How Far are AI-generated Videos from Simulating the 3D Visual World: A Learned 3D Evaluation Approach

Chirui Chang[1] Jiahui Liu[1] Zhengzhe Liu[3] Xiaoyang Lyu[1] Yi-Hua Huang[1]
Xin Tao[2] Pengfei Wan[2] Di Zhang[2] Xiaojuan Qi[1*]
[1]The University of Hong Kong [2]Kling Team, Kuaishou Technology [3]Lingnan University

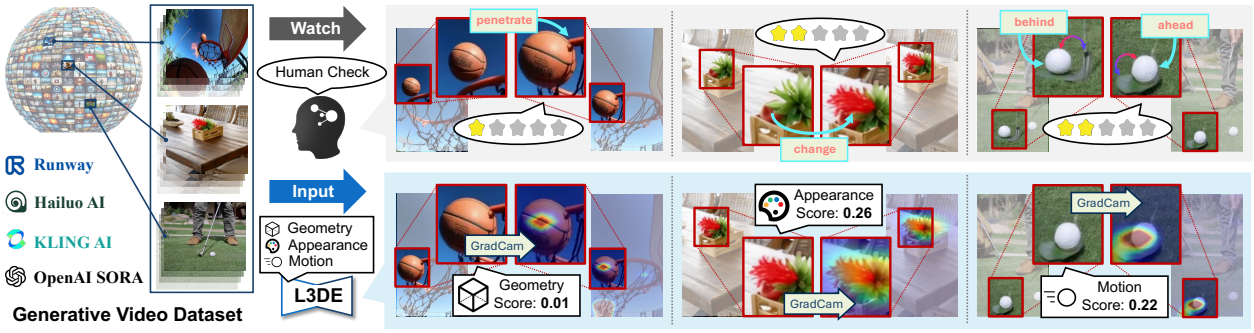

Figure 1. L3DE evaluates videos from any generative model based on 3D visual coherence, assessing appearance, motion, and geometry. Its scores align closely with human perception and can localize regions of 3D simulation failures, similar to human intuition. Examples highlight key failure cases: (1) incorrect occlusion between the basketball and hoop, disrupting geometric consistency, (2) abrupt texture transition in plant leaves, and (3) unnatural relative motion between the golf ball and the golf club, violating real-world motion dynamics.

Abstract

Recent advancements in video diffusion models enable the generation of photorealistic videos with impressive 3D consistency and temporal coherence. However, the extent to which these AI-generated videos simulate the 3D visual world remains underexplored. In this paper, we introduce Learned 3D Evaluation (L3DE), an objective, quantifiable, and interpretable method for assessing AI-generated videos' ability to simulate the real world in terms of 3D visual qualities and consistencies, without requiring manually labeled defects or quality annotations. Instead of relying on 3D reconstruction, which is prone to failure with in-the-wild videos, L3DE employs a 3D convolutional network, trained on monocular 3D cues of motion, depth, and appearance, to distinguish real from synthetic videos. Confidence scores from L3DE quantify the gap between real and synthetic videos in terms of 3D visual coherence, while a gradient-based visualization pinpoints unrealistic regions, improving interpretability. We validate L3DE through extensive experiments, demonstrating strong alignment with 3D reconstruction quality and human judgments. Our evaluations on leading generative models (e.g., Kling, Sora, and MiniMax) reveal persistent simulation gaps and subtle inconsistencies. Beyond generative video assessment, L3DE extends to broader applications: benchmarking video generation models, serving as a deepfake detector, and enhancing video synthesis by inpainting flagged inconsistencies.

1. Introduction

Video diffusion models, such as Sora [5], have recently shown remarkable capabilities in visual simulation, producing photorealistic videos with 3D consistency and temporal coherence that can even deceive human observers. This progress raises a fundamental question: how well do AI-generated videos simulate the 3D visual world? While existing evaluations heavily rely on subjective user studies, a quantifiable and interpretable approach remains missing for assessing 3D visual coherence of generative videos.

3D scene reconstruction [17, 20, 31, 56, 58] is a natural

*Corresponding author.

way to assess whether generative videos preserve 3D visual coherence. The intuition is that if a video enables high-quality 3D reconstruction, it should maintain 3D-consistent appearance, structure, and motion across frames. However, even state-of-the-art reconstruction methods struggle with in-the-wild videos due to challenges such as unreliable pose estimation [9, 55, 58] and the absence of multi-view cues [9, 24, 53, 63], making large-scale evaluation based on reconstruction impractical. To overcome these limitations, inspired by [10], we turn to monocular 3D cues, such as depth and optical flow, which naturally emerge from videos and serve as strong proxies for 3D structure and motion. Thus, we explore leveraging monocular cues from foundation models [37, 51] as an alternative for assessing 3D realism. Specifically, we use RAFT [51] for optical flow estimation and UniDepth [37] for depth prediction, while utilizing DINOv2 [34] to capture high-level appearance features.

