# GeoVideo: Introducing Geometric Regularization into Video Generation Model

**Yunpeng Bai**[1] **Shaoheng Fang**[1] **Chaohui Yu**[2,3] **Fan Wang**[2] **Qixing Huang**[1]

[1]The University of Texas at Austin, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab
https://geovideo.github.io/GeoVideo/

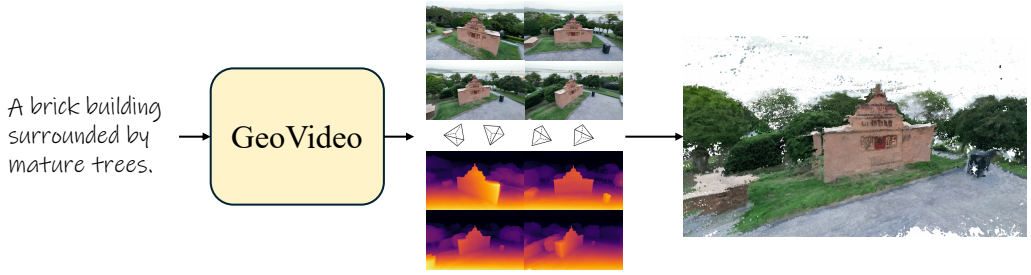


Figure 1: GeoVideo introduces a geometric consistency loss using predicted depths and camera poses to enhance multi-view consistency of the output frames, leading a high-quality 3D reconstruction from the output video frames.

## Abstract

Recent advances in video generation have enabled the synthesis of high-quality and visually realistic clips using diffusion transformer models. However, most existing approaches operate purely in the 2D pixel space and lack explicit mechanisms for modeling 3D structures, often resulting in temporally inconsistent geometries, implausible motions, and structural artifacts. In this work, we introduce geometric regularization losses into video generation by augmenting latent diffusion models with per-frame depth prediction. We adopted depth as the geometric representation because of the great progress in depth prediction and its compatibility with image-based latent encoders. Specifically, to enforce structural consistency over time, we propose a multi-view geometric loss that aligns the predicted depth maps across frames within a shared 3D coordinate system. Our method bridges the gap between appearance generation and 3D structure modeling, leading to improved spatio-temporal coherence, shape consistency, and physical plausibility. Experiments across multiple datasets show that our approach produces significantly more stable and geometrically consistent results than existing baselines.

## 1 Introduction

Video generation has recently made significant strides in creating visually impressive and high-quality clips [4, 28, 69, 29, 53, 77]. Powered by diffusion transformer models [39] and large-scale training datasets [2, 63, 8, 61], these systems are capable of synthesizing realistic videos conditioned on various inputs, such as text prompts [26, 35, 26, 62] or images [12, 43, 57, 38]. Despite these successes, current video generation models often fail to accurately capture the underlying geometry, coherent motions, and physical consistency in dynamic scenes. As a result, the generated videos,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

although plausible at first glance, often exhibit temporal artifacts such as shape deformation, structure flickering, and implausible object interactions.

This limitation stems from the fact that most video generation models operate purely in the 2D pixel/latent space and rely on temporal attention to promote cross-frame coherence. Although effective in maintaining short-term consistency, these approaches lack explicit mechanisms to model 3D structures, leading to violations of object permanence, shape integrity, and motion realism.

This shortcoming highlights a deeper insight: realistic video generation demands more than visual coherence—it requires a structured understanding of the 3D world. After all, videos can naturally encode spatio-temporal observations of real environments. Viewed through this lens, video generation can be reframed as a form of *world modeling*—the construction of continuous, physically grounded representations of the dynamic world. Emerging research on world generation [4, 5] underscores its potential in applications such as 3D scene synthesis [34, 70, 44], robotics [64, 20, 56], and embodied AI [41, 45, 14]. However, achieving physically plausible world modeling requires the generated scenes to maintain consistent geometry over time, which current models struggle to enforce.

To address these limitations, we propose **GeoVideo**, which introduces *geometric regularization* into the video generation process. Specifically, we augment the generative model to predict per-frame depth maps alongside RGB frames and enforce cross-frame depth consistency. This regularization encourages the model to maintain coherent 3D geometry throughout the video. By aligning predicted depth across consecutive frames, the model is guided to better capture the underlying scene structure, resulting in enhanced realism, temporal stability, and physical plausibility. Our key insight is that depth consistency offers implicit geometric supervision that complements appearance-based learning. This helps bridge the gap between 2D frame-level synthesis and 3D-consistent scene modeling, paving the way toward more structured and physically grounded video generation.

