



You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

Baorui Ma*, Huachen Gao*, Haoge Deng*, Zhengxiong Luo, Tiejun Huang, Lulu Tang†, Xinlong Wang†

Beijing Academy of Artificial Intelligence (BAAI)

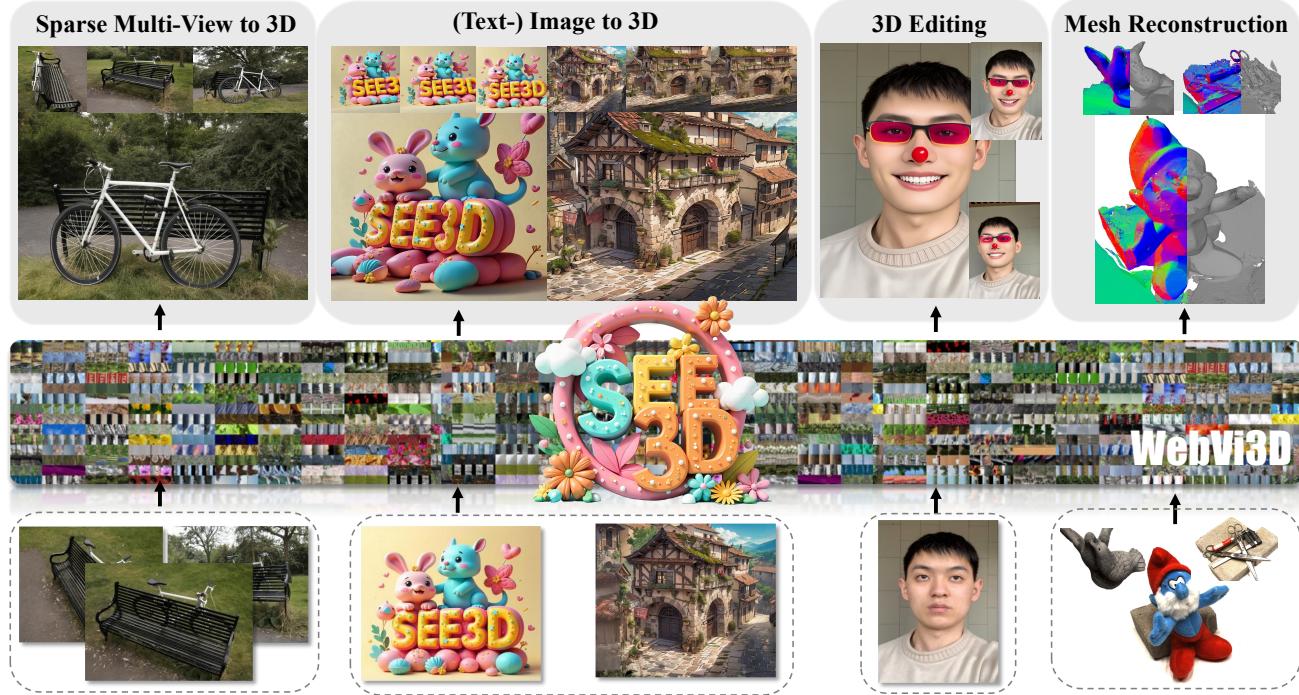


Figure 1. Benefiting from the proposed web-scale dataset WebVi3D, **See3D** enables both object- and scene-level 3D creation, including sparse-view-to-3D, (text-) image-to-3D, and 3D editing. It can also be used for Gaussian Splatting to extract meshes or render images.

Abstract

Recent 3D generation models typically rely on limited-scale 3D ‘gold-labels’ or 2D diffusion priors for 3D content creation. However, their performance is upper-bounded by constrained 3D priors due to the lack of scalable learning paradigms. In this work, we present **See3D**, a visual-conditional multi-view diffusion model trained on large-scale Internet videos for open-world 3D creation. The model aims to **Get 3D** knowledge by solely **Seeing the visual contents from the vast and rapidly growing video data** — *You See it, You Got it*. To achieve this, we first scale up the training data using a proposed data curation pipeline that automatically filters out multi-view inconsistencies and insufficient observations from source videos. This results in a high-quality, richly diverse, large-scale dataset of multi-

view images, termed **WebVi3D**, containing 320M frames from 16M video clips. Nevertheless, learning generic 3D priors from videos without explicit 3D geometry or camera pose annotations is nontrivial, and annotating poses for web-scale videos is prohibitively expensive. To eliminate the need for pose conditions, we introduce an innovative visual-condition - a purely 2D-inductive visual signal generated by adding time-dependent noise to the masked video data. Finally, we introduce a novel visual-conditional 3D generation framework by integrating **See3D** into a warping-based pipeline for high-fidelity 3D generation. Our numerical and visual comparisons on single and sparse reconstruction benchmarks show that **See3D**, trained on cost-effective and scalable video data, achieves notable zero-shot and open-world generation capabilities, markedly outperforming models trained on costly and constrained 3D datasets. Additionally, our model naturally supports other

*Equal contribution. † Correspondence to XW and LT.

image-conditioned 3D creation tasks, such as 3D editing, without further fine-tuning. Please refer to our project page at: <https://vision.baai.ac.cn/see3d>.

1. Introduction

Recent advances in 3D generation are essential for fields like virtual reality, entertainment, and simulation, offering the potential not only to recreate intricate real-world structures but also to expand human imagination. Nevertheless, developing these models is constrained by the scarcity and high costs of accessible 3D datasets. Despite recent industry efforts [94, 117, 125] create extensive proprietary 3D assets, these initiatives come with substantial financial and operational burdens. Currently, building such a large-scale 3D dataset for academia remains prohibitively expensive. This motivates us to pursue a scalable, accessible, and affordable data source that can compete with advanced closed-source solutions, thereby enabling the broader research community to train high-performance 3D generation models.

Human perception of the 3D world does not rely on specific 3D representation (e.g., point clouds[19], voxel grids [39], meshes [98], or neural fields [65]) or precise camera conditions. Instead, our 3D awareness is shaped by multi-view observations accumulated throughout our lives. This raises the question: *Can models similarly learn universal 3D priors from large collections of multi-view images?* Fortunately, Internet videos offer a rich source of multi-view images, captured from various locations with diverse sensors and complex camera trajectories, providing a scalable, accessible, and cost-effective data source. Thus, *how can we effectively learn 3D knowledge from Internet videos?*

The core challenges in achieving this goal are twofold: 1) filtering relevant, 3D-aware video data from raw sources, specifically static scenes with varied camera viewpoints that provide sufficient multi-view observations; and 2) learning generic 3D priors from videos lacking explicit 3D geometry and camera pose annotations (i.e. pose-free videos). This work carefully addresses these challenges and introduces a pose-free, visual-conditional multi-view diffusion (MVD) model, **See3D**, for open-world 3D creation.

Specifically, we establish a novel video data curation pipeline that automatically filters out data with dynamic content or restricted camera viewpoints from source videos. The resulting dataset, termed WebVi3D, comprises 15.99M video clips from 25.48M source videos, totaling 4.41 years in duration—orders of magnitude larger than previous 3D datasets, such as DLV3D (0.01M) [50], RealEstate10K (0.08M) [129], MVIImgNet (0.22M) [122] and Objaverse (0.8M) [15].

MVD models have recently gained widespread attention due to their advantages of integrating the generative capabil-

ities of 2D diffusion models while maintaining consistency across multiple views [51, 56, 80, 88, 128]. Typically, these models rely on precise camera poses [2, 23, 28, 33, 45, 54, 66, 80, 110, 111, 124] or warped images rendered according to camera position [95, 121] as conditional inputs to guide 3D-consistent view generation. We refer to these conditions, derived from pose or 3D annotations, as 3D-inductive conditions. However, annotating web-scale videos is prohibitively costly, or even intractable in some cases, posing significant challenges for scaling. To address this, we propose a novel, pose-free *visual-condition* derived from pixel-space hints within videos. It is a purely 2D-inductive visual signal, created by introducing *time-dependent noise* to masked input videos, free from any 3D-inductive bias. This enables training MVD model at scale, without requiring pose annotations.

Intuitively, the proposed *visual-condition* can generalize effectively to tasks that rely on pixel-space hints distinct from those in videos, such as warping-based 3D generation [12, 81] and mask-based 3D editing [10], without requiring additional training, see Fig. 1. For instance, in warping-based 3D generation, pixels from a reference image are rearranged through warping operations, creating a *specific* visual-condition to indicate camera movement. However, these warped images often exhibit artifacts or distortions, causing a significant domain gap compared to video frames. Whereas, our *visual-condition* functions as a *generic* one, capable of accommodating such unnatural images.

To further harness the potential of **See3D**, we introduce an innovative visual-conditional 3D generation framework utilizing a warping-based pipeline. This framework first constructs the *visual-condition* using **See3D**, then iteratively refines the geometry of novel views to build comprehensive scene observations. Finally, the generated images are used for Gaussian Splatting reconstruction [35, 41], which can be rendered from arbitrary viewpoints or converted into meshes through post-processing [59]. In summary, our key contributions are as follows.

- We present **See3D**, a scalable visual-conditional MVD model for open-world 3D creation, which can be trained on web-scale video collections without pose annotations.
- We curate WebVi3D, a multi-view images dataset containing static scenes with sufficient multi-view observations, and establish an automated pipeline for video data curation to train the MVD model.
- We introduce a novel warping-based 3D generation framework with **See3D**, which supports long-sequence generation with complex camera trajectories.
- We achieve state-of-the-art results in single and sparse views reconstruction, demonstrating remarkable zero-shot and open-world generation capability, offering a novel perspective on scalable 3D generation.

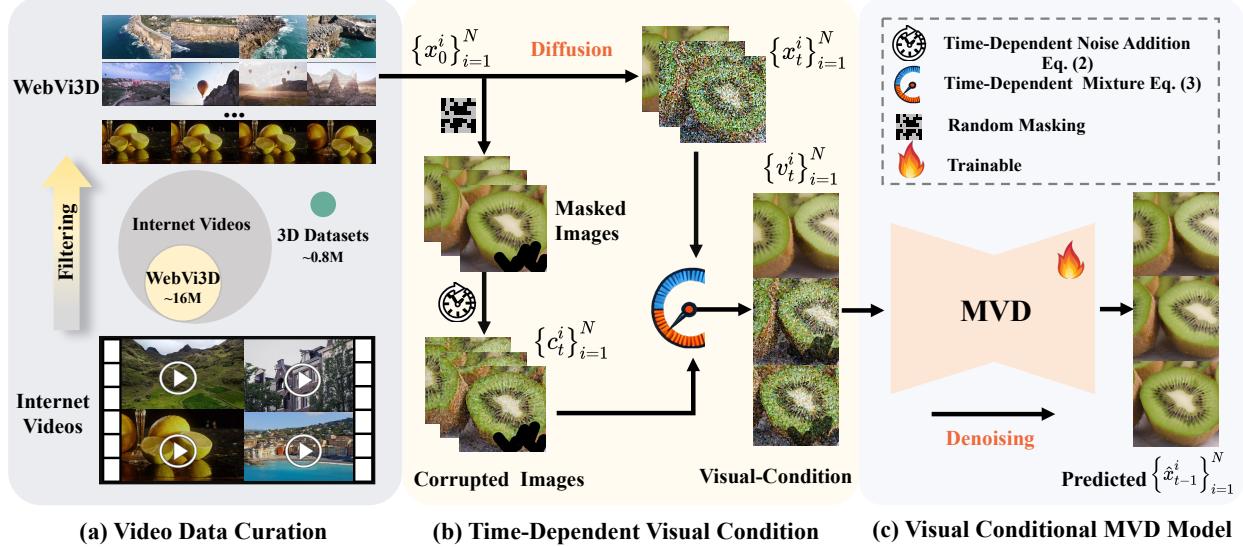


Figure 2. Overview of See3D. (a) We propose a four-step data curation pipeline to select multi-view images from Internet videos, forming the WebVi3D dataset, which includes $\sim 16M$ video clips across diverse categories and concepts. (b) Given multiple views, we corrupt the original data into corrupted images c_t^i at timestep t by applying random masks and time-dependent noise. We then reweight the guidance of c_t^i and the noisy latent x_t^i for the diffusion model to form *visual-condition* v_t^i through a time-dependent mixture. (c) MVD model is capable of training at scale to generate multi-view images conditioned on v_t^i , without requiring pose annotations. Since v_t^i is a task-agnostic visual signal formed through time-dependent noise and mixture, it enables the trained model to robustly adapt to various downstream tasks.

2. Related work

Lifting 2D Generation into 3D. Recent advances in 3D generation have been largely driven by the success of 2D diffusion models [31, 76, 84, 85], which have revolutionized image and video generation. These works typically optimize 3D representations by maximizing the likelihood evaluated by 2D diffusion priors [42, 48, 53, 63, 72, 87, 90, 103, 118]. An alternative approach uses a warping-inpainting pipeline, integrating an offline depth estimator with a 2D diffusion-based inpainting model to iteratively generate 3D content [12, 17, 32, 66, 97, 119, 121]. However, 2D priors do not readily translate into coherent 3D representations. As a result, 2D lifting-based approaches often struggle to preserve high geometric fidelity, leading to issues like multi-view inconsistency and poor global geometry [120].