We collect real and synthetic videos from Pexels [36] and Stable Video Diffusion (SVD) [3], respectively. Pexels provides diverse real-world videos, while SVD is one of the most accessible video generator. We align their visual content by using real video frames as prompts to generate paired synthetic videos. This minimizes content disparities, isolating differences in 3D consistency and helping analyze how generative videos deviate from real ones.

Equipped with 3D proxies and data, the next challenge is measuring the gap between generative and real-world videos. To tackle this, we develop **Learned 3D Evaluation (L3DE)**, a data-driven learning-based tool that uses monocular 3D cues to evaluate generative videos and identify 3D visual simulation failures. L3DE captures intrinsic differences between real and synthetic videos by training a 3D convolutional network with contrastive learning using 3D proxies as inputs. The confidence scores quantify the gap between synthetic and real videos regarding these 3D proxies. Additionally, L3DE enhances interpretability by highlighting key failure regions via a gradient-based method [48] (see Table 3). Finally, by integrating depth, motion, and appearance proxies through a feature fusion module, L3DE provides a stable and comprehensive evaluation of 3D visual coherence in generative videos.

To validate L3DE's effectiveness, we conduct 3D scene reconstruction experiments and user studies. Our results in Sec. 5.1 show that L3DE scores highly correlate with reconstruction quality, with flagged areas aligning with regions of high 3D inconsistency, as confirmed by reconstruction errors. Human studies in Sec. 5.2 further reveal that L3DE scores align closely with human perceptual judgments, with flagged areas consistently rated high by annotators. These results demonstrate L3DE's effectiveness in assessing and analyzing 3D visual coherence in generative videos.

We conduct experiments applying L3DE to videos from leading generative models, including Sora [5], MiniMax [32], Kling [23], and others to benchmark their 3D visual simulation capabilities and analyze their strengths and limitations. With L3DE validated through 3D reconstruction and human evaluation, these results provide insights into how well different models capture 3D realism. As shown in Table 5, models like Sora and Kling achieve higher L3DE scores, particularly in appearance simulation, while all models show room for improvement in motion and geometry consistency. Most generative videos still exhibit noticeable gaps from real ones in 3D visual coherence, as reflected in their lower L3DE scores.

Beyond evaluating 3D visual coherence in AI-generated videos, L3DE can serve as a deepfake detector by applying a confidence score threshold. Despite not being trained on videos from specific sources, L3DE effectively identifies fake videos from Kling and others (see Table 1 in the appendix) with over 0.7 accuracy. Additionally, L3DE's localized failure regions can help improve video synthesis. By inpainting flagged areas (see appendix), we can enhance the 3D visual coherence of generative videos.

Our contributions can be summarized as follows:

- We take the first step in systematically investigating the 3D visual coherence of AI-generated videos across appearance, motion, and geometry—key factors in representing a dynamic 3D world. To facilitate quantitative analysis, we extract monocular clues from foundation models to disentangle these aspects.
- We introduce Learned 3D Evaluation (**L3DE**) that quantifies the 3D visual coherence of a video using confidence scores from models trained on pairing data with contrastive loss. L3DE also highlights spatial and temporal regions as evidence for its assessment. Moreover, we integrate these three aspects to deliver a more robust assessment tool.
- Through controlled user studies and 3D reconstruction experiments on diverse generative videos, we show that L3DE's quantification scores and localized regions align well with user intent and reconstruction quality.
- L3DE can be used for broader applications. Our experiments and studies provide valuable insights and findings about the capabilities of current video generation models.

2. Related Work

Diffusion models for video generation. The success of diffusion models [14, 50] in image synthesis [6, 12, 27, 33, 38, 41, 43] has driven advancements in video generation [4, 11, 13, 15, 16, 21, 29, 57, 62, 64]. Stable Video Diffusion [3] leverages large-scale training for high-quality video synthesis. Sora [5] demonstrate the ability to simulate humans, animals, and environments, highlighting video generation as a potential path towards world simulation. Our work aim to help the community gain more understand-

ing about generative videos, especially their gap from real-world videos in terms of 3D viusal simulation capabilities.

AI-generated video evaluation. Existing metrics for evaluating AI-generated videos include Inception Score (IS) [44], Fréchet Video Distance (FVD) [52], Perceptual Input Conformity(PIC) [57] and CLIPSIM [40], among others. Recent benchmarks, such as VBench [18] and EvalCrafter [26], establish standardized protocols by integrating automated metrics for comprehensive model comparisons. In contrast, our approach identifies differences between real and generative videos using a data-driven yet simple method, complemented by low-level statistical analysis to assess their 3D visual simulation capabilities.

Video feature extraction. Extracting appearance, motion, and geometry information is crucial for evaluating video realism. DINOv2 [34] shows strong image appearance representation, while optical flow estimation methods [7, 19, 51] provide robust motion features. Monocular depth cues encode rich geometric information, with recent methods like UniDepth [37] achieving precise metric depth estimation with excellent video consistency. We leverage these techniques to extract relevant features for our analysis.

