DIMO: Diverse 3D Motion Generation for Arbitrary Objects

Linzhan Mou[1] Jiahui Lei[1†] Chen Wang[1] Lingjie Liu[1] Kostas Daniilidis[1,2]
[1]University of Pennsylvania [2]Archimedes, Athena RC

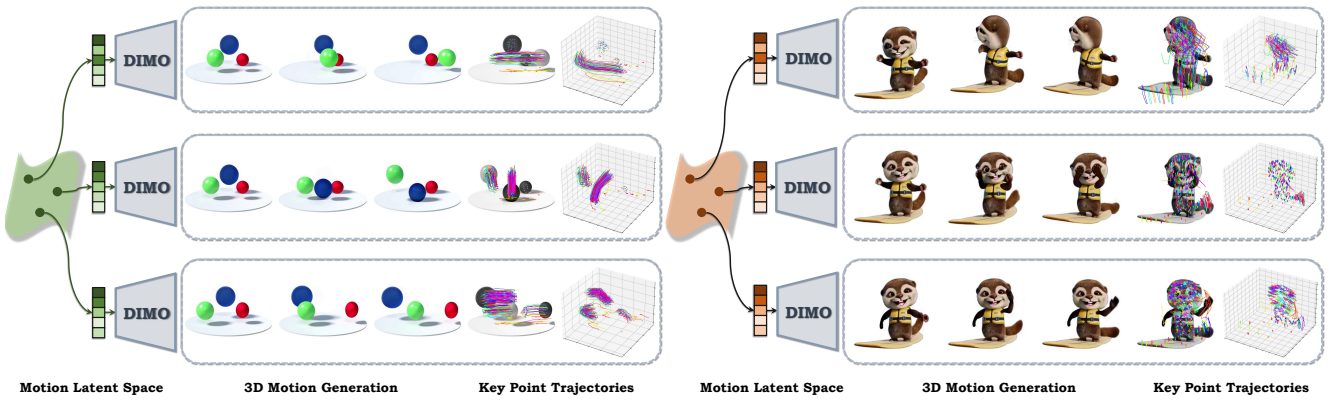

Figure 1. During inference, **DIMO** can ***instantly*** generate ***diverse*** 3D motions and high-fidelity 4D contents in ***a single forward pass*** from ***a single generative model***, by sampling from a continuous motion latent space.

Abstract

*We present **DIMO**, a generative approach capable of generating diverse 3D motions for arbitrary objects from a single image. The core idea of our work is to leverage the rich priors in well-trained video models to extract the common motion patterns and then embed them into a shared low-dimensional latent space. Specifically, we first generate multiple videos of the same object with diverse motions. We then embed each motion into a latent vector and train a shared motion decoder to learn the distribution of motions represented by a structured and compact motion representation, i.e., neural key point trajectories. The canonical 3D Gaussians are then driven by these key points and fused to model the geometry and appearance. During inference time with learned latent space, we can instantly sample diverse 3D motions in a single-forward pass and support several interesting applications including 3D motion interpolation and language-guided motion generation.*

1. Introduction

4D generation should be diverse in terms of motion. Our community desires a generative model capable of producing these dynamic 3D objects, and once generated, with not just a single motion per object but a space of many possible motions. This paper takes the first step toward addressing this challenging problem.

Previously, generating dynamic objects with such rich motion spaces has only been possible for category-specific objects, such as humans and animals, using template models where motion priors are obtained through time-consuming and labor-intensive motion capture. A successful example is the SMPL [33, 37] model for humans, whose generative capabilities are built upon massive pose sequences captured from real human data. However, this paradigm is not scalable to more general objects in the real world. *How can we build an SMPL-like model that can model and generate diverse 3D motions for any dynamic object?*

On the other hand, previous 4D object generative models [3, 18, 27, 31, 36, 43, 44, 48, 55, 69, 70, 73, 75, 76, 78, 83, 84] can work on general objects, but can only generate one motion per object in each expensive inference pass. Generating diverse 3D motions of the same object requires re-running the diffusion model [5, 32, 61, 70] and 4D reconstruction [20, 34] repeatedly, which will easily lose identity and cost extensive computing. *How can we directly output a 4D object with a diverse motion space that can be sampled instantly during inference?*

Our key insight to address these issues is that recent advanced video models [5, 7, 62, 74] already contain rich motion prior knowledge for general objects and can serve the

† Corresponding Author {leijh@cis.upenn.edu}

role of the previous expensive motion capture for category-specific objects like humans. With a proper motion representation, we can distill motion priors to specific objects being generated and build an SMPL-like model for any dynamic object, benefiting from jointly learning numerous motion patterns of the same object together.

