(\bar{z}_0 \mid z_0^{[1:N]})] \preceq \mathbb{E}[\Sigma(\bar{z}_0 \mid z_0^{[I_1]})]. \tag{21}$$

Namely, when conditioned on more views $z_0^{[I_2]}$, the variance of the final output $\bar{z}_0$ reduces in expectation by an amount positively related to the overlap between $\bar{z}_0$ and $z_0^{[I_2]}$[2]. The same applies when $z_0^{[1:N]}$ is replaced by estimates from intermediate $z_t^{[1:N]}$, hence the variance reduction happens during the entire denoising process.

5. Experimental Setup

Autonomous Data Collection. To construct our test set, we deployed an indoor-navigation robot that autonomously explored diverse indoor environments. During each trajectory, the robot captured depth maps with its on-board sensors and logged precise camera poses, producing trajectory-aligned inputs compatible with prior work [15, 23, 37, 42]. We gathered data from 16HM3D [31] and 5 MP3D [6] scenes, for each scene we collected around 10 independent trajectories; detailed trajectory statistics appear in the Appendix. For captioning, we rendered each depth map to a grayscale image and queried the vision–language model `GPT-4o` [5] to generate concise, scene-aware descriptions, yielding 2011 *(depth, pose, caption)* triplets in total.

Baselines. We compare our method with state-of-the-art multi-view generators that leverage geometric priors: MVDiffusion [37], Warp-Inpaint [23], SceneTex [7], and SceneCraft [42]. Although our depth maps can be converted to meshes required by SceneTex, running SceneTex itself to synthesize multi-view images for a single scene takes more than 20 hours, making a full scenes evaluation impractical. Therefore, we run SceneTex on an 8-scene subset and report its results separately for fairness. SceneCraft expects semantic layouts with per-object depth; we derive the necessary bounding-box layouts from our dataset. We additionally assess semantic fidelity against Text2Room [15], a specialist indoor-scene generator.

Evaluation metrics. Our task seeks images that (i) are geometrically consistent across overlapping views, (ii) exhibit high perceptual quality and stylistic coherence, and (iii) faithfully reflect the text prompt. Following MVDiffusion [37], we measure geometric consistency with Warped

[2]The amount reduced represents the variability of $\bar{z}_0$ due to changes in $z_0^{[I_2]}$, which is higher when $z_0^{[I_2]}$ covers more area in $\bar{z}_0$.

Figure 4. **Scene-level reconstruction.** Fusing the multi-view images generated by MOSAIC with a standard TSDF pipeline produces coherent, textured meshes across diverse indoor rooms.

Table 1. Quantitative evaluation with Kernel Inception Distance (KID), CLIPIQA$^+$, CLIPScore (CS) and CLIPConsistency(CC).

Method	KID ↓	CS ↑	CIQA ↑	CC ↑
Warp-Inpaint [23]	0.04646	0.6849	0.6517	29.93
MVDiffusion [37]	0.03640	0.7016	0.6025	29.41
SceneCraft [42]	0.07697	0.7011	0.4531	27.68
Text2Room [15]	0.08594	0.6618	0.4388	27.34
MOSAIC (ours)	**0.03391**	**0.7166**	**0.6526**	**30.85**
SceneTex [7] - small set	0.10536	0.7045	0.6308	28.5740
MOSAIC- small set (ours)	**0.06547**	**0.7141**	**0.7322**	**29.5043**

PSNR [16]—the PSNR after warping one view into another using ground-truth geometry—and its normalized variant against GT, Warped Ratio, where 1.0 indicates perfect alignment. Image quality is assessed with Kernel Inception Distance (KID) [3] and CLIPIQA$^+$ [39]; prompt adherence is quantified by CLIPScore [12] and CLIPConsistency [30].