3D scene reconstruction. Recent advancements in 3D reconstruction, such as NeRF-based [2, 25, 30, 31, 35, 39, 54, 60] and 3D-GS-based [17, 20, 56, 58] methods, have improved static and dynamic scene modeling. Despite the robustness of novel view synthesis (NVS) methods for in-the-wild scenes, unreliable camera pose estimation in such videos limits the feasibility of 3D scene reconstruction as a robust large-scale evaluation tool for assessing the 3D visual simulation capabilities of AI-generated videos.

3. Data Curation

To gain a deeper understanding of the 3D visual simulation capabilities of AI-generated videos, we design a data curation process and compile a dataset that includes both real-world and AI-generated videos, as detailed in Table 1. Our model training, method validation, and subsequent analysis are all conducted using different subsets in this dataset.

In-the-wild real-world videos. We begin by collecting approximately 100,000 real-world, in-the-wild videos from Pexels [36]. These videos encompass a wide range of content, including animals, people, natural scenes, urban landscapes, indoor environments, and more. For raw video processing, we follow the method introduced in [3]. More details on the data processing can be found in the appendix.

Paired generative videos. We employ the open-source generative model Stable Video Diffusion (SVD) [3] to generate synthetic videos. To ensure the focus is on the 3D visual coherence, rather than potential biases in the generated content or color distribution, we condition SVD model using the first frames from real video clips. This enables SVD to generate paired synthetic samples that preserve the same semantic content and color distribution as their real video counterparts. Thus we create a paired generative video dataset, where the video clips share similar visual content to the real videos, minimizing the risk of model bias.

3D reconstruction verification set. To evaluate L3DE's effectiveness, we curate a verification set using videos generated by the commercial model Kling [23], as SVD-generated videos are typically of low quality, hindering 3D reconstruction and rendering. Our verification set consists of two parts: (1) *Generated Videos for In-the-wild Scenes.* Given the low success rate of pose estimation [9, 55, 58] on AI-generated videos, we generate diverse samples conditioned on keyframes from unseen real videos. We then screen the large pool of generated videos and retain 30 that successfully undergo 3D reconstruction. (2) *Twin Videos for Public Scene Datasets.* To analyze the correlation between 3D consistency and L3DE score, we iteratively generate twin videos for 15 scenes from public static datasets (i.e., Mip-NeRF360 [2], Tanks-and-Temples [22]) and dynamic datasets (i.e., Hyper-NeRF [35], Neural 3D Video Synthesis Dataset [25]), ensuring that each scene yields at least one video that successfully undergoes COLMAP and reconstruction. Each twin video pair is generated using one real frame as the start frame and another with sufficient overlap as the end frame, maintaining close alignment with the real 3D content. Videos from (1) and (2) form the 3D reconstruction verification set, totaling 3000 videos. For validation experiments, we use only videos that successfully undergo pose estimation, while the entire set is used in supplementary fake video detection experiments.

3D visual simulation benchmark. We conduct studies using L3DE on generated videos from recent commercial generative models, augmented with data from [61], to assess their ability to simulate the 3D visual world. The dataset includes videos from models such as Sora [5], Kling [23], Runway-Gen3 [42], Luma [28], MiniMax [32], Vidu [49], and CogVideoX [59]. To ensure relevance, we exclude videos with non-realistic content, such as animations. Since all videos are generated with the same set of image or text prompts, this dataset enables a direct and fair comparison of 3D visual simulation capabilities across different models by eliminating prompt-induced variability. Furthermore, we provide 14,000 unseen real video samples as references to establish an empirical upper bound for L3DE scores.

4. Learned 3D Evaluation

Below, we first discuss proxies for representing the 3D visual world, followed by a detailed explanation of the newly proposed Learned 3D Evaluation (L3DE) for assessing the 3D visual simulation capabilities of AI-generated videos.

Source	Synthetic/Real	Number of Videos	Clip Length	Resolution	Frame Rate	Prompt Type
Paired Real/Synthetic Video Set						
Pexels [36]	Real	80,000	4s	Variable	Variable	–
Stable Video Diffusion [3]	Synthetic	80,000	4s	1024*576	7 FPS	I2V
3D Reconstruction Verification Set						
Kling 1.5 [23]	Synthetic	3,000	5s	Variable	30 FPS	I2V & T2V
3D Visual Simulation Benchmark						
Pexels [36]	Real	14,000	4s	Variable	Variable	–
Runway-Gen3 [42]	Synthetic	539	5s	1280*768	24 FPS	I2V & T2V
MiniMax [32]	Synthetic	539	5s	1280*720	25 FPS	I2V & T2V
Vidu [49]	Synthetic	539	3s	Variable	24 FPS	I2V & T2V
Luma Dream Machine 1.6 [28]	Synthetic	539	Variable	Variable	24 FPS	I2V & T2V
Kling 1.5 [23]	Synthetic	539	5s	Variable	30 FPS	I2V & T2V
CogVideoX-5B [59]	Synthetic	539	6s	720*480	8 FPS	I2V & T2V
Sora [5]	Synthetic	539	5s	Variable	30 FPS	I2V & T2V
Kling 2.1 [23]	Synthetic	539	5s	Variable	30 FPS	I2V & T2V

Table 1. Overview of our dataset, which consists of (1) Paired Real/Synthetic Video Set, designed to study the gap between real-world and AI-generated videos; (2) the 3D Reconstruction Verification Set, curated for validating L3DE through 3D reconstruction; and (3) the 3D Visual Simulation Benchmark, which includes videos from multiple generative models to evaluate their 3D visual simulation capabilities.

