GO-Renderer: Generative Object Rendering with 3D-aware Controllable Video Diffusion Models

Zekai Gu[1,2], Shuoxuan Feng[3], Yansong Wang[4], Hanzhuo Huang[5], Zhongshuo Du, Chengfeng Zhao[1], Chengwei Ren[1], Peng Wang[2⋆], and Yuan Liu[1⋆⋆]

[1]Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]VAST, China [3]Nanyang Technological University, Singapore
[4]Tsinghua University, China [5]ShanghaiTech University, China
zekai.gu@connect.ust.hk, yuanly@ust.hk

Abstract. Reconstructing a renderable 3D model from images is a useful but challenging task. Recent feedforward 3D reconstruction methods have demonstrated remarkable success in efficiently recovering geometry, but still cannot accurately model the complex appearances of these 3D reconstructed models. Recent diffusion-based generative models can synthesize realistic images or videos of an object using reference images without explicitly modeling its appearance, which provides a promising direction for object rendering, but lacks accurate control over the viewpoints. In this paper, we propose GO-Renderer, a unified framework integrating the reconstructed 3D proxies to guide the video generative models to achieve high-quality object rendering on arbitrary viewpoints under arbitrary lighting conditions. Our method not only enjoys the accurate viewpoint control using the reconstructed 3D proxy but also enables high-quality rendering in different lighting environments using diffusion generative models without explicitly modeling complex materials and lighting. Extensive experiments demonstrate that GO-Renderer achieves state-of-the-art performance across the object rendering tasks, including synthesizing images on new viewpoints, rendering the objects in a novel lighting environment, and inserting an object into an existing video. Project page: https://igl-hkust.github.io/GO-Renderer.

Keywords: Generative Rendering · Appearance Modeling · Video Diffusion

1 Introduction

Rendering high-fidelity objects within diverse environments is a fundamental pursuit with extensive applications in advertising, film production, and immersive content creation. Achieving realistic object relighting and novel view synthesis requires not only reconstructing precise 3D geometry from 2D observations

⋆ Project leader.
⋆⋆ Corresponding author.

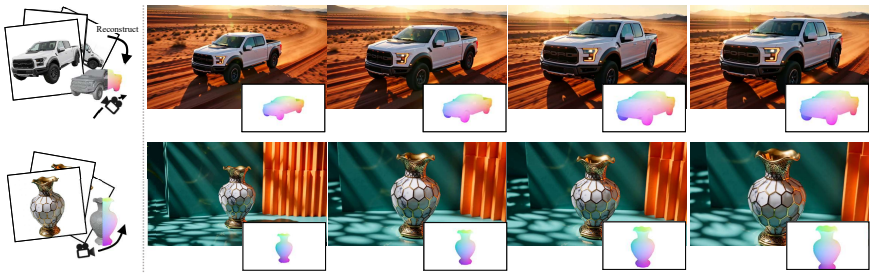

Fig. 1: Results of our GO-Renderer. Driven by multi-view references and 3D poses, GO-Renderer renders multi-view consistent object videos seamlessly integrated into novel environments along arbitrary camera trajectories.

but also accurately simulating complex light-transport phenomena. Fundamentally, this relies on inverse rendering, the process of disentangling intrinsic scene properties (i.e., geometry, illumination, and materials) from 2D images. However, because shape, texture, and lighting are inextricably coupled in the observed radiance, decoupling illumination from surface materials constitutes a highly ill-posed problem. Consequently, acquiring accurate and multi-view consistent Physically Based Rendering (PBR) materials from standard imagery remains a formidable challenge, often limiting the physical correctness of rendered objects under drastic viewpoint and environmental changes.

To acquire a renderable object model suitable for diverse environments, mainstream solutions generally follow a "reconstruct-then-render" paradigm. These approaches can be broadly categorized into traditional mesh-based pipelines and recent radiance-based neural rendering methods. Traditional multi-view reconstruction techniques [30, 34, 35] and contemporary generative 3D models [5, 22, 45, 46] are highly effective at establishing robust geometric scaffolds. However, relying on these geometries to further recover complex Physically-based Rendering (PBR) materials involves brittle, multi-stage optimizations, making it exceedingly difficult to restore accurate, high-fidelity surface materials. Conversely, neural rendering paradigms like NeRF [28] and 3D Gaussian Splatting (3DGS) [17] excel at Novel View Synthesis (NVS) for specific captured scenes. Yet, they inherently bake the original environmental illumination into their representations. Consequently, rendering these objects in novel lighting environments is non-trivial and necessitates complex inverse rendering to disentangle the baked lighting. In essence, while recovering coarse 3D geometry has become increasingly tractable, especially with the emergence of feedforward 3D reconstruction models [6,14,16,21,39,41,42], accurately modeling and decoupling surface appearance remains a critical bottleneck across both traditional and neural reconstruction frameworks, severely hindering the acquisition of fully renderable object models.

A highly promising direction has emerged with reference-based video diffusion models [8, 19, 26, 51]. By leveraging the robust visual priors encapsulated

within foundation models, these approaches can directly synthesize high-fidelity videos of a target object under diverse, complex lighting conditions based simply on a 2D reference image. This elegantly bypasses the notoriously difficult process of explicitly decoupling physical materials and illumination. However, despite their impressive generative capabilities, these models are still far from serving as reliable object renderer. First, they typically rely on a single-view reference; consequently, the appearance of unseen regions is left entirely to the model's unconstrained hallucination, often destroying multi-view consistency. Second, they inherently lack precise viewpoint controllability, making it exceedingly difficult to dictate strict camera trajectories. In summary, while reference-based diffusion models successfully avoid explicit physical modeling and achieve high-quality video synthesis, their inherent lack of spatial determinism and multi-view constraints prevents them from functioning as a controllable object renderer.