Building upon this insight, we introduce DIMO, the first generative model of diverse 3D motions for any general objects. DIMO addresses two key issues: *where* the diverse motions prior knowledge comes from and *how* to model these motions into a unified object-specific motion space. First, to obtain diverse motion knowledge for the target object, we generate numerous videos containing diverse motion patterns from a single-view image with diverse motion text prompts [1, 59] (Sec. 3.1). Second, to jointly model diverse motion patterns, we propose to factorize each 3D motion sequence into explicit and compact neural curves represented by sparse key point trajectories (Sec. 3.2.1). This allows us to embed motions into a shared low-dimension latent space and jointly learn their diverse distributions within a single generative model (Sec. 3.2.2). To further capture the object geometry and appearance, we attach canonical 3D Gaussians [20] to the dynamic neural key points and fuse them for differentiable 4D optimization using only photometric losses (Sec. 3.2.3).

During inference time, we can **instantly** generate diverse 3D motions and 4D contents in a single forward pass by sampling from the motion latent space. We can also generate new 3D motions by interpolating within this space and reconstruct unseen motions by optimizing latent codes. Additionally, it supports the automatic creation of 4D animations that align with natural language descriptions, making motion generation both intuitive and versatile (Sec. 4.3).

In summary, our main contributions include:

1. We propose the first generative approach of *diverse* 3D motions for any *general* objects from a single-view image, by distilling motion priors from video models.
2. We embed each motion pattern into shared latent space and *jointly* learn the diverse motion distributions represented by structured neural key point trajectories.
3. At inference time, we can *instantly* generate *diverse* 3D motions and 4D contents in *a single forward pass* and support applications like latent space motion interpolation and language-guided motion generation.
4. State-of-the-art performance on extensive settings and standard 4D generation benchmarks.

2. Related Works

Video Generative Model. Leveraging Internet-scale image and video datasets, 2D video diffusion models (e.g. T2V and I2V) [5–7, 11, 13, 14, 22, 47, 62, 74] have demonstrated impressive results in generating photo-realistic videos with consistent geometry and plausible motion patterns. Building on the success of these 2D video generation models, a series of works [25, 61] have adapted the latent video diffusion model as the 3D generator to generate novel views of static objects from different viewpoints by fine-tuning it on 3D data. Other works insert additional modules to enable camera control of video diffusion models [2, 4, 24, 66, 71]. Most recent works [29, 67, 70, 78] extend this capability into the 4D domain, which leverages the pre-trained video generation model for 4D generation by incorporating an additional view attention layer to align multi-view images.

Diffusion-Based 4D Generation. Recent diffusion-based 4D generation methods have demonstrated significant advancements in achieving spatio-temporal consistency and motion fidelity. Existing optimization-based approaches [3, 8, 18, 27, 31, 43, 48, 69, 76, 83, 84] leverage pre-trained text-to-image [45], text-to-video [47], and image-to-3D [32] diffusion models to distill a unified 4D representation (deformable neural 3D representation) [20, 34] via score distillation [41, 65]. In contrast, photogrammetry-based methods [36, 55, 70, 73, 75] directly generate multi-view videos of a 4D object and use them as supervision for subsequent 4D reconstruction. Despite these advancements, current 4D generation approaches still focus on per-motion optimization from a single prompt and fail to generate diverse motion patterns during a single inference stage.

3D Motion Modeling and Generation. Modeling the motion patterns of dynamic objects is crucial for behavior analysis and content generation. Recent works [30, 35, 39, 40, 50, 58, 63, 79, 80, 85] have explored learning generative models for 3D human motions, leveraging parametric human shape models such as SMPL [33]. While some studies have concentrated on modeling animal motions using keypoint tracking-based methods [19, 52, 53], others have learned articulated [17, 28, 54, 68] and animatable [46, 72] 3D animals with generative motion templates. However, 3D motions generated by these models are often restricted to a *specific* category or skeleton structure, relying on category-specific template models or requiring extensive annotated data. In contrast, our work aims to model and generate 3D motions for any *general* objects without any pre-defined template model or human-annotated data.

