Animus3D: Text-driven 3D Animation via Motion Score Distillation

QI SUN, City University of Hong Kong, Hong Kong
CAN WANG, City University of Hong Kong, Hong Kong
JIAXIANG SHANG, Central Media Technology Institute, Huawei, China
WENSEN FENG, Central Media Technology Institute, Huawei, China
JING LIAO*, City University of Hong Kong, Hong Kong

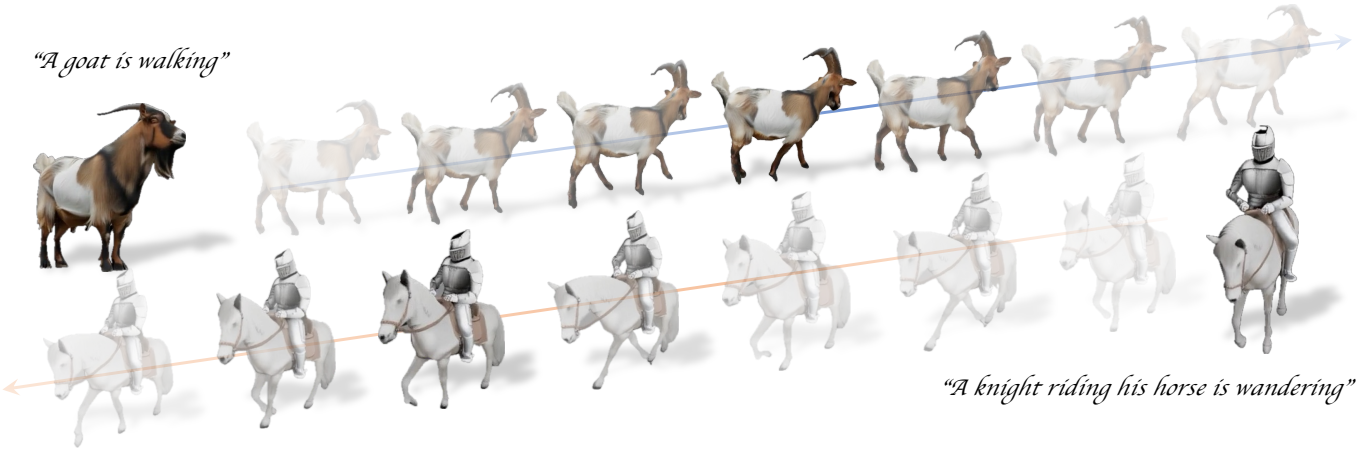

Fig. 1. **Animus3D** transforms a static 3D object to an animated object sequence given text descriptions via motion score distillation.

We present *Animus3D*, a text-driven 3D animation framework that generates motion field given a static 3D asset and text prompt. Previous methods mostly leverage the vanilla Score Distillation Sampling (SDS) objective to distill motion from pretrained text-to-video diffusion, leading to animations with minimal movement or noticeable jitter. To address this, our approach introduces a novel SDS alternative, Motion Score Distillation (MSD). Specifically, we introduce a LoRA-enhanced video diffusion model that defines a static source distribution rather than pure noise as in SDS, while another inversion-based noise estimation technique ensures appearance preservation when guiding motion. To further improve motion fidelity, we incorporate explicit temporal and spatial regularization terms that mitigate geometric distortions across time and space. Additionally, we propose a motion refinement module to upscale the temporal resolution and enhance fine-grained details, overcoming the fixed-resolution constraints of the underlying video model. Extensive experiments demonstrate that *Animus3D* successfully animates static 3D assets from diverse text prompts, generating significantly more substantial and detailed motion than state-of-the-art baselines while maintaining high visual integrity. Code will be released upon acceptance.

CCS Concepts: • **Computing methodologies** → **Animation**; **Machine learning**;

Additional Key Words and Phrases: 3D Animation, Video Diffusion Model, Score Distillation Sampling

ACM Reference Format:
Qi Sun, Can Wang, Jiaxiang Shang, Wensen Feng, and Jing Liao*. 2025. Animus3D: Text-driven 3D Animation via Motion Score Distillation. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3757377.3763916

1 INTRODUCTION

Text-to-3D animation has long been a foundational component of visual storytelling, entertainment, and simulation. Recent advancements demonstrate that large-scale text-to-image and text-to-video diffusion models can effectively learn valuable priors for generating 3D animations [Jiang et al. 2024; Li et al. 2024b; Liang et al. 2024; Sun et al. 2024; Zhang et al. 2024].

A representative class of techniques that leverage such priors are score distillation sampling (SDS)-based methods [Brooks et al. 2024; Chen et al. 2023, 2024; HaCohen et al. 2024; He et al. 2022; Hong et al. 2022; Kong et al. 2024; Wang et al. 2025; Xing et al. 2023; Yang et al. 2024]. The core idea of SDS is to render an image of a 3D scene, add noise to the rendered image, and then use a pre-trained diffusion model to denoise it. The denoising process enables the estimation of gradients, which are then used to update the underlying 3D representation, such as neural radiance fields [Mildenhall et al. 2020] or

*Corresponding Author.
Authors' addresses: Qi Sun, qisun.new@gmail.com, City University of Hong Kong, Hong Kong; Can Wang, cwang355-c@my.cityu.edu.hk, City University of Hong Kong, Hong Kong; Jiaxiang Shang, Central Media Technology Institute, Huawei, China; Wensen Feng, Central Media Technology Institute, Huawei, China; Jing Liao*, jingliao@cityu.edu.hk, City University of Hong Kong, Hong Kong.

SA Conference Papers '25, December 15–18, 2025, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763916

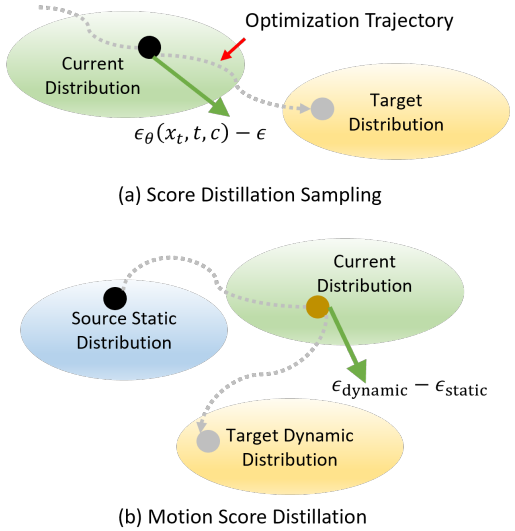

Fig. 2. **Motivation.** We illustrate two distillation sampling procedures: (a) Score Distillation Sampling and our (b) Motion Score Distillation.

Gaussian splatting [Kerbl et al. 2023]. Recent studies have explored the theoretical foundations of Score Distillation Sampling (SDS) and formulated it as a domain transportation problem [McAllister et al. 2024; Yang et al. 2023], where the goal is to shift the current data distribution (i.e., rendered outputs from the 3D representation) toward a target distribution. The estimated gradient guides this transformation, as illustrated in Fig. 2-(a). We observe that existing SDS-based motion generation methods all adopt the original formulation of SDS without modification, inherently following this transportation framework. However, this framework reveals several key limitations for motion distillation: 1) The current distribution lacks a well-defined static source as its starting point. In motion generation, a clear static initialization is typical; the absence of its explicit modeling can obscure the starting point of the optimization trajectory, potentially resulting in limited distilled motion. 2) Motion and appearance are inherently entangled. SDS, however, does not account for this interdependency, and its estimated gradient can lead to the degradation of appearance when the distribution evolves towards the motion target.