Directly Learning 3D Priors. To better preserve geometric features, some works focus on directly learning 3D priors. For instance, feed-forward approaches [8, 11, 25, 33, 46, 47, 55, 60, 78, 89, 91, 94, 100, 106, 115, 116, 131, 132] take single/few views as input and directly output 3D representations using an encoder-decoder architecture, eliminating the need for additional optimization process per instance. Another line of research involves training diffusion models to predict 3D representations, such as point clouds [67, 123], mesh [1, 38, 61], and implicit neural representation [9, 62, 108, 125]. However, these methods generally fo-

cus on object-level generation [15, 91, 109, 125, 132], limiting their applicability to scene-level generation. Although recent research has made strides in building scene-level 3D datasets [3, 13, 43, 50], their scale remains relatively limited. The reliance on costly, limited-scale 3D datasets restricts generalization to open-world or highly imaginative scenarios. In contrast, our approach curates a large-scale, richly diverse dataset of multi-view images from Internet videos. By training the model at scale, it effectively supports both object-level and scene-level 3D creation.

Learning Multi-view Priors for 3D Generation. MVD model inherits the generative capabilities of 2D diffusion models while capturing multi-view correlations, achieving both generalizability and 3D consistency. These merits have made it a focal point in recent 3D generation research [23, 26, 52, 56, 58, 73, 77, 79, 80, 99, 121]. However, as 2D diffusion models are typically trained on 2D datasets, they lack precise control over image pose. To address this, MVD-based approaches often train their models on images paired with camera poses [24, 54, 77, 105, 107], where poses serve as essential conditional inputs, represented by camera extrinsics [77, 80], relative poses [54, 56, 79], or Plücker rays [23, 111]. Yet, pose-conditional models rely heavily on costly pose-annotated data, restricting training to smaller 3D datasets, thereby constraining their adaptability to out-of-distribution scenarios. In contrast, we introduce a novel visual-conditional approach that supports scalable,

pose-free MVD model training for open-world 3D generation.

3. Method

The primary objective of this work is to build a robust 3D generative model from the perspective of dataset scaling-up. Previous works [15, 75, 95] laboriously collect 3D data from designed artists, stereo matching, or Structure from Motion (SfM), which can be costly and sometimes infeasible. In contrast, multi-view images offer a highly scalable alternative, as they can be automatically extracted from the vast and rapidly growing Internet videos. By using multi-view prediction as a pretext task, we demonstrate that learned 3D priors enable various 3D creation applications, including single view generation, sparse views reconstruction, and 3D editing in open-world scenarios.

The following sections outline our approach (Fig.2). Sec. 3.1 details the data curation pipeline, which selects static scenes with sufficient multi-view observations from raw video footage. Sec. 3.2 introduces our visual-conditional multi-view diffusion model, which effectively learns general 3D priors from pose-free videos. Finally, Sec. 3.3 demonstrates a new visual-conditional 3D generation framework utilizing a warping-based pipeline.

3.1. Video Data Curation

High-quality, large-scale video data rich in 3D knowledge is essential for learning accurate and reliable 3D priors. A well-defined 3D-aware video clip should exhibit two key properties: **1) temporally static scene content** and **2) significant viewpoint variation** caused by the camera’s egomotion. The first property ensures consistent 3D geometry across different viewpoints, while dynamic content can distort scene geometry and introduce biases that may degrade generation performance (Fig. 3a-Row1). The second property guarantees sufficient 3D observations from diverse viewpoints. When the model is trained on videos with limited viewpoint variation (Fig. 3a-Row2), it risks focusing on views adjacent to the reference view, rather than developing a comprehensive 3D understanding.

To obtain a massive volume of 3D data, we collect approximately **25.48M** open-sourced raw videos, totaling **44.98 years** from the Internet, covering a wide range of categories, such as landscapes, drones, animals, plants, games, and actions. Specifically, our dataset is sourced from four websites: Pexels [69], Artgrid [36], Airvuz [70], and Skypixel [96]. We follow Emu3 [102] to split the videos with PySceneDetect [7] to identify content changes and fade-in/out events. Additionally, we remove clips with excessive text using PaddleOCR [92]. The detailed composition of our WebVi3D dataset is presented in Tab. 1.

However, identifying 3D-aware videos presents a non-

Website	Domain	# Src. Vids	Total Hrs.	#Fil. Vids	#Fil. Clips	Fil. Hrs.
Pexels	Open	6.18M	101.77K	0.61M	2.65M	9.96K
Artgrid	Open	3.94M	92.49K	0.54M	1.10M	8.77K
Airvuz	Drone Shot	5.10M	94.75K	0.54M	5.87M	8.72K
Skypixel	Landscape	10.27M	105.47K	0.61M	6.37M	8.82K
Total	Open	25.48M	394.48K	2.30M	15.99M	36.27K

Table 1. **WebVi3D Dataset.** Sourced from four open websites, we curate $\sim 2.30M$ videos, which are divided into $15.99M$ clips featuring temporally static scenes with large-range viewpoint.

trivial challenge. As most videos are derived from real-world footage, such videos often contain dynamic scenes or small camera movement. To address this, we propose a pipeline that automatically selects relevant, high-quality 3D-aware data (i.e., multi-view images) by leveraging priors from instance segmentation [29], optical flow [93], and pixel tracking [40]. This pipeline comprises four core steps:

a) Temporal-Spatial Downsampling. To improve data filtering efficiency, we first downsample each video clip both temporally and spatially. The final resolution is set to 480p, and the temporal downsampling rate is set to 2. Note that this downsampling operation is applied only during data curation, not during model training.

b) Semantic-Based Dynamic Recognition. We employ the instance segmentation model, Mask R-CNN [29], to generate motion masks for potential dynamic objects, such as humans, animals, and sports equipment. A threshold is applied to filter out videos based on the proportion of frames containing these objects, as they are more likely associated with dynamic scenes.

c) Non-Rigid Dynamic Filtering. To precisely filter out videos with dynamic regions, we use offline optical flow estimation [93] to obtain dense matching, which enables us to identify non-rigid motion masks in video frames. These masks are then analyzed based on their locations to further determine whether the video contains dynamic content.

d) Tracking-Based Small Viewpoint Filtering. The previous three steps yield videos with static scenes. To further ensure these videos contain multi-view images captured from a larger camera viewpoint, we track the motion trajectory of key points across frames and calculate the radius of the minimum outer tangent circle of the trajectory. Videos with a small trajectory radius are then filtered out. More details about the data curation pipeline are provided in the Appendix B.

Finally, we curate approximately 320M multi-view images from $15.99M$ video clips with static content and sufficient multi-view observations (see Fig.3b). To validate the effectiveness of our data acquisition method, we randomly select 10,000 video clips for human annotation, of which 8,859 were labeled as 3D-aware, representing 88.6% of the total. This indicates that our pipeline effectively identifies 3D-aware videos from massive source videos. As the vol-

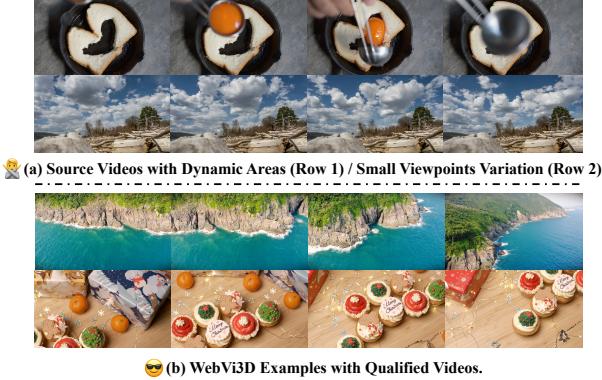


Figure 3. (a-Row1): Dynamic content modifies scene geometry across views; (a-Row2): Limited camera movement provides insufficient multi-view observations; (b) Our WebVi3D comprises static scenes with diverse camera trajectories.

ume of Internet videos continues to grow, this pipeline can continuously acquire more 3D-aware data, allowing for ongoing expansion of our dataset.

3.2. Visual Conditional Multi-View Diffusion Model

Preliminary. Diffusion models [31, 84, 85] operate by perturbing the training data $X_0 \sim q(X_0)$ through a forward diffusion process and learning to reverse it. The forward diffusion process $X_t \sim q_{t|0}(X_t|X_0)$ can be formally represented by $X_t = \sqrt{\bar{\alpha}_t} \mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, where $\bar{\alpha}_t$ is variance schedule used in noise scheduler. In theory, X_t approximates an isotropic Gaussian distribution for sufficiently large timesteps t . The training objective is to learn the reverse process.

Objective. We aim for multi-view prediction: generating novel views along specified camera trajectories from a single or sparse input while ensuring consistency with the input appearance. The MVD model inherits the generalizability of the 2D diffusion model while capturing cross-view consistency, which naturally aligns with our goal. Following this line, we present **See3D**, a pose-free, visual-conditional MVD model trained on Internet videos to enable robust 3D generation, as shown in Fig.2.

Challenge. The main technical challenge lies in learning precise camera control from pose-free videos. Previous works commonly incorporate camera parameters for both input and target views into diffusion models to guide multi-view generation from specified viewpoints. However, training these models generally requires expensive 3D data with precise camera pose annotations, which limits scalability. To address this, we explore an alternative approach that conditions on 2D-inductive visual hints to implicitly control camera movement during training, thereby avoiding the need for hard-to-obtain camera trajectories.

Formulation. Formally, we propose training the MVD model conditioned on 2D-inductive visual signals, referred to as *visual-condition*, without incorporating camera parameters. This task can be formulated as designing a conditional distribution, achieved by a conditional diffusion model that minimizes:

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} \left[\|\epsilon_\theta(X_t, Y_0, V, t) - \epsilon\|_2^2 \right], \quad (1)$$

where X_t denotes the noisy latent. $X_0 = \{x_0^i\}_{i=1}^N$ represents a multi-view observation of 3D content, formed by sampling one clip from *WebVi3D* as described in Section 3.1, with $N = S + L$ being the number of frames in each clip. From X_0 , S frames are randomly selected as reference views, noted as $Y_0 = \{y_0^i\}_{i=1}^S$, while the remaining L frames are treated as target images, denoted $G = \{g^i\}_{i=1}^L$. Our approach focuses on constructing the *visual-condition* V , which guides the diffusion model to generate plausible 3D content estimates from target viewpoints, ensuring consistency with the appearance of Y_0 .

3.2.1 Principle of Visual-Condition

A desirable *visual-condition* should meet the following criteria: a) it can be constructed without the need for additional 3D annotations, b) it is independent of specific downstream tasks, and c) it offers sufficient generalization to support various task-specific visual conditions, enabling precise control of camera movements.

Ideally, this *visual-condition* can be derived from pixel-space hints within the original videos, implicitly guiding the model to learn camera control. Moreover, it should be robust enough to handle domain gaps between task-specific visual cues and pixels extracted from video data. For example, in warping-based generation, warped images often suffer from issues like self-occlusions, artifacts, and distortions, creating a significant gap compared to real video data as shown in Fig.5 and Fig.6.

3.2.2 Time-dependent Visual Condition

Building on the analysis above, we propose constructing the *visual-condition* by applying masks, noise, and mixture to the input video data.

Random Masking: We first corrupt target images G through random irregular masking to reduce reliance on direct pixel-space visual signals, helping the model partially mitigate the domain gap between task-specific visual cues and video data. Meanwhile, we keep the reference images Y_0 clean to provide effective appearance signals.

Time-dependent Noise: We further add noise to video data to approximate a Gaussian distribution. For downstream

tasks, task-specific visual inputs are similarly noised, aligning their distributions with this Gaussian profile and further bridging the gap between video data and task-specific inputs. A key challenge lies in determining the optimal noise level: excessive noise weakens conditional signals, resulting in poor visual quality and inaccurate camera control, whereas insufficient noise preserves too many details from the target images, causing the model to over-rely on visual hints from the video data.

Previous studies [18, 34, 66, 83, 127] have explored modulating noise levels by adding noise to input data. Notably, as pointed out in the previous work [127], diffusion models tend to over-rely on the conditional image at larger time steps, leading to signal leakage. Inspired by this [127], we introduce time-dependent noise to the corrupted target images. In addition, we develop a function $t' = f(t)$ to regulate signal leakage, preventing excessive noise from completely obscuring visual cues and disabling camera control. Specifically, we define:

$$C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)\mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_{t'}}\boldsymbol{\epsilon}. \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (2)$$

Here, f is a strictly monotonically increasing function, ensuring $t' < t$, so that C_t contains at least as much information as X_t at earlier timesteps. $\bar{\alpha}_{t'}$ are the variances used in DDIM [85]. A detailed explanation of $f(t)$ can be found in Appendix C.3.

