VideoFrom3D: 3D Scene Video Generation via Complementary Image and Video Diffusion Models

GEONUNG KIM, POSTECH, Republic of Korea
JANGHYEOK HAN, POSTECH, Republic of Korea
SUNGHYUN CHO, POSTECH, Republic of Korea

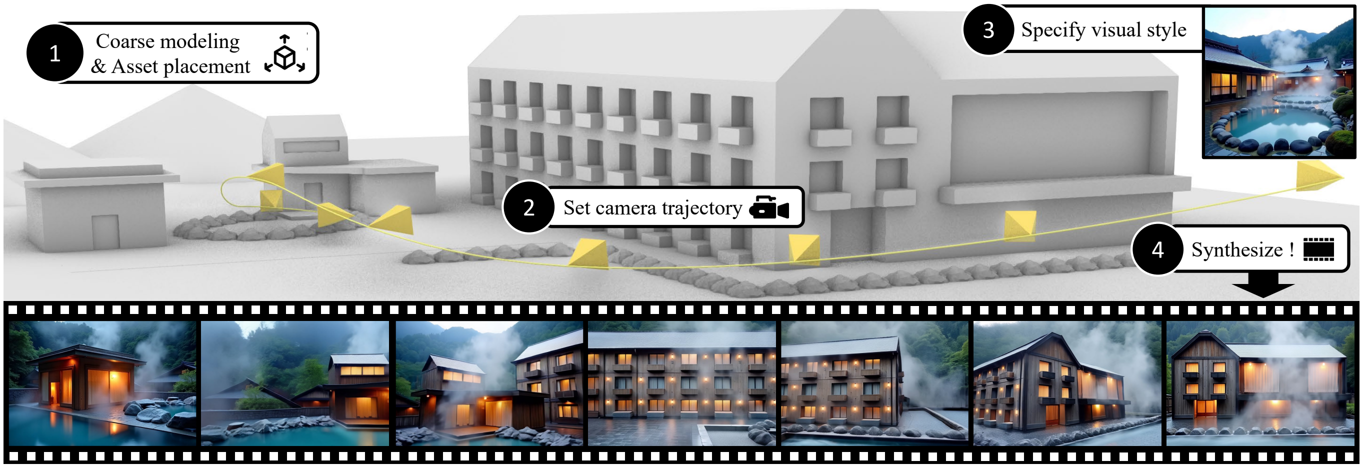


Fig. 1. Overall framework. (1) Users construct a scene using coarse geometry or 3D assets. (2) A camera trajectory and (3) a reference image are provided. (4) The framework then generates a high-quality video reflecting the specified style, structure, and camera motion. The synthesized video sequence shows consistent, high-quality visuals that reflect the input geometry and reference style, including challenging visual elements such as rising steam.

In this paper, we propose VideoFrom3D, a novel framework for synthesizing high-quality 3D scene videos from coarse geometry, a camera trajectory, and a reference image. Our approach streamlines the 3D graphic design workflow, enabling flexible design exploration and rapid production of deliverables. A straightforward approach to synthesizing a video from coarse geometry might condition a video diffusion model on geometric structure. However, existing video diffusion models struggle to generate high-fidelity results for complex scenes due to the difficulty of jointly modeling visual quality, motion, and temporal consistency. To address this, we propose a generative framework that leverages the complementary strengths of image and video diffusion models. Specifically, our framework consists of a Sparse Anchor-view Generation (SAG) and a Geometry-guided Generative Inbetweening (GGI) module. The SAG module generates high-quality, cross-view consistent anchor views using an image diffusion model, aided by Sparse Appearance-guided Sampling. Building on these anchor views, GGI module faithfully interpolates intermediate frames using a video diffusion model, enhanced by flow-based camera control and structural guidance. Notably, both modules operate without any paired dataset of 3D scene models and natural images, which is extremely difficult to obtain. Comprehensive experiments show that our method produces high-quality, style-consistent scene videos under diverse and challenging scenarios, outperforming simple and extended baselines. Code is available at github.com/KIMGEONUNG/VideoFrom3D.

CCS Concepts: • **Computing methodologies** → **Computer graphics**.

ACM Reference Format:
Geonung Kim, Janghyeok Han, and Sunghyun Cho. 2025. VideoFrom3D: 3D Scene Video Generation via Complementary Image and Video Diffusion Models. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3757377.3763871

1 Introduction

3D graphic design refers to the process of creating visually compelling three-dimensional representations for communication, simulation, or artistic purposes. It serves diverse purposes across domains, including conveying design intent in architecture, building immersive worlds in games, generating photorealistic effects in film, and enabling real-time interaction in VR and metaverse applications. Across these domains, the underlying production process typically follows a common sequence of stages. The design workflow usually begins with a concept development phase, where rough visual ideas, a preliminary 3D scene layout, and a camera trajectory are established. This is followed by detailed production steps including modeling, texturing, and lighting, with a focus on the regions that will be visible in the final render. The process culminates in the rendering stage, which produces the final visual output such as images or videos [Bettis 2005; Hamdani and Barreto 2023].

Authors' Contact Information: Geonung Kim, POSTECH, Republic of Korea, k2woong92@postech.ac.kr; Janghyeok Han, POSTECH, Republic of Korea, hjh9902@postech.ac.kr; Sunghyun Cho, POSTECH, Republic of Korea, s.cho@postech.ac.kr.

SA Conference Papers '25, Hong Kong, Hong Kong
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763871

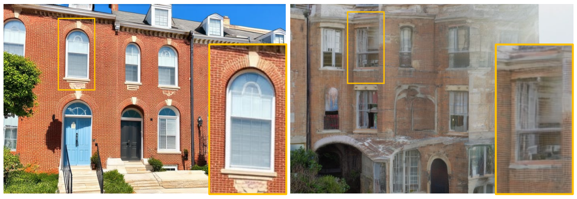

(a) StableDiffusion3-T2I (2B) (b) CogVideoX-T2V (5B)

Fig. 2. Comparison of outputs from image and video diffusion models, conditioned only on a text prompt. Model sizes are noted in parentheses.

In practice, however, this workflow does not proceed in a single pass but involves repeated iterations across stages. Specifically, to refine the design, designers often receive feedback after the rendering stage from clients or collaborators, and return to earlier stages to revise the work. One major challenge is that even minor changes in a single component can require extensive adjustments in multiple stages of the workflow [Bettis 2005; Lord 2024]. For example, when the intended camera trajectory or scene composition is modified, previously detailed modeling, texturing, and lighting may all require updates, as the regions visible to the camera also change. Similarly, changes in the visual concept often necessitate broad adjustments in both texturing and lighting. Because each stage is time-consuming and requires a high level of expertise, even minor revisions can result in significant increases in production cost.

