VideoFrom3D: 3D Scene Video Generation via Complementary Image and Video Diffusion Models [Supplemental Document]

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

"photo-realistic traditional Japanese onsen building engulfed in massive raging flames, fire consuming all, no escape, terrifying inferno, night"

Fig. S1. Fire visual effects example. The prompt shown above is used for post-prompting. Our framework generates a visually consistent and temporally smooth fire animation with camera motion, achieving high-impact VFX results.

We have attached a video named "SuppleVideo.mp4" as Supplementary Material, containing generation results and comparisons illustrated in the main paper and supplemental document.

A PSEUDO-CODE FOR SAG MODULE

ALGORITHM 1: Sparse Anchor-view Generation (SAG)

Input: HED edge-conditioned image diffusion model ϵ_η, total denoising step T, replacement limit n_r, reference image v_{ref}, VAE encoder $\mathcal{E}_I$, VAE decoder $\mathcal{D}_I$, identifier prompt $\mathbf{p}$, structural guidance of start view h_0, structural guidance of end view h_N, optical flow $f_{N\to 0}$,
Output: Anchor-view v_0 and v_N

Func. `DiffusionSampling`$(\epsilon_{\hat{\eta}}, h_N, \mathbf{p})$:
  $z_t \sim \mathcal{N}(0, I)$;
  for $t \leftarrow T$ **to** 1 **do**
    $z_t \leftarrow \text{DenosingOneStep}(z_t, h_N, t, \epsilon_{\hat{\eta}})$;
  return $\mathcal{D}_I(z_t)$
Func. `SparseAppearanceGuidedSampling`$(\epsilon_{\hat{\eta}}, h_N, \mathbf{p}, v_0, f_{N\to 0})$:
  $z_t \sim \mathcal{N}(0, I)$;
  $v_{0\to N}, m \leftarrow \text{BackwardWarp}(v_0, f_{N\to 0})$;
  $\bar{z}_0 \leftarrow \mathcal{E}_I(v_{0\to N})$;
  $\bar{m} \leftarrow \text{Downsample}(m)$;
  for $t \leftarrow T$ **to** 1 **do**
    $z_t \leftarrow \text{DenosingOneStep}(z_t, h_N, t, \epsilon_{\hat{\eta}})$;
    if $T - t \geq n_r$ **then**
      continue;
    $\bar{z}_t \leftarrow \text{ApplyNoiseSchedule}(\bar{z}_0, t)$;
    $z_t \leftarrow \bar{m} \odot \bar{z}_t + (1 - \bar{m}) \odot z_t$;
  return $\mathcal{D}_I(z_t)$
$\epsilon_{\hat{\eta}} \leftarrow \text{DistributionAlignment}(\epsilon_\eta, v_{ref}, \mathbf{p})$; `/* LoRA training */`
$v_0 \leftarrow \text{DiffusionSampling}(\epsilon_{\hat{\eta}}, h_0, \mathbf{p})$
$v_N \leftarrow \text{SparseAppearanceGuidedSampling}(\epsilon_{\hat{\eta}}, h_N, \mathbf{p}, v_0, f_{N\to 0})$

Authors' addresses: 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.

B FIRE VISUAL EFFECTS (VFX)

Figure S1 illustrates a generation result animating a building engulfed in flames, accompanied by camera motion. To produce this animation, the anchor views are generated with post-prompting, where fire-related descriptions are added to the unique identifier prompt. The intermediate frames are then synthesized using our GGI module. As shown in the supplemental video, the generated fire flows dynamically with swirling and rising motion, maintaining temporal smoothness and visual realism throughout the sequence. Achieving this with traditional graphics pipelines requires not only the construction of static scenes but also the integration of physically-based fire simulation, which involves significant complexity and domain expertise. In comparison, our framework provides a highly effective and efficient solution for generating such challenging visual effects with minimal overhead.