Time-dependent Mixture: However, as t decreases, lower noise levels increase the risk of signal leakage, causing a domain gap between the video data and task-specific visual condition distributions. To address this issue, we propose gradually replacing the corrupted data C_t with noisy latent variables X_t as timestep decreases. This encourages the model to rely more on pixel-space signals from video data at larger time steps, and transition to X_t at smaller timesteps. To achieve this, we further introduce a weighting factor $W_t \in [0, 1]$, which decreases monotonically with the timestep t , to combine C_t and X_t . Formally, our final *visual-condition* is defined as:

$$V_t = [W_t * C_t + (1 - W_t) * X_t; M], \quad (3)$$

where $M = \{m^{0:S} \cup m^{S+1:N}\}$, with $m^{0:S}$ as a zero matrix, keeping the reference images Y_0 unmasked, and $m^{S+1:N}$ as random irregular masks applied to the target images G . $V_t = \{v_t^i\}_{i=1}^N$ represents a mixture of C_t and X_t , concatenated with masks M along the channel dimension. In practice, an additional processing step assigns $v_t^{0:S}$ to the reference images Y_0 directly, in order to inject the clean information of Y_0 into the model, facilitating alignment between the predicted images and the reference images. Consequently, Eq.1 can be reformulated as $\mathbb{E}_{X_0, Y_0, \boldsymbol{\epsilon}, t} [\|\boldsymbol{\epsilon}_\theta(X_t, Y_0, V_t, t) - \boldsymbol{\epsilon}\|_2^2]$. A more detailed definition of W_t is provided in Appendix C.3.

3.2.3 Model Architecture

Our model architecture is based on video diffusion model [6]. However, we removed the time embedding, as we aim for the model to control the camera movement purely through visual conditions, rather than inferring movement trends based on temporal cues. To further minimize the effect of temporality, we shuffle the frames in each video clip, treating the data as unordered X_0 . Specifically, we randomly select a subset of frames from a video clip as reference images, with the remaining frames as target images. The number of reference images is randomly selected to accommodate different downstream tasks. The multi-view diffusion model is optimized by calculating the loss only on the target images, as described in Eq.1. Additional details regarding the model architecture, including the design of self-attention layers, Zero-Initialize, trainable parameters, noise schedule, and cross-attention, can be found in the Appendix C.1.

3.3. Visual Conditional 3D Generation

Overview. This section demonstrates the application of **See3D** for domain-free 3D generation, supporting long-sequence novel view synthesis with complex camera trajectories. Starting with one or a few input views, we iteratively generate warped images as visual hints, guided by predefined camera poses and estimated global depth [5]. **See3D** is then utilized to generate novel views along the predefined camera trajectory, conditioned on the proposed *visual-condition*. This iterative pipeline is illustrated in Fig.4, where the brown cameras represent the already generated views, and the gray cameras indicate the target views we aim to generate.

Challenge. Recent warping-based 3D generation approaches [12, 22, 44] rely on monocular depth or point clouds, and perform global point-cloud alignment to recover the actual geometry for subsequent generations. However, as the reference view often provides a limited scene observation, using offline methods tends to suffer from *scale ambiguity* and *geometric estimation errors*. Moreover, previous methods often overlook correcting these geometric errors, leading to distortions and stretching artifacts. These errors accumulate during iterative generation, severely degrading the generation quality. To address this, we propose an iterative strategy with sparse pixel-wise depth alignment, comprising two core steps: pixel-wise depth scale alignment and global metric depth recovery.

Pixel-wise Depth Scale Alignment. We introduce pixel-wise depth scale alignment using sparse keypoints. This approach performs high-degree-of-freedom independent optimization for all keypoints by leveraging multi-view matching priors from anchor views. Each keypoint independently identifies its multi-view correspondences, allowing for the

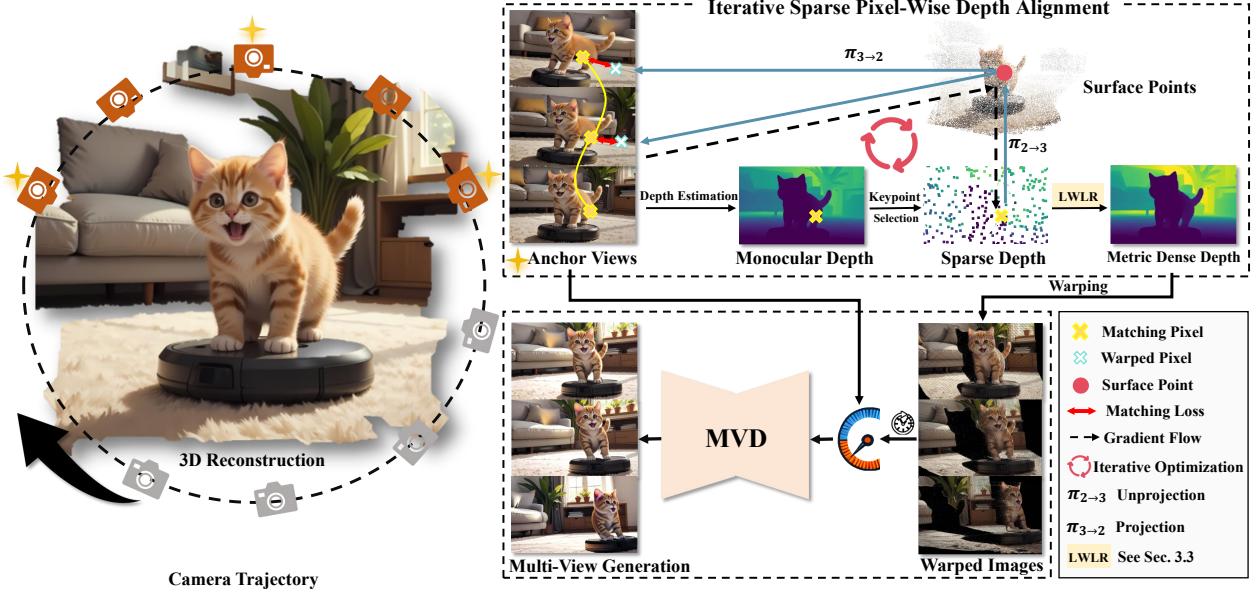


Figure 4. See3D for Multi-View Generation: From iteratively generated views (brown camera), we randomly select a few anchor views (yellow stars) to guide the generation of target views along the gray camera trajectory. Keypoint matching is first performed to establish correspondences between the anchor views. Next, monocular depth estimation is applied to the latest anchor view, followed by our *Iterative Sparse Pixel-Wise Depth Alignment* to refine the depth and recover a dense map. This dense depth is then used to warp images along the gray camera viewpoints. Subsequently, the warped images and anchor images are combined and processed according to Eq.2 and Eq.3, without random masking, forming the *visual-condition*, which guides MVD model to produce 3D-consistent target views. Finally, the gray camera turns to brown, guiding multi-view generation in the next iteration.

recovery of both depth scale and surrounding geometry. The corrected scale is then propagated across the entire depth map using 2D distances between keypoints and their neighbors.

Specifically, denote $\{T_i\}_{i=0}^N$ the predefined camera trajectory. Assuming we have generated n images $\{I_i\}_{i=0}^n$, we now proceed to generate the next m views using the warped image from the last anchor view I_n , referred to as the source view. We first utilize the pre-trained MoGe [101] to estimate the affine-invariant depth \hat{D}_n of I_n . Inspired by [112], we perform sparse alignment with 1024 pairs of matching keypoints $\{\mathbf{m}_n, \mathbf{m}_i\}_k$, obtained by the pre-trained extractor SuperPoint [16] and feature matcher LightGlue [49]. For each matched point, we optimize the corresponding scale α^k and shift β^k parameters, where $k \in [0, 1024]$. Our core idea is to recover the depth scaling by minimizing the L_2 distance of re-projection between matching points. For each iteration, the warping operation $\Pi_{n \rightarrow i}$ transforms pixels from the source image's coordinate frame to the target image's coordinate frame, formulated as: $\Pi_{n \rightarrow i}(\hat{d}_n) = \hat{d}_n K_i T_i T_n^{-1} K_n^{-1}$, where K_i, K_n, T_i, T_n represent the intrinsic and extrinsic parameters of the source and target frames, respectively. The alignment for each pair is performed using normalized coordinates, ensuring that the warping aligns with the matching prior:

$$\alpha^{k*}, \beta^{k*} = \underset{\alpha^k, \beta^k}{\operatorname{argmin}} \| \hat{d}_n^{k*} K_i T_i T_n^{-1} K_n^{-1} m_n^t - m_i^t \|_2^2, \quad (4)$$

where the recovered depth of k th pixel is $\hat{d}_n^{k*} = \alpha^k \odot \hat{d}_n^k + \beta^k$, the \odot is the pixel-wise Hadamard Product. We minimize the matching loss via gradient descent to obtain best scale α^{k*} and shift parameters β^{k*} for each pixel. By performing individual scale recovery and geometry correction, we decouple the depth correlation among different points, achieving accurate single-view reconstruction.

Global Metric Depth Recovery. After that, we set these recovered positions as sparse guidance \hat{d}_n^* , and introduce Locally Weighted Linear Regression [112] (marked as LWLR in Fig.4) to recover the whole depth map based on the locations between guided points and the other target points. Denote (u, v) represent the 2D positions of the remaining target points, their depth \hat{D}_n can be fitted to the sparse guided points by minimizing the squared locally weighted distance, which is reweighed by the diagonal weight matrix as:

$$\begin{aligned} \mathbf{W}_{u,v} &= \operatorname{diag}(w_1, w_2, \dots, w_m), \\ w_i &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\operatorname{dist}_i^2}{2b^2}\right), \end{aligned} \quad (5)$$

where b is the bandwidth of Gaussian kernel, and dist is the Euclidean distance between the guided point and the under-estimated target point. Denote \mathbf{X} the homogeneous representation of \hat{D}_n , the scale map \mathbf{S}_{scale} and shift map \mathbf{S}_{shift} of target points can be calculated by iterating every location

on the whole image, which can be formulated as:

$$\begin{aligned} \min_{\beta_{u,v}} & (\hat{d}_n^* - \mathbf{X}\boldsymbol{\beta}_{u,v})^\top \mathbf{W}_{u,v} (\hat{d}_n^* - \mathbf{X}\boldsymbol{\beta}_{u,v}) + \lambda S_{shift}^2, \\ \hat{\boldsymbol{\beta}}_{u,v} &= (\mathbf{X}^\top \mathbf{W}_{u,v} \mathbf{X} + \lambda)^{-1} \mathbf{X}^\top \mathbf{W}_{u,v} \hat{d}_n^*, \\ \boldsymbol{\beta}_{u,v} &= [S_{scale}, S_{shift}]_{u,v}^\top, \\ \mathbf{D}_n &= \hat{d}_n^* \oplus S_{scale} \odot \hat{\mathbf{D}}_n + S_{shift}, \end{aligned} \quad (6)$$

where \mathbf{D}_n is the scaled whole depth map, \oplus is the concatenation operator, λ is a l_2 regularization hyperparameter used for restricting the solution to be simple. Besides, the explicit constraint of the source frame with the target frames allows each novel view to maintain contextual consistency from preceding generations.

Novel View Generation. After obtaining the aligned depth \mathbf{D}_n , we generate target visual hints through warping as $\hat{I}_j = \Pi_{n \rightarrow j}(\mathbf{D}_n)$. The warped images $\{\hat{I}_j\}_{j=n}^{n+m}$ contain unfilled regions, as indicated by the binary warping mask $\{M_j\}_{j=n}^{n+m}$, providing strong visual hints for **See3D** to perform novel view generation. To ensure strong multi-view consistency between the newly generated sequence and the previous content, we randomly select k anchor views $\{I_k\}$, $k \in [1, N]$ from the earlier generated frames to guide subsequent generation. The generation process is formulated as: $I_j = \text{See3D}(\hat{I}_j, M_j, \{I_0, I_k\})$. We iteratively perform depth estimation, alignment, warping, and generation until all predefined multi-view images are obtained.

3D Reconstruction. We reconstruct the 3D scene using 3D Gaussian Splatting (3DGS) [41]. The training objective is to minimize the sum of photometric loss and SSIM loss, consistent with the original 3DGS approach. Additionally, we introduce a perceptual loss (LPIPS [126]) to mitigate subtle *inter-frame* discrepancies in multi-view generated images during 3DGS reconstruction. LPIPS emphasizes higher-level semantic consistency between Gaussian-rendered and generated multi-view images, rather than focusing on minor high-frequency differences. Furthermore, the potential *inner-frame* diversity may lead to inconsistencies with the corresponding camera poses. Following [20], we implement joint pose-Gaussian optimization, treating camera parameters as learnable variables alongside Gaussian attributes, thereby reducing gaps between generated viewpoints and their corresponding camera poses.