To address these challenges, we propose *Animus3D*, a text-driven 3D animation method. At the core of our approach is a novel Motion Score Distillation strategy as depicted in Fig. 2-(b), which consists of two key components. First, we define a source static distribution as a canonical space, modeled using a video diffusion model enhanced with Low-Rank Adaptation (LoRA) [Hu et al. 2022], capable of generating static video frames. Then, we introduce a noise inversion technique. This method estimates deterministic noise for gradient computation, thereby enabling effective control of motion direction while maintaining the integrity of the appearance. We demonstrate that our Motion Score Distillation can predict more accurate transportation directions, enabling more reasonable and substantial motions while preserving the object's appearance. Beyond the distillation process, we find motion field regularization to be crucial. To further enhance performance, we introduce temporal and spatial regularization terms into our method, which helps mitigate geometric distortions across time and space. Additionally, due to the fixed temporal resolution of video diffusion, the motion details of animated objects are constrained. To address this, we propose an extension of our work through a motion detailization module, which extends the temporal length and enhances motion detail.

To conclude, we summarize our main contributions as follows:

- ***Animus3D* Framework:** We propose *Animus3D*, a text-driven 3D animation framework capable of generating high-quality motion for static 3D assets from diverse text prompts.
- **Motion Score Distillation:** We introduce a novel Motion Score Distillation strategy that models the source static distribution using a LoRA-enhanced video diffusion model. Additionally, we adopt an appearance-preservation noise estimation technique, ensuring that the estimated noise during distillation does not affect the original 3D appearance.
- **Temporal and Spatial Regularizations:** We introduce temporal and spatial regularization terms to enhance motion field regularization, effectively reducing geometric distortions across time and space.
- **Motion Refinement Module:** We address the limitation of fixed temporal resolution in video diffusion by introducing a motion refinement module, which extends the temporal length and enhances motion details for animated objects.

2 RELATED WORK

2.1 Learning-based 3D Motion Generation

A growing body of work [Jiang et al. 2024; Liang et al. 2024; Pan et al. 2024; Ren et al. 2024; Sun et al. 2024; Wu et al. 2024a; Xie et al. 2024; Zeng et al. 2024] explores data-driven approaches for 3D motion generation by leveraging diffusion models to synthesize temporally consistent multi-view images, followed by pixel-wise optimization to recover coherent 3D representations with motion fields. For instance, SV4D [Xie et al. 2024] introduces temporal layers into multi-view diffusion [Voleti et al. 2024], enabling spatio-temporal modeling from monocular video inputs and supporting orbital-view synthesis of dynamic objects. Animate3D [Jiang et al. 2024] extends AnimateDiff [Guo et al. 2024] by incorporating multi-view images, generating temporally synchronized video sequences through 3D object rendering. Although these methods achieve impressive results on general object motion and are typically fast, they typically depend on large-scale training datasets, such as multi-view captures or densely sampled videos of dynamic scenes, which are often costly and difficult to obtain. In contrast, our method focuses on enhancing SDS to bridge the gap between 2D generative priors and 3D animation. By distilling motion knowledge from powerful 2D diffusion models, our approach enables motion generations without requiring extensive multi-view or temporal supervision.

2.2 SDS for 3D Generation

SDS [Poole et al. 2022; Wang et al. 2023a] was introduced for 3D content generation and image/video editing [Hertz et al. 2023; Jeong et al. 2024] by distilling supervision from pre-trained 2D diffusion models [Ho et al. 2022; Rombach et al. 2022]. A common issue with early SDS-based methods [Lin et al. 2023; Poole et al. 2022; Wang

et al. 2023a] is over-smoothing and lack of fine geometric or textural details. These methods often rely on high classifier-free guidance (CFG ∼ 100) [Ho and Salimans 2021] to reduce output variance, which tends to cause over-saturation and unnatural results. ProlificDreamer [Wang et al. 2023b] significantly improved fidelity by introducing a second diffusion model that is overfit to the current 3D estimate, allowing high-quality outputs with standard CFG values (e.g., 7.5). LucidDreamer [Liang et al. 2023] and SDI [Lukoianov et al. 2024] mitigates SDS's over-smoothing by replacing the random noise term with one obtained via DDIM inversion and applying multi-step denoising. Recent analyses such as SDS-Bridge [McAllister et al. 2024] and LODS [Yang et al. 2023] further explore theoretical foundations and architectural optimizations for SDS-based 3D generation. Distinct from these works, our approach distills motion from a pre-trained video diffusion model by optimizing a motion field for a given static 3D object.

2.3 SDS for Motion Generation

Recent advances in video diffusion models [Brooks et al. 2024; Chen et al. 2023, 2024; HaCohen et al. 2024; He et al. 2022; Hong et al. 2022; Kong et al. 2024; Wang et al. 2025; Xing et al. 2023; Yang et al. 2024] have inspired a growing line of research that distills dynamic 3D scenes evolving over time from pre-trained video diffusion models. MAV3D [Singer et al. 2023] is one of the earliest works in text-to-dynamic object generation, introducing a hexplane representation to model scene dynamics. Some approaches [Bahmani et al. 2024b; Zhao et al. 2023; Zheng et al. 2024] uses a hybrid SDS pipeline that multi-stage optimization between supervision from text-to-image and multi-view diffusion models, improving both geometric consistency and motion fidelity. Other methods [Li et al. 2024a; Ling et al. 2024; Wimmer et al. 2025] explore novel 3D representations to better capture motion. AYG [Ling et al. 2024] employs 3D Gaussian Splatting [Cui et al. 2025; Huang et al. 2024; Kerbl et al. 2023] for efficient and high-fidelity motion representation, while Text2Life [Wimmer et al. 2025] introduces a training-free autoregressive approach to generate consistent video guidance across viewpoints, enhancing the quality of the distilled dynamics. Several approaches also incorporate explicit motion priors to constrain or regularize the motion fields. TC4D [Bahmani et al. 2024a] uses parameterized object trajectories (e.g., translation and rotation) as motion priors. AKD [Li et al. 2025] further extends this idea by incorporating articulated skeletal structures into score distillation, guided by rigid-body physics simulators. However, these methods largely adopt the original SDS formulation without modification and do not explicitly address its limitations in motion generation. In contrast, we propose a novel Motion Score Distillation strategy tailored for motion optimization, and further introduce a motion refinement module to reduce distortion caused by score distillation, resulting in more stable training and improved motion fidelity.

3 PRELIMINARY

In this section, we first introduce the parametric 3D representation with motion fields. We then provide an overview of SDS.