6. Results

6.1. Evaluation Against Baselines

Qualitative Comparisons. As illustrated in Fig. 3, MOSAIC generates scenes that are simultaneously more photorealistic and more cross-view–coherent than all competing methods when provided with the same depth inputs. MVDiffusion [37] exhibits clear view-to-view inconsistencies and is unable to accommodate stylistic prompts such as "in Van Gogh style." The warp-and-inpainting pipeline [23] produces plausible first frames, yet progressive style drift and error accumulation degrade subsequent views. SceneCraft [42], hindered by its NeRF backbone, fails to reconstruct complex layouts and introduces pronounced blur. Text2Room [15], which also relies on warp-and-inpainting, shows patchwork seams and depth misalignment. SceneTex [7] requires over 20 hours to process a single scene, whereas MOSAIC finishes in just a few minutes. Moreover, SceneTex's score distilling optimization often converges to a single appearance mode, yielding nearly identical textures across seeds; our DDIM sampler preserves diversity under identical prompts. Overall, MOSAIC preserves depth fidelity and stylistic intent across viewpoints and eliminates boundary artifacts through its depth-weighted projection mechanism.

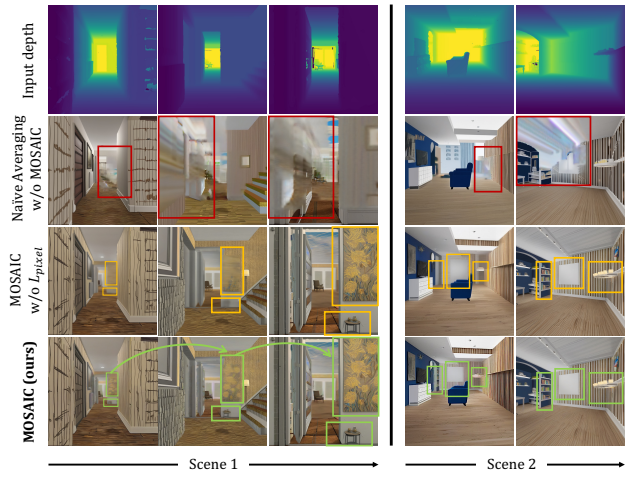

Figure 5. **Qualitative ablation.** Columns show matched views for two scenes; rows compare naïve averaging, MOSAIC without L_{pixel}, and full MOSAIC. Boxes indicate identity objects across different viewpoints along generation. Red: blur/ghosts artifacts; orange: inconsistent texture drift; green: full model corrects both.

Table 2. Cross-view geometric consistency analysis.

Method	Warped PSNR ↑	Warped Ratio ↑
GT	25.45	1.00
MVDiffusion [37]	13.58	0.53
Warp-Inpaint [23]	22.00	0.86
MOSAIC (ours)	**25.30**	**0.99**

Quantitative Comparisons. Tab. 1 quantitatively compares our approach with Warp-and-Inpainting [23], MVDiffusion [37], SceneCraft [42], Text2Room [15], and SceneTex [7]. MOSAIC achieves the lowest KID (↓0.0391) and the highest CS (↑0.7166), CIQA (0.6526), and CC (30.85), outperforming every baseline across all perceptual metrics; it also surpasses SceneTex on the 8-scene subset. Tab. 2 further highlights our superior geometric consistency, recording a Warped PSNR of 25.30 and a Warped Ratio of 0.99—virtually matching ground truth (25.45, 1.00) and eclipsing all alternatives. Although MVDiffusion attains a competitive KID of 0.0364, its Warped PSNR is only 13.58, underscoring its limited multi-view consistency in 3D. In short, MOSAIC delivers state-of-the-art performance in both image fidelity and cross-view alignment.