In this paper, by combining explicit 3D representations with expressive video generation models, we introduce GO-Renderer (Generative Object Renderer), a novel framework designed to construct a fully controllable object renderer from sparse images. Given a few reference images of a target object, a desired relative camera trajectory, and a novel environmental illumination (specified via texts, images, or videos), GO-Renderer synthesizes high-fidelity, multi-view consistent videos for the object.

The core insight of our approach lies in leveraging a fast, albeit coarse, geometric reconstruction, like ReconViaGen [5] or VGGT [39], to build a structural 3D proxy. This 3D proxy not only acts as a robust geometric guidance signal to enforce precise, spatially deterministic viewpoint control within a video diffusion model. Meanwhile, it also functions as a spatial bridge to seamlessly aggregate features from multiple reference images, effectively mitigating the hallucination issues of single-view generation and ensuring strict multi-view consistency. By marrying the explicit spatial constraints of a 3D proxy with the powerful visual priors of generative models, GO-Renderer elegantly bypasses the notoriously ill-posed physical modeling of complex materials and lighting. Our method enables highly controllable, photorealistic novel view and lighting synthesis, establishing a highly promising new direction for generative object rendering.

To explicitly realize this conceptual 3D proxy within a generative framework, our primary technical contribution lies in formulating the geometric guidance of 3D proxy as object-centric coordinate maps. Specifically, we render the coarse 3D reconstruction into dense coordinate maps for both the input reference views and the target camera trajectory. In these coordinate maps, the color of each pixel directly encodes its 3D coordinates in the object's coordinate system. We then channel-wise concatenate the target coordinate map with the latent noise, while similarly pairing the reference images with their corresponding reference coordinate maps. By feeding these spatially aligned conditions into the diffusion model, we establish a dense, pixel-level correspondence. The diffusion model essentially learns to "look up" the correct appearance from the reference images based on shared 3D coordinates. This elegant design not only enforces strict

spatial constraints for precise viewpoint control but also guarantees highly consistent texture transfer across varying camera poses.

Finally, training a robust object renderer requires large-scale data containing multi-view reference images paired with target videos under varying environmental lighting and known camera poses. Existing datasets (e.g., CO3D [32], OmniObject3D [44]) typically feature static illumination and lack the necessary lighting variations. To overcome this critical data scarcity, we constructed a comprehensive large-scale dataset. By leveraging advanced video generation models and synthetic rendering pipelines, we synthesized 57k high-quality video clips with diverse lighting conditions and corresponding proxy geometries, providing a robust foundation for training our generative renderer.

We extensively evaluate GO-Renderer across two challenging settings: object video synthesis under novel environmental illumination and novel view synthesis (NVS) under original lighting. Our method significantly outperforms current state-of-the-art diffusion [1] or reconstruction methods [5, 16]. Furthermore, we demonstrate the compelling practical utility of GO-Renderer through downstream applications like seamless video object insertion, where our method naturally harmonizes the target object with complex, dynamic backgrounds while maintaining strict geometric and illumination consistency.

2 Related Work

2.1 Reconstruction and Rendering

While recent image-domain methods have explored 3D integration—such as RefAny3D [13], which uses dual-branch perception to condition image generation on 3D asset point maps, AGP [20], which fine-tunes 2D generation via viewpoint-specific coordinate maps, thereby maintaining the pure generative nature of appearance synthesis without explicit 3D optimization, and IC-Light [47], which parses lighting priors to relight 3D scenes via Gaussian Splatting—mainstream object-centric video rendering still heavily relies on explicitly reconstructing the 3D model of the object first.

3D Object Reconstruction. 3D reconstruction methods have long dominated novel-view synthesis. NeRF [28] pioneered the use of continuous volumetric scene functions optimized via MLPs for view synthesis, while 3D Gaussian Splatting (3DGS) [17] revolutionized the field by enabling real-time rendering through interleaved optimization of 3D Gaussians with anisotropic covariance. To eliminate the need for costly per-scene optimization, feed-forward networks have been proposed. Models like AnySplat [16] and VOLSplat [41] predict Gaussian primitives directly from uncalibrated image collections in a single forward pass. Similarly, VGGT [39] and DepthAnything3 [25] leverage transformer backbones to infer 3D attributes (camera parameters, point maps) from arbitrary visual inputs, and Pi3 [42] introduces a permutation-equivariant architecture to predict affine-invariant poses without requiring a fixed reference view. On the generative side [22, 45, 46] synthesize high-fidelity 3D meshes and structured latent

representations from images or text. Additionally, ReconViaGen [5] integrates multi-view reconstruction priors with diffusion-based generation to address the hallucination constraints and inconsistencies of pure generative models.

Despite these advancements, a critical limitation remains: while explicit methods guarantee multi-view geometric consistency, the reconstruction process inevitably causes a loss of precision, texture details, and material properties. For instance, explicit representations like 3DGS [17] struggle with dynamic relighting and interactive editing, often leading to degradation of high-frequency texture details. Conversely, purely generative reconstruction methods [5] often produce outputs with inevitable appearance and geometric inconsistencies.

Rendering and Relighting. Once a 3D model is obtained, generative methods are widely employed to render the subject in novel scenes with distinct strategies: DiffusionRenderer [23] estimates G-buffers for photorealistic forward rendering; Diffusion as Shader [11] leverages 3D tracking videos to impose spatial constraints; UniLumos [27] and Light-A-Video [54] focus on geometry-guided correction and training-free temporally smooth relighting, respectively; while RelightMaster [3] and RelightVid [10] utilize Multi-plane Light Images and in-the-wild augmentations to tackle paired-data scarcity. Ultimately, the performance of these cascaded approaches remains strictly bottlenecked by the precision of the preceding 3D reconstruction stage.