4.1. Proxies for Representing 3D Visual World

Reconstructing and rendering in-the-wild videos to assess 3D world simulation is challenging, primarily due to issues such as unreliable camera pose estimation [46, 47, 58]. Beyond reconstructing a scene in 3D space, the realism of the 3D visual world is shaped by multiple perceptual factors. Inspired by [10, 45], we identify three key aspects: **1) Appearance:** Visual attributes of video frames, including color, texture, and lighting; **2) Motion:** Temporal dynamics and changes within the video; and **3) Geometry:** The spatial structure and shape of objects in the frames. These cues reflect the consistency of a video's 3D structure and can be reliably estimated from videos using foundation models, which we leverage as proxies for the 3D visual world. We extract these cues using the following foundation models:

- **Appearance representation**: Instead of simply using the original RGB information, we extract per-frame visual feature with DINOv2 [34] as the appearance representation. Its features are capable of cross-image dense and sparse matching [8, 34], which enhances the potential to capture cross-frame appearance consistency.
- **Motion representation**: We leverage optical flow, which is well-studied to represent motion, to examine the motion pattern differences between synthetic and real videos. To be more specific, we employ RAFT [51], a state-of-the-art optical flow estimation model, to extract optical flow between the adjacent frames.
- **Geometry representation**: To investigate the geometric properties of generative videos, we leverage the per-frame depth as the geometry representation. Depth conveys many 2.5D geometric cues, such as occlusion, spatial relationships, scales, and so on. In detail, considering the cross-frame scale consistency, we adopt metric depth from UniDepth [37] as it has a uniform scale and provides better consistency across frames, which aids in perceiving changes in the geometric structure of the video.

4.2. Design of L3DE

With the prepared data and extracted 3D visual proxies, we develop L3DE. The model first trains a classifier on the paired real/synthetic video dataset in Table 1, enabling it to learn to distinguish them based on the three proxies. This is achieved with a contrastive learning objective, which enhances the discriminative power of the learned features. Additionally, we integrate Grad-CAM [48] to enable L3DE to identify simulation traits. Finally, we design a fusion module that combines all three 3D proxies to produce a more comprehensive evaluation score for video assessment.

Classifier construction and training. Based on the 3D proxies outlined in Sec. 4.1, we design a 3D convolutional network with multiple layers interleaved with ReLU activation functions. It predicts the confidence score evaluating whether a sample belongs to real or synthetic videos. Further details are provided in the appendix. The penultimate layer features are used to construct the contrastive loss. For any input generative video feature $\mathbf{f}_{\text{gen}}$, the loss encourages pushing apart its closest real video feature, thereby making

real video feature more distinguishable. It is computed as:

$$\mathcal{L}_{\text{contrastive}} = \sum_i \exp\left(-\left\|\mathbf{f}_{\text{gen}}^{(i)} - \mathbf{f}_{\text{real}}^{(j(i))}\right\|_2^2\right), \quad (1)$$

where $\mathbf{f}_{\text{real}}^{(j(i))}$ is the closest real video feature to $\mathbf{f}_{\text{gen}}^{(i)}$ in Euclidean distance. The total loss function combines the classification loss $\mathcal{L}_{\text{cls}}$ and contrastive loss $\mathcal{L}_{\text{contrastive}}$ as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{contrastive}}. \quad (2)$$

As the network learns to distinguish between real and synthetic videos, its confidence scores serve as a quantitative metric for assessing how closely an input video resembles real-world videos in 3D visual coherence. To interpret its predictions and understand the underlying evidence, we apply Grad-CAM [48], which generates a class-discriminative localization map by backpropagating gradients to the last convolutional layer. This map highlights the video regions that mostly influence the model's decision (see Fig. 4).

Feature fusion for comprehensive scores. Since video content inherently combines appearance, motion, and geometry, we design a feature fusion module for a more robust and comprehensive evaluation. Within the network, we concatenate features from these three aspects:

$$\mathbf{f}_{\text{fused}} = \text{Concat}\left(\mathbf{f}_{\text{app}},\ \mathbf{f}_{\text{mot}},\ \mathbf{f}_{\text{geo}}\right), \quad (3)$$

where $\mathbf{f}_{\text{app}}$, $\mathbf{f}_{\text{mot}}$, and $\mathbf{f}_{\text{geo}}$ represent the features for appearance, motion, and geometry. The fused representation in our Fusion variant of L3DE produces an overall score, jointly accounting for all three aspects. This holistic evaluation provides a more comprehensive measure of 3D visual coherence, complementing single-aspect assessments.