In this paper, we propose VideoFrom3D, a novel framework for synthesizing high-quality 3D scene videos from coarse geometry. By leveraging generative models, our approach streamlines the 3D graphic design pipeline, offering a faster and more flexible alternative to the traditional, labor-intensive workflow. Fig. 1 illustrates our framework step by step. Firstly, (1) a user constructs a scene by modeling coarse geometry or by assembling a scene using pre-existing 3D assets. Then, (2) a camera trajectory and (3) a reference image representing the desired visual concept are provided. Given these inputs, (4) our framework synthesizes a high-quality video that reflects the specified style, structure, and camera motion through a generative process. This addresses the aforementioned inefficiencies in two ways. First, by relying on coarse modeling and asset placement instead of detailed modeling, the framework allows flexible adaptation to changes in scene layout and concept. Second, the generative synthesis strategy based on a reference style enables efficient adaptation to changes in visual style or camera trajectory without redoing time-consuming texturing and lighting. Consequently, our framework can be employed in early-stage design development by enabling rapid iteration and visual exploration prior to labor-intensive asset production. Alternatively, for visualization-only purposes, the generated output can serve directly as the final deliverable.

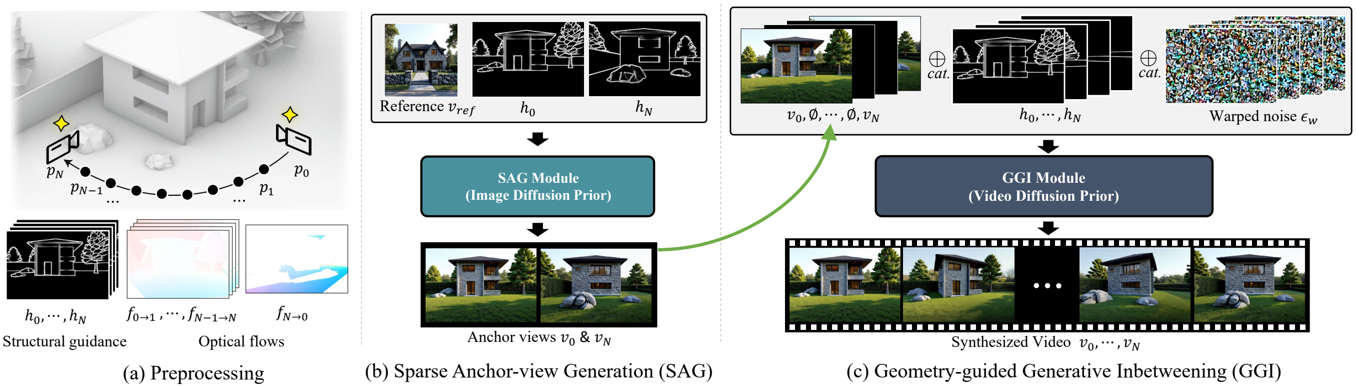
Based on the recent success of video diffusion models in 3D scene generation, a naïve solution to this problem would be to condition a video diffusion model on geometric information, such as with a depth-based ControlNet [Zhang et al. 2023]. However, this naïve solution faces a core limitation: *video diffusion models are fundamentally limited in handling complex scenes compared to image diffusion models*. Fig. 2 compares the outputs of image and video diffusion models on a complex scene. While the image model produces realistic building details, the video models generate distorted structures with lower visual quality, despite having far more parameters. This limitation is also evident in the quantitative comparison in Table 1. The primary reason lies in the inherent challenges of video synthesis. Unlike image diffusion models, which focus exclusively on generating high-quality still frames, video models must simultaneously learn to synthesize individual frames, ensure realistic motion, and maintain temporal coherence across video frames. This added complexity makes it harder for video diffusion models to match the individual image quality achieved by image diffusion models.

Table 1. Image aesthetics (CLIP-A) and quality (MUSIQ) are compared across 1,000 generated samples (parameter size in parentheses). Prompts are auto-generated by GPT to describe complex outdoor scenes. For video models, only the first frame of each video is evaluated.

	StableDiffusion3 (2B)	CogvideoX (2B)	CogvideoX (5B)
CLIP-A↑	**5.942**	5.119	5.144
MUSIQ↑	**67.04**	56.36	58.20

To address this issue, we leverage the complementary strengths of image and video diffusion models. Specifically, image diffusion models are highly effective at generating high-quality frames with fine spatial detail, while video diffusion models excel at maintaining temporal consistency across sequences. Building on this insight, our key idea is to first generate a set of high-quality, multi-view-consistent anchor frames using an image diffusion model, and then interpolate the anchor frames using a video diffusion model to synthesize temporally coherent intermediate frames, instead of synthesizing complex scenes from scratch. To realize this, we introduce two key modules: a Sparse Anchor-view Generation (SAG) module and a Geometry-guided Generative Inbetweening (GGI) module. The SAG module produces high-quality, multi-view-consistent anchor views using an image diffusion model. A major challenge at this stage is preserving multi-view consistency. To overcome this, we introduce Sparse Appearance-guided Sampling, which adopts a distribution alignment strategy and leverages appearance guidance from a warped adjacent view to generate consistent results. Interpolating these anchor views, the GGI module generates consistent intermediate frames using a video diffusion model. To ensure natural interpolation and precise trajectory alignment, we incorporate flow-based camera control and structural guidance into the GGI module. Notably, our modules achieve high visual quality without relying on any paired dataset of 3D scene models and natural images, which are typically unavailable in practice.

A potential alternative to generate a video from the input geometry is to synthesize textures for a mesh model using recent texturing techniques [Chen et al. 2023b; Richardson et al. 2023; Yu et al. 2024b; Zeng et al. 2024; Zhang et al. 2024], and then render the result. However, our method differs in two important ways. First, texturing-based approaches require detailed geometry to produce natural-looking results. When applied to coarse geometry, this often leads to visual artifacts, e.g., flowers or grass appearing unnaturally flattened onto planar ground surfaces. Second, texture maps are inherently static and cannot capture dynamic, view-dependent effects such as reflections, flickering flames, or flowing streams. By directly



(a) Preprocessing (b) Sparse Anchor-view Generation (SAG) (c) Geometry-guided Generative Inbetweening (GGI)

Fig. 3. Overall pipeline. (a) Preprocessing extracts structural edges and optical flows from geometry and camera trajectory. (b) SAG module generates high-quality anchor views v_0 and v_N using an image diffusion prior. (c) GGI module interpolates intermediate frames with a video diffusion prior.

generating video, our method naturally models such variations. These advantages come with certain limitations: our framework does not support real-time navigation or enforce pixel-level consistency across views. Nonetheless, it offers a compelling alternative for fast, stylized scene video generation from minimal input.

Our main contributions can be summarized as follows:

- We propose VideoFrom3D, a novel framework that synthesizes a high-quality 3D scene video given coarse geometry, a camera trajectory, and a reference image.
- We propose a two-stage approach that leverages the complementary strengths of image and video diffusion models, where the SAG module uses image diffusion for anchor view generation, and the GGI module applies video diffusion to interpolate the anchor views.
- Extensive experiments show that our method robustly synthesizes high-fidelity videos under diverse and challenging scenarios, outperforming naïve and extended baselines.

2 Related Work

Geometry-guided Video Generation. Structure-conditioned video generation methods [Alhaija et al. 2025; Jiang et al. 2025] provide a simple baseline for our problem by conditioning video outputs on depth or edge maps rendered from the input geometry. However, their results often suffer from poor visual quality in complex scenes, due to the limited prior of video diffusion models. Another line of work explores 3D generation methods conditioned on geometry [Chen et al. 2024a, 2023a; Dong et al. 2024; Metzer et al. 2023; Shi et al. 2023; Wang et al. 2024], whose outputs can be rendered into videos. Yet, they are limited to object-level generation. In contrast, Urban Architect [Lu et al. 2024] and ControlRoom3D [Schult et al. 2024] propose scene-scale 3D generation methods conditioned on semantic proxy geometry. However, the former produces blurry results due to the limited guidance from the SDS-based image prior, while the latter is limited to rectangular room layouts. To the best of our knowledge, we are the first to enable geometry-guided generation of high-quality, large-scale, and versatile 3D scenes.