3D Gaussian Splatting (3D-GS) [Kerbl et al. 2023] uses millions of learnable 3D Gaussians to explicitly represent a scene. Each Gaussian is defined by its center, rotation, scale, opacity, and view-dependent color encoded via spherical harmonics. The scene is rendered through a differentiable splatting-based renderer $\mathcal{R}_{\text{cam}}$ given camera parameters: $x = \mathcal{R}_{\text{cam}}(\mathcal{G})$. 4D Gaussian Splatting (4D-GS) [Wu et al. 2024b] extends 3D-GS by introducing a motion field to a canonical 3D representation. In our approach, we first reconstruct the static 3D object using 3D-GS, denoted as the canonical space $\mathcal{G}_c$. The motion field is modeled using a multi-resolution HexPlane with MLP-based decoders [Cao and Johnson 2023]. During training, we keep $\mathcal{G}_c$ fixed and optimize only the motion field. At each timestamp, the model queries the HexPlane using a 4D coordinate (x, y, z, τ) and decodes the resulting feature into deformation values for position and rotation. By querying the motion field at each timestamp $\tau \in \{0, \ldots, T-1\}$, we generate a sequence of deformed Gaussians $\mathcal{G}_{0:T-1}$. Given camera parameters, we render the resulting T-frame video as $x^o = \mathcal{R}_{\text{cam}}(\mathcal{G}_{0:T-1})$.

Score Distillation Sampling (SDS) leverages the knowledge from pretrained text-to-image diffusion models to optimize a parametric representation like 3D-GS. Given an output sample x_0, (e.g., a rendered image from a 3D-GS), SDS operates as follows: *stochastic* Gaussian noise ϵ is added to x_0 at a randomly sampled timestep t:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t$ is a noise schedule coefficient. After that, a pretrained denoising model $\epsilon_\theta(x_t, t, c)$ predicts the noise in x_t, conditioned on the timestep t and a text prompt c. SDS uses the difference between the predicted noise and the sampled stochastic noise as the gradient to update the parameterized representation:

$$\nabla \mathcal{L}_{\text{SDS}} = \mathbf{E}_{t,\epsilon}[w(t)(\epsilon_\theta(x_t, t, c) - \epsilon)\frac{\partial x_t}{\partial \phi}], \tag{2}$$

where $w(t)$ is the weighting function. Recent works [McAllister et al. 2024; Yang et al. 2023] formulate SDS as a domain transportation problem, aiming to find the optimal transport from the current data distribution $\mathcal{D}_c$ to the target distribution $\mathcal{D}_t$. Here, the rendered sample x_0 is drawn from $\mathcal{D}_c$ as $x_0 \sim \mathcal{D}_c$, while the text condition c describes the target distribution $\mathcal{D}_t$. SDS approximates the optimal transport step ϵ^* between $\mathcal{D}_c$ and $\mathcal{D}_t$ at a given timestep t by:

$$\epsilon^* = \epsilon_\theta(x_t, t, c) - \epsilon \tag{3}$$

Here $\epsilon_\theta(x_t, t, c)$ is a projection of the noised image x_t onto the target distribution and ϵ is a random gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

4 METHOD

Given the canonical 3D-GS $\mathcal{G}_c$ of a 3D object and a text prompt c describing the desired motion, our method aims to automatically predict a motion field $f(\phi)$ for $\mathcal{G}c$. This produces a Gaussian sequence $\mathcal{G}_{0:T-1}$ that exhibits substantial, photorealistic motion while preserving the object's appearance. To achieve this, as illustrated in Fig.3, we first introduce Motion Score Distillation (§4.1), an enhanced SDS framework tailored for motion learning. It incorporates dual distribution modeling (§4.1.1) and a appearance preservation noise estimation (§4.1.2) to better guide motion generation. We further propose temporal and spatial regularization terms (§4.2) to constrain the deformation fields and a Motion refinement method to extends the temporal length and enhances motion detail (§4.3), as shown in Fig. 5.

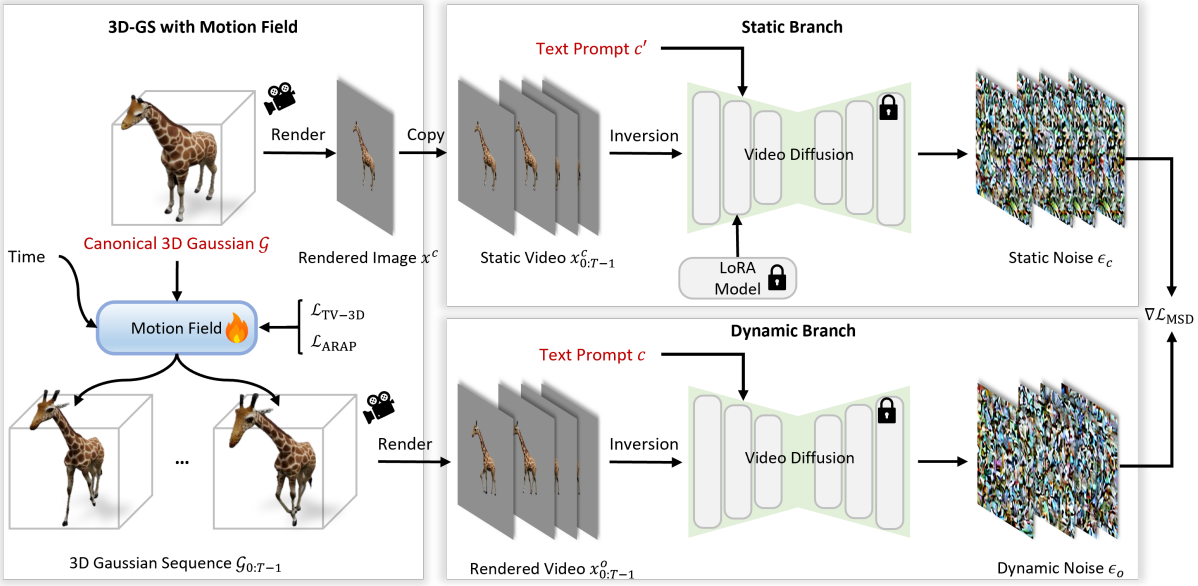

Fig. 3. **Framework of motion generation.** Given canonical 3D Gaussian $\mathcal{G}$, the motion field predicts the offset of Gaussian properties in each timestamp, obtaining the Gaussian sequence $\mathcal{G}_{0:T-1}$. Then given a camera parameter, we can render the image and video from $\mathcal{G}$ and $\mathcal{G}_{0:T-1}$. We use video diffusion and video diffusion with LoRA to model the dynamic distribution and static distribution, respectively. Given dynamic text prompt c and static text prompt c', the loss gradient is computed with two predicted noises. The gradient will guide the optimization of the motion field. We further design temporal and spatial regularization terms for the motion field to improve the performance.

4.1 Motion Score Distillation

Building on the explanation of SDS in the previous section, we propose a novel approach called Motion Score Distillation (MSD). MSD aims to estimate the optimal transport from a static source distribution to a dynamic target distribution. Different from SDS, MSD approximates the optimal motion step $\epsilon^*_{\text{motion}}$ between the dual distributions at a given timestep t as follows:

$$\epsilon^*_{motion} = \epsilon_{\text{dynamic}} - \epsilon_{\text{static}}. \tag{4}$$

Thus, our MSD is formulated as follows:

$$\nabla\mathcal{L}_{\text{MSD}} = \mathbf{E}_t[w(t)(\epsilon_{\text{dynamic}} - \epsilon_{\text{static}})\frac{\partial x_t}{\partial t}]. \tag{5}$$

We demonstrate that $\epsilon_{\text{dynamic}} - \epsilon_{\text{static}}$ serves as an effective gradient when both the source static and target dynamic distributions are well expressed. Next, we detail the definitions of $\epsilon_{\text{dynamic}}$ and ϵ_{static}.