2.2 Video Diffusion Models

The rapid evolution of video diffusion models encompasses both foundational base models and methods tailored for controllable video generation. Recent breakthroughs in video foundation models have significantly pushed the boundaries of generation through the widespread adoption of Diffusion Transformer (DiT) architectures. Various models have emerged with distinct and complementary capabilities: Wan [38] demonstrates the scaling laws of video generation with up to 14 billion parameters for state-of-the-art visual performance; LongCat [37] and Open-Sora-Plan [24] excel in efficient, high-resolution long video generation; Hunyuan Video [18] optimizes consumer-grade efficiency while maintaining highly coherent motion; and LTX-Video [12] extends the generative scope by introducing native synchronized audiovisual generation.

Geometry-aware Controllable Generation. Building upon these foundation models, recent frameworks have increasingly incorporated 3D geometric conditions to achieve precise spatial and camera control. For instance, GEN3C [33] integrates 3D-informed representations to maintain world consistency under explicit camera trajectories. ViewCrafter [48] leverages point cloud renderings to guide video diffusion models, facilitating high-fidelity Novel View Synthesis (NVS). Similarly, Diffusion as Shader (DaS) [11] formulates the diffusion process as a neural shader, utilizing 3D-aware conditions (e.g., geometry and lighting) to control versatile video generation.

While these approaches demonstrate capabilities in scene-level novel view synthesis, they are sub-optimal for object-centric generative rendering. These methods predominantly condition on single images or temporally adjacent frames,

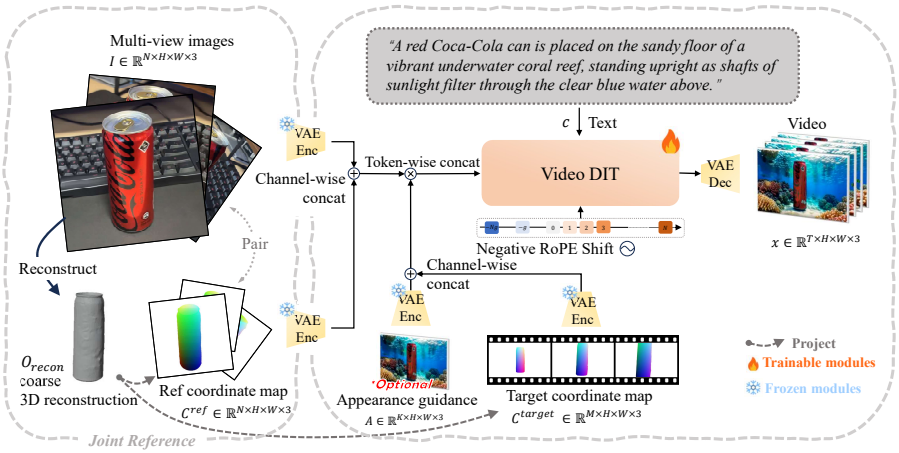

Fig. 2: Overview of GO-Renderer. Given multi-view reference images, we first establish a coarse 3D geometry via 3D reconstruction to render reference and target coordinate maps. These maps bridge explicit 3D geometry with the multi-view reference images via video diffusion model. Specifically, the reference images and their coordinate maps are paired via channel-wise concatenation, while the target coordinate maps are combined with optional appearance guidance. These spatial conditions are then jointly processed by a text-conditioned Video Diffusion Transformer to render high-fidelity object videos with precise viewpoint control.

limiting their capacity to extract and fuse comprehensive appearance priors from unconstrained multi-view references. Consequently, when synthesizing a specific object under complex, arbitrary camera trajectories within entirely novel environments, they struggle to maintain strict multi-view consistency and preserve fine-grained textures without resorting to explicit, lossy appearance modeling.

Subject-driven Generation. Subject-driven generation seeks to end-to-end synthesize videos of a specific subject in novel environments, typically conditioned on a single object view. Recent advances highlight various unique capabilities: Phantom [26] aligns text-image-video triplets to prevent multi-subject confusion; MAGREF [8] utilizes region-aware masked guidance to alleviate copy-paste artifacts; and StoryMem [51] maintains cross-shot storytelling consistency via a visual memory bank and negative RoPE shifts. However, relying purely on single-view references inherently restricts their ability to precisely control viewpoint transformations. Due to the limited 3D geometric comprehension of foundational 2D models, generating dynamic viewpoint changes in these methods frequently results in severe geometric distortions and anatomically incorrect content.

3 Methodology

3.1 Overview

We aim to construct an object renderer for rendering high-fidelity object videos with precise viewpoint control from sparse visual inputs. Explicit 3D reconstruction pipelines can establish robust geometry but struggle with the extraction of complex surface materials, whereas reference-based video diffusion models can synthesize realistic illumination but inherently lack multi-view consistency. To overcome these limitations, we propose a hybrid framework that uses a coarse 3D proxy to provide explicit spatial constraints while leveraging the powerful visual priors of generative models for high-fidelity appearance synthesis.

Specifically, given a set of reference images $\mathcal{I} = \{I_j\}_{j=1}^{N}$ of an object, a sequence of target poses $\mathcal{P}^{\text{target}} = \{P_k^{\text{target}}\}_{k=1}^{M}$, text prompts c and optional appearance guidance videos $\mathcal{A}$ which describe the lighting environment, our target is to render a video x of the object corresponding to the given target poses in the given lighting environment

$$x = \mathcal{R}_{\text{GO-Renderer}}(\mathcal{I}, \mathcal{P}^{\text{target}}, \mathcal{A}, c). \tag{1}$$

To achieve this, we first run a fast 3D reconstruction method, e.g. ReconViaGen [5] or VGGT [39], to get the camera poses $\mathcal{P}$ for all reference images and a coarse 3D reconstruction $\mathcal{O}_{\text{recon}}$ which is either a mesh or a 3D point cloud. Then, we will construct a 3D proxy from the 3D reconstruction (Sec. 3.2) and utilize the constructed 3D proxy as condition to guide a video diffusion model to generate the corresponding object videos (Sec. 3.3).