Few-shot 3D Reconstruction. Few-shot 3D reconstruction recovers a 3D scene from a small number of input views. Similar to our GGI module, recent approaches interpolate keyframes across sparse input views using video diffusion models. To this end, most approaches first construct an intermediate point-based representation from sparse views, which is then used to condition the video diffusion model for intermediate frame synthesis [Cao et al. 2025; Chen et al. 2024b; Liu et al. 2024; Ma et al. 2024; Ren et al. 2025; Yu et al. 2024a]. For example, MVSplat360 [Chen et al. 2024b] builds a coarse 3D Gaussian Splatting [Kerbl et al. 2023] via feedforward prediction to guide video generation. However, such methods often fail to construct reliable structure under wide keyframe baselines, resulting in severe artifacts. In other lines of work, LVSM [Jin et al. 2024] performs regression-based interpolation but lacks strong generative priors, leading to failure under complex transitions. SEVA [Zhou et al. 2025] leverages multi-view diffusion models with Plucker embeddings to condition the camera trajectory. However, this explicit pose representation suffers from scale ambiguity, making it difficult to follow the intended path, and also causes temporal flickering inherent to multi-view diffusion. In contrast, our GGI module leverages a video diffusion prior and structure-aware conditioning, enabling accurate, smooth, and robust interpolation.

3 Method

We define the VideoFrom3D task as follows. Let $\mathbb{M}$ be a 3D mesh model, and v_{ref} be a reference image representing the desired style. Additionally, let $\mathbb{P} = \{p_0, \ldots, p_N\}$ be a camera trajectory, where p_i is the camera pose at the i-th frame, such that $0 \leq i \leq N$. Given these inputs, the goal of VideoFrom3D is to generate a sequence of images $\mathbb{V} = \{v_0, \ldots, v_N\}$ that are synthesized from the corresponding camera poses in $\mathbb{P}$, faithfully reflect the geometry $\mathbb{M}$, and are consistently stylized according to v_{ref}.

Fig. 3 illustrates the overall pipeline of VideoFrom3D, which consists of three stages. The preprocessing stage extracts structural guidance $\{h_0, \ldots, h_N\}$ and optical flows $\{f_{0\to1}, \ldots, f_{N-1\to N}, f_{N\to0}\}$ from an input 3D model $\mathbb{M}$ and a camera trajectory $\mathbb{P}$, where $f_{i\to j}$ denotes the optical flow from view i to view j. The structural guidance constrains the SAG and GGI modules to synthesize images

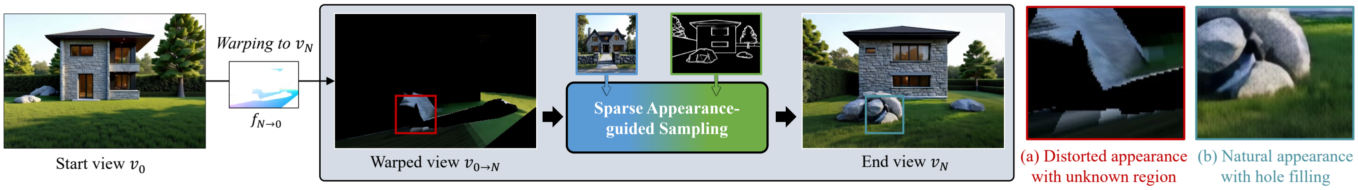

Warping to v_N
$f_{N\to 0}$
Start view v_0
Warped view $v_{0\to N}$
Sparse Appearance-guided Sampling
End view v_N
(a) Distorted appearance with unknown region
(b) Natural appearance with hole filling

Fig. 4. To generate a multi-view coherent v_N from v_0, Sparse Appearance-guided sampling uses the distorted appearance of the warped image as guidance during sampling, achieving the successful generation of a coherent and high-quality v_N.

accurately reflecting the geometry of $\mathbb{M}$. $f_{N\to 0}$ is used by the SAG module to enforce cross-view consistency between v_0 and v_N, while $\{f_{0\to 1}, \ldots, f_{N-1\to N}\}$ is used by the GGI module to provide guidance on the camera trajectory. In the next stage, the SAG module generates high-quality anchor views, v_0 and v_N, using an image diffusion prior. The generation of v_N is conditioned on v_0 for cross-view consistency. Finally, the GGI module interpolates between the anchor views using a video diffusion prior, producing the full sequence $\mathbb{V}$. To support longer camera trajectories, the pipeline can be iteratively applied by treating v_N as the next starting frame v_0.

In the following, we describe each stage of our method and the training strategy for the GGI module in detail.

3.1 Preprocessing

To compute the optical flow $f_{i\to j}$, we first backproject a coordinate-encoded color map from p_j onto the 3D mesh, where each pixel encodes its own image-space coordinate. We then reproject the coordinate-encoded mesh onto the image plane of p_i, allowing us to establish dense correspondences based on color information, which are then used to derive the optical flow.

As structural guidance h_i, we employ a 2D edge map, which represents the shapes of the input 3D model $\mathbb{M}$ as projected at camera pose p_i. Specifically, we empirically select four types of geometry-based edges: silhouette, crease, object boundary, and intersection, provided by Blender (see h_0 and h_N in Fig. 3). For more details on edge extraction, refer to the supplemental document.

While depth and normal maps can serve as structural guidance, they are generally less reliable compared to edge maps in preserving geometric structures. Depth maps suffer from scale inconsistencies across different 3D models. Even within a single scene, some geometric details, such as windows on a building's surface, may have only slight depth differences from their surroundings. These issues make it difficult for diffusion models to faithfully reflect the geometric guidance in depth maps. Normal maps, while invariant to scale, are less effective at representing geometric boundaries when distinct surfaces that should be separated exhibit similar normal values. In contrast, edge maps remain scale-invariant and precisely define object boundaries, ensuring robust shape preservation.

3.2 Sparse Anchor-view Generation (SAG)

The SAG module synthesizes high-quality anchor views, v_0 and v_N, using FLUX-dev [Labs 2024], a state-of-the-art text-to-image diffusion model. For high-quality anchor view synthesis, the SAG module needs to satisfy three criteria: reflect the structural conditions h_0 and h_N, match the visual style of the reference image v_{ref}, and maintain cross-view consistency between v_0 and v_N. We describe how each of these criteria is addressed in the following.

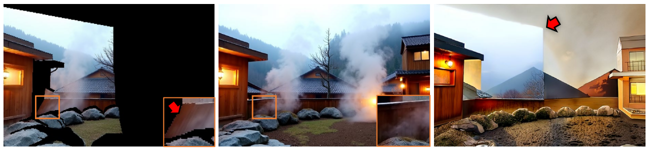
To incorporate structural guidance, the SAG module adopts ControlNet [Zhang et al. 2023] as the conditioning mechanism. To this end, rather than training a ControlNet specifically on our structural guidance, we adopt a pretrained ControlNet using edges from the HED edge detector [Xie and Tu 2015], which extracts perceptually-aligned edges from 2D images[1]. Although HED edges do not perfectly match those in our structural guidance, we empirically found that this approach performs effectively. More importantly, using HED edges eliminates the need for a specialized dataset of 3D-model-derived edges paired with natural images for training, which is extremely difficult to obtain.