5. Validation of L3DE

To validate L3DE's reliability in evaluating 3D visual coherence, we employ two complementary strategies: 3D reconstruction and human perceptual judgment. 3D reconstruction objectively assesses how well AI-generated videos preserve spatial structure and motion realism. However, pose estimation often fails on in-the-wild videos, meaning that only a subset of videos– those where camera parameters can be reliably estimated– can be reconstructed for validation. Within this subset, we use reconstruction to precisely verify L3DE's predicted scores and detected regions. Beyond this subset, human perception provides a more flexible and perceptually grounded evaluation of 3D visual coherence, as it is not constrained by camera estimation failures. This allows us to confirm that L3DE remains effective across a wider range of generative videos.

5.1. Validation using 3D Reconstruction

We conduct 3D reconstruction experiments in two controlled settings to assess the correlation between L3DE scores and 3D rendering quality. Additionally, we examine whether L3DE's detected inconsistencies align with reconstruction errors by comparing its localized regions to rendering-based discrepancy maps. These experiments utilize the 3D reconstruction verification dataset (Table 1).

Correlation with L3DE	Fusion	Appearance	Motion	Geometry
3D Reconstruction Quality	0.7566	0.7181	0.6669	0.3142
Human Ratings	0.6460	0.5643	0.4617	0.3479

Table 2. Spearman correlation between L3DE scores and different reference evaluations. The first row shows correlation with 3D reconstruction quality, while the second shows correlation with human ratings on the same verification dataset.

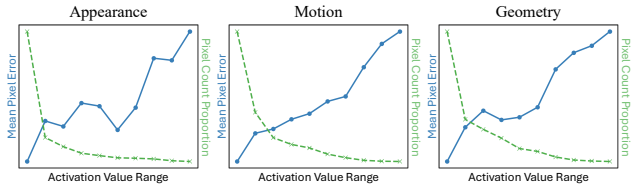

Figure 2. Illustration of the statistics of activation value, pixel value error and the distribution of pixel number for each proxy.

L3DE score v.s. 3D rendering quality. We evaluate the correlation between L3DE scores and 3D reconstruction quality by optimizing a 3D representation, such as 3D-GS [20], across all video frames to reconstruct each scene. To ensure a more adaptive evaluation, we use the 'Twin Videos for Public Scene Datasets' from the 3D reconstruction verification set, which provides real and synthetic videos of the same content for fair comparisons. For static scenes, we assess L3DE's appearance and geometry scores by measuring visual fidelity and spatial accuracy. For dynamic scenes, we focus on validating the motion score by analyzing temporal coherence and movement realism. Specifically, we use 3D-GS [20] for static scenes and SC-GS [17] for dynamic scenes. Rendering quality is quantified using Peak Signal-to-Noise Ratio (PSNR). To compensate for content-dependent variations in PSNR, we normalize the rendering quality of synthetic videos $Q_{\text{synthetic}}$ relative to that of real videos Q_{real}. This normalization mitigates scene-specific biases, leading to a more robust assessment. The normalized quality difference is defined as :

$$\Delta Q = \max\left(Q_{\text{real}} - Q_{\text{synthetic}}, 0\right). \quad (4)$$

To quantify the disparity between real and synthetic videos, we define the simulation gap G based on L3DE scores S:

$$G = 1 - S. \quad (5)$$

We then evaluate L3DE's ability to capture 3D rendering quality by computing the correlation between ΔQ and G. As shown in Table 2, L3DE scores are *positively correlated*