C DETAILS OF VISUAL COMPARISON BETWEEN IMAGE AND VIDEO DIFFUSION MODELS

In this section, we provide additional details on the comparison between image and video diffusion models, as presented in Figure 2 and Table 1 of the main paper. For the comparison, we select text-conditioned generation models, as they serve as foundational generative priors that are widely used in downstream tasks via fine-tuning. This implies that their performance reflects an upper bound on generative quality. To evaluate the models, we generate 1,000 prompts describing complex scenes using OpenAI's ChatGPT, and use them to synthesize 1,000 images and 1,000 videos via text-to-image and text-to-video diffusion models, respectively. For qualitative comparison in Figure 2, we use the prompt 'brick townhouse with arched doorways overlooking the ocean'.

One might argue that the output of the video diffusion model in Fig. 2b of the main paper exhibits lower visual quality compared to generated videos commonly seen online, raising concerns about the validity of the comparison. However, many high-quality online videos are not generated by text-to-video models, but rather by image-to-video models conditioned on high-quality first frames, which gives them a significant advantage. Without such a strong

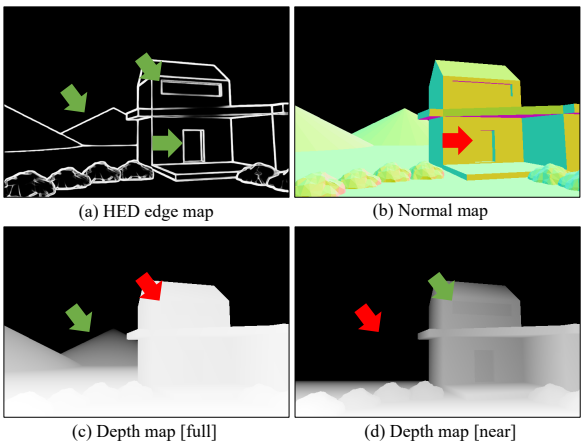


Fig. S2. Visual comparison of structural guidance, including (a) HED edge map, (b) normal map, (c) depth map covering the full range, and (d) depth map focusing on near-range geometry.

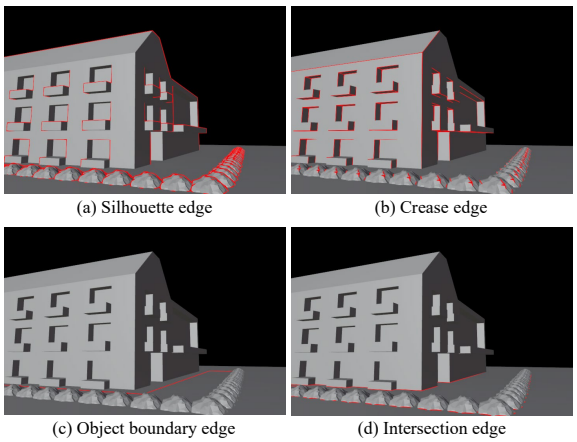


Fig. S3. Visualization of edge types extracted during preprocessing.

visual prior, the per-frame visual quality of video generation significantly degrades, especially for complex scenes. Moreover, even when starting from a high-quality first frame, generating scenes along a long camera trajectory eventually requires the model to synthesize entirely new content. From that point, the generation quality of video diffusion models often begins to degrade.

D TYPES OF STRUCTURAL GUIDANCE

Fig. S2 shows a visual comparison of structural guidance, including the HED edge, normal, and depth maps. The normal map fails to capture geometric structure in regions where perceptually adjacent surfaces exhibit similar surface normals, often omitting important structural cues (Fig. S2b). A ground-truth depth map, when visualized over the full range, captures the global geometry of the scene but suppresses local structures such as window details (Fig. S2c). Conversely, a depth map that focuses on near-range structures better preserves fine details, such as window edges, but discards distant structures (Fig. S2d). As such, it is challenging to find a normalization scheme that robustly represents all geometric structures from depth maps rendered from 3D models, making it less effective for structural guidance. In contrast, the edge map remains effective across both global and local contexts, robustly preserving geometric structure without these limitations (Fig. S2a).