3.2 Pose Representation via 3D Proxy

In this section, we construct a 3D proxy that serves as the condition for the subsequent video diffusion model. This constructed 3D proxy fulfills a critical dual purpose. First, it acts as a robust geometric guidance signal to enforce precise, spatially deterministic viewpoint control over the generated video sequence. Second, it functions as a spatial bridge to seamlessly aggregate appearance features from the multi-view reference images $\mathcal{I}$, effectively mitigating the hallucination issues inherent in single-view generation and ensuring strict multi-view consistency across the rendered frames.

Specifically, we formulate this 3D proxy using object-centric coordinate maps. Given the coarse 3D reconstruction $\mathcal{O}_{\text{recon}}$ and the estimated reference camera poses $\mathcal{P} = \{P_j\}_{j=1}^{N}$, we rasterize $\mathcal{O}_{\text{recon}}$ to render a set of reference coordinate maps $\mathcal{C}^{\text{ref}} = \{C_j^{\text{ref}}\}_{j=1}^{N}$. Similarly, we project $\mathcal{O}_{\text{recon}}$ onto the desired target camera trajectory $\mathcal{P}^{\text{target}} = \{P_k^{\text{target}}\}_{k=1}^{M}$ to obtain the corresponding target coordinate maps $\mathcal{C}^{\text{target}} = \{C_k^{\text{target}}\}_{k=1}^{M}$. In these coordinate maps, the RGB color value of each valid foreground pixel directly encodes its normalized (x, y, z) spatial coordinates within the object's local coordinate system.

3.3 Architecture

In this section, we introduce a Joint Reference Structure to integrate the aforementioned conditional variables into the video diffusion model. By leveraging the correspondence between the reference images $\mathcal{I}$ and the coordinate maps $\mathcal{C}^{\text{ref}}$, this mechanism accurately transfers the appearance features of the references $\mathcal{I}$ to the rendered video x across different target coordinate maps $\mathcal{C}^{\text{target}}$. Additionally, we introduce a Negative RoPE Shift mechanism to temporally isolate the spatial references from the target video sequence x, preventing transition artifacts.

Joint Reference Structure. As shown in the right part of Fig. 2, to construct the generation conditions using these coordinate maps, we pair the multi-view reference images $\mathcal{I}$ with their corresponding reference coordinate maps $\mathcal{C}^{\text{ref}}$ along the channel dimension. Concurrently, during the iterative denoising process, the target coordinate maps $\mathcal{C}^{\text{target}}$ and the optional appearance videos $\mathcal{A}$ describing the lighting environment are concatenated with the target latent noise along the channel dimension. By feeding these spatially aligned conditions into the diffusion model, we establish a dense, pixel-level correspondence between the source references and the target generation. Because the object's local coordinates are invariant to camera transformations, the diffusion model learns to "*look up*" the appearance and texture details from the reference views $\mathcal{I}$ based on the shared 3D coordinates encoded in $\mathcal{C}^{\text{ref}}$ and $\mathcal{C}^{\text{target}}$. This mechanism ensures consistent texture transfer across varying camera poses without requiring explicit physical modeling of surface materials.

Negative RoPE Shift. Video diffusion models normally apply temporal positional embeddings (PE) to the input sequence. If the reference sequence and the main sequence share a continuous PE formulation, two issues arise. First, the continuous temporal indices force the model to interpret the reference images as immediate historical frames. Consequently, the first generated frame tends to exhibit transition artifacts as it attempts to interpolate from the reference views. Second, the shared contiguous PE scale degrades the model's capability to distinguish between the conditioning reference priors and the target generation sequence. To address these issues, and inspired by Ominicontrol [36, 51], we modify the 3D Rotary Positional Embedding (RoPE) by assigning negative, discrete temporal indices to the reference latents. Given $\mathcal{N}$ reference views and $\mathcal{M}$ target frames, the temporal positional indices are formulated as:

$$\{-N \cdot g, -(N-1) \cdot g, \ldots, -g, 0, 1, \ldots, M-1\} \tag{2}$$

where g is a predefined temporal interval gap. By utilizing negative indices with an explicit gap, we preserve the zero-based temporal encoding for the target video x. Concurrently, the reference views are explicitly separated and embedded as global spatial conditions, preventing the network from treating them as adjacent historical frames.

3.4 Dataset Construction

Training GO-Renderer requires a dataset that provides multi-view reference images of objects, high-quality target video sequences with camera motions conforming to real-world distributions, and accurate 3D pose annotations. However, existing datasets exhibit structural limitations that hinder their direct application to this task.

Current general video datasets, such as OpenVid [29] and Koala36M [40], provide dynamic scenes but lack 3D proxy annotations and predominantly feature stationary cameras. Meanwhile, existing real-world object-centric datasets (*e.g.*, CO3D [32], OmniObject3D [44], and MVImageNet [49]) provide multi-view observations but typically operate under static lighting conditions with simple orbital camera motions, diverging from the varied trajectories observed in standard videos. Furthermore, 3D asset repositories such as Objectverse [7] and Texverse [52] contain detailed geometries and textures, but they consist of uncurated static models that necessitate extensive filtering and rendering pipelines to formulate the paired video sequences. Recently, multi-camera synthetic datasets like SynCamVideo [2] and OmniWorld [53] have been proposed to offer multi-view content with annotated camera trajectories; however, their primary focus lies on human subjects or large-scale scenes rather than object-centric scenarios.

To address this data scarcity, we construct a mixed dataset comprising data from multiple domains, including synthetic rendering from 3D assets, real-world extraction and high-fidelity AI-generated. This multi-source strategy mitigates the biases inherent in single-domain data, ensuring a balanced distribution of object categories, variable environmental lighting, and diverse camera trajectories.