To incorporate the style reference, we adopt a distribution alignment strategy. Specifically, we add LoRA [Hu et al. 2022] layers to both the image diffusion model and ControlNet, and train them using the reference image v_{ref} with a unique identifier prompt before synthesizing anchor views. This strategy aligns the target distribution of the diffusion model to the reference style, enabling style-consistent anchor view generation. As a result, the start view v_0 is synthesized using the style-aligned diffusion model, guided by the identifier prompt and the structural condition h_0.

Sparse Appearance-guided Sampling. To generate the end view v_N while maintaining cross-view consistency with v_0, we propose a Sparse Appearance-guided Sampling strategy (Fig. 4). Our strategy first obtains a sparse observation $v_{0\to N}$ by warping v_0 to the end view v_N using the optical flow $f_{N\to 0}$. The observed regions in $v_{0\to N}$ often exhibit distortions due to excessive warping (Fig. 4a). Nevertheless, they still retain useful semantic and appearance information that supports cross-view consistency. To exploit this information, we replace the latent of v_N with that of $v_{0\to N}$, in the regions observed in $v_{0\to N}$, during the diffusion sampling process [Ryu et al. 2025].

Specifically, we first compute the latent of $v_{0\to N}$, denoted as $\bar{z}_0$, using the encoder of the image diffusion model. Additionally, we generate a binary mask m to indicate the observed regions in $v_{0\to N}$, and obtain a downsampled version, $\bar{m}$, according to the size of the latent $\bar{z}_0$. We then randomly initialize the latent of v_N, denoted as z_T, where T represents the total number of diffusion timesteps. The standard diffusion process iteratively denoises the latent z_t from $t = T$ to $t = 0$. To guide the diffusion process to synthesize an image consistent with $v_{0\to N}$, we perform a replacement operation before

[1] https://huggingface.co/XLabs-AI/flux-controlnet-hed-v3



(a) Warped view $v_{0\to N}$ (b) End view v_N w/ distribution alignment (c) End view v_N w/o distribution alignment

Fig. 5. Effect of distribution alignment in generating the end view v_N using Sparse Appearance-guided sampling.

denoising at each timestep t, defined as:

$$z_t \leftarrow \bar{m} \odot \bar{z}_t + (1 - \bar{m}) \odot z_t, \tag{1}$$

where $\odot$ is element-wise multiplication. In Eq. (1), $\bar{z}_t$ is the latent of $v_{0\to N}$ at timestep t, obtained by adding noise to $\bar{z}_0$ following the noise scheduling of the image diffusion model. We apply the replacement operation only for early timesteps to reflect only the semantic and color information without distorted details in $v_{0\to N}$.

Through this process, we can synthesize natural-looking content across both observed and unobserved regions. Thanks to the replacement operation, the details generated for observed areas adhere closely to the semantic structures provided by the warped image $v_{0\to N}$. Meanwhile, the synthesis of unobserved regions maintains consistency with the visual characteristics of the observed areas, ensuring spatial coherence throughout the final output (Fig. 4b). We use 25 diffusion steps to generate each anchor view, and apply the replacement operation for the first 12 steps.

It is noteworthy that the proposed approach is made possible thanks to the distribution alignment using the style reference image performed before synthesizing anchor views. Without the distribution alignment, the aforementioned approach fails to produce coherent results in observed regions. This is because unknown regions typically occupy a much larger area than observed ones, making it difficult for the model to generate consistent content based on limited guidance. As a result, as shown in Fig. 5(c), the model often produces entirely different content with noticeable seams and inconsistency across the boundary. In contrast, the distribution alignment process narrows the solution space toward the reference style, enabling coherent synthesis, as shown in Fig. 5(b). The supplemental document provides additional discussions on the difference against inpainting approaches and multi-view-diffusion-based approaches as well as a detailed pseudocode.

Style Variation. Depending on the application, multiple reference styles may be required in a single scene. For example, when a scene includes transitions between indoor and outdoor areas, each region exhibits different structural and appearance characteristics, necessitating separate style references. However, training individual LoRA models for each style is cumbersome. To address this, we train a single LoRA model, assigning a unique identifier prompt to each reference image instead of training separate models for different styles. During anchor view generation, we selectively apply the desired style by using the identifier prompt corresponding to the target reference. In another scenario, users may want to apply global style variations such as seasonal changes or tonal shifts. In a similar vein, to avoid training additional LoRA models, we adopt a post-prompting strategy in which style variation is introduced at inference time by modifying the text prompt. Specifically, anchor views are generated using a unique identifier prompt combined with an additional style description, such as ‘winter’ or ‘cozy’.

3.3 Geometry-guided Generative Inbetweening (GGI)

The GGI module synthesizes a high-quality video frames $\mathbb{V}$ from the anchor views v_0 and v_N, by leveraging a video diffusion prior. To effectively perform the inbetweening task, we build upon a pretrained Image-to-Video (I2V) diffusion model, CogVideoX-5B-1.0 [Yang et al. 2024]. To condition on both endpoints, we encode the start and end frames v_0 and v_N using the VAE encoder $\mathcal{E}$. Zero-valued latents $\emptyset$ are used for the intermediate frames, resulting in $V = [\mathcal{E}(v_0), \emptyset, \cdots, \emptyset, \mathcal{E}(v_N)]$, where $[\cdot]$ denotes stacking along the temporal dimension. The feature V is concatenated with the noisy latent along the channel dimension. Additional implementation details on encoding the conditions with the 3D causal VAE of CogVideo-X are provided in the supplemental document.

To condition on the camera trajectory $\mathbb{P}$, we adopt a flow-based camera control approach similar to Go-with-the-Flow [Burgert et al. 2025; Jin et al. 2025]. Specifically, we obtain a warped noise volume, denoted as ϵ_w, that implicitly encodes the camera motion. To this end, we sample the initial noise for the first frame and recursively warp it using the consecutive optical flows $\{f_{0\to 1}, \cdots, f_{N-1\to N}\}$ while preserving Gaussianity. To reflect the motion information in the generation process, we employ the pretrained flow-aware LoRA module from Go-with-the-Flow [2025].

While the warped noise provides approximate guidance for the overall camera motion, it is insufficient for accurately capturing the intended motion trajectory, for a couple of reasons. First, the warped noise volume is constructed in a downsampled latent space, e.g., 8× smaller spatially and 4× temporally, inherently limiting the granularity of motion guidance. In addition, to preserve gaussianity during the noise warping process, Gaussian noise is continually re-injected, which results in the flow information being only implicitly encoded. This makes precise camera control challenging and often leads to structural distortions. To address this, we additionally concatenate the VAE-encoded HED edge maps $\mathcal{E}([h_0, \cdots, h_N])$ to the latent feature as structural guidance.