The main contributions of this work are:

- **Explicit Geometry Modeling in Video Generation.** We introduce per-frame depth prediction into latent diffusion-based video generation models, enabling explicit modeling of 3D scene structure throughout the generation process.
- **Geometric Regularization.** We propose a cross-frame consistency loss that lifts predicted depths into a global 3D point cloud and supervises them via multi-view reprojection alignment, encouraging globally coherent geometry.
- **Improved Spatiotemporal Coherence.** Our approach significantly enhances structural consistency, motion stability, and geometric plausibility in generated videos, as demonstrated on several benchmarks.

## 2 Related Work

### 2.1 Diffusion Models for Video Generation

The remarkable success of diffusion models in image generation [42, 47, 46] has recently inspired their extension to video generation [25, 65, 13], where they have quickly become the dominant paradigm. In particular, latent diffusion [55, 46] has emerged as a widely adopted strategy: a VAE [27] module first encodes video data into a compact latent space, and the diffusion process is performed within this lower-dimensional representation. Current state-of-the-art methods [69, 29, 76] utilize a 3D Variational Autoencoder [69] in combination with a Diffusion Transformer (DiT) [39] backbone, achieving highly realistic and high-fidelity video synthesis. Despite these advances, existing diffusion-based video generation models [18, 3] often struggle to accurately capture geometric structures, coherent temporal motions, and physical consistency. To mitigate these limitations, recent research has explored the introduction of additional priors to guide the generation process toward better alignments with real-world dynamics.

For example, Track4Gen [22] jointly models video generation and point tracking across frames within a single network, providing enhanced spatial supervision over diffusion features and improving both motion and structural consistency. Similarly, VideoJAM [6] learns a joint appearance-motion representation that instills an effective motion prior to the video generator. Other contemporary approaches have also taken advantage of motion representations to improve motion coherence in image-to-video generation tasks [50, 59]. Meanwhile, OmniVDiff [66] models appearance, depth,

Canny edges, and semantic segmentation simultaneously, enabling multi-modal video generation and multi-modal conditional generation. However, it does not explicitly impose priors on the generated auxiliary signals. IDOL [73] proposes a human-centered joint video-depth generation framework, but it does not introduce explicit priors and is limited to human-centered scenarios. WVD [75] supports the simultaneous generation of appearance and point representations but is restricted to image-to-video translation in small static environments. Complementary to these efforts, Yue et al. [72] proposed to lift the semantic characteristics of each frame into a 3D Gaussian representation, demonstrating that fine-tuning a foundation model with these 3D-aware characteristics leads to better performance across downstream tasks. Building on these insights, our work seeks to jointly model geometry during video generation and introduce an explicit geometric regularization loss to further improve the quality, consistency, and realism of synthesized videos.

### 2.2 Geometry-Related Tasks with Pretrained Diffusion Models

Another line of work uses pre-trained generative models [46, 3, 67] for geometry-related tasks. A pioneering effort in this direction is Marigold [24], which first proposed treating depth as an image-like modality. By encoding depth maps into the same latent space as RGB images using a latent diffusion VAE, Marigold demonstrated that image generation models can be repurposed into depth estimators. This idea has inspired a series of subsequent works [10, 1] that exploit the strong priors of diffusion models to estimate geometric properties such as depth and surface normals. Following this direction, DepthCrafter [19] and Depth Any Video [68] adapt video generation models to perform video-based depth estimation, effectively extending Marigold's latent-space strategy from images to videos. Similarly, DiffusionRender [32] leverages video generation models for inverse rendering tasks, jointly recovering not only scene geometry but also materials and lighting from video sequences. Building on these advancements, Geo4D [23] further expands the use of video generators to tackle 4D scene reconstruction, capturing dynamic scene structures over time.

### 2.3 3D Scene Generation

The generation of 3D content has also emerged as a highly active area of research. The early text-to-3D scene generation methods [17, 9] mainly relied on image generation inpainting models and progressively completed a scene through multiple iterations, resulting in limited efficiency and quality. Subsequent methods design specialized models for text-to-3D scene generation. Director3D [30] proposes a text-to-3D generation framework capable of synthesizing both real-world 3D scenes and adaptive camera trajectories. Recently, some methods like SplatFlow [11] and Bolt3D [52] have also leveraged intermediate 3D representations to directly generate scenes in the form of 3D Gaussian Splatting, either through multiview diffusion or flow matching models. Prometheus [71] introduces a multiview diffusion model based on an RGB-D latent space to generate 3D Gaussian scenes. However, since Prometheus relies primarily on image-based models, the quality and fidelity of the generated 3D content remain limited. Apart from the methods mentioned above, LDM3D [51] is a previous work similar to ours that re-trains Stable Diffusion, but its scope is limited to the latent space of RGB-D images. Our method incorporates geometric regularization into the video generation model, enabling the direct extraction of high-quality 3D scenes from the generated videos.