Synthetic Rendering. We construct a synthetic dataset by stochastically composing diverse assets via Blender. Specifically, the 3D object models are sourced from Google Scanned Objects [9] and TexVerse [52], while the environmental backgrounds are sampled from a curated collection of over 40 scene templates and 100 high-dynamic-range imaging (HDRI) maps. This rendering pipeline explicitly outputs object videos x featuring random camera motions, alongside the perfectly aligned multi-view reference images and their precise spatial poses.

Real-World Extraction and AI-Generated. To capture complex real-world dynamics, we extract an object-centric subset from the OpenVidHD [29] and CO3D [32] dataset. For each target video, we employ SAM3 [4] to segment the object of interest and utilize ReconViaGen to reconstruct its corresponding 3D proxy. Subsequently, FoundationPose is applied to estimate the frame-wise spatial configurations. To construct the multi-view reference images, we heuristically select a subset of video frames that provide the most comprehensive viewpoint coverage of the target object. Furthermore, we also utilize Wan2.2 to synthesize high-quality, highly diverse object-centric videos. These generated videos are then processed through the identical extraction and annotation pipeline as the real-world data (i.e., SAM3 [4], ReconViaGen [5], and FoundationPose [43]) to obtain the reference images, corresponding 3D proxy, and object poses.

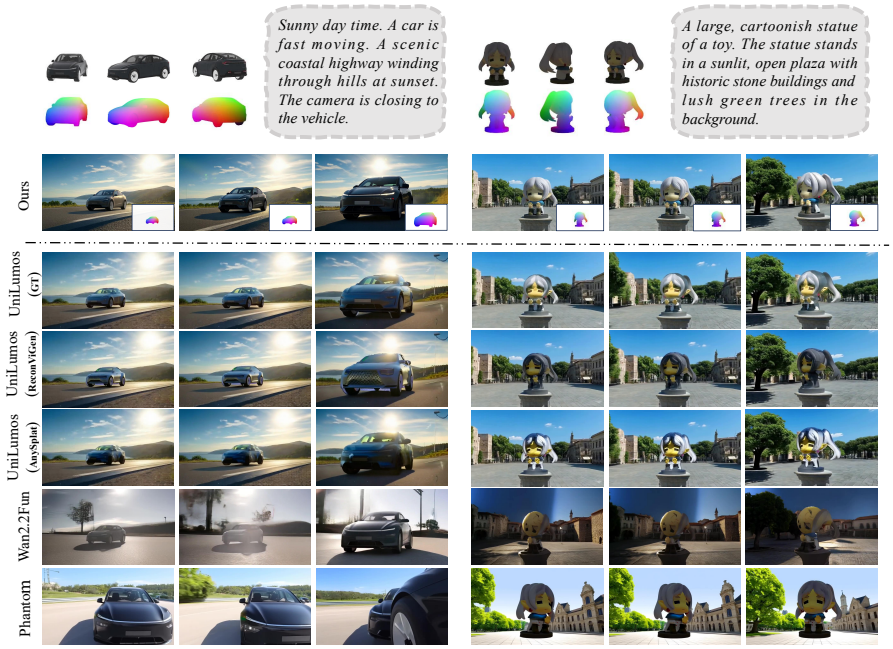

Fig. 3: Comparision of rendering quality and lighting consistency.

4 Experiments

The experimental evaluation validates GO-Renderer on diverse synthetic and real-world benchmarks. We assess the framework across several key dimensions: Rendering Quality and Lighting Consistency (Sec. 4.2), and the Multi-view Consistency of rendered objects (Sec. 4.3). Furthermore, we perform ablation studies to verify the efficacy of specific architectural components (Sec. 4.4) and demonstrate practical downstream applications (Sec. 4.5). Extensive additional experiment and ablations are deferred to the supplementary material.

4.1 Implementation Details

We implement GO-Renderer based on the pre-trained *Wan2.2Fun 5B Ref Control* [1] architecture. The training process is conducted on 16 NVIDIA A800 GPUs for a total of 20,000 steps with a batch size of 32, utilizing the AdamW optimizer. Specifically, the model is trained on video sequences featuring a spatial resolution of 480×832 and a temporal length of 81 frames. During the training phase, to construct the multi-view reference $\mathcal{I}$, we randomly sample 3 to 8 images from the reference sources. To further enhance the model's robustness and prevent overfitting to specific spatial arrangements, we apply several targeted augmentation strategies: we randomly shuffle the sequence order of the reference views and introduce slight random scaling as well as spatial translation to

the positional coordinates of the 3D proxy. On a 48GB RTX 4090, GO-Renderer generates a 480×832 video with 81 frames in about 58 seconds, with a peak GPU memory usage of 31.8GB.

4.2 Rendering Quality and Lighting Consistency

This section evaluates the proposed method against state-of-the-art approaches in terms of rendering quality and lighting consistency under novel viewpoints and environments. Given the absence of identical end-to-end frameworks for this specific task, we construct representative proxy baselines categorized into mainstream two-stage "reconstruct-then-render" pipelines and subject-driven video generation models.

Baselines. We select the baselines based on their relevance to object rendering and spatial composition, categorizing them into mainstream two-stage pipelines and subject-driven generative models. For the two-stage pipelines, which explicitly separate 3D modeling from environmental relighting, we select UniLumos [27] as the standard compositing and relighting backbone. To decouple reconstruction errors from relighting performance, we construct three variants: (1) **UniLumos(GT)**, which utilizes ground-truth geometry and foreground textures to establish the upper bound of the relighting module; (2) **UniLumos(ReconViGen [5])**, coupled with generative 3D priors; and (3) **UniLumos(AnySplat [16])**, integrating feed-forward 3DGS for explicit reconstruction. To represent the alternative paradigm of 2D reference-based generation, we employ Phantom [26] and *Wan2.2Fun 5B Ref Control* [1]. These subject-driven models utilize visual conditioning to inject object features into new backgrounds. As they lack intrinsic support for multi-view inputs, we select a single optimal reference view from $\mathcal{I}$ as their visual condition.