6.2. Ablation Study

Effect of MOSAIC Objective and Multi-Channel Test-Time Optimization. We build a naive ablation by extend-

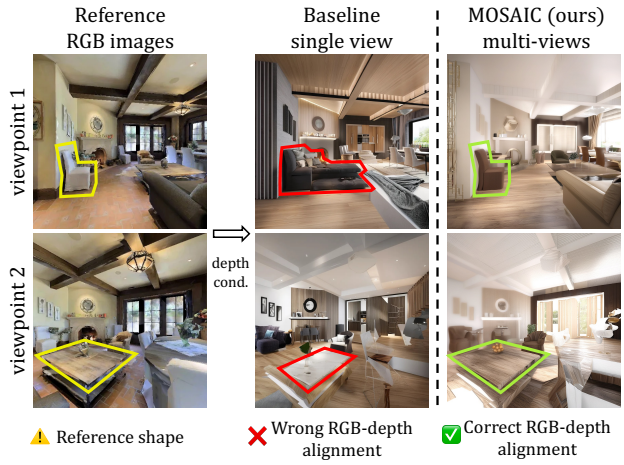

Figure 6. **RGB–depth alignment** Left: reference images at two viewpoints (yellow). Middle: single-view baseline producing geometric artifacts (red). Right: MOSAIC conditioned on multiple depth maps recovers accurate geometry (green).

ing [18] to scene-level multi-view image generation, averaging each denoising latent prediction z_0 without our multi-channel inference time optimization for MOSAIC objectives. As reported in Tab. 3, this variant records the worst KID and the weakest consistency scores. Fig. 5 (second row) further exposes severe blur and cross-view misalignment (red boxes). Conversely, the full MOSAIC model achieves the best KID, the strongest consistency, and visually tight alignment across views (green boxes), confirming that MOSAIC is critical for large viewpoint changes naive averaging collapses both appearance and geometry.

Effect of Pixel Space Refinement. Adding the pixel-space loss L_{pixel} markedly improves performance (Tab. 3), lowering KID and pushing consistency scores close to ground truth. In Fig. 5, boxes track identical objects across different viewpoints: without L_{pixel} (third row) the flower vanishes, the stool disappears, and decorative shapes deform (orange boxes). With L_{pixel} (fourth row) textures remain coherent from different viewpoints (green boxes), demonstrating the benefit of pixel-level supervision.

Effect of Trajectory Key Frame Selection. Tab. 3 contrasts our region-coverage sampling with uniform trajectory sampling. Although overall image quality and consistency are similar, text–image alignment degrades under uniform sampling. Uniformly selected frames often face blank walls, leaving the VLM with insufficient semantic cues; it is also likely to cause pretrained ControlNet [46] branches to hallucinate textures. Our key-frame strategy strikes a balance between information-rich viewpoints and sufficient overlap, yielding more accurate text-conditioned results.

Multi-Views for better RGB-Depth Alignment. Traditional scene-level generation approaches employ warp-and-inpainting strategies that suffer from error accumulation—initial RGB-depth misalignments compound through

Table 3. Ablation: Image Quality and 3D Consistency

	Image Quality				Geometry Consistency	
	KID ↓	CS ↑	CIQA ↑	CC ↑	PSNR ↑	Ratio ↑
w/o test-time optimization	0.06715	**0.7285**	0.5725	31.08	15.74	0.6185
w/o key frame selection	0.03920	0.6746	0.6733	29.25	24.67	0.9692
w/o L_{Pixel}	0.04273	0.7249	**0.6797**	**31.26**	23.02	0.9045
MOSAIC (ours full)	**0.03391**	0.7166	0.6526	30.85	**25.30**	**0.9940**

Table 4. Assessing the effect of number of depth views, NMSE stands for normalized MSE.

Views	NMSE ↓	CS ↑	KID ↓
single view	0.018630	0.6874	0.06123
two views	0.013453	0.6883	**0.06022**
three views	0.011961	0.6986	0.06709
four views	**0.011269**	**0.7100**	0.07833

sequential generation. By contrast, Eq. 21 shows that conditioning on multiple depth views analytically reduces variance, producing geometry closer to ground truth. To validate this, we select key frames whose viewpoints overlap by more than 50% of their pixels and vary the number N of such views during generation. Tab. 4 corroborates this: the normalized MSE between generated depths (estimated with Depth-Anything [41]) and ground-truth depths decreases monotonically as the number of views increases. Fig. 6 visualizes the trend: single-view conditioning (red boxes) hallucinates structures that diverge from ground truth (yellow boxes), whereas multi-view conditioning (green boxes) preserves geometric accuracy. Our analysis further reveals an inflection point in the quality-viewpoint curve, with text-image alignment steadily improving across viewpoints, while perceptual quality peaks at 2-3 highly overlapping views. This finding suggests an optimal operating point that balances computational efficiency with generation quality.