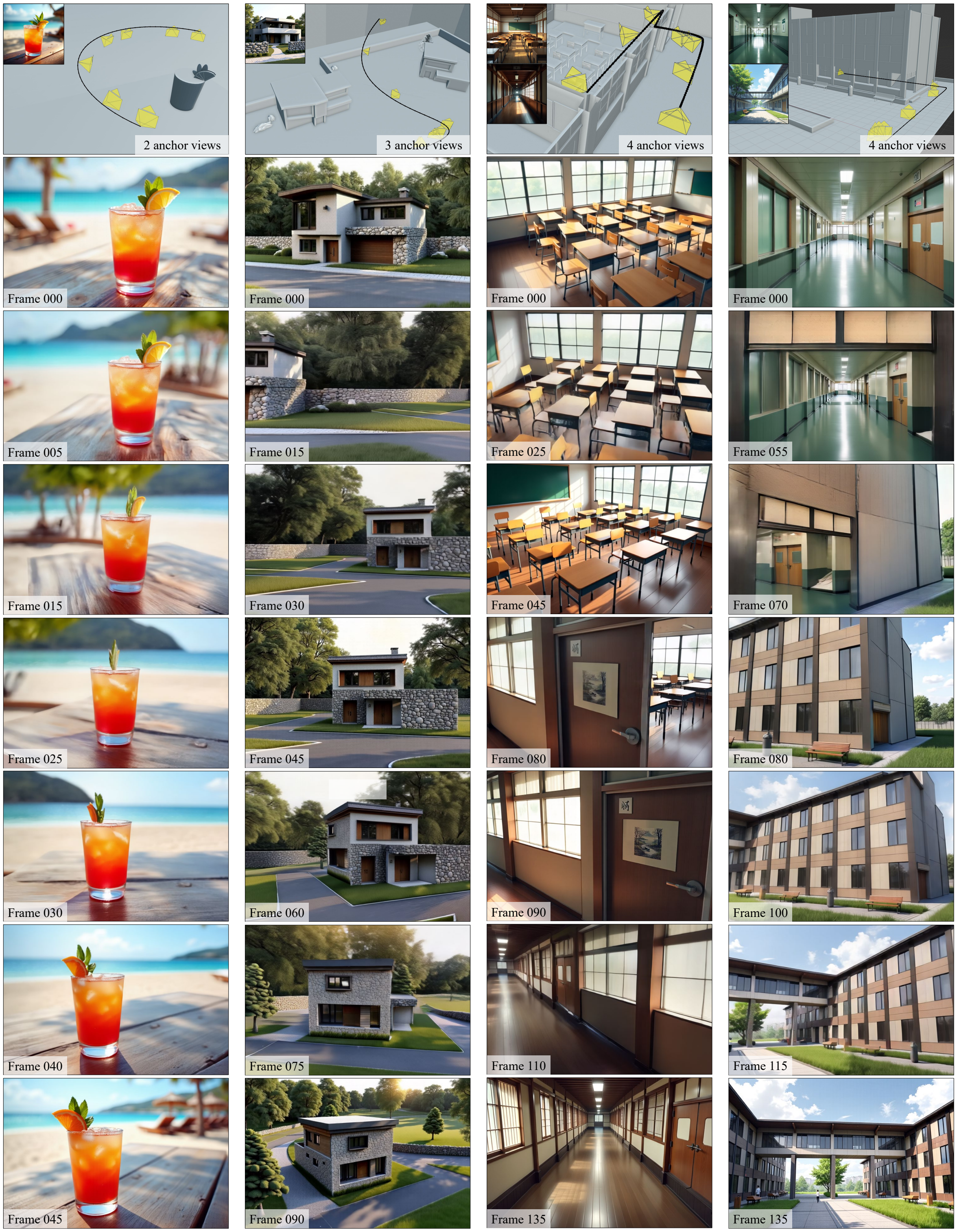
Finally, the diffusion sampling step of the GGI module is represented as:

$$\epsilon_{\Theta,\pi} \left(Z_t \oplus V \oplus \mathcal{E}\left([h_0, \cdots, h_N]\right),\ t \right) \mapsto Z_{t-1}, \tag{2}$$

where $\epsilon_{\Theta,\pi}$ denotes the diffusion sampling operation with parameters Θ for the base video diffusion model and π for the flow-aware LoRA. Z_t is the noisy latent at timestep t, initialized with ϵ_w, and $\oplus$ indicates channel-wise concatenation.

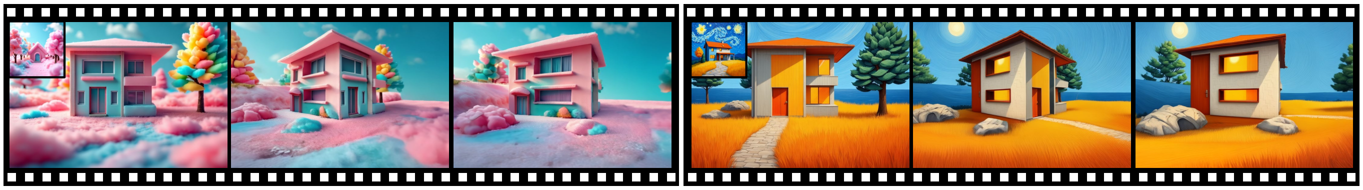
3.4 GGI Module Training

Training the GGI module ideally requires coarse geometry, camera trajectories, and their corresponding high-quality multi-view images, but such datasets are rarely available. To approximate this setting, we use the DL3DV-10K [Ling et al. 2024] dataset, which provides various videos of static scenes. Specifically, for each training video X, we compute the optical flows using RAFT [Teed and Deng



(a) Object (b) Outdoor (c) Indoor (d) Indoor-outdoor transition

Fig. 6. Qualitative results across various scenarios. The first row illustrates scene information: the style reference (top-left), the camera trajectory (black line), camera positions corresponding to each generated view shown below (yellow), and the number of anchor views (bottom-right). The input geometry in (c) and (d) is from TurboSquid (©Okhey).



(a) Animation style

(b) Painting style

Fig. 7. Non-photorealistic generation results. The top-left image is the style reference image.

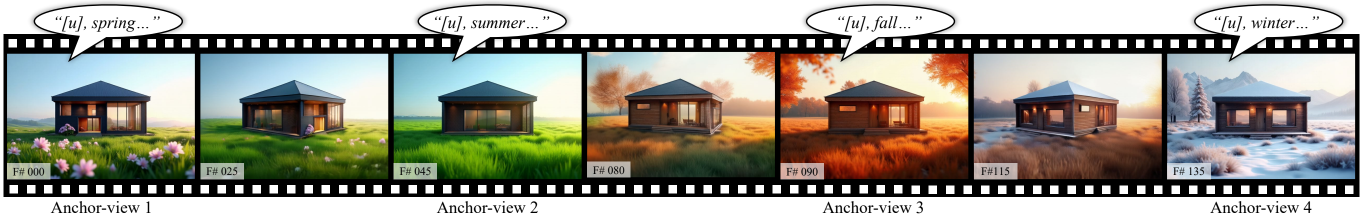


Fig. 8. Visual effect example showing simultaneous changes in camera motion and temporal context, such as seasonal appearance. Speech bubbles indicate text prompts used in the SAG module, where *[u]* denotes the identifier prompt used in LoRA training.