E EDGE EXTRACTION FROM 3D GEOMETRY

As described in the main paper, we extract four types of edges to acquire structural guidance h_i: silhouette, object boundary, crease, and intersection edges. In this section, we detail the extraction process for each edge type and describe its utility. First, the silhouette edge is a surface edge shared by one front-facing and one back-facing polygon with respect to the viewing direction. Formally, given a viewing direction vector $\mathbf{v}_i$ from a camera pose p_i, an edge shared by two adjacent faces with normals $\mathbf{n}_1$ and $\mathbf{n}_2$ is classified as a silhouette edge if $\text{sign}(\langle \mathbf{n}_1, \mathbf{v}_i \rangle) \neq \text{sign}(\langle \mathbf{n}_2, \mathbf{v}_i \rangle)$. As shown in Fig. S3(a), silhouette edges outline the primary contour of the object from a specific viewpoint. The crease edge is defined as one where the angle between adjacent face normals exceeds a threshold θ, i.e., $\arccos(\langle \mathbf{n}_1, \mathbf{n}_2 \rangle) > \theta$. As shown in Fig. S3(b), it captures structurally important edges that are not part of the silhouette. We set the crease angle threshold θ to 40° in our experiments. The object boundary edge depicts the topological boundary of open surfaces. As illustrated in Fig. S3(c), they are useful when structure needs to be guided using a simple planar surface. Additionally, a scene typically consists of multiple mesh objects rather than a single one, and when these objects intersect, structurally significant edges may not be captured by the previously defined edge types. To address this, we extract intersection edges to capture such cases, as shown in Fig. S3(d). After extracting all four types of edges, we combine them to form the final edge map. Unlike standard binary edge maps, the structural guidance used during training is derived from HED edge maps, which exhibit soft, brush-like edges. To reduce the domain gap between the extracted edge map and the style of HED outputs, we render the combined edge map as an image and pass it through a pretrained HED edge detector [Xie and Tu 2015], resulting in the final structural guidance h_i.

F OPTICAL FLOW EXTRACTION

To compute correspondences between multi-view images, prior approaches [Burgert et al. 2025; Jin et al. 2025] adopt a depth-based reprojection scheme. This technique estimates correspondences by projecting pixels from one view to another using the predicted depth map and camera pose. While simple and effective in many cases, these methods inherently struggle to handle occlusions, as they rely on a single depth value per pixel and do not explicitly model visibility. In our setting, occlusions are common, particularly in large-baseline correspondences such as $f_{N\to 0}$. To handle occlusions more effectively, we adopt a coordinate-map reprojection approach. Fig. S5 illustrates the process of acquiring correspondences using this method. As shown in the figure, this approach successfully handles occluded regions even under large viewpoint changes, resulting in more reliable correspondence maps.