4.1.1 Dual distribution modeling. Given a time sequence $\{0 : T-1\}$, we define the static video rendered from the static 3D-GS $\mathcal{G}_c$ as $x^s_{0:T-1}$, and the dynamic video rendered from the dynamic 3D-GS $\mathcal{G}_{0:T-1}$ as $x^d_{0:T-1}$. Similar to SDS, the target dynamic distribution can be approximated using a pretrained latent video diffusion model:

$$\epsilon_{\text{dynamic}} = \epsilon_{\text{dynamic}}(x^d_t, t, c) = \epsilon_\theta(x^d_t, t, c). \tag{6}$$

Here, the text prompt c describes the motion of the object, such as "*A walking <object>*". However, modeling the source static distribution is non-trivial. An effective static distribution should satisfy two key properties: 1) it should preserve the original semantics of the reconstructed object in the canonical space; 2) it should represent a video sequence without introducing any motion. An intuitive solution is to model the static distribution as:

$$\epsilon_{\text{static}} = \epsilon_{\text{static}}(x^s_t, t, c') = \epsilon_\theta(x^s_t, t, c'), \tag{7}$$

where c' is a static text description. However, we observe that even when conditioned on a static description, the video diffusion model does not consistently generate videos of truly static objects, thus violating the second requirement. To address this, we propose to efficiently derive a static denoiser by adapting Low-Rank Adaptation (LoRA) [Hu et al. 2022]:

$$\epsilon_{\text{static}} = \epsilon_{\text{static}}(x^s_t, t, c') = \epsilon_{\text{lora}}(x^s_t, t, c'), \tag{8}$$

where ϵ_{lora} is the lora-fined denoiser. The LoRA parameters are trained using the static video $x^s_{0:T-1}$, with the loss function:

$$\mathcal{L}_{\text{diff}} = \mathbf{E}_{t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon_{\text{lora}}(x^s_t, t, c') - \epsilon||^2]. \tag{9}$$

4.1.2 Appearance preserved faithful noise estimation. In SDS and its variants [Bahmani et al. 2024a; Ling et al. 2024], we observe that it is difficult to preserve the original appearance of the static object. In some cases, the object even drifts into the background. SDS's noise estimation entangles motion and appearance, but we only optimize the motion field. This means that the appearance loss has to be compensated by geometric distortion, causing artifacts, which we refer as *motion-appearance entanglement*. We find that this issue is strongly correlated with the stochastic noise ϵ added during the diffusion process in Eq. 1. We have assessed this observation in Fig. 4. Therefore, instead of adding stochastic noise, we adopt DDIM inversion [Lukoianov et al. 2024; Song et al. 2022] to obtain deterministic and faithful noise. Given a noised input x_t, we first

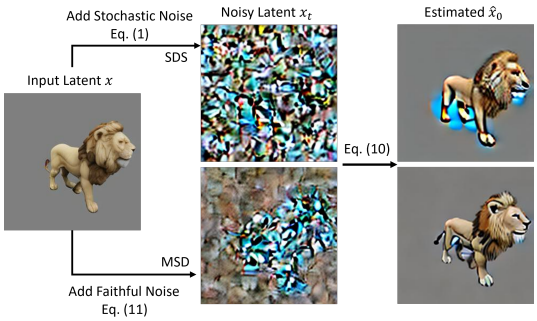

Fig. 4. **Motivation of faithful noise.** Given an input image, we first add noise using the noise in SDS and our MSD, and then denoise it to obtain an estimated image. We find that the denoised image using SDS is significantly different from the original, exhibiting large appearance changes and background noise. In contrast, our MSD better preserves the appearance and maintains a clearer background. All latents are decoded into pixel space for visualization. We use t=600 in this case.

predict the noise $\epsilon_\theta(x_t, t, c)$ using the pretrained diffusion model. We then estimate the corresponding denoised image $\hat{x}_0$ as:

$$\hat{x}_0(x_t, t, c) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, c)}{\sqrt{\bar{\alpha}_t}}. \tag{10}$$

Subsequently, we apply deterministic forward noising steps to obtain x_{t+1} iteratively, continuing until the predefined timestep t:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\hat{x}_0(x_t, t, c) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(x_t, t, c). \tag{11}$$

DDIM inversion provides a deterministic noise estimate, producing a denoised output $\hat{x}_0$ that is faithfully consistent with the input video. This facilitates appearance preservation during the optimization process. We then apply this method to x_t^d in Eq. 6 and to x_t^s in Eq. 8.

$$\begin{aligned} x_t^d &= \sqrt{\bar{\alpha}_t}\hat{x}_0(x_{t-1}^d, t, c) + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_{t-1}^d, t, c), \\ x_t^s &= \sqrt{\bar{\alpha}_t}\hat{x}_0(x_{t-1}^s, t, c') + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_{t-1}^s, t, c'). \end{aligned} \tag{12}$$

4.2 Motion Regularization

To further improve performance, our method incorporates both temporal and spatial regularization terms.

4.2.1 Gaussian trajectory total variation (TV-3D) for temporal regularization. Inspired by traditional 2D total variation (TV) losses applied in pixel space [Chan et al. 2005], we propose a TV-3D loss to encourage temporal smoothness in motion. This loss directly penalizes abrupt changes in the 3D positions of Gaussians across consecutive timesteps. Specifically, it computes the $\mathcal{L}_1$ norm of the positional differences for each Gaussian between adjacent frames:

$$\mathcal{L}_{\text{TV-3D}} = \frac{1}{N \cdot (T-1)} \sum_{\tau=1}^{T-1} \sum_{i=0}^{N} ||\mathbf{x}_{i,\tau} - \mathbf{x}_{i,\tau-1}||_1, \tag{13}$$

Here, $x_{i,\tau}$ denotes the position of the i-th 3D Gaussian at timestep τ. By operating in the 3D Gaussian space, this constraint effectively enforces temporal consistency in the underlying geometric motion.

4.2.2 As-rigid-as-possible (ARAP) for spatial regularization. To facilitate the learning of rigid motion dynamics while preserving the high-fidelity appearance of the static reference model, we employ an ARAP [Sorkine and Alexa 2007] regularization term for spatial smoothness. For each gaussian point p_j, its ARAP contribution $\mathcal{L}_{\text{ARAP}}(p_j)$ is computed over the sequence of frames $\tau \in \{0, \cdots, T-1\}$ as:

$$\mathcal{L}_{\text{ARAP}} = \sum_{j=1}^{N} \sum_{\tau=0}^{T-1} \sum_{k \in \mathcal{N}_j} w_{jk} \left\| (p_j^\tau - p_k^\tau) - R_j^\tau (p_j^c - p_k^c) \right\|^2, \tag{14}$$

where $p_j^\tau, p_k^\tau \in \mathbb{R}^3$ are the spatial positions of point j and its neighbor k at the current frame τ. $p_j^c, p_k^c \in \mathbb{R}^3$ are their corresponding positions in the canonical 3D-GS. $\mathcal{N}_j$ denotes the set of neighboring point indices for p_j^c, defined in the reference configuration (e.g., points within a fixed radius of p_j^c). $R_j^\tau \in SO(3)$ is the optimal local rigid rotation for point p_j at frame τ. This ARAP loss enforces spatial consistency by encouraging locally rigid deformations, thereby promoting realistic motion while preserving geometric fidelity.