Evaluation. We employ dataset-specific evaluation metrics. For synthetic data with available ground truth, we compute PSNR and SSIM to evaluate pixel-level fidelity. For real-world captures lacking ground truth, we utilize VBench++ [15] to assess perceptual quality, temporal consistency, and illumination naturalness.

Results. As shown in Tab. 1 and Fig. 3, GO-Renderer consistently outperforms all baselines across both physical-structural and perceptual metrics. Visually, two-stage pipelines exhibit distinct limitations in complex relighting: UniLumos(GT) presents observable color shifts despite using ground-truth foregrounds, while variants relying on practical 3D reconstruction suffer from further degradation due to error accumulation. For the subject-driven generation paradigm, Phantom leverages the priors of video diffusion models to yield competitive smoothness and aesthetic scores. However, lacking explicit 3D geometric constraints, it cannot maintain strict multi-view consistency or adhere to precise camera trajectories. This spatial deviation directly contributes to its lower PSNR and SSIM scores. By integrating explicit 3D proxy guidance with 2D visual priors, GO-Renderer avoids the structural deviations of 2D models and the lighting artifacts of the two-stage pipelines, ensuring consistent high-fidelity rendering.

Table 1: Quantitative comparison of Rendering Quality and Lighting Consistency. **Bold** number indicate the best performance.

Method	PSNR↑	SSIM↑	Text Align↑	Motion Smoothness↑	Aesthetic Quality↑	Imaging Quality↑
Phantom [26]	11.400	0.514	24.620	0.984	0.568	0.709
UniLumos(GT) [27]	12.060	0.585	25.630	0.978	0.569	0.669
UniLumos(ReconViGen [5])	11.010	0.338	25.870	0.975	0.571	0.666
UniLumos(AnySplat [16])	11.160	0.362	25.930	0.975	0.557	0.624
Wan2.2Fun	14.300	0.641	25.250	0.984	0.530	0.601
Ours	**18.260**	**0.684**	**27.720**	**0.988**	**0.573**	**0.712**

Table 2: Quantitative comparison of Multi-view Consistency and Reconstruction. **Bold** numbers indicate the best performance.

Method	CLIP		DINO	
	Img/Avg.↑	Img/Max.↑	Avg.↑	Max.↑
AnySplat	0.861	0.934	0.191	0.323
ReconViaGen	0.796	0.927	0.448	0.727
Ours	**0.888**	**0.956**	**0.725**	**0.860**

Fig. 4: Qualitative comparison of Multi-view Consistency.

4.3 Multi-view Consistency

To evaluate multi-view consistency while decoupling it from the confounding factors of environmental relighting, we isolate the object rendering process against canonical blank backgrounds. This controlled setup strictly assesses the preservation of intrinsic object appearance across varying viewpoints.

Baselines. We construct the baselines to represent two primary paradigms of 3D reconstruction suitable for novel view synthesis. Specifically, we select

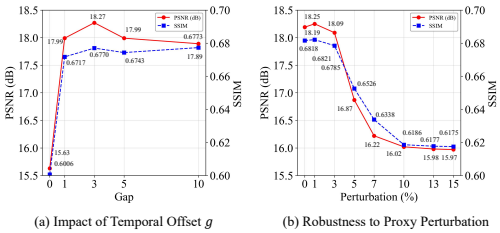

Fig. 5: Quantitative Ablation Results. (a) Impact of the temporal offset g on rendering quality. A sufficient gap ($g = 3$) effectively isolates spatial priors. (b) Model robustness under varying degrees of spatial perturbations.

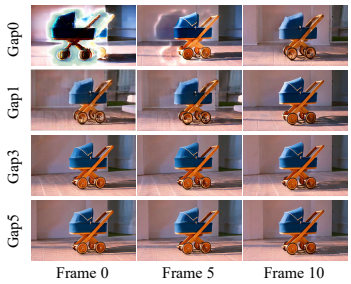

Fig. 6: Qualitative ablation results of temporal offset.

AnySplat [16] as the representative of feed-forward explicit 3DGS, and ReconViaGen [5] to represent generative 3D modeling. For both baselines, the multi-view reference images are utilized to reconstruct the underlying 3D representations. The reconstructed models are subsequently rendered along the target pose sequence to synthesize the video frames. This setup establishes a direct comparison between conventional 3D reconstruction-based pipelines and our generative rendering framework.

Evaluation. We employ a comprehensive set of metrics to quantify both the perceptual and physical fidelity of the synthesized novel views. To evaluate semantic and structural consistency, we calculate the CLIP [31] similarity to measure the high-level semantic alignment between the rendered frames and the multi-view references, while utilizing the DINO [50] score to assess fine-grained structural preservation and spatial layout correspondences. Furthermore, we introduce PSNR and SSIM to evaluate pixel-level reconstruction accuracy and geometric fidelity under canonical setups. Detailed calculation procedures and formulations for all utilized metrics are provided in the supplementary material.

Results. As shown in Tab. 2. GO-Renderer achieves the best performance across all evaluated metrics. AnySplat exhibits a noticeable degradation across the structural (DINO) and physical fidelity metrics (PSNR and SSIM). As shown in Fig. 4, the 3DGS reconstruction suffers from blurriness and distortion. Conversely, while ReconViaGen demonstrates distinct performance patterns, it falls short in overall semantic alignment and pixel-level accuracy. The synthesized geometry and texture details lack strict consistency with the reference inputs. By integrating explicit 3D proxy guidance with 2D diffusion priors, our method circumvents the blurriness of explicit 3DGS reconstruction and the geometric-texture inconsistencies of generative pipelines, thereby maintaining robust identity preservation and high-fidelity rendering.