4. Experiments

In Sec. 4.1 and Sec. 4.2, we present the single view and sparse views reconstruction with **See3D** as prior. Next, we conduct ablation experiments in Sec. 4.3 to validate the effectiveness of the proposed modules. Additional implementation details, more results on open-world 3D creation, and further ablation experiments are provided in the Appendix.

4.1. Single View to 3D

Experimental Setting. **See3D** supports multi-view generation from a single input view. Following prior work [121], our evaluation is conducted on the test split of three real-world datasets with various camera trajectories, including Tanks-and-Temples [43], RealEstate10K [129], CO3D [75]. We follow the approach in ViewCrafter [121] for constructing easy/hard evaluation sets based on different sampling rates applied to the original videos. We reimplement ViewCrafter using the official code released by [121] to validate our easy/hard set splitting, with results shown as ViewCrafter* in Tab. 2. We conduct comparisons with warping-based baselines, including LucidDreamer [12], camera-conditional video generation model MotionC-trl [104], warp-image conditional ViewCrafter [121], and multi-view diffusion model ZeroNVS [77]. We use the same point cloud rasterization as proposed in ViewCrafter [121] instead of depth-based warping to generate visual conditions for fair comparisons. Following [121], we evaluate only the visual quality of images generated by multi-view diffusion without rendering novel views through 3D reconstruction. We report PSNR, SSIM, and LPIPS [126] as evaluation metrics. Among these, PSNR is a traditional pixel-level metric that measures image similarity, which is significantly affected by viewpoint shifts. As such, PSNR reflects the accuracy of viewpoint control provided by our proposed *visual-condition* in multi-view generation.

Results. The quantitative comparison results are presented in the top rows of Tab. 2. Only average metrics for the easy and hard sets are reported here, detailed values are available in the Appendix D.1. The results for ViewCrafter* are comparable to those reported in its original paper, confirming successful alignment between our method and the baselines. Numerically, our approach outperforms all baseline methods across all metrics. Specifically, compared to the re-implemented ViewCrafter, our approach achieves a 4.63 dB improvement, demonstrating its capability to generate high-quality novel views. PSNR further demonstrates significant gains, indicating our proposed *visual-condition* enables precise camera control. Qualitative results are shown in the top rows of Fig. 5. **See3D** generates high-quality, realistic content within minutes. Despite limited visual cues provided by the warped images, our method produces more reliable and realistic results with fewer artifacts.

4.2. Sparse Views to 3D

Experimental Setting. We extend our model to the sparse-view reconstruction task, evaluating it on three datasets: LLFF [64], DTU [37], and Mip-NeRF 360 [3]. We compare our method against several few-shot 3D reconstruction baselines, including optimization-based method MuRF [113], FSGS [130], and BGGS [27]; diffusion-based meth-

Methods	Tanks-and-Temples [43]			RealEstate10K [129]			CO3D [75]		
Single View	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
LucidDreamer [12]	13.11	0.314	0.485	15.24	0.545	0.357	13.90	0.412	0.473
ZeroNVS [77]	13.38	0.344	0.525	15.37	0.556	0.397	14.23	0.444	0.495
MotionCtrl [104]	14.31	0.405	0.436	16.30	0.596	0.363	16.16	0.515	0.418
ViewCrafter [121]	19.66	0.609	0.238	21.93	0.797	0.161	20.17	0.664	0.283
ViewCrafter* [121]	19.13	0.616	0.255	20.49	0.802	0.183	19.07	0.678	0.339
Ours	23.76	0.735	0.191	25.36	0.854	0.146	24.28	0.765	0.251
Sparse Views (3 Views)	LLFF [64]			DTU [37]			MipNeRF-360 [3]		
Zip-NeRF † [4]	17.23	0.574	0.373	9.18	0.601	0.383	12.77	0.271	0.705
MuRF [113]	21.34	0.722	0.245	21.31	0.885	0.127	-	-	-
FSGS [130]	20.31	0.652	0.288	17.34	0.818	0.169	-	-	-
BGGS [27]	21.44	0.751	0.168	20.71	0.862	0.111	-	-	-
ZeroNVS † [77]	15.91	0.359	0.512	16.71	0.716	0.223	14.44	0.316	0.680
DepthSplat [114]	17.64	0.521	0.321	15.59	0.525	0.373	13.85	0.254	0.621
ReconFusion [107]	21.34	0.724	0.203	20.74	0.875	0.124	15.50	0.358	0.585
CAT3D [23]	21.58	0.731	0.181	22.02	0.844	0.121	16.62	0.377	0.515
Ours	23.23	0.768	0.135	28.04	0.884	0.073	17.35	0.442	0.422

Table 2. **Quantitative Comparison of Single/Sparse Views Generation.** The top rows are results given single view as input, where ViewCrafter* indicates our re-implemented result. The bottom rows are novel view rendering quality given 3 views as input, where Zip-NeRF † and ZeroNVS † are modified versions with sparse views input as reported in CAT3D.

ods CAT3D [23], ZeroNVS (modified to handle multi-view input) [77], and ReconFusion [107]; as well as the feed-forward method DepthSplat [114]. Following the evaluation protocols from [68, 107, 130], we use 3, 6, and 9 views as input. For few-shot reconstruction, dense multi-view images are generated from sparse views, similar to CAT3D [23], and 3DGS reconstruction is performed with pose optimization to render test views for evaluation. We report PSNR, SSIM, and LPIPS [126] to evaluate novel view synthesis performance.

Results. Qualitative and quantitative results are presented in Tab. 2 and Fig. 5, respectively, with additional comparisons for 3, 6, and 9 input views available in Appendix D.2. The 3DGS model, trained on dense multi-view images generated by **See3D**, surpassed state-of-the-art reconstruction models in novel view rendering. This indicates its ability to provide high-quality, consistent multi-view support for 3D reconstruction without imposing additional constraints. Compared to ReconFusion [107] and CAT3D [23], which also leverage diffusion priors for sparse-view reconstruction, our model exhibits effective scalability. Qualitative comparisons in Figure 5 reveal that NVS results produced by **See3D** exhibit fewer floating artifacts, suggesting its capability to generate more consistent and high-fidelity multi-view images.

4.3. Ablation Study

Scaling up Data. We investigate the impact of training data by ablating different proportions of our training dataset. The model is trained with 10%, 20%, 40%, 80%, and 100% of the training set, and its single-view generation performance is evaluated on RealEstate10K, achieving PSNR val-

ues of 19.32, 21.04, 22.57, 24.08, and 25.01, respectively. Additionally, training with unfiltered data results in generated content that often exhibits movement or deformation, leading to a substantial performance drop with a PSNR of 19.55. We analyze that this degradation likely stems from the lack of stationary and geometrically invariant properties in much of the source video content, which undermines multi-view consistency. In summary, these findings highlight the critical importance of data quality and diversity for effectively training large-scale MVD models.

Visual-condition. Excluding the benefits of data scaling, we investigate the effectiveness of our *visual-condition* on pose-free data. Previous work [121] has demonstrated that warped images can serve as a pivot condition to guide the model to generate the target viewpoint. However, due to the reliance on the annotated camera to control the projection and unprojection, warp-based conditions are inherently unscalable. Therefore, we compare the model’s ability to control cameras conditioned on pose-free *visual-condition* and conditioned on warped images. Specifically, we extract a subset of MVImageNet [122] for training and testing.

For each multi-view sequence in the training set, we select the point cloud of the first frame and render it into the subsequent 5 camera planes along the camera trajectory, based on the 3D annotations in the dataset. We obtain warped images and form pairs with the ground-truth multi-views to train an MVD model, referred to as MV-Posed. With the same experimental settings (training set, network architecture, batch size and predicted sequence length), we train an additional model without any 3D annotations, except for the modification of warp condition to the time-

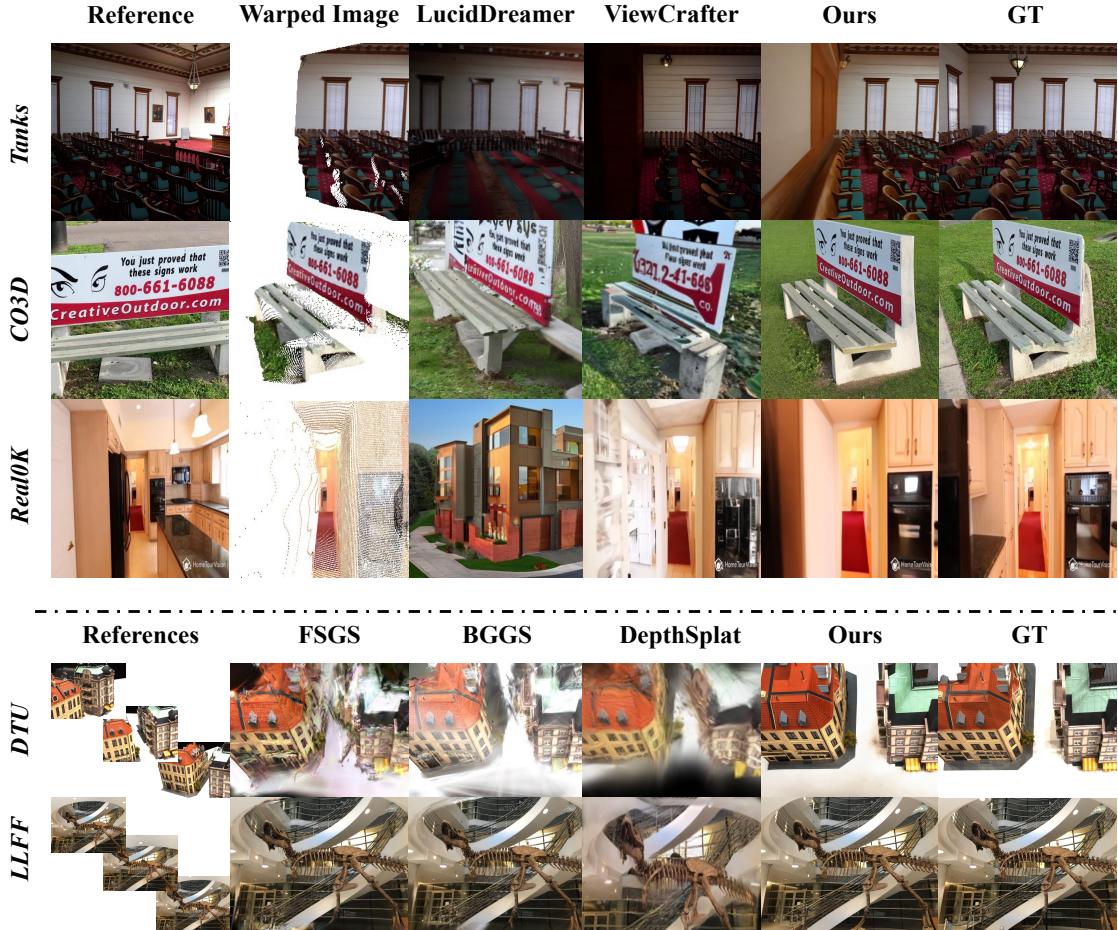


Figure 5. **Qualitative Comparison of Single/Sparse View Generation.** The top three rows are results with a single view input. The bottom two rows are novel view renderings from 3DGs, where Ours is trained on dense multi-view generation given 3 views as input. Our method outperformed other baselines in capturing high-frequency details, such as text and stairs.

dependent *visual-condition* V_t described in Sec.3.2, called MV-UnposeT. Meanwhile, we employ randomly masked multiple views as condition to train the model as an additional baseline, called MV-UnposeM.

Model	LPIPS ↓	PSNR ↑	SSIM ↑
MV-Posed	0.182	26.21	0.822
MV-UnposeM	0.443	16.14	0.521
MV-UnposeT	0.194	25.56	0.811

Table 3. **Ablation Study on Visual-condition.**

The results are reported in Tab.3 and Fig.6, where the performance of MV-Posed and MV-UnposeT is comparable. In contrast, MV-UnposeM struggles to handle the gap between the warped image and masked images, in the case of geometric distortion and self-obscurcation. These findings indicate that the *visual-condition* offers a viable alternative to 3D-reliant warped conditions. Despite a significant domain gap between V_t and warp images as shown in Fig.6, our model robustly handles this discrepancy, thanks to the

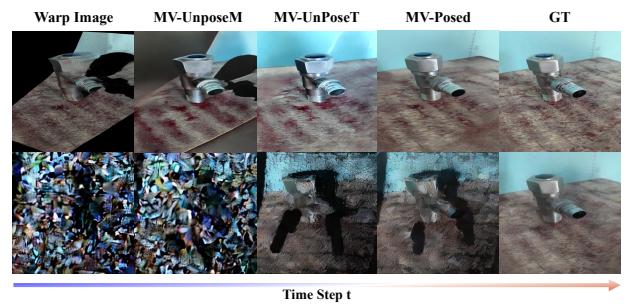


Figure 6. Top: Qualitative ablation of *visual-condition*; Bottom: As timestep decreases, visualize the trend of *visual-condition*.

time-dependent nature of the proposed condition.