## 3 Approach

We begin by reviewing the preliminaries of the video diffusion models in Section 3.1. Section 3.2 and Section 3.3 then describe the proposed RGB-D video generation model and our novel geometric regularization loss, respectively. Finally, Section 3.4 introduces the training procedure.

### 3.1 Preliminaries: Video Diffusion Models

Our method is based on latent diffusion-based video generation models, particularly those with 3D VAE backbones such as CogVideoX [69], and transformer-based denoisers such as DiT [39]. These models encode video frames into a compact latent space and apply denoising diffusion to synthesize temporally coherent video sequences. Formally, let $\mathbf{x}_{1:T}$ denote a video clip of $T$ frames, and let $\mathbf{z} = E(\mathbf{x}_{1:T})$ be its latent representation encoded by a VAE. Latent diffusion models define a forward noising process over $\mathbf{z}$ by gradually adding Gaussian noise: $q(\mathbf{z}^{(t)} \mid \mathbf{z}^{(t-1)}) =$

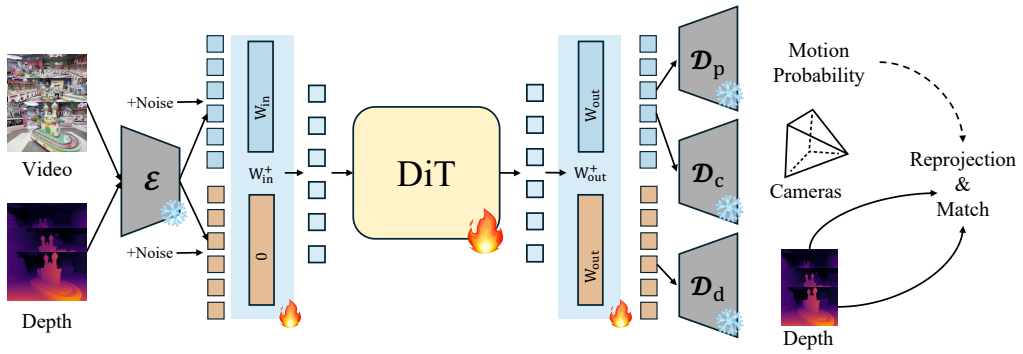


Figure 2: **Overview of the proposed method.** The **orange-yellow** modules are additionally added components. $\mathcal{D}_p$, $\mathcal{D}_c$, and $\mathcal{D}_d$ respectively denote the heads for decoding motion probability, camera pose, and depth from the generated features. We use their outputs to define a geometric consistency loss to regularize the video generative model.

$\mathcal{N}(\mathbf{z}^{(t)}; \sqrt{\alpha_t}\mathbf{z}^{(t-1)}, (1-\alpha_t)\mathbf{I})$, where $t = 1, \ldots, S$ indexes the diffusion timestep and $\alpha_t$ is a variance schedule. The model learns a reverse process to sample denoised latents: $p_\theta(\mathbf{z}^{(t-1)} \mid \mathbf{z}^{(t)}) = \mathcal{N}(\mathbf{z}^{(t-1)}; \mu_\theta(\mathbf{z}^{(t)}, t), \Sigma_\theta(t))$, which is parameterized by a spatio-temporal transformer (DiT). After obtaining the denoised latent $\mathbf{z}^{(0)}$, the final RGB video is reconstructed via the VAE decoder.

### 3.2 RGBD Video Modeling

To incorporate explicit geometric structure into the video generation process, we choose to represent scene geometry using per-frame depth maps. This choice is motivated by two key factors. First, recent advances [7, 19] in video depth estimation have yielded robust and high-quality predictions, making it feasible to obtain reliable depth supervision even in unconstrained settings. Second, depth maps have a natural image-like structure and can be efficiently encoded into the same latent space as RGB frames using a shared VAE encoder. This design has been validated in prior works such as Marigold [24] and DepthCrafter [19]. Using this modality compatibility, we can extend existing latent video diffusion models to jointly generate RGB and depth with minimal architectural modifications.

As shown in Figure 2, given a pair of RGB and depth frames $(\mathbf{x}^{\text{RGB}}_{1:T}, \mathbf{x}^{\text{D}}_{1:T})$, we encode them in a shared latent space:

$$\mathbf{z} = [\mathbf{z}^{\text{RGB}}; \mathbf{z}^{\text{D}}] = [E(\mathbf{x}^{\text{RGB}}_{1:T}); E(\mathbf{x}^{\text{D}}_{1:T})],$$

where $E(\cdot)$ denotes the VAE encoder and $[\cdot;\cdot]$ denotes channel-wise concatenation. The diffusion model operates on the latent sequence $\mathbf{z}_{1:T}$ and learns to jointly denoise both modalities. The generated latent is then decoded by the VAE decoder into video frames and depth maps.