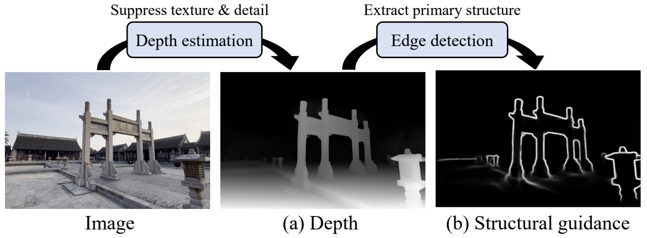


Fig. 9. Structural guidance simulation during GGI module training to reduce the domain gap between training and inference. The image is from DL3DV dataset [Ling et al. 2024].

2020] to generate the warped noise ϵ_w. For each frame, we extract the HED edge map h_i for structural guidance.

While training the GGI module requires edge maps derived from 3D models for structural guidance, the DL3DV-10K dataset lacks such 3D models. Thus, instead of using 3D models, we synthesize edge maps from training videos as illustrated in Fig. 9. The 3D-model-derived edge maps in our scenario exhibit two key characteristics: they contain no appearance information, such as texture, and they are derived from coarse geometry. To replicate these properties during training, we first estimate depth maps from training videos using an off-the-shelf depth estimator [Ranftl et al. 2020], and apply the HED edge detector [Xie and Tu 2015] to the estimated depth maps. Since depth maps inherently lack textures, and the HED detector selectively extracts strong structural contours, ignoring weak edge signals, this approach produces edge maps that closely resemble inference-time structural guidance, effectively reducing the domain gap between training and inference.

Finally, the training objective of the GGI module is defined as:

$$\underset{\Theta}{\operatorname{argmin}}\ \mathbb{E}_{X,t}\left[\left\|\epsilon_w - \epsilon_{\Theta,\pi}\left(Z_t \oplus V \oplus H,\ t\right)\right\|^2\right], \qquad (3)$$
$$H = \mathcal{E}\left([\mathcal{A}_e(\mathcal{A}_d(x_0)),\ \cdots,\ \mathcal{A}_e(\mathcal{A}_d(x_N))]\right),$$

where $\mathcal{A}_e$ denotes the HED edge estimator, $\mathcal{A}_d$ denotes the depth estimator, and x_i denotes the i^{th} frame of the video X.

4 Experiments

Implementation Details. In the SAG module, the LoRA layers (rank 16) are applied to the first 23 transformer blocks and trained for 400 iterations using a reference image and its HED edge map with the Adam optimizer [Kingma and Ba 2014]. In the GGI module, the resolution of the output video is 720×480. The last frame index N is set to 45, generating 46 frames. The module is trained for 1,300 iterations with a batch size of 16 using AdamW optimizer [Loshchilov and Hutter 2017]. During training and inference of the GGI module, we provide a text prompt generated from the first frame using BLIP [Li et al. 2022], as the base model, CogVideoX, requires an input text prompt.

4.1 Video Generation Results

Fig. 6 shows qualitative results of our method in various scenarios. The first example shows that our method works reliably in a simple object-centric scene. The second one highlights our robustness to dynamic camera motions involving large translations and rotations. The third and last columns demonstrate robust performance even in complex spatial transitions, across rooms and hallways, and between indoor and outdoor spaces, respectively. In addition, Fig. 7 presents results on artistic styles, demonstrating effectiveness in generating non-photorealistic scenes such as animations and paintings.

Fig. 8 shows an example where the style changes over time. To achieve this, each anchor view is assigned a distinct style via post-prompting with a different seasonal description. To enable natural scene-style transition, we intentionally omit the replacement operation in the Sparse Appearance-guided Sampling. Finally, the GGI module smoothly interpolates between distinct style frames, enabling challenging animation.

Table 2. Quantitative comparisons with Depth-I2V, VACE, and SAG-augmented MVSplat360, LVMS, and SEVA.

	Visual fidelity			Structural fidelity	Visual quality		Style consistency		
	PSNR↑	SSIM↑	LPIPS↓	PSNR-D↑	CLIP-A↑	MUSIQ↑	CLIP-I↑	SC↑	BC↑
Depth-I2V	-	-	-	**20.696**	6.136	65.240	0.787	0.864	0.920
VACE [Jiang et al. 2025]	-	-	-	18.850	6.189	65.318	0.787	0.856	0.914
SAG + MVSplat360 [Chen et al. 2024b]	13.163	0.374	0.315	13.881	5.714	50.524	0.788	0.797	0.894
SAG + LVSM [Jin et al. 2024]	15.103	0.472	0.280	15.222	5.680	50.323	0.804	0.843	0.917
SAG + SEVA [Zhou et al. 2025]	14.014	0.437	0.261	16.598	**6.782**	66.359	0.834	0.884	0.939
SAG + GGI (Ours)	**16.739**	**0.554**	**0.236**	19.754	6.730	**68.615**	**0.840**	**0.891**	**0.942**

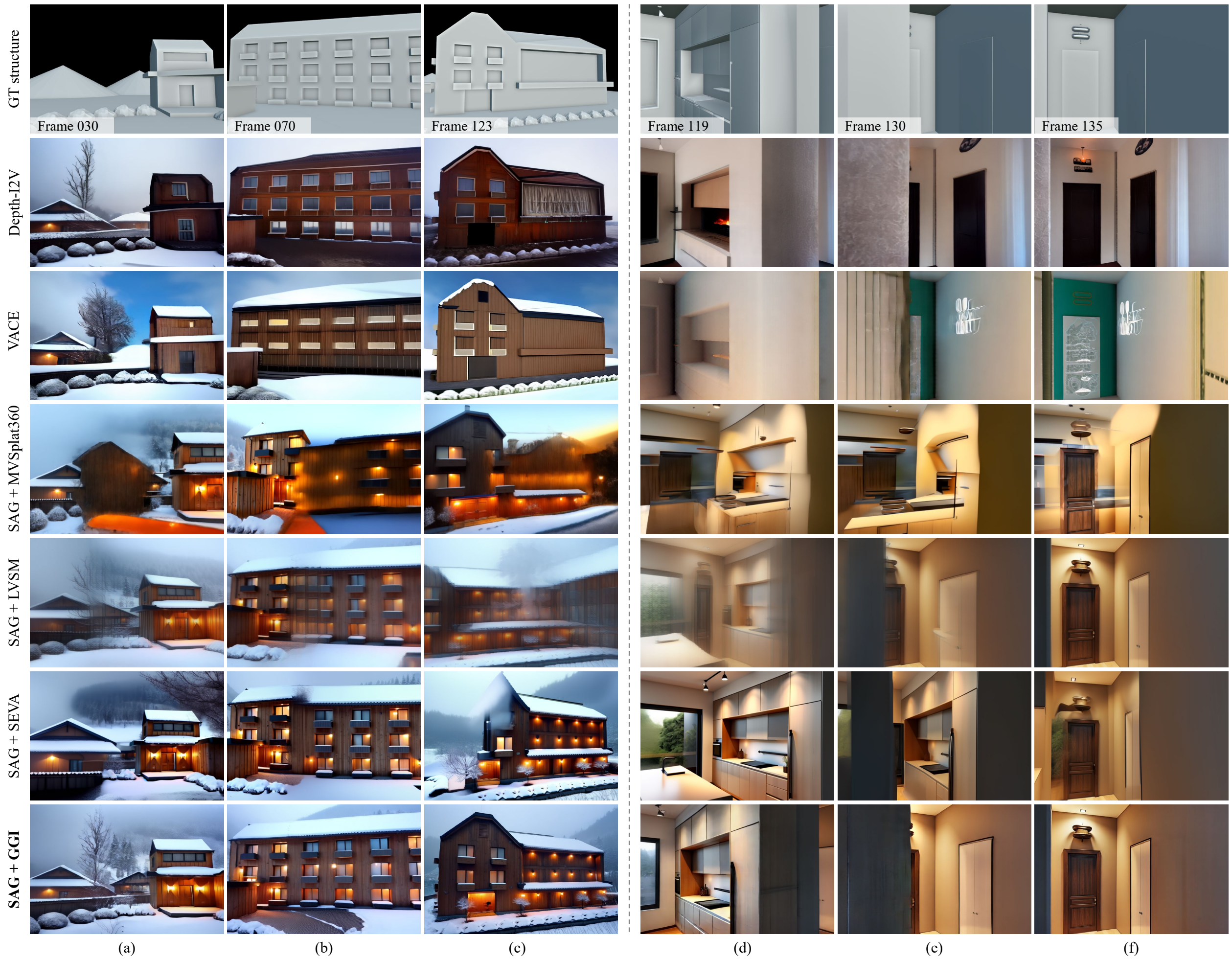
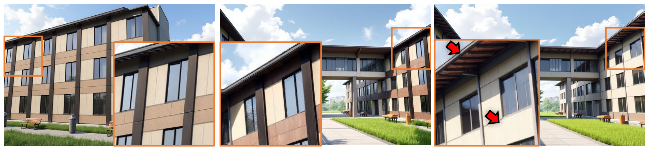