5. Conclusion

We propose a scalable 3D generation framework from the perspective of dataset scaling, offering a systematic solution that includes: 1) a new dataset, WebVi3D, curated via

an automated pipeline, with the potential to evolve with the growing volume of Internet data. 2) a new model, **See3D**, capable of scalable training without pose annotations, aligning with the concept of ‘Get 3D by solely Seeing’. 3) a novel **See3D**-based 3D generation framework that supports long-sequence view generation with complex camera trajectories. We show that the 3D priors learned by **See3D** enable a range of 3D creation applications, including single-view generation, sparse view reconstruction, and 3D editing in open-world scenarios. We believe **See3D** provides a new direction to advancing the upper bound of 3D generation through dataset scaling. We hope our efforts will encourage the 3D research community to pay more attention to large-scale unposed data, bypassing the costly 3D data barrier and chasing parity with powerful closed-source 3D solutions.

Acknowledgments. We thank Wenyuan Zhang and Yu-Shen Liu from Tsinghua University, as well as Yance Jiao, Hua Zhou, Liao Zhang, Yaohui Chen, Jinxin Xie, Yiwen Shao, and other colleagues from BAAI, for their valuable support and contributions to the See3D project.

References

- [1] Antonio Alliegro, Yawar Siddiqui, Tatiana Tommasi, and Matthias Nießner. Polydiff: Generating 3d polygonal meshes with diffusion models. *arXiv preprint arXiv:2312.11417*, 2023. 3
- [2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 2
- [3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 3, 8, 9
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 9, 25
- [5] Shariq Farooq Bhat, Reiner Birk, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 6
- [6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6
- [7] Brandon Castellano. Pyscenedetect. <https://github.com/Breakthrough/PySceneDetect/>. [Online; accessed 13-Oct-2024]. 4
- [8] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 3
- [9] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2416–2425, 2023. 3
- [10] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21476–21485, 2024. 2
- [11] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2025. 3
- [12] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2, 3, 6, 8, 9, 20, 23, 24
- [13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3
- [14] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 22
- [15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3, 4, 22
- [16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabenovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 7, 19
- [17] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024. 3
- [18] Martin Nicolas Everaert, Athanasios Fotsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the signal-leak bias in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4025–4034, 2024. 6
- [19] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a

- single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2
- [20] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2, 2024. 8
- [21] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 19
- [22] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 6
- [23] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 3, 9, 25
- [24] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3
- [25] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 3
- [26] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. In *European Conference on Computer Vision*, pages 333–350. Springer, 2025. 3
- [27] Liang Han, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Binocular-guided 3d gaussian splatting with view consistency for sparse view synthesis. *arXiv preprint arXiv:2410.18822*, 2024. 8, 9, 25
- [28] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2
- [29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 18
- [30] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 22
- [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5
- [32] Lukas Höller, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 3
- [33] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3
- [34] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 6
- [35] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [36] Specializes in royalty-free digital content. <https://artlist.io/stock-footage/>. [Online; accessed 15-Aug-2024]. 4
- [37] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 8, 9
- [38] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3
- [39] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017. 2
- [40] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *Proc. ECCV*, 2024. 4, 19
- [41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 8
- [42] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual editing. *Advances in Neural Information Processing Systems*, 36:7323–73257, 2023. 3
- [43] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3, 8, 9, 23, 24
- [44] Jiabao Lei, Jiapeng Tang, and Kui Jia. Rgb2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8422–8434, 2023. 6
- [45] Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *arXiv preprint arXiv:2406.08659*, 2024. 2
- [46] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3
- [47] Mengfei Li, Xiaoxiao Long, Yixun Liang, Weiyu Li, Yuan Liu, Peng Li, Xiaowei Chi, Xingqun Qi, Wei Xue, Wenhan

- Luo, et al. M-lrm: Multi-view large reconstruction model. *arXiv preprint arXiv:2406.07648*, 2024. 3
- [48] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3
- [49] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 7
- [50] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 3, 22
- [51] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 2
- [52] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3
- [53] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [54] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 3
- [55] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, pages 37–53. Springer, 2025. 3
- [56] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3
- [57] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 18
- [58] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3
- [59] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 2
- [60] Longfei Lu, Huachen Gao, Tao Dai, Yaohua Zha, Zhi Hou, Junta Wu, and Shu-Tao Xia. Large point-to-gaussian model for image-to-3d generation. *arXiv preprint arXiv:2408.10935*, 2024. 3
- [61] Zhaoyang Lyu, Jinyi Wang, Yuwei An, Ya Zhang, Dahua Lin, and Bo Dai. Controllable mesh generation through sparse latent point diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 271–280, 2023. 3
- [62] Qi Ma, Yue Li, Bin Ren, Nicu Sebe, Ender Konukoglu, Theo Gevers, Luc Van Gool, and Danda Pani Paudel. Shapesplat: A large-scale dataset of gaussian splats and their self-supervised pretraining. *arXiv preprint arXiv:2408.10906*, 2024. 3
- [63] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 3
- [64] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 8, 9
- [65] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [66] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kortschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 2, 3, 6
- [67] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3
- [68] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 9
- [69] Provider of stock photography and stock footage. <https://www.pexels.com/search/videos/videos/>. [Online; accessed 13-Oct-2024]. 4
- [70] Premiere online destination for drone pilots. <https://www.airvuz.com/collections/>. [Online; accessed 29-Sept-2024]. 4

- [71] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 19
- [72] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [73] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 3
- [74] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 22
- [75] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 4, 8, 9, 22, 23, 24
- [76] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [77] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Hermann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3, 8, 9, 24, 25
- [78] QiuHong Shen, Zike Wu, Xuanyu Yi, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *arXiv preprint arXiv:2403.18795*, 2024. 3
- [79] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3, 20
- [80] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3, 19, 22
- [81] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 2
- [82] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 19
- [83] Vedant Singh, Surgan Jandial, Ayush Chopra, Siddharth Ramesh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. On conditioning the input noise for controlled image generation with diffusion models. *arXiv preprint arXiv:2205.03859*, 2022. 6
- [84] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3, 5
- [85] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5, 6, 22
- [86] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 20
- [87] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 3
- [88] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 2
- [89] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023. 3
- [90] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3
- [91] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 3
- [92] Paddle Team. Paddle ocr. <https://github.com/PaddlePaddle/PaddleOCR/>. [Online; accessed 13-Oct-2024]. 4
- [93] Zachary Teed and Jia Deng. 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, pages 402–419. Springer, 2020. 4, 18
- [94] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2, 3
- [95] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 2, 4

- [96] Videos and photos shot by DJI devices. <https://www.skypixel.com/>. [Online; accessed 29-Aug-2024]. 4
- [97] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multi-view consistent images for single-view scene reconstruction. *arXiv preprint arXiv:2410.16892*, 2024. 3
- [98] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2
- [99] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3
- [100] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3
- [101] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024. 7
- [102] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyi Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4
- [103] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [104] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 8, 9, 24
- [105] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3
- [106] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 3
- [107] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 3, 9, 25
- [108] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 3
- [109] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan, Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [110] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024. 2
- [111] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2, 3
- [112] Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, and Feng Zhao. Frozenrecon: Pose-free 3d scene reconstruction with frozen depth models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9276–9286. IEEE, 2023. 7
- [113] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20050, 2024. 8, 9, 25
- [114] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 9, 21, 25
- [115] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3
- [116] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3
- [117] Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiaceng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024. 2
- [118] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 3
- [119] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 3
- [120] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 3

- [121] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. [arXiv preprint arXiv:2409.02048](#), 2024. [2](#), [3](#), [8](#), [9](#), [20](#), [23](#)
- [122] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 9150–9161, 2023. [2](#), [9](#), [22](#), [25](#)
- [123] Yaohua Zha, Naiqi Li, Yanzi Wang, Tao Dai, Hang Guo, Bin Chen, Zhi Wang, Zhihao Ouyang, and Shu-Tao Xia. Lcm: Locally constrained compact point cloud model for masked point modeling. [arXiv preprint arXiv:2405.17149](#), 2024. [3](#)
- [124] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. [arXiv preprint arXiv:2411.05003](#), 2024. [2](#)
- [125] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. [ACM Transactions on Graphics \(TOG\)](#), 43(4):1–20, 2024. [2](#), [3](#)
- [126] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In [Proceedings of the IEEE conference on computer vision and pattern recognition](#), pages 586–595, 2018. [8](#), [9](#)
- [127] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. [arXiv preprint arXiv:2406.15735](#), 2024. [6](#)
- [128] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. [arXiv preprint arXiv:2411.02319](#), 2024. [2](#)
- [129] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. [ACM Trans. Graph. \(Proc. SIGGRAPH\)](#), 37, 2018. [2](#), [8](#), [9](#), [22](#), [23](#), [24](#)
- [130] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In [European Conference on Computer Vision](#), pages 145–163. Springer, 2025. [8](#), [9](#), [25](#)
- [131] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. [arXiv preprint arXiv:2410.12781](#), 2024. [3](#)
- [132] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. [arXiv preprint arXiv:2312.09147](#), 2023. [3](#)

Contents

1. Introduction	2
2. Related work	3
3. Method	4
3.1. Video Data Curation	4
3.2. Visual Conditional Multi-View Diffusion Model	5
3.2.1 Principle of Visual-Condition	5
3.2.2 Time-dependent Visual Condition	5
3.2.3 Model Architecture	6
3.3. Visual Conditional 3D Generation	6
4. Experiments	8
4.1. Single View to 3D	8
4.2. Sparse Views to 3D	8
4.3. Ablation Study	9
5. Conclusion	10
A Broader Impact and Limitations	18
B Video Data Curation	18
C Technical Implementations	19
C.1. Model Architecture	19
C.2. Training Details	20
C.3. Definition of $f(t)$ and W_t	23
D More Experimental Results	23
D.1. Single View to 3D	23
D.2. Sparse Views to 3D	24
D.3. 3D Editing	24
E Additional Ablation Studies	24
E.1. Effectiveness of Pixel-level Depth Alignment	24
E.2. Efficacy of Scaling up Data	25
F Additional Visualizations	27

Appendix

A. Broader Impact and Limitations

Broader Impact: Our model facilitates open-world 3D content creation from large-scale video data, eliminating the need for costly 3D annotations. This can make 3D generation more accessible to industries like gaming, virtual reality, and digital media. By leveraging visual data from the rapidly growing Internet videos, it accelerates 3D creation in real-world applications. However, careful consideration of ethical issues, such as potential misuse in generating misleading or harmful content, is crucial. Ensuring that the data used is curated responsibly to avoid bias and privacy concerns is vital for safe deployment.

Limitations: While our model excels at long-sequence generation, it comes with some limitations regarding: 1) Inference Speed: The model requires several minutes for inference, making it challenging for real-time applications. Future work should aim to improve inference speed for real-time generation. 2) Focus on 3D Generation: The current model focuses only on 3D generation, avoiding the modeling of object motion. Future research could extend the model to simultaneously generate 3D and 4D content for dynamic scenes. 3) Model Scalability: While the data scaling approach is effective, the scalability of the model itself has not been explored. Expanding the model’s architecture could enhance its capability to handle more complex and diverse 3D content.

B. Video Data Curation

Our WebVi3D dataset is sourced from Internet videos through an automated four-step data curation pipeline. In this section, we provide further details on this pipeline process.

Step 1: Temporal-Spatial Downsampling. To enhance data curation efficiency, we downsample each video both temporally and spatially. Temporally, we retain one frame for every two by downsampling with a factor of two. Spatially, we adjust the downsampling factor according to the original resolution to ensure consistent visual appearance across different video aspect ratios. The final resolution is standardized to 480p in our experiment.

Step 2: Semantic-Based Dynamic Recognition We perform content recognition on each frame to identify potential dynamic regions. Following [57], we utilize the off-the-shelf instant segmentation model Mask R-CNN [29] to generate coarse motion masks \mathcal{M}_m for potential dynamic objects, including humans, animals, and sports activities. If motion masks are present in more than half of the video frames, the sequence is deemed likely to contain dynamic regions and excluded from further processing.

Step 3: Non-Rigid Dynamic Filtering After filtering out videos with common dynamic objects, we implement a precise strategy to identify and exclude videos containing non-rigid dynamic regions, such as drifting water and swaying trees. Following [57], we use the pretrained RAFT [93] to compute the optical flow between consecutive frames. Based on the optical flow, we calculate the Sampson Distance, which measures the distance of each pixel to its corresponding epipolar line. Pixels exceeding a predefined threshold are marked to create a non-rigid motion mask \mathcal{M}_s . The number of pixels in \mathcal{M}_s serves as an indicator of the likelihood of motion in the current frame.