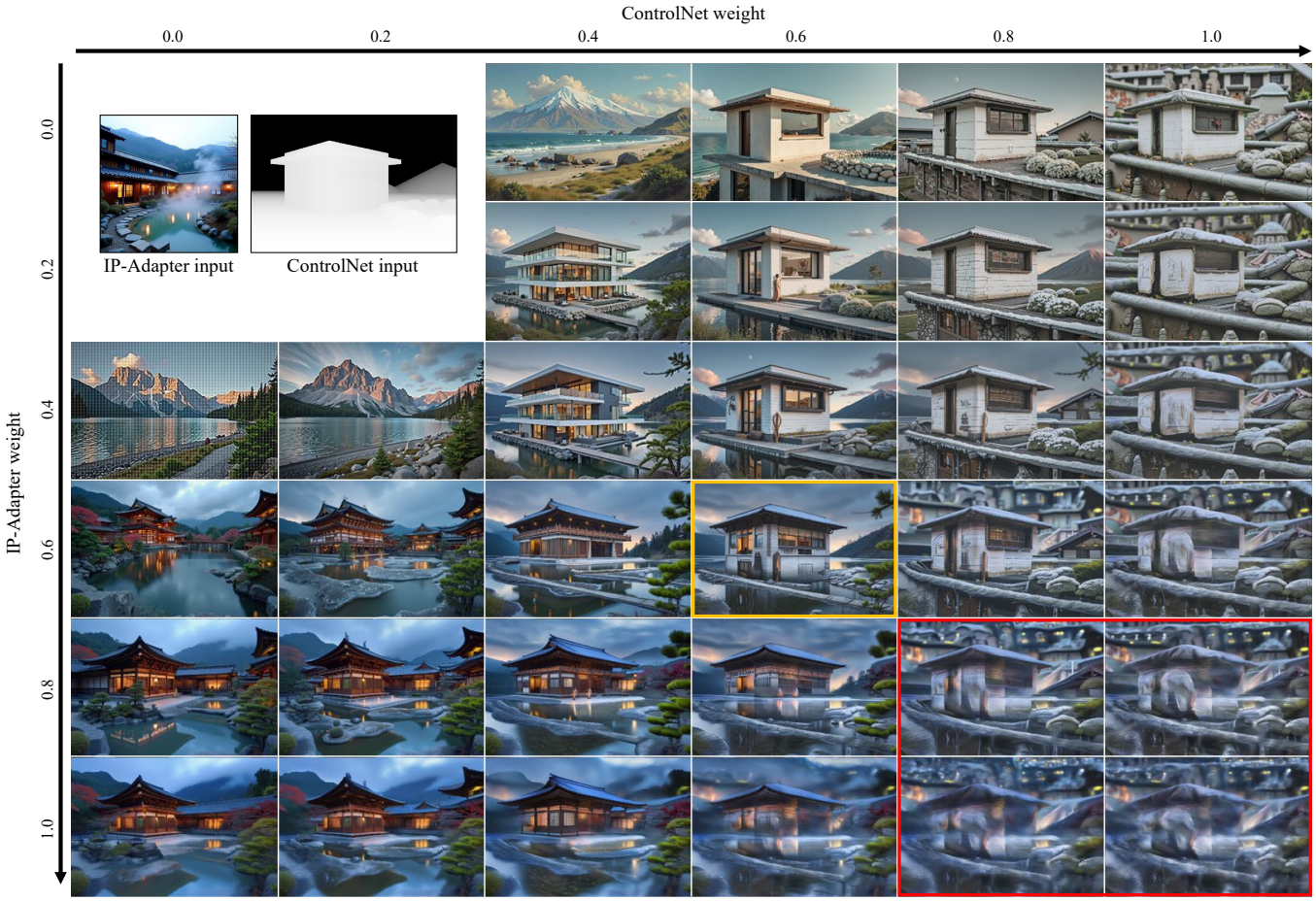


Fig. S4. Effect of IP-Adapter [Ye et al. 2023] weight (rows) and ControlNet [Zhang et al. 2023] weight (columns) on the generated results.

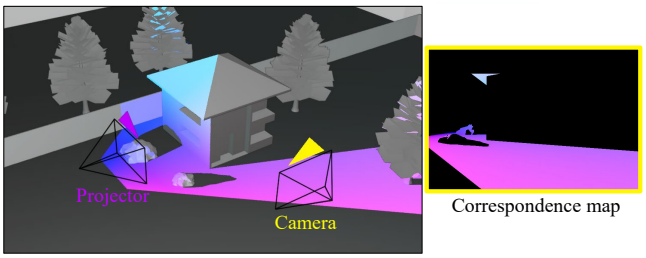


Fig. S5. Visualization of a projector-camera setup in a 3D scene (left) and the corresponding pixel-wise correspondence map (right).

G DISCUSSION ON IP-ADAPTER

As noted in the limitation paragraph of the Introduction section, our distribution alignment strategy results in a latency bottleneck within the overall generation process. To address this, IP-Adapter [Ye et al. 2023] could be a promising solution, since it introduces an amortized inference method for distribution alignment and thus significantly reduces the computational overhead. However, in our unusual setting involving ControlNet [Zhang et al. 2023], applying IP-Adapter proves incompatible, as shown in Fig. S4. Specifically, assigning conventionally used weights to both adapters introduces severe visual artifacts (red box). We speculate that this arises from uncoordinated features being simultaneously injected by the two independently trained adapters during inference. While moderate weights for both adapters produce reasonable visual quality, the resulting style deviates from the reference because the style imposed by ControlNet weakens that provided by the IP-Adapter (orange box).

H MULTI-VIEW CONSISTENCY AMONG ANCHOR VIEWS

When generating a new anchor view, there may be cases where it shares correspondences with non-adjacent anchor views. In such cases, we identify overlapping regions with all previously generated anchor views and apply warping to those regions, thereby ensuring multi-view consistency with non-adjacent anchor views. A concrete example is illustrated in Fig. S6. The camera starts indoors, passes through an outdoor scene, and then returns to a location similar to

1st anchor view (Indoor) 2nd anchor view (Outdoor) 3rd anchor view (Indoor)

Fig. S6. Multi-view consistency across three anchor views. The input geometry is from TurboSquid (©Okhey).