Fig. 10. Qualitative comparisons with Depth-I2V (a depth-conditioned I2V diffusion model), VACE [Jiang et al. 2025], and SAG-augmented variants of MVSplat360 [Chen et al. 2024b], LVMS [Jin et al. 2024], and SEVA [Zhou et al. 2025]. The input geometry of (d-f) is from TurboSquid (©3D LT).

4.2 Baseline Comparisons

To validate our method, we compare it with several baselines. For the video diffusion-only approach, we compare with VACE [Jiang et al. 2025] and a depth-conditioned I2V model, denoted as Depth-I2V. For VACE inference, we use depth maps as the structural clue, since they yield the best performance compared to other types of clues. Depth-I2V is trained on DL3DV-10K [Ling et al. 2024] by concatenating depth maps to the latent input, and is initialized from I2V-CogVideoX-5B-1.0. We also compare with state-of-the-art few-shot reconstruction models, each representing a distinct paradigm: MVSplat360 [Chen et al. 2024b] (video diffusion-based), LVSM [Jin et al. 2024] (regression-based), and SEVA [Zhou et al. 2025] (multi-view diffusion-based). These models take as input the anchor-view images generated by our SAG module.



(a) Start view v_0 (b) End view v_N w/ appearance guidance (c) End view v_N w/o appearance guidance

Fig. 11. Qualitative results with and without the Sparse Appearance-guided Sampling. Orange boxes indicate the same regions of the input geometry. The input geometry is from TurboSquid (©Okhey).

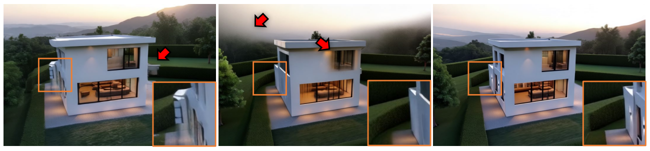
Table 3. Quantitative comparison of different structural conditions of GGI.

Condition	PSNR↑	SSIM↑	LPIPS↓	PSNR-D↑	CLIP-A↑	MUSIQ↑
-	14.160	0.417	0.282	16.780	6.622	62.562
HED	15.935	0.545	0.238	19.655	6.580	67.066
HED-S	**16.739**	**0.554**	**0.236**	**19.754**	**6.730**	**68.615**

To measure visual fidelity, we use PSNR, SSIM, and LPIPS [Zhang et al. 2018]. Since ground-truth (GT) intermediate frames are unavailable, we construct pseudo-GT frames by warping the anchor frames v_0 and v_N to obtain $v_{0\to i}$ and $v_{N\to i}$, where i represents the target frame index. We then composite these warped results, and compute the metrics only on the known regions. For structural fidelity, we compute PSNR between the GT depth maps rendered from input 3D models, and depth maps estimated by a monocular depth estimator from the corresponding synthesized videos (PSNR-D). To compensate for nonlinear errors and scene-dependent scale variations inherent in monocular depth estimation, we apply histogram equalization before computing PSNR. For visual quality, we report CLIP-A [LAION-AI 2023] and MUSIQ [Ke et al. 2021] scores. For style similarity, we measure CLIP image similarity [Radford et al. 2021] (CLIP-I) with the reference style image, as well as Subject Consistency (SC) and Background Consistency (BC) [Huang et al. 2024], which compute feature similarity between each frame and both the first and adjacent frames using DINO [Zhang et al. 2022] and CLIP, respectively.

For the test dataset, we construct 16 3D models, either manually modeled or sourced from open-source 3D assets [TurboSquid [n. d.]]. The dataset includes 4 object-centric, 2 indoor, 8 outdoor, and 2 indoor-outdoor transition scenes. For each model, we synthesize three different styles using either a distinct reference style image or post-prompting, resulting in a total of 48 generated videos.

Fig. 10 shows qualitative comparisons with the baselines. Depth-I2V and VACE generally produce low-quality results with insufficient details, due to the limited generative capability of the video diffusion model. MVSplat360 often produces severe artifacts due to frequent failures in reconstructing intermediate 3D representations when the distance between anchor views is large. LVSM generates blurry outputs in regions that require strong generative priors. SEVA often fails under challenging trajectories, mainly due to scale ambiguity arising from its reliance on explicit camera poses (Fig. 10c). Even in simpler cases, it deviates significantly from the GT structure (Fig. 10d). In contrast, our method achieves higher visual quality and structural fidelity even under challenging conditions. Table 2 shows quantitative comparisons with the baselines. Our method achieves the best performance across most metrics and ranks second-best in a few, demonstrating its overall effectiveness.



(a) w/o structure condition (b) w/ HED (c) w/ HED-S

Fig. 12. Qualitative comparison of different structural conditions of GGI.

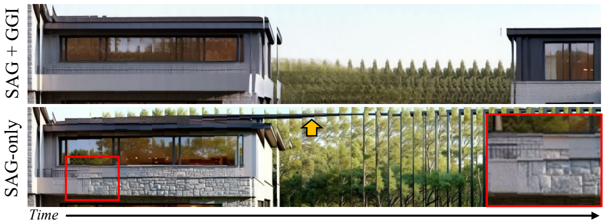


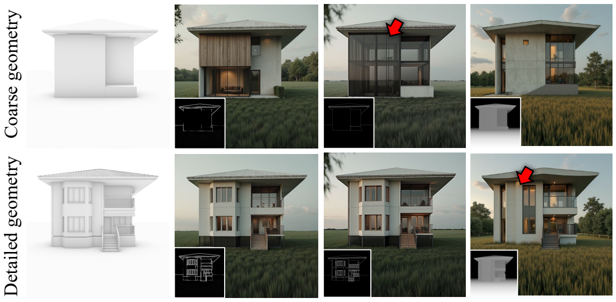
Fig. 13. Dense view generation using only the SAG module causes severe flickering (red) and accumulated warping errors (yellow).