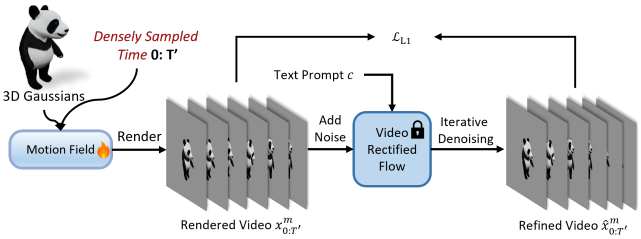

Fig. 5. **Motion refinement.** We proposed a motion refinement method to obtain fine-grained motion details.

4.3 Motion Refinement

The fixed frame length in video diffusion models limits the ability of SDS-generated 3D animations to capture fine-grained motion details. To address this, we introduce a motion refinement module that leverages a high-capacity, pre-trained rectified flow-based text-to-video model to generate long, detailed animation sequences, illustrated in Fig. 5. Given a time sequence $\{0 : T-1\}$, we interpolate it to produce $T' = 2T-1$ frames. We then render the 3D-GS with the motion field, resulting in a higher-resolution video $x^m_{0:T'-1}$ with spatial dimensions $H' \times W' \times T'$. Following the SDEdit [Meng et al. 2022] framework, we add noise to $x^m_{0:T'}$ and apply iterative denoising to generate a refined video $\hat{x}^m_{0:T'-1}$. This process preserves the original motion while enhancing temporal consistency and motion details. Finally, we optimize the motion field using an $\mathcal{L}1$ loss between the refined video $\hat{x}^m_{0:T'-1}$ and the initial input $x^m_{0:T'-1}$. This results in a motion field capable of producing longer and more detailed animations than the original.

5 EXPERIMENTS

5.1 Experimental Setup

Implementation details We implement our method using threestudio [Guo et al. 2023]. We use ModelScopeT2V [Wang et al. 2023c] as the base video model. The image resolution is set to 256, with 16 frames per video. All experiments are conducted on a single 24G GPU, with approximately 3k iterations for LoRA (rank=4, alpha=4) fine-tuning with learning rate 1e-5, 5k iterations for motion distillation, and 100 iterations for motion detailization.

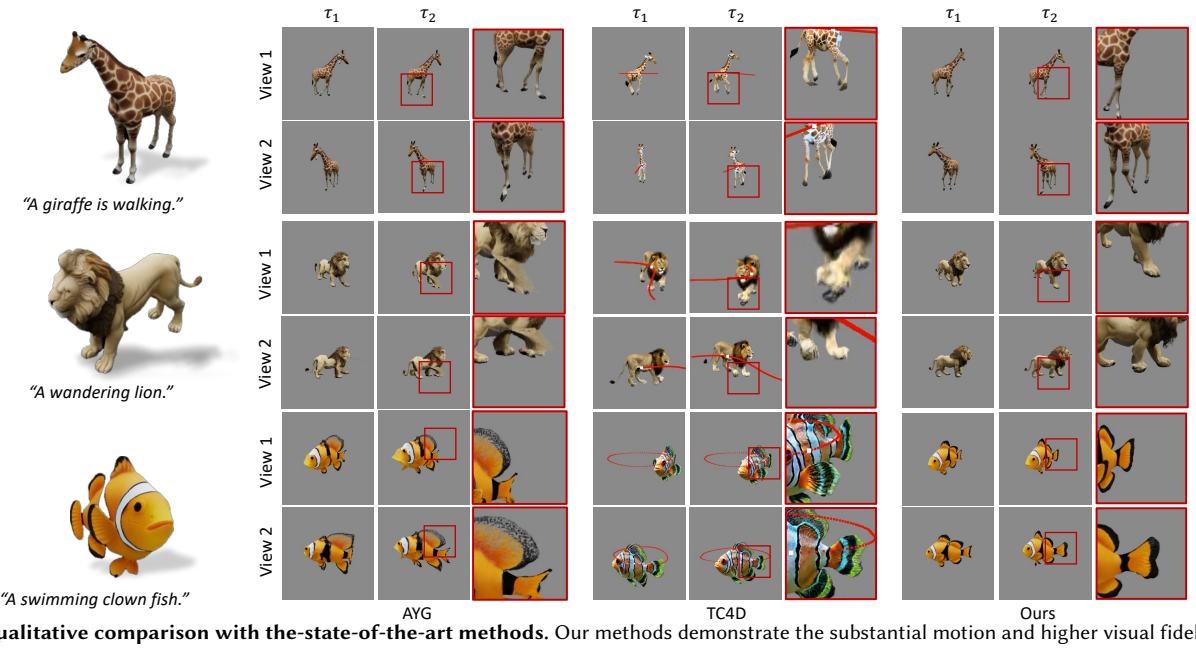

Fig. 6. **Qualitative comparison with the-state-of-the-art methods.** Our methods demonstrate the substantial motion and higher visual fidelity of 3D animation. It is recommend to watch the demo for better visualization.

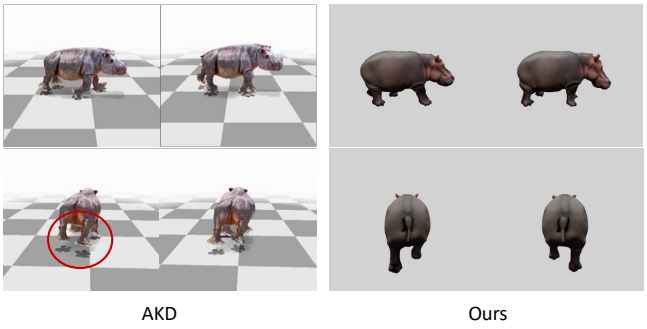

Fig. 7. **Comparison with the concurrent work [Li et al. 2025].** AKD utilizes skeleton-based motion, which tends to result in perceptible stiffness in its animations.

Table 1. **Quantitative comparison.**

Methods	CLIP-Image↑	CLIP-Text ↑	FID ↓	FVD ↓
AYG [Ling et al. 2024]	91.75	44.31	105.33	647.6
TC4D [Bahmani et al. 2024a]	90.99	50.02	179.14	340.0
Ours	**93.04**	**51.05**	**88.50**	**204.1**

Table 2. **User study.**

Method Preference	Ours preferred (%)	AYG preferred (%)	TC4D preferred (%)
Overall Quality	**65.9**	13.4	20.7
Appearance Preservation	**69.5**	11.0	19.5
Motion Dynamism	**75.5**	14.2	10.3
Motion Text Alignment	**55.3**	7.1	36.5
Motion Realism	**70.1**	9.7	19.2

Evaluation metrics. Following previous works [Bahmani et al. 2024a; Ling et al. 2024], we utilize CLIP-image to evaluate the semantic similarity between the rendered canonical 3D-GS and the 3D-GS with the motion field. For this, we render 8 views evenly spaced around the azimuth for calculation. Additionally, we use CLIP-text to assess the alignment of the rendered object with the given text prompt. To evaluate the overall quality of the rendered videos, we compute FID [Heusel et al. 2017] and FVD [Unterthiner et al. 2018].