However, relying solely on this metric is unreliable, as most data are captured in real shots, where dynamic objects of interest are often concentrated near the center of the imaging plane. These moving regions may not occupy a significant portion of the frame. Therefore, we also consider the spatial location of the dynamic mask and propose a dynamic score \mathcal{S} to evaluate the motion probability for each frame. Let H, W denote the height and width of an image, respectively. We define the central region as starting at $W' = 0.25 \times W, H' = 0.25 \times H$. The proportions of the mask occupying the entire image, Θ_i , and the central area Θ_c are calculated as:

$$\Theta_i = \frac{\sum_{u,v=0}^{W,H} \mathcal{M}_s(u,v)}{H \times W}, \quad \Theta_c = \frac{\sum_{u,v=W',H'}^{W-W',H-H'} \mathcal{M}_s(u,v)}{H/2 \times W/2}. \quad (7)$$

The dynamic score \mathcal{S} can be formulated as:

$$\mathcal{S}_i = \begin{cases} 2, & \Theta_i \geq 0.12 \text{ } \& \text{ } \Theta_c \geq 0.35 \\ 1.5, & \Theta_i \geq 0.12 \text{ } \& \text{ } 0.2 \leq \Theta_c < 0.35 \\ 1, & \Theta_i < 0.12 \text{ } \& \text{ } 0.2 \leq \Theta_c < 0.35 \\ 0.5, & \Theta_i < 0.12 \text{ } \& \text{ } \Theta_c < 0.2 \end{cases}. \quad (8)$$

This strategy targets the dynamic regions near the image center, enhancing data filtering accuracy. The final dynamic score \mathcal{S} for the entire sequence is calculated as:

$$\mathcal{S} = \sum_{i=0}^N \mathcal{S}_i, \quad (9)$$

where N represents the total number of extracted frames. If $\mathcal{S} >= 0.25 \times N$, the sequence is classified as dynamic and subsequently excluded.

Step 4: Tracking-Based Small Viewpoint Filtering. The previous steps produced videos with static scenes. We require videos that contain multi-view images captured from a wider camera viewpoint. To achieve this, we track the motion trajectory of key points across frames and calculate the radius of the minimum outer tangent circle for each trajectory. Videos with a substantial number of radii below a defined threshold are classified as having small camera trajectories and are excluded. This procedure includes keypoint extraction, trajectory tracking, and circle fitting using RANSAC (Random Sample Consensus) [21].

Keypoint Extraction. To reduce computational complexity, we downsample the extracted video frames by selecting every fourth frame. SuperPoint [16] is then used to extract keypoints $\mathbf{K} \in \mathbb{R}^{N \times 2}$ from the first frame, where $N = 100$ represents the number of detected keypoints used to initialize tracking.

Trajectory Tracking. Keypoints are tracked across all frames using the pretrained CoTracker [40], which generates trajectories and visibility over time as:

$$\mathbf{T}_{\text{pred}}, \mathbf{V}_{\text{pred}} = \text{CoTracker}(\mathbf{I}, \text{queries} = \mathbf{K}). \quad (10)$$

Here, \mathbf{I} denotes the input frames, $\mathbf{T}_{\text{pred}} \in \mathbb{R}^{1 \times T \times N \times 2}$ represents the tracked positions of each keypoint over time, and $\mathbf{V}_{\text{pred}} \in \mathbb{R}^{1 \times T \times N \times 1}$ indicates the visibility of each point.

Circle Fitting. For each tracked keypoint, a circle fitting method is applied to its trajectory, selecting only frames where the keypoint is visible ($\mathbf{V}_{\text{pred}} = 1$). Let $\mathbf{T}_{\text{visible}} \in \mathbb{R}^{M \times 2}$ be the filtered points, where M is the number of visible points. We then use the RANSAC-based circle fitting algorithm on $\mathbf{T}_{\text{visible}}$ to determine the circle's center $\mathbf{c} = (c_x, c_y)$ and radius r :

$$\mathbf{c}, r = \text{RANSAC}(\mathbf{T}_{\text{visible}}). \quad (11)$$

The RANSAC algorithm selects random subsets of three points to define candidate circles, computes the inliers, and optimizes for the circle with the highest inlier count and smallest radius. Finally, we count the number of circles with a radius smaller than a specified threshold, $r \leq 20$:

$$\text{count} = \sum_{i=1}^N \mathbb{I}(r_i \leq 20), \quad (12)$$

where \mathbb{I} is the indicator function. The mean radius is also computed to provide an overall measure of circular motion. If the number of small-radius circles exceeds 40 and the average circular motion is less than 5, we classify this video as having small camera trajectories.

User Study. To verify the effectiveness of our data curation pipeline, we conducted a user study with a randomly selected set of 10,000 video clips before filtering. We require our users to evaluate videos based on two aspects: *static content* and *large-baseline trajectories*. Only videos meeting both criteria are classified as 3D-aware videos. Among these, 1,163 videos met our criteria for 3D-aware videos, accounting for 11.6% of the total validation set. After applying our data screening pipeline, we randomly selected 10,000 video clips for annotation. In this filtered set, 8,859 videos were identified as 3D-aware, yielding a ratio of 88.6%, representing a 77% improvement compared to the previous set. These results demonstrate the efficacy of our pipeline in filtering 3D-aware videos from large-scale Internet videos.

C. Technical Implementations

C.1. Model Architecture

The main backbone of **See3D** model is based on the structure of 2D diffusion models but integrates 3D self-attention to connect the latents of multiple images, as shown in prior work [80]. Specifically, we adapt the existing 2D self-attention layers of the original 2D diffusion model into 3D self-attention by inflating different views within the self-attention layers. To incorporate visual conditions, we introduce the necessary convolutional kernels and biases using Zero-Initialize [82]. The model is initialized from a pretrained 2D diffusion model [71] and fine-tuned with all parameters, leveraging FlashAttention

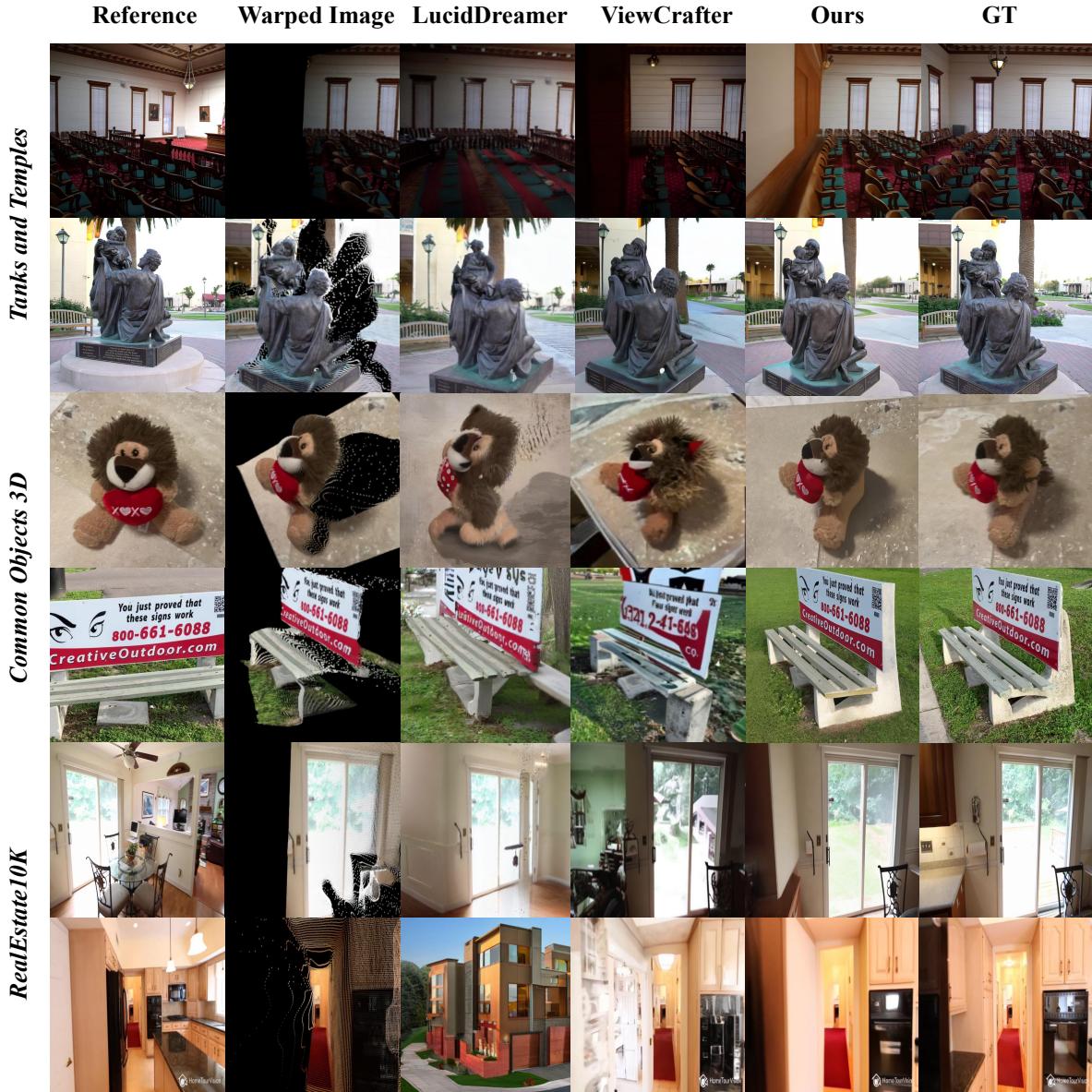


Figure 7. **Single-view to 3D.** Compared with LucidDreamer [12] and ViewCrafter [121], which are also conditioned on warped images, our model can consistently generate high-fidelity views with detailed texture and structural information.

for acceleration. In accordance with prior work [79], switching from a scaled-linear noise schedule to a linear schedule is essential for achieving improved global consistency across multiple views. Additionally, we implement cross-attention between the latents of multiple views and per-token CLIP embeddings of reference images using a linear guidance mechanism [86]. For training, we randomly select a subset of frames from a video clip as reference images, with the remaining frames serving as target images. The number of reference images is randomly chosen to accommodate different downstream tasks. The multi-view diffusion model is optimized by calculating the loss only on the target images, as outlined in Eq. 1.

C.2. Training Details

Brightness Control. We observe that the *visual-condition* effectively guides camera movement but cannot control brightness changes, posing a significant limitation. Determining the light source position is particularly challenging with limited observations from single or sparse views. In our real-world test data, camera movement often causes random highlighting or