7. Conclusion

We proposed MOSAIC, the first training-free multi-view consistent image generation pipeline that operates at scene level. MOSAIC handles arbitrary numbers of views along any trajectory, adapting to extensive camera viewpoint changes. It employs a depth-weighted projection loss and a pixel-space refinement process to maintain visual coherence across complex multi-room environments, making it well-suited for privacy-sensitive applications. We developed a novel multi-channel inference-time optimization procedure that minimizes cross-view projection errors, which we mathematically prove to be equivalent to the ideal learning objective. We showed theoretically and demonstrated experimentally that by intelligently fusing multi-view information, MOSAIC substantially improves alignment between generated RGB images and ground truth depth and significantly outperforms state-of-the-art methods—addressing the error accumulation issues of current autoregressive multi-view image generation pipelines.

Acknowledgements

We are grateful to Yanbo Xu, Zhipeng Bao, Yifan Pu, and Zongtai Li for their helpful comments and discussion. The project was partly supported by NSF IIS-2112633.

References

[1] Sunday Amatare, Gaurav Singh, Raul Shakya, Aavash Kharel, Ahmed Alkhateeb, and Debashri Roy. Dt-radar: Digital twin assisted robot navigation using differential ray-tracing, 2024. 1

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 6

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017. 6

[7] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors, 2023. 2, 6, 7

[8] 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. 1, 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2

[10] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36:39897–39914, 2023. 1, 2

[11] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models, 2024. 2

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[15] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 1, 2, 6, 7

[16] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. 7

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1

[18] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions, 2024. 2, 8

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 3, 5

[20] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions, 2023. 2

[21] Wenrui Li, Fucheng Cai, Yapeng Mi, Zhe Yang, Wangmeng Zuo, Xingtao Wang, and Xiaopeng Fan. Scenedreamer360: Text-driven 3d-consistent scene generation with panoramic gaussian splatting, 2024. 2

[22] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion, 2023. 2

[23] Mikonvergence. Controlnetinpaint, 2023. Accessed: March 2025. 2, 6, 7

[24] 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, 2020. 1

[25] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020. 1

[26] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Ditto: Diffusion inference-time t-optimization for music generation, 2024. 2

[27] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing, 2023. 2

[28] 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, 2023. 2

[29] Mihir Prabhudesai, Tsung-Wei Ke, Alexander C. Li, Deepak Pathak, and Katerina Fragkiadaki. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback, 2023. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[31] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. 6

[32] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living, 2021. 1

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[34] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. Controlroom3d: Room generation using semantic proxy rooms, 2023. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3

[36] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Yang Zhao. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture, 2023. 2

[37] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 6, 7

[38] Yun-Yun Tsai, Fu-Chen Chen, Albert Y. C. Chen, Junfeng Yang, Che-Chun Su, Min Sun, and Cheng-Hao Kuo. Gda: Generalized diffusion for robust test-time adaptation, 2024. 2

[39] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 7

[40] Qi Wang, Ruijie Lu, Xudong Xu, Jingbo Wang, Michael Yu Wang, Bo Dai, Gang Zeng, and Dan Xu. Roomtex: Texturing compositional indoor scenes via iterative inpainting, 2024. 2

[41] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. 8

[42] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems*, 37: 82060–82084, 2025. 6, 7

[43] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024. 2

[44] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image, 2024. 1, 2

[45] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. Wonderjourney: Going from anywhere to everywhere, 2024. 1, 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 3, 4, 8

[47] Yiqin Zhao, Sheng Wei, and Tian Guo. Privacy-preserving reflection rendering for augmented reality, 2022. 1

[48] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2