We model the joint distribution over RGB and depth frames as:

$$P(\mathbf{x}^{\text{RGB}}_{1:T}, \mathbf{x}^{\text{D}}_{1:T}) = P(\mathbf{z}) \prod_{t=1}^{T} P(\mathbf{x}^{\text{RGB}}_t, \mathbf{x}^{\text{D}}_t \mid \mathbf{z}),$$

where each pair is generated from the same underlying latent representation. This formulation enables a tight coupling between appearance and geometry throughout the generation process.

### 3.3 Introducing Geometric Regularization

Although per-frame depth maps provide localized 3D cues, they do not guarantee cross-frame consistency. To enforce coherent 3D structure over time, we introduce a geometric regularization loss that lifts predicted depths into world coordinates using known camera intrinsics and extrinsics. Since depth and appearance are decoded from the same underlying features, applying supervision on the depth prediction allows us to enhance the geometric consistency of the underlying shape without needing to account for view-dependent appearance differences across views.

**Global point cloud construction.** Since the built-in 3D VAE in video generation models is computationally heavy, we additionally train a lightweight decoder $\mathcal{D}_d$ to convert the depth latent $\mathbf{z}^{\text{D}}$ into depth map $\mathbf{D} \in \mathbb{R}^{H \times W}$ of the same resolution as the RGB frame, enabling more fine-grained and accurate geometric regularization. In parallel, and inspired by VGGT [58], we also predict the camera intrinsics and extrinsics for each frame using a camera head from generated $\mathbf{z}^{\text{RGB}}$. Let $\mathbf{D}_i$ be the predicted depth map for frame $i$, and $\mathbf{P}_i \in \text{SE}(3)$ its camera pose. Using the intrinsic matrix $K$, we backproject depth into the camera space and transform it into the world space:

$$\mathbf{X}_i = \mathbf{P}_i \cdot \pi^{-1}(\mathbf{D}_i, K), \tag{1}$$

where $\pi^{-1}$ denotes backprojection from depth to 3D coordinates. The global point cloud is then obtained by aggregating:

$$\mathcal{X}_{\text{global}} = \bigcup_{i=1}^{T} \mathbf{X}_i. \tag{2}$$

We denoise $\mathcal{X}_{\text{global}}$ using voxel grid downsampling and statistical outlier removal to improve robustness and computational efficiency.

**Depth reprojection consistency.** To supervise consistency, we reproject the global point cloud back to each frame and compare its depth with the predicted depth map. For frame $i$, we project:

$$\hat{\mathbf{D}}_i(\mathbf{u}) = \pi_z(\mathbf{P}_i^{-1} \cdot \mathbf{x}), \quad \mathbf{x} \in \mathcal{X}_{\text{global}}, \tag{3}$$

where $\pi_z(\cdot)$ denotes depth value after projection into image coordinates $\mathbf{u}$. We then compute the loss:

$$\mathcal{L}_{\text{geo}} = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{|\mathcal{V}_i|} \sum_{\mathbf{u} \in \mathcal{V}_i} \mathbb{1}(|\hat{\mathbf{D}}_i(\mathbf{u}) - \mathbf{D}_i(\mathbf{u})| < \delta) \cdot |\hat{\mathbf{D}}_i(\mathbf{u}) - \mathbf{D}_i(\mathbf{u})|, \tag{4}$$

where $\mathcal{V}_i$ is the set of valid pixels and $\delta$ is a tolerance threshold set to 0.05. This encourages global shape consistency by penalizing depth discrepancies only when local reprojections are reliable. When multiple points project to the same pixel, we use the average of these points to compute loss. For dynamic videos, we introduce an additional head $\mathcal{D}_p$ that predicts an object movement probability map [31] from the generated video features $\mathbf{z}^{\text{RGB}}$, representing pixels that correspond to dynamic content based on multi-frame information. For dynamic pixels identified in the probability map, we only align them with points of similar probability in adjacent frames.

### 3.4 Parameters Initialization and 2-stage Fine-tuning with Geometric Regularization

**Parameters Initialization.** To enable the pretrained video generation model to support dual-modality inputs (RGB and Depth), we modify the input and output projection layers of the transformer. Let $W_{\text{in}} \in \mathbb{R}^{C_v \times C_t}$ be the input projection matrix, where $C_v$ is the input feature dimension and $C_t$ is the transformer token dimension. Inspired by ControlNet [74], we extend it by vertically concatenating a zero matrix of the same shape, resulting in $W_{\text{in}}^{+} \in \mathbb{R}^{2C_v \times C_t}$. The associated input bias $b_{\text{in}} \in \mathbb{R}^{C_t}$ is left unchanged. On the output side, let $W_{\text{out}} \in \mathbb{R}^{C_t \times C_v}$ denote the output projection matrix. To accommodate dual outputs, we horizontally concatenate a copy of $W_{\text{out}}$, resulting in $W_{\text{out}}^{+} \in \mathbb{R}^{C_t \times 2C_v}$. The output bias $b_{\text{out}} \in \mathbb{R}^{C_v}$ is duplicated to form $b_{\text{out}}^{+} \in \mathbb{R}^{2C_v}$. The initialization is summarized as:

$$\begin{aligned} W_{\text{in}}^{+} &= \begin{bmatrix} W_{\text{in}} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{2C_v \times C_t}, \quad b_{\text{in}}^{+} = b_{\text{in}} \in \mathbb{R}^{C_t}, \\ W_{\text{out}}^{+} &= [W_{\text{out}} \quad W_{\text{out}}] \in \mathbb{R}^{C_t \times 2C_v}, \quad b_{\text{out}}^{+} = \begin{bmatrix} b_{\text{out}} \\ b_{\text{out}} \end{bmatrix} \in \mathbb{R}^{2C_v}. \end{aligned} \tag{5}$$

This initialization ensures that the depth modality is initially a zero-influence pathway, allowing the model to begin fine-tuning from the pretrained RGB state without disrupting performance. During fine-tuning, the model progressively learns to represent depth channels.


Figure 3: **Comparisons on the down-stream 3D reconstruction task.** The detailed prompt inputs are provided in the supp. materials. The top three rows show samples of video generation results of each method, while the corresponding 3D reconstructions are show in the bottom row.The videos generated by our method offer complete and high-quality 3D reconstructions.

**Stage 1: RGB-D Joint Generation.** In the first stage, we fine-tune the model to generate RGB and depth frames simultaneously. To ensure stable learning, we apply a gradually increasing weight to the depth loss:

$$\lambda_{\text{depth}}(t) = \min(1.0, 0.1 + \alpha t), \tag{6}$$

where $t$ is the training step and $\alpha$ is set to 0.0001. This gradual increase allows the model to adapt to the new depth supervision without destabilizing RGB generation. The loss for this stage is:

$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{\text{diff}}^{\text{RGB}} + \lambda_{\text{depth}}(t) \cdot \mathcal{L}_{\text{diff}}^{\text{D}}. \tag{7}$$

Here, $\mathcal{L}_{\text{diff}}^{\text{RGB}}$ and $\mathcal{L}_{\text{diff}}^{\text{D}}$ are formulated using the v-prediction [48] strategy.

**Stage 2: Geometric Regularization.** Once the model can generate perceptually and structurally coherent depth maps, we introduce the geometric regularization term $\mathcal{L}_{\text{geo}}$ (described in Section 3). This loss encourages cross-frame depth consistency via reprojection-based supervision.

The final training objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}}^{\text{RGB}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{diff}}^{\text{D}} + \lambda_{\text{geo}} \cdot \mathcal{L}_{\text{geo}}. \tag{8}$$

Here, $\lambda_{\text{geo}}$ is set to 0.5. This staged training process allows the model to gradually learn to incorporate geometry without sacrificing visual fidelity.

## 4 Experimental Results

**Implementation details.** Our experiments are primarily based on CogVideoX-5B [69], a popular and advanced diffusion-based video generation model built on the DiT architecture. We conducted experiments for both text-to-video (T2V) and image-to-video (I2V) generation tasks. The experiments

Table 1: **Multi-view geometric consistency evaluation on the DL3DV dataset.** We use VGGT to predict multi-frame depth and pose, and evaluate consistency using our proposed Multi-View Consistency Score (MVCS) and reprojection error. MVCS measures frame-to-frame depth consistency, ↑: higher is better; while Reprojection Error evaluates how well the globally reconstructed 3D structure aligns with original views, ↓: lower is better.

| **Method** | **MVCS ↑** | **Reproj. Error ↓** |
|---|---|---|
| CogVideoX-5B (T2V) | 61.2 | 4.58 |
| CogVideoX-5B (T2V, finetuned on DL3DV) | 66.4 | 3.91 |
| Ours w/o $\mathcal{L}_{\text{geo}}$ (T2V) | 71.3 | 3.36 |
| **Ours (T2V)** | **77.2** | **2.52** |