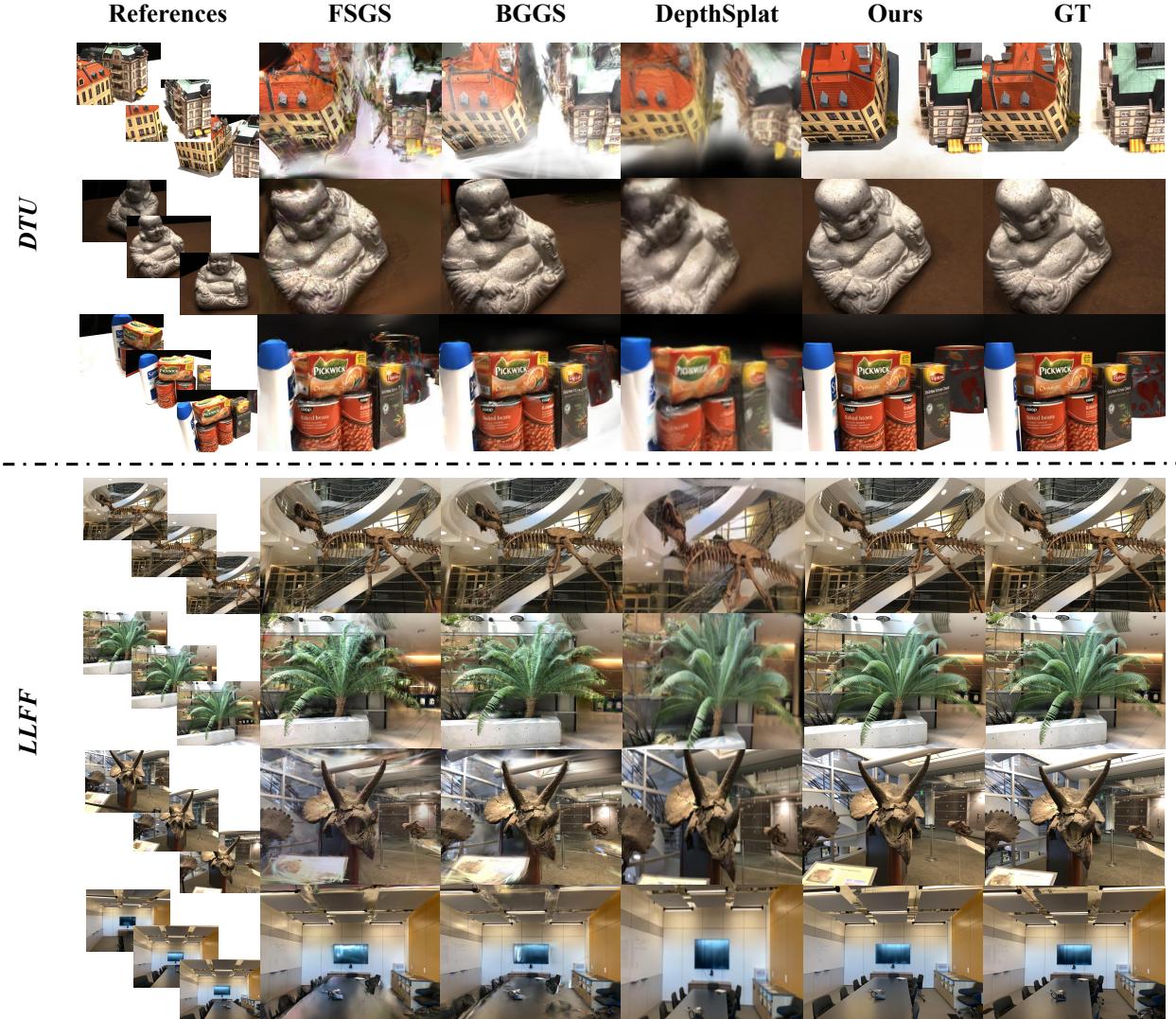
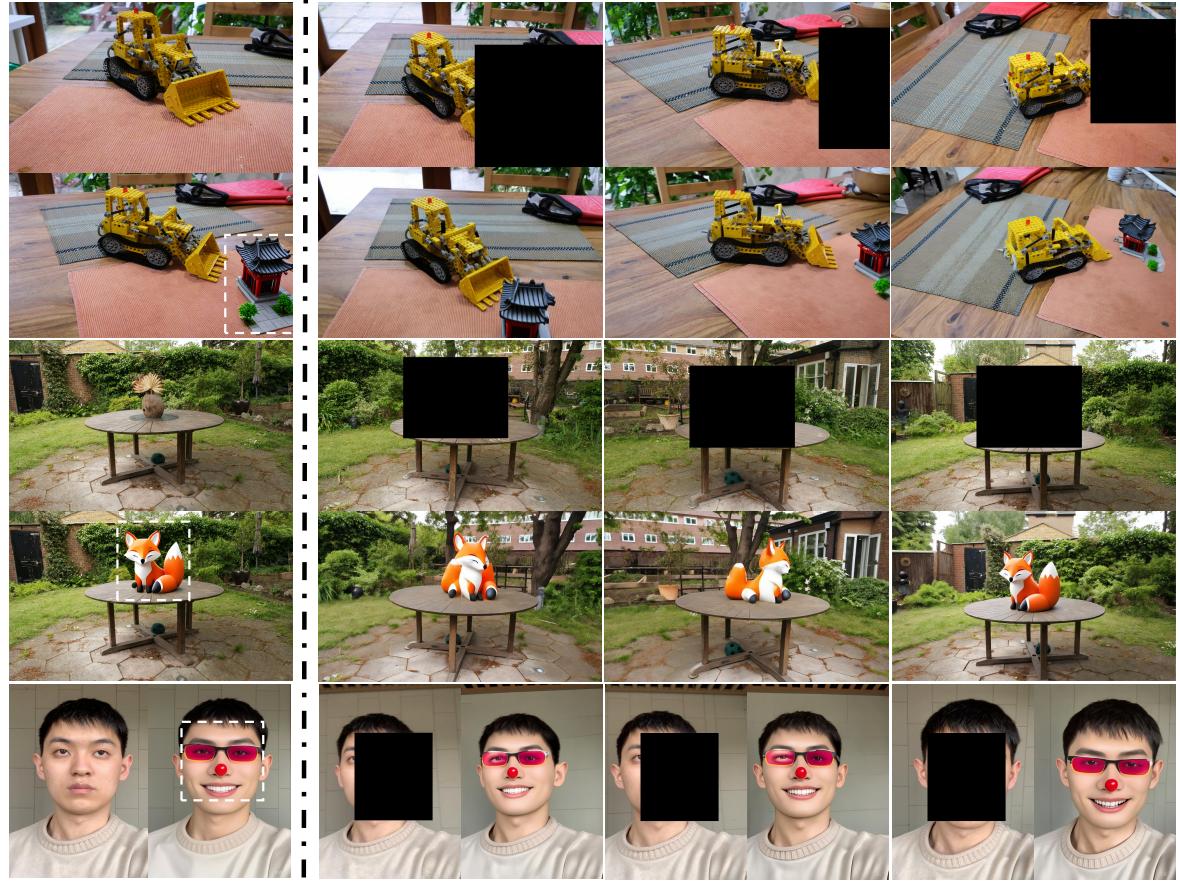


Figure 8. **Sparse-views to 3D**. Given 3 input views, our model generates clear, high-fidelity novel views that closely match the ground truth (GT), without artifacts or blurring. Note that the results from DepthSplat [114] are cropped and resized following the same data processing as the official source code.

darkening in some regions of scenes, which has a significant impact on pixel-level metrics like PSNR. This issue highlights a key problem: the inability to control brightness undermines the reliability of pixel-level metrics, as brightness variations affect these metrics more than the actual quality of the generated content. To achieve illumination control, 1) we preprocess the training data by converting corrupted images into HSV format, which represents hue, saturation, and brightness. 2) We define a $w \times h$ window and calculate the average brightness difference within this window between the ground truth image and the corrupted data. Using this difference, we apply a scaling factor to the brightness channel of the corrupted data while preserving hue and saturation, before converting the image back to RGB. This ensures brightness adjustment in the *visual-condition* without leaking color or content from the ground truth.

During training, we randomly drop this preprocessing with a probability of 0.5, enabling the model to infer lighting changes on its own during inference when brightness control is not required. In our evaluation experiments, brightness scaling is applied to the unmasked regions of warped images to align with ground truth, reducing the impact of brightness, and thus yielding a higher correlation between the generated content and pixel-level metrics. Meanwhile, keeping hue and saturation



Ori. & 2D Ref. view

Masks & 3D Editing Results

Figure 9. Examples of Open-world 3D Editing. (a) Occlusion-free Editing: An Asian-style attic is added, and novel views are generated realistically. (b) Full Replacement Editing: A vase is replaced with a toy fox, seamlessly integrated into the scene from various viewpoints. (c) Occluded Editing: Hidden regions in the masked areas are inferred and completed to produce novel views.

unchanged to avoid content or color leakage. Additionally, the model enables user-controlled brightness adjustments for specific regions in multi-view generation by modifying the *visual-condition* as needed.

Training Configuration. We initialize the **See3D** model from MVDream [80] and employ a progressive training strategy. First, the model is trained at a resolution of 512×512 with a sequence length of 5. This phase involves 120,000 iterations, using 1 reference view and 4 target views. Due to the relatively small sequence length, a larger batch size of 560 is used to enhance stability and accelerate convergence. Next, the sequence length is increased to 16, and the model is trained for 200,000 iterations with 1 or 3 reference views and 15 or 13 target views, maintaining the resolution of 512×512 . In this phase, the batch size is reduced to 228. Finally, a multi-view super-resolution model is trained using the same network structure. It takes the multi-view predictions from **See3D** as input and outputs target images with multi-view consistency at a resolution of 1024×1024 , using a batch size of 114. In all stages, all parameters of the diffusion model are fine-tuned with a learning rate of 1e-5. Additionally, we render some multi-views or extract clips from datasets such as Objaverse [15], CO3D [75], RealEstate10k [129], MVIImgNet [122], and DL3DV [50] datasets, forming a supplemental 3D dataset with fewer than 0.5M samples, please refer to Section E.2 for details on analysis and ablation. During training, this supplemental data is randomly sampled and incorporated into our WebVi3D dataset ($\sim 16M$). To enhance training efficiency, we utilize FlashAttention [14] alongside DeepSpeed with ZeRO stage-2 optimizer [74] and bf16 precision. We also implement classifier-free guidance (CFG) [30] by randomly dropping visual conditions with a probability of 0.1. The **See3D** model is trained on 114 \times NVIDIA-A100-SXM4-40GB GPUs over approximately 25 days using a progressive training scheme. During inference, a DDIM sampler [85] with classifier-free guidance is employed.

C.3. Definition of $f(t)$ and W_t

Definition for $f(t)$. In Eq.2, C_t is formulated as $C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)\mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_{t'}}\boldsymbol{\epsilon}$, where $\alpha_{t'}$ is a composite function that depends on α and t' , with $t' = f(t)$ and $f(t) = \beta \cdot t$. In our experiments, we set the hyper-parameter $\beta = 0.2$, which controls the noise level added to C_t . A larger β increases the noise in C_t . As β approaches 1, C_t converges toward a Gaussian distribution, improving robustness but reducing the correlation between C_t and \mathbf{X}_0 , thereby weakening camera control. Conversely, as β approaches 0, the distributions of C_t and \mathbf{X}_0 become more similar, improving controllability. However, for downstream tasks, a very small β creates a significant domain gap between task-specific visual cues and the video data, compromising robustness. Thus, β serves as a trade-off parameter, balancing camera control and robustness.

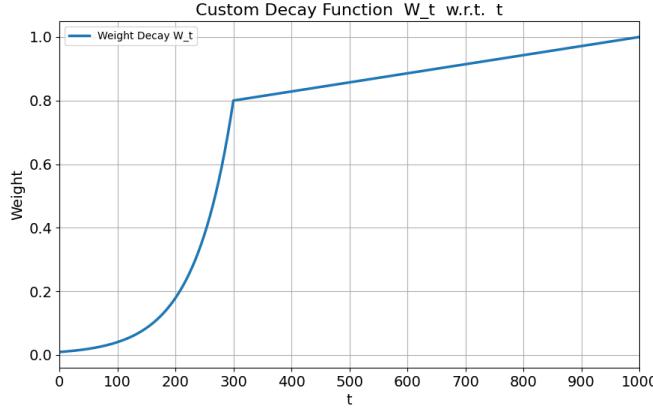


Figure 10. **Piecewise Function W_t** , showing linear decay for timesteps t between 300 and 1000, and a monotonically decreasing concave behavior for $t < 300$.

Formulation for W_t . Recapping Eq.3 from the main manuscript, $V_t = [W_t * C_t + (1 - W_t) * X_t; M]$, where W_t is defined as a piecewise function of t .

$$W_t = \begin{cases} v_{\text{decay_end}} \cdot e^{-b \cdot (t_{\text{decay_end}} - t)}, & \text{if } t < t_{\text{decay_end}}, \\ 1 - (1 - v_{\text{decay_end}}) \cdot \frac{t_{\text{peak}} - t}{t_{\text{peak}} - t_{\text{decay_end}}}, & \text{if } t \geq t_{\text{decay_end}}, \end{cases}$$

where $t_{\text{peak}} = 1000$, $t_{\text{decay_end}} = 300$, $v_{\text{decay_end}} = 0.8$, and $b = 0.075$. To ensure that W_t remains within the range $[0, 1]$, it is clamped as: $W_t = \text{clamp}(W_t, 0, 1)$. As shown in Figure 10, 1) For t between 300 and 1000, W_t decreases linearly as t decreases; 2) For $t < 300$, W_t transitions to a monotonically decreasing concave function of t .

The rationale behind this design is to ensure that when C_t has significant noise, it exerts a stronger influence on V_t , thus affecting MVD generation. Conversely, as the noise in C_t diminishes, X_t rapidly replaces C_t , reducing the risk of information leakage from C_t and improving the robustness of task-specific visual cues. The formulation of W_t enables flexible parameter tuning, such as $v_{\text{decay_end}}$ and b , to control its monotonic behavior. Smaller parameter values emphasize the impact of C_t on MVD, while larger values prioritize robustness.

D. More Experimental Results

Leveraging the developed web-scale dataset WebVi3D, our model supports both object- and scene-level 3D creation tasks, including single-view-to-3D, sparse-view-to-3D, and 3D editing. Additional experimental results for these tasks are presented below.

D.1. Single View to 3D

Table 4 presents a quantitative comparison of zero-shot novel view synthesis performance on the Tanks-and-Temples [43], RealEstate10K [129], and CO3D [75] datasets. Our method consistently outperforms all others on both easy and hard sets, achieving the best results in every evaluation metric. Qualitative results are shown in Figure 7. Compared to warping-based competitors such as LucidDreamer [12] and ViewCrafter [121], our approach more effectively captures both geometric

structure and texture details, producing more realistic 3D scenes. These results highlight the robustness and versatility of our method in synthesizing high-quality novel views across diverse and challenging scenarios.