4.3 Ablation & Analysis

Ablation on SAG Module. Fig. 11 presents anchor-view generation results with and without the Sparse Appearance-guided Sampling. The orange boxes indicate corresponding regions in the input geometry. In Fig. 11(c), where the guided sampling is not applied, the generated details such as the roof, windows, and facade color patterns significantly deviate from those in the start frame. In contrast, with the guided sampling applied (Fig. 11b), these details remain visually consistent, demonstrating the effectiveness of the method.

Ablation on GGI Module. Fig. 12 presents inference results with different GGI modules trained under varying structural conditions. Without any structural condition, severe distortions frequently occur (Fig. 12a). Using HED edges directly extracted from the RGB image results in missing details (Fig. 12b). In contrast, our simulated structural condition, denoted as HED-S, accurately preserves structure and avoids detail loss. This qualitative observation aligns well with the quantitative comparison in Table 3.

Dense View Generation using SAG. One might wonder whether the SAG module alone could be used to generate intermediate views, rather than relying on the GGI module, since it already produces plausible novel views for the anchor frames. To investigate this, we compare temporal profiles in Fig. 13, which visualize a fixed 160×20 pixel region over time, comparing our full method and the SAG-only approach. In the SAG-only setting, frames are generated along the camera trajectory solely using the SAG module. As shown in the red box, the inherent randomness of the generation process leads to severe flickering and temporal inconsistency. This highlights the necessity of the GGI module for consistent video synthesis.



(a) 3D model (b) HED (c) Canny-edge (d) Depth

Fig. 14. FLUX ControlNet generation results using (b) HED edge map, (c) Canny-edge map and (d) depth map.

Table 4. Latency measurements for each process using an A100-80GB GPU.

	Preprocessing	Distribution alignment	SAG module	GGI module
Latency	12 sec/traj.	27 min./style	20 sec./view	145 sec./traj.

Structural Condition for Anchor-view Generation. Fig. 14 shows generation results using Flux ControlNet [2024] with different types of structural conditions, applied to both coarse and detailed geometry. The Canny-edge condition yields visually monotonous results under coarse geometry due to a mismatch between fine-texture training edges and sparse test-time inputs. Conversely, the depth condition tends to ignore weak signals in the depth map, making it less effective for guiding detailed geometry. In contrast, HED edge conditioning generalizes well to both coarse and detailed cases, as its estimator is trained on sparse, human-annotated edge maps that closely resemble the distribution of 3D-model-derived edges.

Latency. Table 4 shows the latency of each component. After LoRA training, generating a single trajectory takes 197 seconds.

5 Conclusion

In this paper, we introduce VideoFrom3D, a novel framework synthesizing high-quality 3D scene videos given coarse geometry, camera trajectories, and reference images. By combining the complementary strengths of image and video diffusion models through the SAG and GGI modules, our method produces style-consistent, natural, and geometrically faithful videos. Extensive evaluations demonstrate its effectiveness across diverse and challenging scenarios.

Limitations. VideoFrom3D does not support real-time interactive camera control. In addition, temporal inconsistency may occur due to the inherent randomness of diffusion models. Our method requires LoRA training, which requires a significant amount of computation time, as shown in Table 4. Addressing these limitations would be an interesting future direction.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH), No. RS-2024-00457882, AI Research Hub Project, No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. 2025. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492* (2025).

Dane Edward Bettis. 2005. *Digital production pipelines: examining structures and methods in the computer effects industry.* Ph. D. Dissertation. Texas A&M University.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. *arXiv preprint arXiv:2501.08331* (2025).

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. 2025. Uni3C: Unifying Precisely 3D-Enhanced Camera and Human Motion Controls for Video Generation. *arXiv preprint arXiv:2504.14899* (2025).

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023b. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision.* 18558–18568.

Hansheng Chen, Ruoxi Shi, Yulin Liu, Bokui Shen, Jiayuan Gu, Gordon Wetzstein, Hao Su, and Leonidas Guibas. 2024a. Generic 3d diffusion adapter using controlled multi-view editing. *arXiv preprint arXiv:2403.12032* (2024).

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision.* 22246–22256.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. 2024b. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *arXiv preprint arXiv:2411.04924* (2024).

Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. 2024. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. In *ACM SIGGRAPH 2024 Conference Papers.* 1–10.

Abdelilah Hamdani and Carlos Barreto. 2023. *3D Environment Design with Blender: Enhance your modeling, texturing, and lighting skills to create realistic 3D scenes.* Packt Publishing Ltd.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. 2024. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503* (2024).

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598* (2025).

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. 2024. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242* (2024).

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. 2025. FloVD: Optical Flow Meets Video Diffusion Model for Enhanced Camera-Controlled Video Synthesis. *arXiv preprint arXiv:2502.08244* (2025).

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision.* 5148–5157.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

LAION-AI. 2023. Aesthetic Predictor. https://github.com/LAION-AI/aesthetic-predictor. Accessed: 2025-05-01.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning.* PMLR, 12888–12900.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22160–22169.

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

Francois Lord. 2024. An OpenUSD Production Pipeline with Very Little Coding: Empowering 3D artists with a parallel workflow using off-the-shelf software. In *ACM SIGGRAPH 2024 Talks.* 1–2.

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

Fan Lu, Kwan-Yee Lin, Yan Xu, Hongsheng Li, Guang Chen, and Changjun Jiang. 2024. Urban architect: Steerable 3d urban scene generation with layout prior. *arXiv preprint arXiv:2404.06780* (2024).

Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. 2024. You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale. *arXiv preprint arXiv:2412.06699* (2024).

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12663–12673.

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751* (2025).

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*. 1–11.

Nuri Ryu, Jiyun Won, Jooeun Son, Minsu Gong, Joo-Haeng Lee, and Sunghyun Cho. 2025. Elevating 3D Models: High-Quality Texture and Geometry Refinement from a Low-Quality Model. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.

Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, et al. 2024. Controlroom3d: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6201–6210.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023).

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

TurboSquid. [n. d.]. TurboSquid - 3D Models for Professionals. https://www.turbosquid.com.

Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. 2024. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. *arXiv preprint arXiv:2409.11406* (2024).

Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*. 1395–1403.

XLabs. 2024. x-flux. https://github.com/XLabs-AI/x-flux.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024a. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).

Xin Yu, Ze Yuan, Yuan-Chen Guo, Ying-Tian Liu, Jianhui Liu, Yangguang Li, Yan-Pei Cao, Ding Liang, and Xiaojuan Qi. 2024b. Texgen: a generative diffusion model for mesh textures. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–14.

Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. 2024. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4252–4262.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* (2022).

Hongkun Zhang, Zherong Pan, Congyi Zhang, Lifeng Zhu, and Xifeng Gao. 2024. Texpainter: Generative mesh texturing with multi-view consistency. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. 2025. STABLE VIRTUAL CAMERA: Generative View Synthesis with Diffusion Models. *arXiv preprint arXiv:2503.14489* (2025).