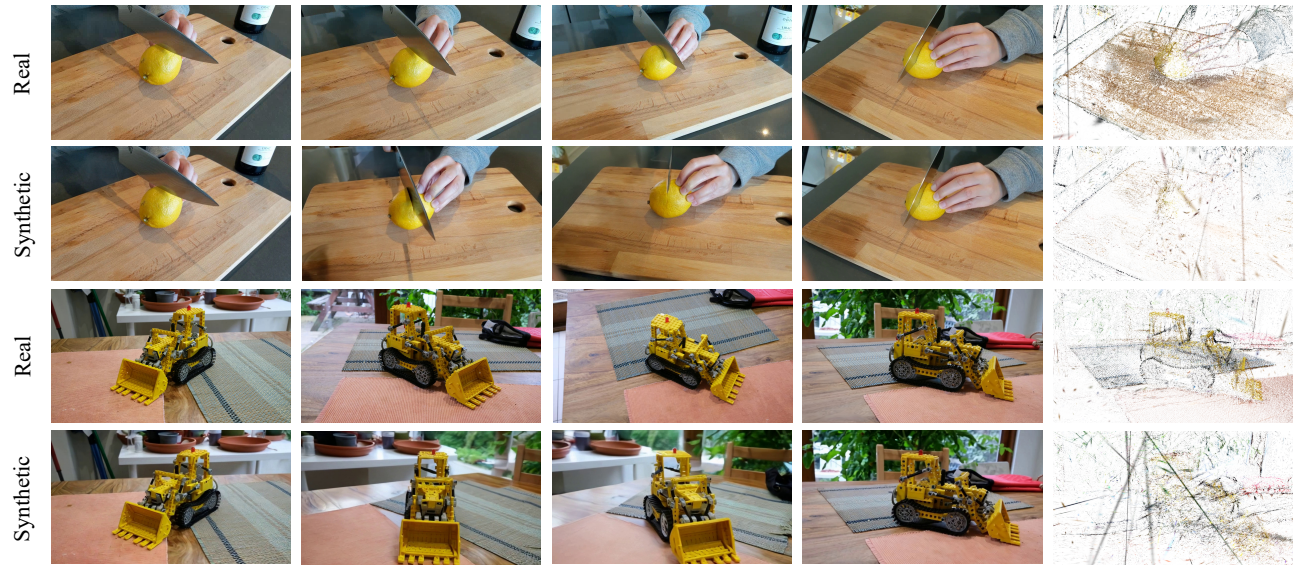

Figure 3. Frames and reconstruction results of twin videos. Even though synthetic videos appear plausible, they do not achieve the same level of 3D scene reconstruction accuracy as real videos (see the Shrunken Gaussians in the rightmost column). This discrepancy underscores a key limitation: current generative videos are not yet adept at faithfully simulating the world in terms of 3D visual coherence.

Figure 4. Illustration of 3D inconsistencies identified by L3DE. From left to right: (a) AI-generated video frame; (b) rendered frame with 3D reconstruction with pose aligned with the original view; (c) pixel-level difference between (a) and (b); (d) Grad-CAM result from the L3DE network, which closely aligns with (c); (e) Blue solid line: large (normalized) activation value in (d) is highly aligned with large mean pixel value error in (c). Green dashed line: areas with high (normalized) activation values cover only a small portion of the entire frame. L3DE identifies key artifacts in the cases: (1) unnatural hand motion in the first case, reflected in a low motion score of 0.4642; (2) abrupt geometric deformation of the marked object in the second case, with a geometry score of 0.637; and (3) sudden texture changes in the chair and table in the third case, resulting in an appearance score of 0.2578.

with 3D rendering quality, indicating that higher L3DE scores correspond to the superior rendering fidelity. Notably, our L3DE fusion model achieves the highest correlation of 0.7566, demonstrating strong alignment with the reconstruction-based evaluation.

L3DE localized region vs. inconsistent region. We assess L3DE's ability to localize 3D-inconsistent regions in AI-generated videos using the 'Generated Videos for In-the-wild Scenes' from the 3D reconstruction verification dataset. Grad-CAM [48] highlights the regions L3DE focuses on for real-fake classification. To establish reference 3D-inconsistent regions, we split the dataset into training and test sets, ensuring discrepancies are measured only

from test viewpoints to mitigate overfitting effects in GS-based reconstruction. We then quantify the alignment between L3DE-detected regions and rendering-based discrepancy maps. Fig. 2 presents the quantitative correlation results, demonstrating strong alignment between L3DE-detected and rendering-inconsistent regions. Qualitative comparisons are shown in Fig. 4.

5.2. Validation using Human Judgment

To complement reconstruction-based validation, we conduct human evaluations to assess whether L3DE scores and detected regions align with human perception judgments. This ensures that L3DE not only correlates with objective reconstruction quality but also reflects subjective judgments of 3D visual coherence.

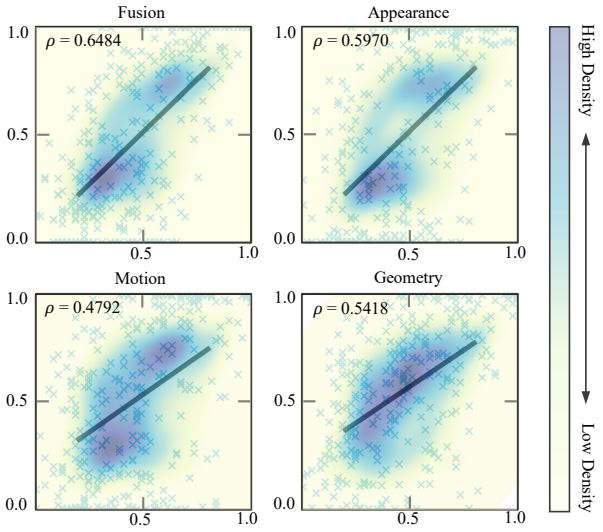

Figure 5. The correlation between L3DE scores and human ratings. The X-axis represents the average human ratings and the Y-axis represents the L3DE scores.