Table S1. Quantitative comparison of video generation using the SAG module alone and the combination of SAG and GGI modules.

	Temporal Flickering ↑	Motion Smoothness ↑
SAG-only	89.67	94.59
SAG+GGI	**94.54**	**98.51**

Table S2. Quantitative comparison of anchor view generation with and without the Sparse Appearance-guided sampling.

	PSNR ↑	SSIM ↑	LPIPS ↓
w/o appearance guidance	8.746	0.276	1.801
w/ appearance guidance	**17.878**	**0.629**	**1.725**

the starting point. As shown in the results, the view consistency between the first and third anchor views is effectively preserved. However, as indicated by the orange boxes, new content may appear that was absent from the previous anchor views, owing to the inherent randomness of diffusion models.

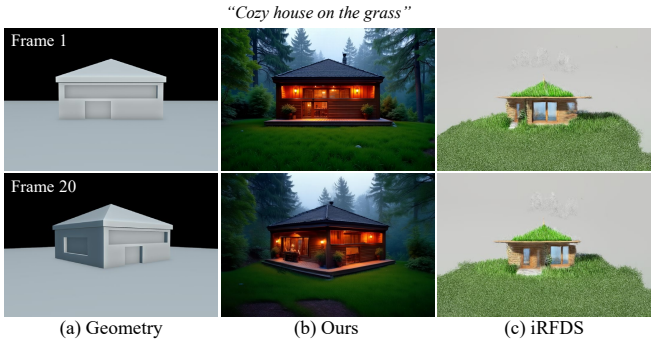
I ADDITIONAL ABLATION RESULTS

Table S1 presents a quantitative comparison of video generation using the SAG module alone versus the combination of SAG and GGI modules. The evaluation employs two metrics, Temporal Flickering and Motion Smoothness, which assess temporal motion quality as introduced in VBench++ [Huang et al. 2024]. As shown in the table, relying solely on the image diffusion model (SAG-only) fails to preserve temporal consistency, leading to severe flickering and lower motion quality. Table S2 presents a quantitative comparison of anchor-view generation with and without Sparse Appearance-guided sampling. For evaluation, the start views are warped to the end views, and the generated end views are compared with the warped images within the valid warping regions. As the table shows, the absence of appearance guidance leads to a complete failure of view consistency.

J COMPARISON WITH SDS-BASED METHOD

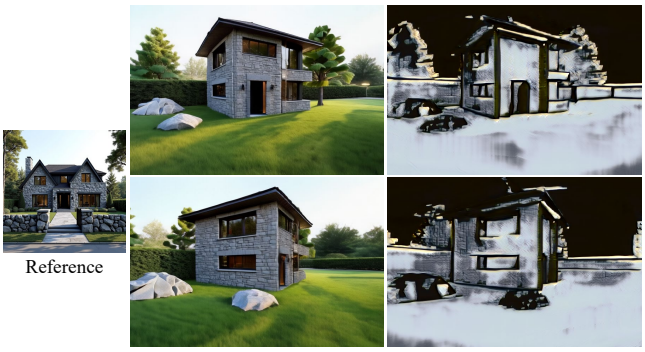
Fig. S7 illustrates a qualitative comparison between our method and a state-of-the-art SDS-based approach, iRFDS [Yang et al. 2024]. To incorporate geometric guidance, which is not natively supported in iRFDS, we augment its optimization process with a depth ControlNet[1]. As shown in the figure, iRFDS produces visually lower-quality results, struggling to synthesize background regions such as the sky and failing to accurately follow geometric guidance. Furthermore,

[1]https://huggingface.co/InstantX/SD3-Controlnet-Depth



(a) Geometry (b) Ours (c) iRFDS

Fig. S7. Qualitative comparison with iRFDS [Yang et al. 2024], a state-of-the-art SDS-based approach.