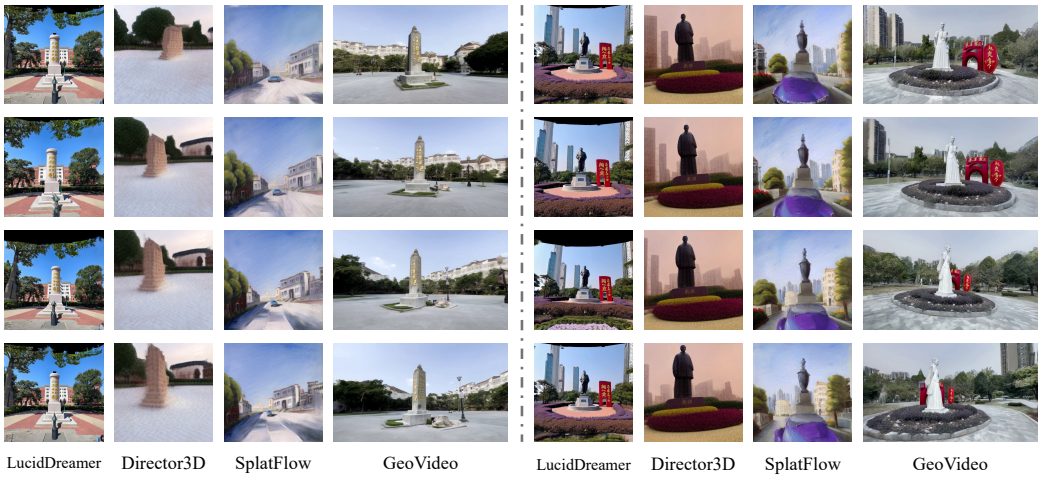


Figure 4: **Comparisons with text to 3d scene generation methods.** The detailed prompt inputs are provided in the supp. materials.

are categorized into two types: videos of static scenes and videos of dynamic scenes. For static scenes, we train on the DL3DV-10K [33] dataset. For dynamic videos, we collect a large-scale dataset of approximately 200,000 videos from online sources such as Pexels [40]. We use 544 and 1000 videos from the two datasets for evaluation, respectively. We train the model with a learning rate of 2e-5 on 8 H20 GPUs for 20,000 steps. The batch size is 1 per GPU, with 15K steps for stage 1 and 5K steps for stage 2. The video resolution is set to 768×1360 with 81 frames, following the standard configuration supported by CogVideoX. The depth labels for videos are estimated using Video Depth Anything [7], while for dynamic videos, we estimate camera poses using MegaSaM [31]. For video captions, we use CogVLM [60], as adopted in CogVideo. $\mathcal{D}_p$, $\mathcal{D}_c$, and $\mathcal{D}_d$ are distilled from the corresponding outputs of MegaSaM [31], VGGT [58], and Video Depth Anything [7], respectively. After distillation, their parameters are kept fixed during fine-tuning to provide supervision.

**Scene generation results.** The original video generation model already possesses some ability to generate coherent scenes. However, due to the typically small range of viewpoint changes, it is difficult to extract sufficient multi-view information from the generated videos to reconstruct a meaningful 3D scene (as shown in the first row of Figure 3). To specialize the model for scene-level video generation, we fine-tune the video generation model on the DL3DV dataset [33]. However, since the base model relies purely on 2D modeling, it struggles to maintain geometric consistency under rapid camera viewpoint changes. The second row of Figure 3 shows the results of CogVideoX after being fine-tuned on DL3DV, where such inconsistencies are still apparent. In contrast, after integrating our proposed geometric modeling framework, we are able to maintain consistent multi-view structures even under complex viewpoint variations (Figure 3, third row). To further demonstrate the geometric consistency of the generated results, we perform structure-from-motion (SfM) [49] reconstruction on videos generated by different methods. The original CogVideoX outputs fail for reconstruction due to a lack of sufficient multi-view cues. The fine-tuned model only achieves partial

Table 2: **Quantitative results on the DL3DV dataset** for text-to-3D scene generation. We compare our method against Director3D, LucidDreamer, and SplatFlow across multiple perceptual metrics. ↑: higher is better; ↓: lower is better.

| Method | FID ↓ | CLIPScore ↑ | NIQE ↓ | BRISQUE ↓ |
|---|---|---|---|---|
| LucidDreamer [9] | 79.96 | 31.25 | 11.23 | 44.52 |
| Director3D [30] | 90.20 | 30.04 | 13.79 | 51.67 |
| SplatFlow [11] | 86.77 | 31.42 | 14.15 | 48.85 |
| **Ours** | **72.78** | **33.84** | **8.53** | **36.41** |


Figure 5: **Comparison with original video generation method & ablation study.** The original video generation model, as well as naive joint modeling of depth, struggle to maintain geometric consistency of objects throughout their motion.

reconstruction with limited consistency among a few consecutive frames. In contrast, our method enables high-quality and structurally complete 3D scene reconstructions.