Dataset	Easy set			Hard set		
	Method	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑
Tanks-and-Temples						
LucidDreamer [12]	0.413	14.53	0.362	0.558	11.69	0.267
ZeroNVS [77]	0.482	14.71	0.380	0.569	12.05	0.309
MotionCtrl [104]	0.400	15.34	0.427	0.473	13.29	0.384
ViewCrafter	0.194	21.26	0.655	0.283	18.07	0.563
ViewCrafter*	0.221	20.39	0.648	0.289	17.86	0.584
Ours	0.167	25.01	0.756	0.214	22.52	0.714
RealEstate10K						
LucidDreamer [12]	0.315	16.35	0.579	0.400	14.13	0.511
ZeroNVS [77]	0.364	16.50	0.577	0.431	14.24	0.535
MotionCtrl [104]	0.341	16.31	0.604	0.386	16.29	0.587
ViewCrafter	0.145	21.81	0.796	0.178	22.04	0.798
ViewCrafter*	0.164	20.59	0.825	0.201	20.40	0.778
Ours	0.125	26.54	0.872	0.167	24.18	0.837
CO3D						
LucidDreamer [12]	0.429	15.11	0.451	0.517	12.69	0.374
ZeroNVS [77]	0.467	15.15	0.463	0.524	13.31	0.426
MotionCtrl [104]	0.393	16.87	0.529	0.443	15.46	0.502
ViewCrafter	0.243	21.38	0.687	0.324	18.96	0.641
ViewCrafter*	0.331	20.12	0.703	0.348	18.02	0.653
Ours	0.225	25.23	0.781	0.276	23.33	0.748

Table 4. Zero-shot Novel View Synthesis (NVS) on Tanks-and-Temples[43], RealEstate10K[129] and CO3D[75] dataset.

D.2. Sparse Views to 3D

Quantitative comparisons using 3, 6, and 9 input views are presented in Table 5. The 3DGS model trained on multi-view images generated by **See3D** outperformed state-of-the-art models in novel view rendering, demonstrating its ability to provide consistent multi-view support for 3D reconstruction without additional constraints. Qualitative comparisons in Figure 8 reveal fewer floating artifacts in the NVS results, indicating **See3D** generates higher-quality and more consistent multi-view images.

D.3. 3D Editing

Our model, trained on large-scale videos, naturally supports open-world 3D editing without the need for additional fine-tuning. Figure 9 illustrates three distinct editing scenarios: a) *Occlusion-free Editing*. An Asian-style attic is placed next to a toy bulldozer in the original image, which serves as the reference view. Our model generates highly realistic images containing the Asian-style attic from various new viewpoints. b) *Full Replacement Editing*. The vase in the original image is completely replaced with a toy fox. Our model generates new scenes from different viewpoints, seamlessly incorporating the toy fox into the designated area with no residual traces of the vase. c) *Occluded Editing*. Given an occluded edited image as a reference view, our model can generate multiple novel views within the specified masked regions, inferring and filling in the hidden details of the occluded parts.

E. Additional Ablation Studies

E.1. Effectiveness of Pixel-level Depth Alignment

We conducted additional ablation experiments to validate the effectiveness of the proposed pixel-level depth alignment. Specifically, we enabled and disabled pixel-level depth alignment when generating novel views through warping and visualized the warped results at a specific generation step. As shown in Figure 11, the left image shows the reference GT image, the

Dataset	3-view			6-view			9-view		
	Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow
LLFF									
Zip-NeRF* [4]	17.23	0.574	0.373	20.71	0.764	0.221	23.63	0.830	0.166
MuRF [113]	21.34	0.722	0.245	23.54	0.796	0.199	24.66	0.836	0.164
FSGS [130]	20.31	0.652	0.288	24.20	0.811	0.173	25.32	0.856	0.136
BGGS [27]	21.44	0.751	0.168	24.84	0.845	0.106	26.17	0.877	0.090
ZeroNVS* [77]	15.91	0.359	0.512	18.39	0.449	0.438	18.79	0.470	0.416
DepthSplat [114]	17.64	0.521	0.321	17.40	0.499	0.340	17.26	0.486	0.341
ReconFusion [107]	21.34	0.724	0.203	24.25	0.815	0.152	25.21	0.848	0.134
CAT3D [23]	21.58	0.731	0.181	24.71	0.833	0.121	25.63	0.860	0.107
Ours	23.23	0.768	0.135	25.32	0.820	0.104	26.19	0.844	0.098
DTU									
Zip-NeRF* [4]	9.18	0.601	0.383	8.84	0.589	0.370	9.23	0.592	0.364
MuRF [113]	21.31	0.885	0.127	23.74	0.921	0.095	25.28	0.936	0.084
FSGS [130]	17.34	0.818	0.169	21.55	0.880	0.127	24.33	0.911	0.106
BGGS [27]	20.71	0.862	0.111	24.31	0.917	0.073	26.70	0.947	0.052
ZeroNVS* [77]	16.71	0.716	0.223	17.70	0.737	0.205	17.92	0.745	0.200
DepthSplat [114]	15.59	0.525	0.373	15.061	0.523	0.406	14.87	0.478	0.451
ReconFusion [107]	20.74	0.875	0.124	23.62	0.904	0.105	24.62	0.921	0.094
CAT3D [23]	22.02	0.844	0.121	24.28	0.899	0.095	25.92	0.928	0.073
Ours	28.04	0.884	0.073	29.09	0.900	0.066	29.99	0.911	0.059
Mip-NeRF 360									
Zip-NeRF* [4]	12.77	0.271	0.705	13.61	0.284	0.663	14.30	0.312	0.633
DepthSplat [114]	13.85	0.254	0.621	13.82	0.260	0.636	14.48	0.288	0.602
ZeroNVS* [77]	14.44	0.316	0.680	15.51	0.337	0.663	15.99	0.350	0.655
ReconFusion [107]	15.50	0.358	0.585	16.93	0.401	0.544	18.19	0.432	0.511
CAT3D [23]	16.62	0.377	0.515	17.72	0.425	0.482	18.67	0.460	0.460
Ours	17.35	0.442	0.422	19.03	0.517	0.365	19.89	0.542	0.335

Table 5. Quantitative Comparison of Sparse-view 3D Reconstruction

middle image corresponds to warping with pixel-level aligned depth, and the right one depicts warping without pixel-level aligned depth. The results demonstrate that pixel-level depth alignment not only effectively restores the scale of the depth map but also significantly corrects errors in monocular depth estimation (e.g., the toy’s neck and the tabletop). Consequently, integrating our proposed 3D generation pipeline improves generation quality.

E.2. Efficacy of Scaling up Data



Figure 11. Ablation on Pixel-level Depth Alignment.

Model	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow
MV-UnposeT	0.194	25.56	0.811
MV-UnposeT-10%	0.187	25.95	0.817
MV-UnposeT-20%	0.183	26.19	0.820
MV-UnposeT-60%	0.181	26.14	0.819
MV-Posed	0.182	26.21	0.822

Table 6. Ablation on Supplementary 3D Data.

In the main manuscript, we conducted an ablation study on the 3D dataset MVImageNet [122] to evaluate the effectiveness of the proposed *visual-condition*. Table 3 shows that: 1) When conditioned on purely masked images, the MV-UnPoseM model performed the worst, struggling with the domain gap issue. 2) When conditioned on pose-guided warped images, the MV-Posed model achieved the best results, benefiting from pose annotations. 3) Our MV-UnposeT model, conditioned on

the time-dependent *visual-condition*, demonstrated performance very close to that of the MV-Posed model.

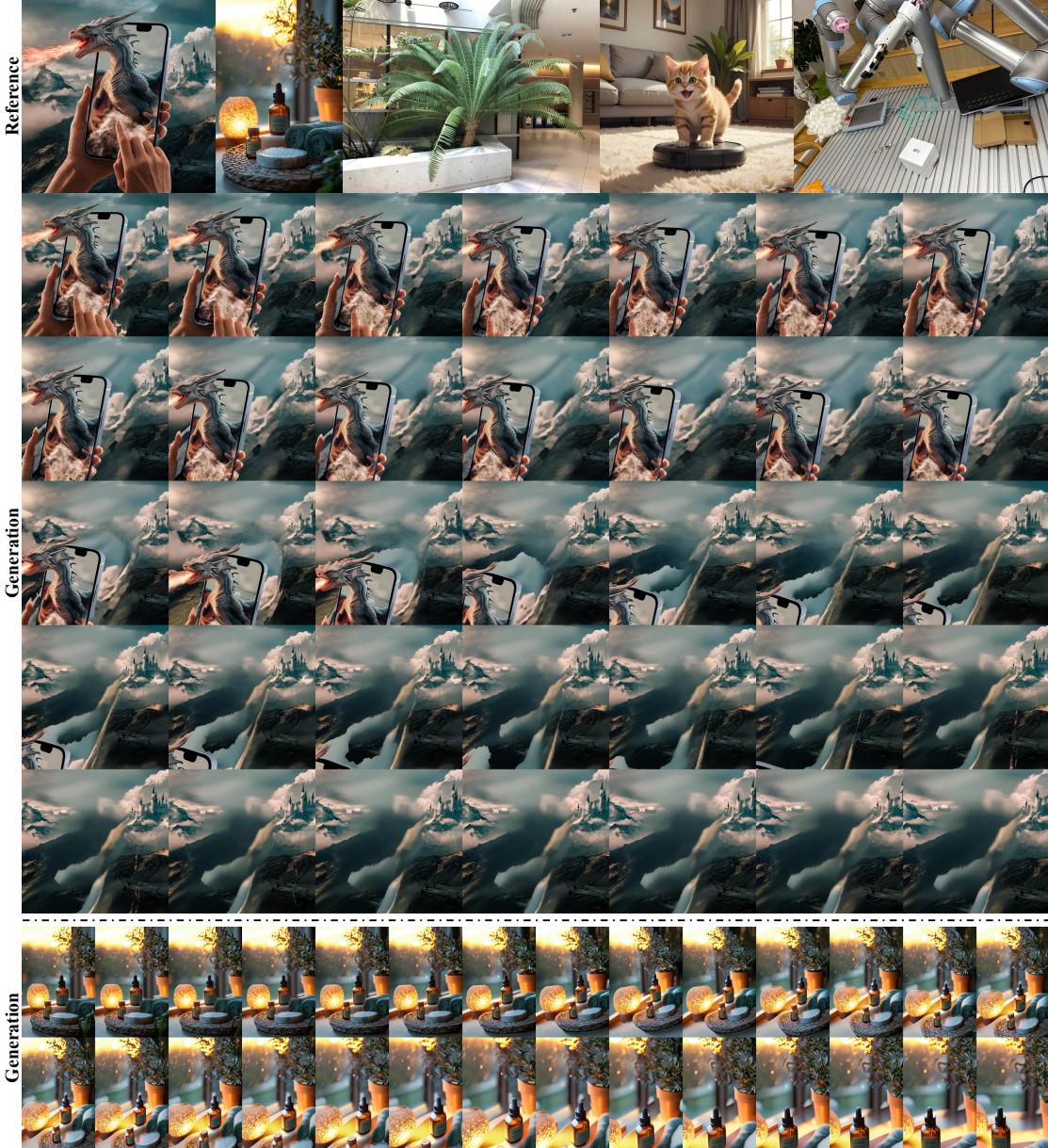


Figure 12. **Examples of Long-sequence Generation.** High-quality novel views generated along complex camera trajectories, maintaining spatial consistency and visual realism across extended sequences.

Intuitively, models trained entirely on 3D data tend to achieve optimal performance at a specific data scale, establishing an upper bound at that scale. When the volume of video data matches that of 3D data, models trained on 3D still set the performance ceiling. However, as video data is virtually unlimited, scaling up the dataset can intuitively raise this upper bound.

Following the same settings in Table 3, we further investigate the impact of supplementing multi-view data with 3D annotations on model performance. We conduct an ablation study using the MV-UnposeT model, trained on unposed multi-view data with *visual-condition*. In this study, we progressively introduce 3D pose annotations at levels of 10%, 20%, 60%, and 100% into the training set. When the training data is entirely composed of 3D annotations, the model configuration is equivalent to the MV-Posed model. The results in Table 6 indicate that our MV-UnposeT model, initially trained on unposed data, improves steadily as 3D annotations are introduced. For instance, with only 20% 3D data (MV-UnposeT-20%), the

model’s performance closely approaches that of the fully 3D-annotated MV-Posed model. This suggests that even a small amount of 3D data in a largely unposed multi-view dataset can significantly boost model performance, approaching the models trained on fully annotated 3D datasets.

This insight is essential because unposed multi-view data is cost-effective and can be easily collected in large quantities. By incorporating a small volume of high-quality 3D data, we can achieve performance comparable to models trained on large, expensive 3D datasets. Therefore, in our proposed WebVi3D dataset (16M samples), we incorporated a small portion (0.5M samples) of 3D data to optimize model performance.

F. Additional Visualizations

Open-world 3D Generation with Long Sequences. We manually configured complex camera trajectories, including rotation, translation, zooming in, zooming out, focus distance adjustments and various random combinations, as shown in Figure 12 and Figure 13. Our model consistently generates high-quality, continuous novel views along these trajectories. Experimental visualizations demonstrate that the model effectively preserves spatial consistency and visual realism across long sequences. This highlights its robustness in handling intricate camera paths, including rapid transitions and diverse perspectives, making it highly applicable to open-world scenarios.



Figure 13. More Examples of Long-sequence Generation.