L3DE scores v.s. human ratings. First, We validate the correlation between L3DE scores and human evaluations through a user study involving 15 participants who provided 4,500 annotations on 300 randomly selected AI-generated videos, rating their realism in terms of 3D visual coherence. More details on the study setup are provided in the appendix. For each video, we compute the average participant rating as the human rating. We then evaluate L3DE scores across appearance, geometry, motion, and their fusion. We then compute the Spearman correlation between these scores and the human ratings. As shown in Fig. 5, L3DE scores exhibit a strong positive correlation with human evaluations, confirming their reliability in assessing generative videos. Notably, the fusion score achieves the highest correlation, underscoring the effectiveness of our fusion strategy. Additionally, we analyze human ratings for the videos used in the rendering quality experiments and compute their correlation with L3DE scores. It shows that L3DE consistently aligns with both reconstruction quality and human judgment on the same dataset (see Table 2).

	Appearance	Motion	Geometry
Average score	0.8600	0.7200	0.7400
Spearman's ρ	0.4894	0.4026	0.4317

Table 3. Average human plausibility scores on the Grad-CAM visualization and correlation between L3DE localized region and human-annotated region from different aspects.

L3DE localized region v.s. human plausibility. We further validate L3DE's localized regions through an additional user study. 10 volunteers are shown highlighted regions from both L3DE and randomly generated maps, without disclosure of their source to prevent bias. Each participant rates the plausibility of the highlighted regions on a 1–5 scale, with scores subsequently normalized. As shown in Table 3, L3DE achieves a significantly higher score (0.7–0.8) compared to random maps (average **0.21**, minimum 0.2). To reinforce our validation, we conduct a second experiment where 10 participants annotate unrealistic regions in 30 unseen videos. The correlation between these annotations and L3DE-detected regions (Table 3) further confirms that L3DE effectively aligns with human perception of unrealistic content.

6. Analysis and Applications of L3DE

6.1. Comparison with Existing Metrics

While existing methods such as VBench [18] and EvalCrafter [26] provide general-purpose video evaluation, they do not specifically assess 3D visual coherence. To compare them with L3DE, we select relevant metrics from each benchmark that focus on spatial and temporal consistency. We evaluate them based on their correlation with human judgments, following the standard approach for validating evaluation methods [18, 26]. As shown in Table 4, L3DE achieves a stronger correlation with human ratings than existing metrics, demonstrating its effectiveness in assessing 3D realism in generative videos. Beyond correlation analysis, L3DE also introduces unique capabilities, such as identifying unrealistic areas—an aspect missing from existing metrics—which enhances interpretability and provides actionable insights for improving generative models.

6.2. Benchmarking Video Generation Models

Given that L3DE effectively evaluates the 3D visual coherence of generative videos, we expand video generation model benchmarking by introducing 3D visual simulation capabilities as a new assessment dimension, which has been

Metric	Method	Spearman's ρ
Subject Consistency	VBench [18]	3.90
Background Consistency	VBench [18]	20.68
Motion Smoothness	VBench [18]	19.99
Temporal Consistency	EvalCrafter [26]	13.85
L3DE Fusion Score	L3DE	**64.84**

Table 4. Correlation of L3DE scores and automatic metrics from different baselines with human ratings.

Generators	Fusion	Appearance	Motion	Geometry
Runway-Gen3 [42]	0.7162	0.6946	0.5768	0.6739
MiniMax [32]	0.7932	0.7714	0.6098	0.7251
Vidu [49]	0.7052	0.6406	0.6228	0.7615
Luma 1.6 [28]	0.5062	0.4950	0.5853	0.6800
Kling 1.5 [23]	0.7518	0.7247	0.5926	0.6927
CogVideoX-5B [59]	0.6104	0.5893	0.6203	0.7539
Sora [5]	0.8895	**0.8394**	0.6467	0.7458
Kling 2.1 [23]	**0.8904**	0.8129	**0.6735**	**0.7623**
Real Videos	0.9999	0.9950	0.8321	0.8435

Table 5. Benchmarking results of generative models. The Fusion column, highlighted as the primary L3DE ranking, represents the overall 3D visual coherence. Real videos achieve near-perfect scores, serving as an empirical upper bound for L3DE.

largely overlooked in existing benchmarks. Using the data outlined in Sec. 3, we evaluate leading generative models based on their ability to simulate the 3D visual world and present our findings below.