To further evaluate 3D consistency across frames, we introduce two metrics based on VGGT [58]: the **Multi-View Consistency Score (MVCS)** and the **Reprojection Error**. Given the predicted depth maps and camera poses from VGGT, we project each frame into a shared 3D space to compute cross-view consistency. MVCS measures the alignment of depth maps across views by warping each depth map to neighboring frames and comparing it against the corresponding predicted depth. In contrast, Reprojection Error evaluates the pixel-wise distance between original image coordinates and those reprojected from the reconstructed global point cloud, serving as a direct indicator of geometric alignment accuracy. Table 1 reports MVCS and Reprojection Error for different models on the DL3DV dataset. Our method significantly outperforms CogVideoX-5B and its finetuned variant. Notably, removing the geometric loss $\mathcal{L}_{\text{geo}}$ leads to a clear performance drop, demonstrating the effectiveness of our geometric regularization in preserving 3D structural consistency across views.

Furthermore, we compare our method with three representative text-to-3D scene generation baselines: **Director3D** [30], **LucidDreamer** [9], and **SplatFlow** [11]. The qualitative comparison of generation results can be found in Figure 4. We then follow Director3D and use the following metrics for quantitative evaluation. We use **CLIPScore** [15] to assess the alignment between the generated content and the text prompt. For perceptual evaluation, we adopt the **Fréchet Inception Distance (FID)**[16], a standard metric for assessing visual fidelity and diversity. In addition, we use two *no-reference image quality metrics*—**Natural Image Quality Evaluator (NIQE)**[37] and **Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)** [36] to directly assess perceptual quality based on image statistics, without relying on ground truth references. Table 2 summarizes the quantitative results in the DL3DV dataset. Our method consistently outperforms all baselines in all metrics. In particular, we achieve the lowest FID and the highest CLIPScore, demonstrating superior semantic consistency

Table 3: **Quantitative comparison with CogVideoX-5B on T2V and I2V tasks.** We evaluate on **CLIPScore**, **FVD** (Fréchet Video Distance), and metrics from **VBench**: **Subject Consistency (SC)**, **Background Consistency (BC)**, **Motion Smoothness (MS)**, **Spatial Relationship (SR)**, and **Video-Image Subject Consistency (VISC)**. ↑: higher is better; ↓: lower is better.

| Method | CLIPScore ↑ | FVD ↓ | SC ↑ | BC ↑ | MS ↑ | SR ↑ | VISC ↑ |
|---|---|---|---|---|---|---|---|
| CogVideoX-5B (T2V) | 32.30 | 145.3 | 93.8 | 95.1 | 93.2 | 79.4 | - |
| Ours w/o $\mathcal{L}_{\text{geo}}$ (T2V) | 33.25 | 134.2 | 94.3 | 96.0 | 95.4 | 87.4 | - |
| Ours w/o MP (T2V) | 33.83 | 131.6 | 95.8 | 96.3 | 96.7 | 88.2 | - |
| **Ours (T2V)** | **34.25** | **122.7** | **97.2** | **97.8** | **98.1** | **90.3** | - |
| CogVideoX-5B (I2V) | 33.42 | 139.8 | 94.6 | 96.4 | 95.9 | 80.5 | 95.2 |
| Ours w/o $\mathcal{L}_{\text{geo}}$ (I2V) | 34.13 | 128.0 | 95.0 | 97.1 | 96.3 | 86.3 | 96.3 |
| Ours w/o MP (I2V) | 34.77 | 126.5 | 96.9 | 97.6 | 96.9 | 88.8 | 96.8 |
| **Ours (I2V)** | **35.02** | **120.5** | **98.1** | **98.5** | **98.6** | **91.1** | **97.6** |

and perceptual quality. Moreover, our NIQE and BRISQUE scores are also lower, indicating a closer match to natural image distributions compared to prior methods. Compared to these 3D generation methods, our approach achieves higher visual fidelity by leveraging the strong priors from pretrained video generation models.

**Video generation results.** Table 3 presents a quantitative comparison between our method and CogVideoX-5B on both text-to-video and image-to-video generation using the 1000 evaluation videos. We report **CLIPScore** [15] to measure semantic alignment, **FVD** [54] to assess overall visual quality and temporal consistency, and multiple metrics from **VBench** [21], including **Subject Consistency (SC)**, **Background Consistency (BC)**, **Motion Smoothness (MS)**, **Spatial Relationship (SR)**, and **Video-Image Subject Consistency (VISC)**. Our method achieves superior performance across all metrics, clearly outperforming CogVideoX-5B in both settings. Notably, we observe substantial improvements in motion quality (MS), subject coherence (SC), and Spatial Relationship (SR).

**Ablation Studies.** Our method is based on two core components: (1) explicitly modeling depth and (2) applying supervision on the jointly generated depth. We conduct ablation studies targeting these two aspects. As shown in Figure 5, we present an image-to-video (I2V) result demonstrating that simply modeling both depth and appearance already improves the video quality to some extent. This is because the temporal consistency of the depth labels themselves imposes a form of constraint across frames. However, this alone is insufficient to ensure global geometric consistency throughout the video. When our proposed geometric loss $\mathcal{L}_{\text{geo}}$ is added, the generated videos exhibit significantly improved frame-to-frame continuity and structural coherence. The corresponding metric improvements are shown in Table 3 and Table 1. In addition, we also study the impact of incorporating the motion probability (MP) map in dynamic videos. Table 3 shows that ignoring the MP map leads to noticeable performance drops, particularly in the T2V setting.