Reference

(a) Ours (b) StyleRF

Fig. S8. Qualitative comparison with StyleRF [Liu et al. 2023], a state-of-the-art 3D style transfer approach.

it remains restricted to forward-facing view generation, making it unsuitable for large-scale scene generation, unlike our approach.

K DISCUSSION ON 3D STYLE TRANSFER

3D style transfer methods [Chiang et al. 2022; Huang et al. 2021; Liu et al. 2023] aim to convert a 3D representation of a realistic scene into a stylized one based on an input style image. These methods could be adapted to our task by training a NeRF with multi-view images rendered from an untextured mesh and using a reference image for style guidance. While conceptually plausible, they fail to produce realistic scenes because they assume photorealistic multi-view inputs, which are unavailable in our setting, and are inherently tailored for artistic style transfer rather than photorealistic scene synthesis. Fig. S8 illustrates this limitation by comparing with StyleRF [Liu et al. 2023], a representative style transfer method. As shown, StyleRF produces grayish results that fail to reflect the reference style.

L DISCUSSION ON SAG MODULE ALTERNATIVES

Beyond our approach, alternative methods could be explored for anchor view synthesis. One option is to apply multi-view diffusion models [Gao et al. 2024; Zhou et al. 2025] to generate multi-view-consistent anchor views. While this is plausible, it presents two

major drawbacks. First, it requires high-cost training with large-scale multi-view datasets to learn multi-view consistency capability. Even with high-cost training, multi-view diffusion models must learn to preserve multi-view consistency, which can lead to compromised visual quality, as seen in video diffusion models. In contrast, our approach fully harnesses the generative power of an image diffusion model by integrating ControlNet for structural guidance and distribution alignment with a style reference, achieving high generation quality without compromising image quality or requiring large-scale training. The other option is to apply inpainting techniques [Lugmayr et al. 2022] to warped view $v_{0\to N}$. However, distortions present in the warped input often persist in the final output v_N, leading to error-prone results. In addition, inpainting methods are typically designed to fill relatively small missing regions, and thus may struggle with large unobserved areas.

M 3D CAUSAL VAE DETAILS

CogVideoX employs a 3D causal VAE to encode video volumes into a latent space. The term *causal* indicates that 3D convolution operations used during encoding do not access future frames. To ensure this constraint, the VAE encodes the first video frame independently and uses it as front padding during the encoding process. As a result, the encoder input and output shapes are $(1+4K)\times 3\times 8H\times 8W$ and $(1+K)\times 16\times H\times W$, respectively, where $1+4K$ is the number of input frames and $1+K$ is the temporal length of the latent representation. Here, 16, H, and W denote the channel, height, and width dimensions in the latent space, respectively. Conversely, the decoder follows the reverse dimensional mapping, transforming latent representations of shape $(1+K)\times 16\times H\times W$ back to video outputs of shape $(1+4K)\times 3\times 8H\times 8W$. In some scenarios, such as image-to-video (I2V) generation, only a single reference frame need to be encoded. To support this, the encoder defines a valid embedding for a single image input $\mathcal{E}(x)$ by replicating it temporally: $\mathcal{E}(x) := \mathcal{E}([x,x,x,x])$, where x is a single image input.

When designing the interpolation model based on this VAE encoding mechanism, a minor issue arises. As described in the main paper, the endpoint guidance V is defined as $[\mathcal{E}(v_0), \emptyset, \cdots, \emptyset, \mathcal{E}(v_N)]$. Unlike the first frame, decoding $\mathcal{E}(v_N)$ produces four output frames due to the structure of the decoder. Since these frames are duplicates, we remove the last three to obtain the final video sequence. In the same vein, when structural guidance is provided, we append the final frame four times to consider the encoding and decoding mechanism. For example, the VAE encoding of structural guidance is given by $\mathcal{E}([h_0, \cdots, h_{N-1}, h_N, h_N, h_N, h_N])$.

REFERENCES

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).

Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. 2022. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1475–1484.

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

Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. 2021. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13869–13878.

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).

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).

Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 2023. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8338–8348.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.

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

Xiaofeng Yang, Cheng Chen, Xulei Yang, Fayao Liu, and Guosheng Lin. 2024. Text-to-image rectified flow as plug-and-play priors. *arXiv preprint arXiv:2406.03293* (2024).

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

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.

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).