Comparison setting. We compare our method with two state-of-the-art SDS-based 3D motion generation methods: AYG [Ling et al. 2024] and TC4D [Bahmani et al. 2024a]. For TC4D, since the authors did not release the animated 3D objects, we were unable to use the same 3D assets for animation. Instead, we extracted screenshots from their results, then used Trellis [Xiang et al. 2024] to generate comparable 3D objects for evaluation. Since AYG did not release their code, we opted to use the self-reproduced results for a fairer comparison. To ensure consistency across evaluations, we used the same 3D object generation approach as with AYG. We also include comparison with 4dfy [Bahmani et al. 2024b] and Dream-in-4D [Zheng et al. 2024] in the supplementary materials. For all comparisons, we used the same motion description. For a fair comparison, we do not employ motion refinement (§4.3).

5.2 Comparisons

We report the quantitative results in Table 1. Our method achieves a better CLIP-Image score, as it incorporates an appearance-preserving faithful noise estimation that prevents appearance degradation during distillation. Moreover, our method outperforms others in the CLIP-Text score, demonstrating that our MSD can better generate motions that align with the user's intension. Furthermore, the improved FID and FVD values confirm that our method produces high-quality results. Fig. 6 presents a visual comparison. AYG struggles to generate semantic-level and large motions due to its use of a simple SDS variant, as seen in the giraffe example. Furthermore, the motion generated by AYG is not smooth and exhibits noticeable

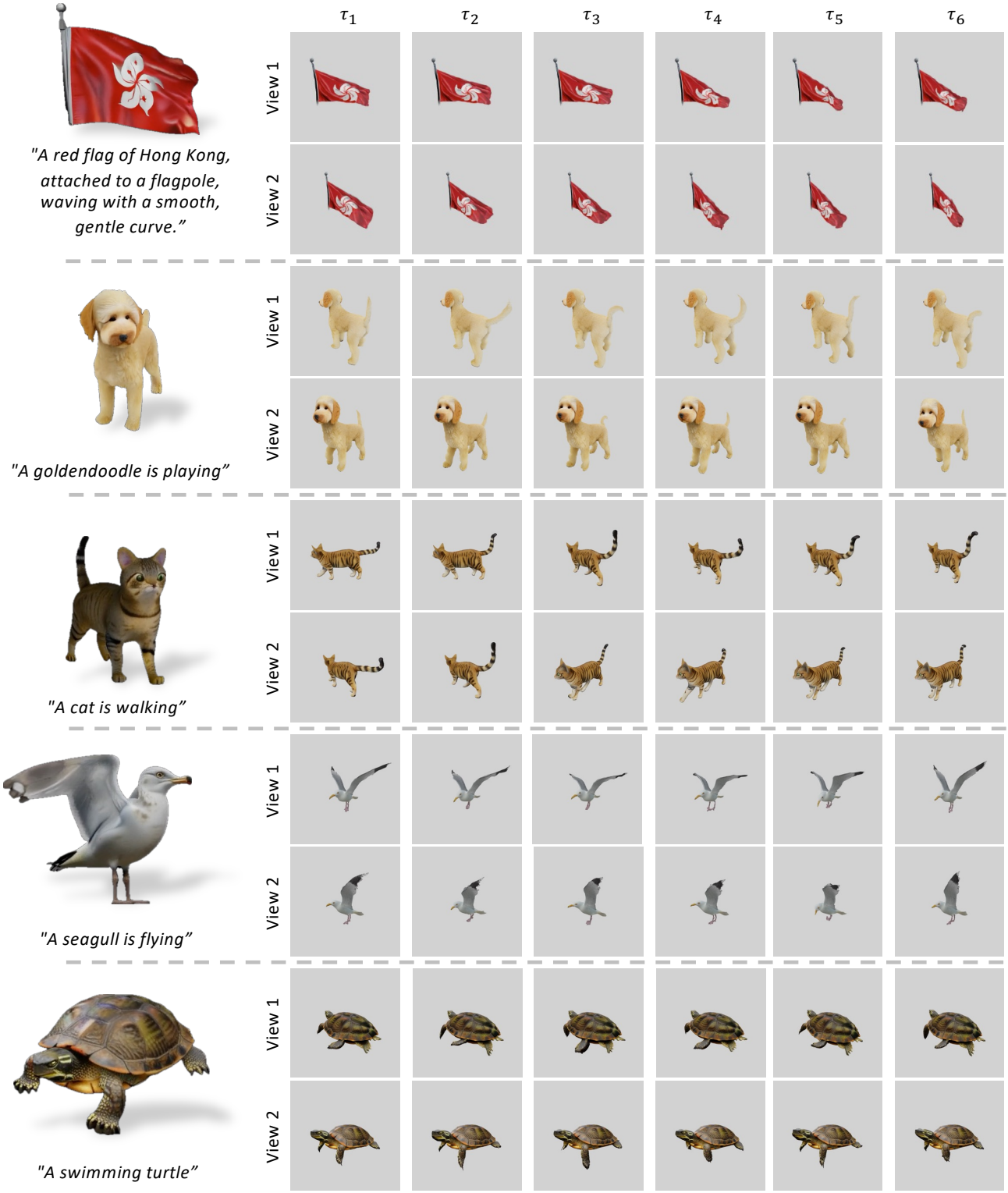

Fig. 8. **Results of our generated 3D animations.** Utilizing diverse text prompts and various 3D assets, our framework generates objects faithful to their original appearance and exhibiting substantial motion. Better watch video demo to see the motion.

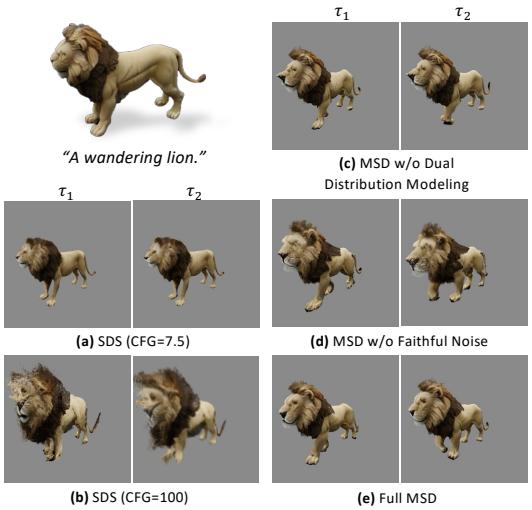

Fig. 9. **Comparison between different score distillation methods.**

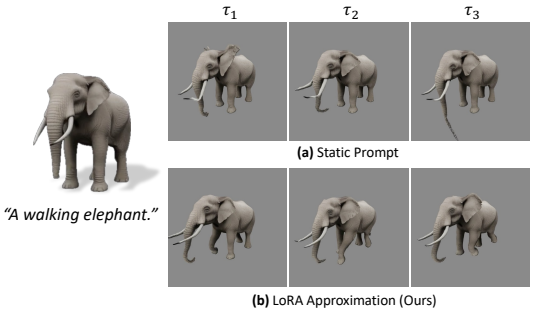

Fig. 10. **Ablation on different static distribution modeling methods.**

flickering, particularly evident in the fish case. TC4D shows simple translations along the trajectory with minimal skeleton movement. It also suffers from significant distortions and appearance changes, as demonstrated in the lion example. While both AYG and TC4D can animate the 3D object, they fail to produce natural and realistic motions. In contrast, our framework, leveraging MSD along with regularizations, generates more substantial and realistic 3D motions.