## 5 Conclusions

In this work, we proposed to augment pre-trained video generation models with *geometric regularization* by introducing a per-frame depth prediction and enforcing cross-frame depth consistency. This approach leverages the natural spatio-temporal cues encoded in video to guide the model toward learning stable and physically grounded scene representations. Our method provides implicit geometric supervision through depth alignment, allowing for more accurate modeling of object permanence, spatial relationships, and scene dynamics. Through extensive experiments, we demonstrate that our framework significantly improves the geometric fidelity and temporal stability of generated videos, outperforming prior baselines in both qualitative and quantitative evaluations. We believe that this represents an important step toward bridging video generation and 3D world modeling, opening new possibilities for downstream applications in simulation, robotics, and embodied intelligence.

**Acknowledgements.** This work was supported by Damo Academy through Damo Academy Innovative Research Program.

## References


[1] Yunpeng Bai and Qixing Huang. Fiffdepth: Feed-forward transformation of diffusion-based generators for detailed depth estimation. *arXiv preprint arXiv:2412.00671*, 2024.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021.

[3] 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.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[5] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

[7] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025.

[8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[9] 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.

[10] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

[11] Hyojun Go, Byeongjun Park, Jiho Jang, Jin-Young Kim, Soonwoo Kwon, and Changick Kim. Splatflow: Multi-view rectified flow model for 3d gaussian splatting synthesis. *arXiv preprint arXiv:2411.16443*, 2024.

[12] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[14] Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. Pre-trained video generative models as world simulators. *arXiv preprint arXiv:2502.07825*, 2025.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[19] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

[20] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[22] Hyeonho Jeong, Chun-Hao Paul Huang, Jong Chul Ye, Niloy Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. *arXiv preprint arXiv:2412.06016*, 2024.

[23] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[25] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[26] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024.

[27] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[28] KlingAI. Kling AI, 2024.

[29] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[30] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *Advances in Neural Information Processing Systems*, 37:75125–75151, 2024.

[31] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024.

[32] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *arXiv preprint arXiv:2501.18590*, 2025.

[33] 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.

[34] 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.

[35] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7038–7048, 2024.

[36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[37] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[38] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9015–9025, 2024.

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[40] Pexels. Pexels — Free Stock Videos and Photos. `https://www.pexels.com/`.

[41] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[43] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

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

[45] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. *arXiv preprint arXiv:2501.09781*, 2025.

[46] 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.

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[48] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[49] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[50] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[51] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023.

[52] Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. *arXiv preprint arXiv:2503.14445*, 2025.

[53] Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

[54] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[55] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.

[56] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. This&that: Language-gesture controlled video generation for robot planning. *arXiv preprint arXiv:2407.05530*, 2024.

[57] Cong Wang, Jiaxi Gu, Panwen Hu, Yuanfan Guo, Xiao Dong, Hang Xu, and Xiaodan Liang. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[59] Shijie Wang, Samaneh Azadi, Rohit Girdhar, Saketh Rambhatla, Chen Sun, and Xi Yin. Motif: Making text count in image animation with motion focal loss. *arXiv preprint arXiv:2412.16153*, 2024.

[60] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

[61] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. 2023.

[62] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8414–8424, 2024.

[63] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[64] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

[65] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[66] Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. *arXiv preprint arXiv:2504.10825*, 2025.

[67] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024.

[68] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024.

[69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[70] 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.

[71] Yang Yuanbo, Shao Jiahao, Li Xinyang, Shen Yujun, Geiger Andreas, and Liao Yiyi. Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. *arxiv:2412.21117*, 2024.

[72] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *European Conference on Computer Vision (ECCV)*, 2024.

[73] Yuanhao Zhai, Kevin Lin, Linjie Li, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, David Doermann, Junsong Yuan, Zicheng Liu, and Lijuan Wang. Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation. In *European Conference on Computer Vision*, pages 134–152. Springer, 2024.

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

[75] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. *arXiv preprint arXiv:2412.01821*, 2024.

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

[77] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly explain how our method introduces geometric regularization into video generation models.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss this issue in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our method does not make strong theoretical contributions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed implementation information in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: At this stage, we do not release the code but provide generated video results from our model.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, these details are thoroughly discussed in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, please refer to the tables in the main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, these details are thoroughly discussed in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have complied with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we addressed this in the introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not identify such a risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we have added explanations in the relevant sections.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our method does not rely on large language models (LLMs) at its core.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.