4.4 Ablation Study

Impact of Temporal Offset (g). To evaluate the sensitivity of the negative temporal indices g, we test various values of $g \in \{0, 1, 3, 5, 10\}$. As illustrated in

Fig. 7: Application of our GO-Renderer.

Fig. 5(a), setting $g = 0$ results in a severe performance drop (PSNR drops to 15.63). As shown in Fig 6 without a sufficient temporal buffer, the pre-trained DiT natively attempts to interpolate between the static reference view and the first target frame, leading to noticeable smearing artifacts. However, simply introducing a gap of $g = 1$ significantly recovers the performance. We observe that $g = 3$ achieves the optimal quantitative results, effectively isolating the spatial priors while maintaining strong conditional guidance.

Robustness to Proxy Quality. To evaluate the robustness of GO-Renderer, we simulate tracking inaccuracies by injecting varying levels of random spatial perturbations into the target coordinate maps during inference. As shown in Fig. 5(b), the model maintains stable performance under minor spatial noise, demonstrating its capacity to absorb slight geometric jitters. As the perturbation level increases, the evaluation metrics exhibit a graceful degradation rather than a sharp collapse. Under severe noise injection, the performance stabilizes into a distinct plateau. This trend suggests that when the 3D proxy becomes heavily corrupted, the model reduces its reliance on unreliable geometric cues and instead depends on the 2D diffusion prior to synthesize a plausible appearance, thereby avoiding severe rendering failures.

4.5 Applications

We demonstrate the practical utility of GO-Renderer in two downstream scenarios: offline rendering and object insertion, as depicted in Fig. 7. On the the first row, we deploy GO-Renderer as an offline renderer plug-in within Blender. It synthesizes high-fidelity object animations driven by text prompts c and explicit camera trajectories exported directly from the 3D editor software. On the second row, we showcase the seamless insertion of 3D models into real-world video sequences. The inserted objects naturally blend into the physical scenes, exhibiting highly realistic physical interactions, including accurate reflections and plausible shadows.

5 Limitation

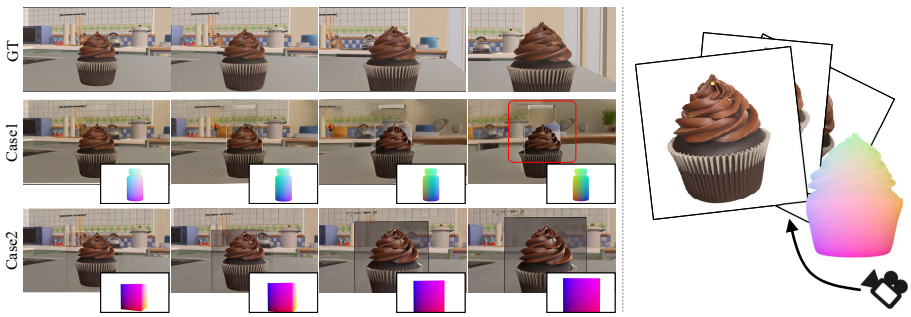

Fig. 8: Failure cases. The top row (GT) displays the ground-truth sequence. In Case 1 and Case 2, the model is conditioned on the multi-view appearances of a cupcake, but provided with the explicit spatial coordinate maps of a bottle and a cube, respectively. This geometric inconsistency prevents the model from reconciling the conflicting spatial and appearance conditions, leading to structural artifacts and degraded rendering fidelity.

Although our ablation studies demonstrate that GO-Renderer exhibits robustness to spatial perturbations by leveraging 2D diffusion priors, the accuracy and structural fidelity of the underlying 3D proxy still impact the final rendering results. Specifically, as illustrated in Fig. 8, utilizing an incorrect or structurally mismatched 3D proxy compromises the generative output. Under such conditions, the model struggles to reconcile the conflicting spatial conditions with the reference appearances, leading to visual artifacts and a quantifiable degradation in viewpoint control accuracy. Therefore, acquiring corresponding 3D geometries remains a prerequisite for strictly controlled generation, which we aim to address in future work by integrating more advanced proxy extraction modules.

6 Conclusion

We propose GO-Renderer, a unified framework integrating the reconstructed 3D proxies to guide the video generative models to achieve high-quality object rendering on arbitrary viewpoints under arbitrary lighting conditions. Our method not only enjoys the accurate viewpoint control using the reconstructed 3D proxy but also enables high-quality rendering in different lighting environments using diffusion generative models without explicitly modeling complex materials and lighting.

References

1. aigc apps: Videox-fun: A more flexible framework that can generate videos at any resolution and creates videos from images (2024)

2. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. arXiv preprint arXiv:2412.07760 (2024)
3. Bian, W., Shi, X., Huang, Z., Bai, J., Wang, Q., Wang, X., Wan, P., Gai, K., Li, H.: Relightmaster: Precise video relighting with multi-plane light images. arXiv preprint arXiv:2511.06271 (2025)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
5. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. arXiv preprint arXiv:2510.23306 (2025)
6. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-view Images, p. 370–386. Springer Nature Switzerland (Oct 2024). https://doi.org/10.1007/978-3-031-72664-4_21
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. arXiv preprint arXiv:2212.08051 (2022)
8. Deng, Y., Yin, Y., Guo, X., Wang, Y., Fang, J.Z., Yuan, S., Yang, Y., Wang, A., Liu, B., Huang, H., Ma, C.: Magref: Masked guidance for any-reference video generation with subject disentanglement. arXiv preprint arXiv:2505.23742 (2025)
9. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. arXiv preprint arXiv:2204.11918 (2022)
10. Fang, Y., Sun, Z., Zhang, S., Wu, T., Xu, Y., Zhang, P., Wang, J., Wetzstein, G., Lin, D.: Relightvid: Temporal-consistent diffusion model for video relighting. arXiv preprint arXiv:2501.16330 (2025)
11. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., Wang, W., Liu, Y.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847 (2025)
12. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
13. Huang, H., Bao, Q., Gu, Z., Du, Z., Lin, C., Liu, Y., Yang, S.: Refany3d: 3d asset-referenced diffusion models for image generation. arXiv preprint arXiv:2601.22094 (2026)
14. Huang, J., Yang, Y., Yang, B., Ma, L., Ma, Y., Liao, Y.: Gen3r: 3d scene generation meets feed-forward reconstruction. arXiv preprint arXiv:2601.04090 (2026)
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. arXiv preprint arXiv:2311.17982 (2023)

16. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., Lin, D., Dai, B.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. arXiv preprint arXiv:2505.23716 (2025)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. arXiv preprint arXiv:2308.04079 (2023)
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2025)
19. Li, D., Fei, Z., Li, T., Dou, Y., Chen, Z., Yang, J., Fan, M., Xu, J., Wang, J., Gu, B., Chang, M., Cai, W., Xie, Y., Mao, B., Zhang, Y., Pang, N., Zhang, H., Jin, Y., Xu, Z., Lin, D., Chen, G., Zhou, Y.: Skyreels-v3 technique report. arXiv preprint arXiv:2601.17323 (2026)
20. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. arXiv preprint **abs/2310.02596** (2023)
21. Li, X., Wang, T., Gu, Z., Zhang, S., Guo, C., Cao, L.: Flashworld: High-quality 3d scene generation within seconds. arXiv preprint arXiv:2510.13678 (2025)
22. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., Cao, Y.P.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv:2502.06608 (2025)
23. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, Z.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., Wang, Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. arXiv preprint arXiv:2501.18590 (2025)
24. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., Jia, T., Zhang, J., Tang, Z., Pang, Y., She, B., Yan, C., Hu, Z., Dong, X., Chen, L., Pan, Z., Zhou, X., Dong, S., Tian, Y., Yuan, L.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
25. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
26. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
27. Liu, R., Yuan, H., Dong, B., Xing, J., Wang, J., Zhao, R., Xing, Y., Chen, W., Wang, F.: Unilumos: Fast and unified image and video relighting with physics-plausible feedback. arXiv preprint arXiv:2511.01678 (2025)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. arXiv preprint arXiv:2003.08934 (2020)
29. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2025)
30. Pan, L., Barath, D., Pollefeys, M., Schönberger, J.L.: Global Structure-from-Motion Revisited. In: European Conference on Computer Vision (ECCV) (2024)

31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
32. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. arXiv preprint arXiv:2109.00512 (2021)
33. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. arXiv preprint arXiv:2503.03751 (2025)
34. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
35. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
36. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2025)
37. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., Zhang, T.: Longcat-video technical report. arXiv preprint arXiv:2510.22200 (2025)
38. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
39. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. arXiv preprint arXiv:2503.11651 (2025)
40. Wang, Q., Shi, Y., Ou, J., Chen, R., Lin, K., Wang, J., Jiang, B., Yang, H., Zheng, M., Tao, X., Yang, F., Wan, P., Zhang, D.: Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. arXiv preprint arXiv:2410.08260 (2025)
41. Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Zhu, Z., Chen, D.Y., Zhuang, B.: Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. arXiv preprint arXiv:2509.19297 (2025)
42. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.:
pi^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
43. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. arXiv preprint arXiv:2312.08344 (2024)
44. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. arXiv preprint arXiv:2301.07525 (2023)
45. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. arXiv preprint arXiv:2412.01506 (2025)

46. Yang, X., Shi, H., Zhang, B., Yang, F., Wang, J., Zhao, H., Liu, X., Wang, X., Lin, Q., Yu, J., Wang, L., Xu, J., He, Z., Chen, Z., Liu, S., Wu, J., Lian, Y., Yang, S., Liu, Y., Yang, Y., Wang, D., Jiang, J., Guo, C.: Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. arXiv preprint arXiv:2411.02293 (2025)
47. Ye, J., Zhuang, J., Mu, L., Zheng, W., Hu, J., Zou, X., Wang, J., Hu, H.: Training-free multi-view extension of ic-light for textual position-aware scene relighting. arXiv preprint arXiv:2511.13684 (2025)
48. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
49. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., Chen, G., Cui, S., Han, X.: Mvimgnet: A large-scale dataset of multi-view images. arXiv preprint arXiv:2303.06042 (2023)
50. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
51. Zhang, K., Jiang, L., Wang, A., Fang, J.Z., Zhi, T., Yan, Q., Kang, H., Lu, X., Pan, X.: Storymem: Multi-shot long video storytelling with memory. arXiv preprint arXiv:2512.19539 (2025)
52. Zhang, Y., Zhang, L., Ma, R., Cao, N.: Texverse: A universe of 3d objects with high-resolution textures. arXiv preprint arXiv:2508.10868 (2025)
53. Zhou, Y., Wang, Y., Zhou, J., Chang, W., Guo, H., Li, Z., Ma, K., Li, X., Wang, Y., Zhu, H., Liu, M., Liu, D., Yang, J., Fu, Z., Chen, J., Shen, C., Pang, J., Zhang, K., He, T.: Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. arXiv preprint arXiv:2509.12201 (2025)
54. Zhou, Y., Bu, J., Ling, P., Zhang, P., Wu, T., Huang, Q., Li, J., Dong, X., Zang, Y., Cao, Y., Rao, A., Wang, J., Niu, L.: Light-a-video: Training-free video relighting via progressive light fusion. arXiv preprint arXiv:2502.08590 (2025)