We conduct a user preference study to evaluate performance in Table 2. Users are asked to evaluate five key aspects of the generated dynamic object. First, *overall quality* provides a general evaluation of the rendered object. *Appearance preservation* focuses on detecting any undesirable appearance deformations. *Motion dynamism* assesses the extent of the object's movement, with a preference for larger motions. *Motion-text alignment* measures how well the generated motion corresponds to the text prompt. Finally, *motion realism* evaluates the naturalness of the generated motion. For each aspect, users are asked to select their preferred option from AYG, TC4D, and our method. We received 17 valid responses and present the user preference rates for each aspect. Our method achieves the highest preference rate across all aspects, demonstrating that it generates more natural and realistic motions.

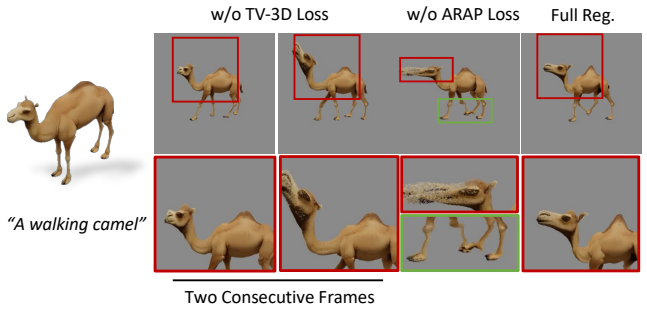

Fig. 11. **Effectiveness of motion regularization.** Without TV-3D loss, the camel demonstrate the large object deformation in consecutive frames, while without ARAP loss, the camel would perform non-rigid motion.

5.3 Ablation Studies

5.3.1 Comparison between different score distillation methods. We compare our motion distillation approach across several variants: (a) SDS with CFG=7.5, (b) SDS with CFG=100, (c) our method without faithful noise, (d) our method without dual distribution modeling, and (e) our full MSD approach. No motion regularization is applied in these experiments to ensure a fair comparison. From variant (c), we observe that generating substantial and large motion is difficult without explicitly modeling the static distribution. While variant (d) produces sufficient motion, the lack of faithful noise significantly compromises the original appearance fidelity and introduces notable background artifacts. Without our MSD, the approach degrades to conventional SDS, which either fails to generate noticeable motion with a small CFG (a) or results in meaningless motion and severe appearance distortion with a large CFG (b). In contrast, our full MSD method is uniquely effective in generating realistic motion while robustly preserving the original appearance fidelity.

5.3.2 Denoising with approximated static distribution. As discussed in §4.1.1, an alternative approach to defining the source static distribution is to use a static text prompt, such as "*low motion, static statue, not moving, no motion, <object>*", as described in Eq.7. However, we observe that even when conditioned on such a static description, the video diffusion model does not consistently generate videos of truly static objects and often leads to the introduction of local distortions during distillation, as shown in Fig.10. In contrast, by using a LoRA-enhanced model to model the static distribution, we can effectively induce large, reasonable motions.

5.3.3 The effectiveness of motion regularization. As illustrated in Fig. 11 (a), the TV-3D loss encourages temporal smoothness in Gaussian points, and removing this loss results in large displacements between consecutive frames. In panel (b), we show that the ARAP loss provides a crucial spatial constraint for maintaining rigid motion; without this loss, the object experiences significant distortion or may even break apart. Therefore, we incorporate both temporal and spatial regularization terms in our method to further enhance the results of MSD.

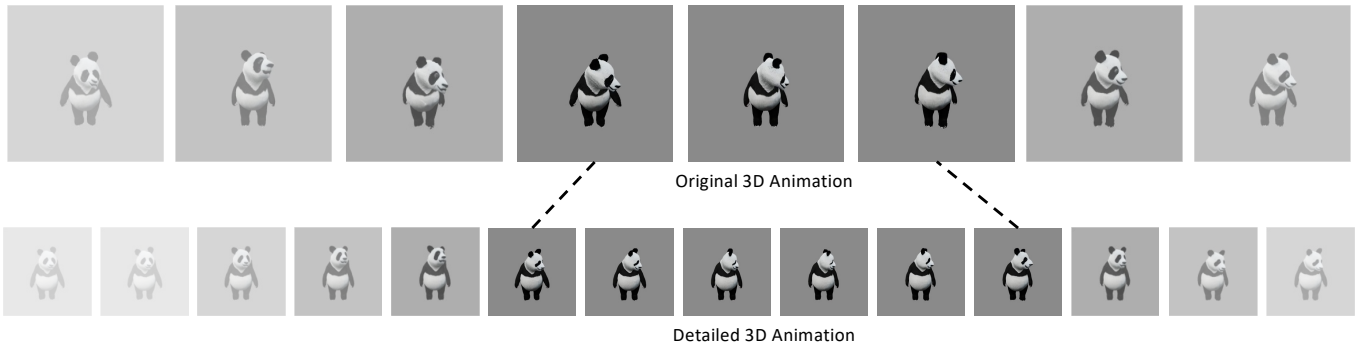

Fig. 12. **Demonstration of motion refinement.** Limited by the frame length of the underlying video diffusion, SDS could only generate fixed-length animation, leading to non-continuous motion (large motion change in consecutive frames) in the original 3D animation; Our motion refinement could generate more fine-grained motion (longer 3D animation) benefited from additional larger video diffusion model.

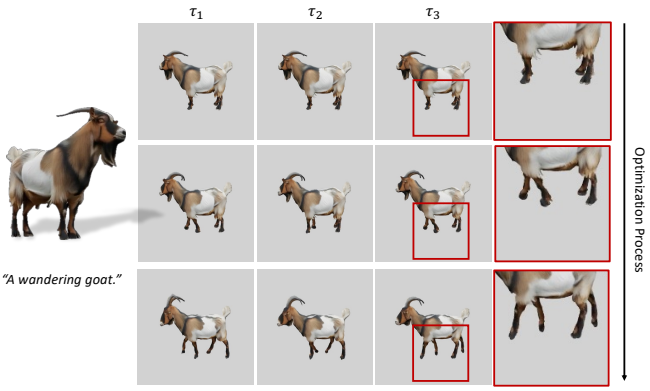

Fig. 13. **Visualization of motion generation process.** With increasing training iterations, 3D object could generate more substantial motion with MSD optimization.

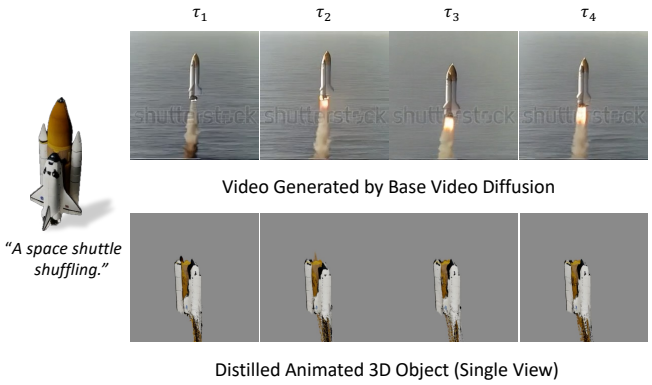

Fig. 14. **Failure case**. Our motion could only deform the given 3D object, which can hardly model fluid ejected from inside a rocket.