Quantitative Studies. To benchmark generative models, we compute the average L3DE score across all generated videos for each model. The fusion score represents the model's overall evaluation, while individual scores for appearance, motion, and geometry are also reported. The evaluation results are shown in Table 5 and the model rankings strongly correlate with large-scale human-preference benchmarks [1] (see appendix), confirming the robustness and generalizability of L3DE. Based on the overall fusion score, Kling 2.1 [23] and Sora [5] produces the highest-quality videos in terms of 3D visual simulation assessment. While these models excel in appearance simulation, their motion and geometry scores remain significantly lower, with minimal variation among models. As a reference, we calculate L3DE scores for a large set of 14,000 real video clips and they achieve an average fusion score of 0.9999, reaffirming the reliability of L3DE. Kling's and Sora's fusion and appearance scores exceed 0.8, but their motion and geometry scores are notably lower, indicating potential areas for improvement. These findings indicate that:

- While some videos generated by leading models achieve high L3DE scores, most still exhibit significant gaps in 3D visual coherence compared to real videos.
- The primary distinction among video generation models lies in their ability to simulate appearance, whereas their motion and geometry performance remains notably lower, lacking the fidelity of real-world videos.

Qualitative Studies. We analyze the Grad-CAM results from the fusion version of L3DE and observe that, while it provides less direct interpretability compared to individual aspects, it effectively captures more complex artifacts. For instance, Fusion Grad-CAM effectively identifies physically implausible interactions, such as issues with liquid, glass, and human scaling. For more qualitative studies, please refer to the supplementary. These findings indicate that integrating multiple cues in L3DE enhances its capability to detect higher-level inconsistencies beyond individual appearance, motion, or geometry assessments.

6.3. Applications

We further demonstrate several downstream applications of L3DE, including fake video detection by applying a threshold on the prediction score and enhancing generative video quality by inpainting regions identified by L3DE. More details on these applications can be found in the appendix.

7. Conclusion and Discussion

We present Learned 3D Evaluation (L3DE), a robust and interpretable framework for assessing the 3D visual coherence of generative videos. By leveraging monocular 3D cues—motion, depth, and appearance—from foundation models, L3DE provides an objective and quantifiable measure of discrepancies between real and synthetic videos. Extensive experiments demonstrate L3DE's effectiveness in evaluating videos from generative models, revealing significant 3D simulation gaps and subtle inconsistencies that are often overlooked by human observers. L3DE aligns well with reconstruction quality and human judgment, validating its role as an analytical tool and deepfake detector. Beyond evaluation, L3DE's insights can inform video synthesis improvements, offering a promising avenue for enhancing the realism of AI-generated content. Overall, L3DE presents a powerful tool for advancing our understanding of AI's capabilities in simulating the 3D visual world, with broad applications in video generation and evaluation.

Acknowledgments: This work has been supported by Kuaishou Technology, Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 17202422), RGC Matching Fund Scheme (RMGS), Lingnan University Start-Up Grant fund code: SUG-001/2526, and Faculty Research Grant fund code:106106. Part of the research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Artificial Analysis. Video Arena Leaderboard. https://artificialanalysis.ai/text-to-video/arena?tab=Leaderboard. 8

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 3

[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. 2, 3, 4

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[5] 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. 1, 2, 3, 4, 8

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 3

[8] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 4

[9] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2, 2024. 2, 3

[10] James A Ferwerda. Three varieties of realism in computer graphics. In *Human vision and electronic imaging viii*, pages 290–297. SPIE, 2003. 2, 4

[11] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[12] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[17] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4220–4230, 2024. 1, 3, 5

[18] 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. *arXiv preprint arXiv:2311.17982*, 2023. 3, 7, 8

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[20] 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. 1, 3, 5

[21] 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. 2

[22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 3

[23] Kuaishou. Kling ai. *https://klingai.kuaishou.com/*, 2024.06. 2, 3, 4, 8

[24] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[25] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5521–5531, 2022. 3

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 3, 7, 8

[27] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or, and Chi-Wing Fu. Object-level scene deocclusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[28] LumaLabs. Dream machine. *https://lumalabs.ai/dream-machine*, 2024.06. 3, 4, 8

[29] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 2

[30] Xiaoyang Lyu, Chirui Chang, Peng Dai, Yang-Tian Sun, and Xiaojuan Qi. Total-decom: Decomposed 3d scene reconstruction with minimal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20860–20869, 2024. 3

[31] 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. 1, 3

[32] MiniMax. Hailuo ai. *https://hailuoai.com/video*, 2024.09. 2, 3, 4, 8

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4

[35] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[36] Pexels. https://www.pexels.com/, 2023. 2, 3, 4

[37] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. *arXiv preprint arXiv:2403.18913*, 2024. 2, 3, 4

[38] 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. 2

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[41] 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. 2

[42] Runway. Gen-3. *https://runwayml.com/*, 2024.06. 3, 4, 8

[43] 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. 2

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[45] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28140–28149, 2024. 4

[46] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[47] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[48] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 2, 4, 5, 6

[49] ShengShu-AI. Vidu. *https://www.vidu.studio/*, 2024.07. 3, 4, 8

[50] 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. 2

[51] 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. 2, 3, 4

[52] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 3

[53] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 2

[54] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 3

[56] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 1, 3

[57] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2, 3

[58] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 1, 2, 3, 4

[59] 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. 3, 4, 8

[60] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 3

[61] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227*, 2024. 3

[62] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2

[63] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 2

[64] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2