5.4 Visual Results

We illustrate the motion distillation process in Fig. 13, where our framework progressively optimizes the motion field of the 3D object (goat) over multiple optimization steps. This allows the 3D object to appear "moving" while preserving its original appearance. Due to the frame length limitation of the underlying video diffusion, SDS can only generate fixed-length animations, resulting in non-continuous motion (i.e., large motion changes between consecutive frames) in the original 3D animation. In contrast, our motion refinement approach generates more fine-grained motion (longer 3D animations) by leveraging a larger video diffusion model. As shown in Fig. 12, our motion refinement transforms non-continuous motion, caused by the fixed-length video diffusion, into more natural and smooth animation sequences. We present additional results in Fig. 8. Our method supports a variety of motion descriptions, such as "playing," "walking," "flying," and "swimming." Furthermore, our approach can animate not only animals but also general objects, as demonstrated with the red flag of Hong Kong.

6 CONCLUSION

In this paper, we introduce Motion Score Distillation (MSD) for text-driven 3D animation. Our method formulates score distillation as distribution transportation, enhancing conventional techniques through dual distribution modeling and faithful noise. Specifically, we tackle the challenge of static video distribution modeling by using LoRA-enhanced video diffusion, and we perform appearance-preserving faithful noise estimation to mitigate the appearance changes often encountered in SDS. Additionally, we integrate spatial-temporal geometric motion regularizations and apply motion detailization using large video models to ensure scalability. Experimental results on text-driven 3D animation, along with comprehensive ablation studies, demonstrate that our method outperforms current state-of-the-art approaches and validates its effectiveness.

In Fig. 14, we illustrate a limitation of our method: it struggles to model new content appearing in the scene, such as ejected fluid. Instead of generating this new content, the model tends to introduce distortions as a form of compensation. This issue could potentially be addressed in the future by designing new particle generation and modeling strategies. Another challenge is the optimization time—a common drawback of score distillation methods—which typically requires several hours. This inefficiency could be mitigated through techniques such as amortized training [Lorraine et al. 2023] or the adoption of more efficient data structures [Müller et al. 2022].

ACKNOWLEDGMENTS

We thank all anonymous reviewers and area chairs for their valuable comments. This work was supported by Central Media Technology Institute, Huawei [Project No. TC20240927042] and a GRF grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China [Project No. CityU 11208123].

REFERENCES

Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. 2024a. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*. Springer, 53–72.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 2024b. 4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

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. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

Ang Cao and Justin Johnson. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 130–141.

Tony Chan, Selim Esedoglu, Frederick Park, Yip, A, et al. 2005. Recent developments in total variation image restoration. *Mathematical Models of Computer Vision* 17, 2 (2005), 17–31.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512 [cs.CV]

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047 [cs.CV]

Xiao Cui, Weicai Ye, Yifan Wang, Guofeng Zhang, Wengang Zhou, Tong He, and Houqiang Li. 2025. StreetSurfGS: Scalable Urban Street Surface Reconstruction with Planar-based Gaussian Splatting. *IEEE Transactions on Circuits and Systems for Video Technology* (2025), 1–1.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *ICLR*.

Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. 2024. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103* (2024).

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent Video Diffusion Models for High-Fidelity Long Video Generation. (2022). arXiv:2211.13221 [cs.CV]

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. 2023. Delta Denoising Score. (2023).

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

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

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868* (2022).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Article 32, 11 pages.

Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. 2024. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *European Conference on Computer Vision*. Springer, 358–376.

Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. 2024. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398* (2024).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

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

Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. 2024b. Vivid-ZOO: Multi-View Video Generation with Diffusion Model. arXiv:2406.08659

Xuan Li, Qianli Ma, Tsung-Yi Lin, Yongxin Chen, Chenfanfu Jiang, Ming-Yu Liu, and Donglai Xiang. 2025. Articulated Kinematics Distillation from Video Diffusion Models. *arXiv preprint arXiv:2504.01204* (2025). https://arxiv.org/abs/2504.01204

Zhiqi Li, Yiming Chen, and Peidong Liu. 2024a. DreamMesh4D: Video-to-4D Generation with Sparse-Controlled Gaussian-Mesh Hybrid Representation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. 2024. Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models. *arXiv preprint arXiv:2405.16645* (2024).

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 2023. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284* (2023).

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. 2024. Align your Gaussians: Text-to-4D with dynamic 3D gaussians and composed diffusion models. In *CVPR*.

Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. 2023. ATT3D: Amortized Text-to-3D Object Synthesis. *The International Conference on Computer Vision (ICCV)* (2023).

Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Campagnolo Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin Solomon. 2024. Score Distillation via Reparametrized DDIM. arXiv:2405.15891 [cs.CV]

David McAllister, Songwei Ge, Jia-Bin Huang, David W. Jacobs, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. 2024. Rethinking Score Distillation as a Bridge Between Image Distributions. In *Advances in Neural Information Processing Systems*.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. 2024. Efficient4D: Fast Dynamic 3D Object Generation from a Single-view Video. *arXiv preprint arXiv 2401.08742* (2024).

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. 2024. L4GM: Large 4D Gaussian Reconstruction Model. *arXiv preprint arXiv:2406.10324* (2024).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023. Text-to-4D dynamic scene generation. *arXiv preprint arXiv:2301.11280* (2023).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG]

Olga Sorkine and Marc Alexa. 2007. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing* (Barcelona, Spain) *(SGP '07)*. 109–116.

Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. 2024. EG4D: Explicit Generation of 4D Object without Score Distillation. *arXiv preprint arXiv:2405.18132* (2024).

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A

new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2024. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008* (2024).

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *CVPR*. 12619–12629.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023c. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Thomas Wimmer, Michael Oechsle, Michael Niemeyer, and Federico Tombari. 2025. Gaussians-to-Life: Text-Driven Animation of 3D Gaussian Splatting Scenes. In *International Conference on 3D Vision (3DV)*.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024b. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20310–20320.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2024a. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613* (2024).

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506* (2024).

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2024. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470* (2024).

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. (2023). arXiv:2310.12190 [cs.CV]

Xiaofeng Yang, Yiwen Chen, Cheng Chen, Chi Zhang, Yi Xu, Xulei Yang, Fayao Liu, and Guosheng Lin. 2023. Learn to optimize denoising scores for 3d generation: A unified and improved diffusion prior on nerf and 3d gaussian splatting. *arXiv preprint arXiv:2312.04820* (2023).

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. 2024. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939* (2024).

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 2024. 4Diffusion: Multi-view Video Diffusion Model for 4D Generation. *arXiv preprint arXiv:2405.20674* (2024).

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. 2023. Animate124: Animating one image to 4D dynamic scene. *arXiv preprint arXiv:2311.14603* (2023).

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. 2024. A Unified Approach for Text- and Image-guided 4D Scene Generation. In *CVPR*.