3. Method

Overview. Given a single-view image of a *general* object, our goal is to model and generate its *diverse* 3D motions. The core idea is to distill the rich motion priors from well-trained video models and embed them into a shared latent space. As shown in Fig. 2, we first distill diverse motion patterns for the target object from video models (Sec. 3.1). Next, we introduce a motion latent space to jointly model the underlying motion distributions (Sec. 3.2). To ensure efficient and robust training, we adopt a coarse-to-fine optimization strategy to jointly learn the diverse motion space

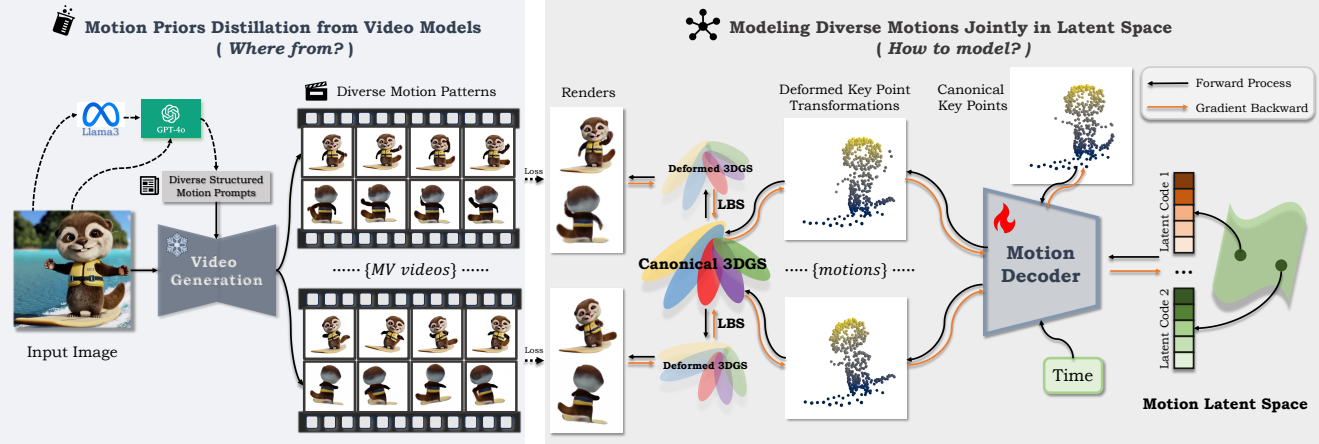

Figure 2. **Pipeline Overview.** Given a single-view image of any general object, DIMO first *distills rich motion priors* from video models (Sec. 3.1). We then represent each motion as structured neural key point trajectories (Sec. 3.2.1). During training, we embed each motion sequence into a latent code in motion latent space and *jointly model diverse motion patterns* using a shared motion decoder (Sec. 3.2.2). The decoded key point transformations are used to drive canonical 3DGS for 4D optimization with only photometric losses (Sec. 3.2.3).

and complex object geometry (Sec. 3.3).

3.1. Motion Priors Distillation from Video Models

A fundamental challenge in diverse 3D motion generation for general objects is the limited motion data from labor-intensive and time-consuming capture [33]. However, current well-trained video models [7, 74] have already encoded diverse and plausible motion patterns learned from extensive Internet-scale data. With this insight, we first distill rich motion priors for the target object from these video models.

Sourcing Diverse Motion Prompts from LLMs. To generate videos with diverse motions, we first need a collection of motion prompts based on the reference image which embeds a partial knowledge about the object motion we intend to explore. Since this information is incomplete, we rely on prior knowledge within the multi-modal large language model GPT-4o [1] to generate diverse motion prompts with an auto-prompting technique. To align all motions with the same initial object geometry state, we first employ a fine-tuned Llama3 [59] to create a structured "meta" prompt which contains *a detailed description of appearance, expression, and an initial action state of the object*. Then we feed the reference image and the "meta" prompt to GPT-4o and ask for the description of subsequent potential motions.

Automatic Motion-Rich Video Generation. With motion prompts and reference image based appearance prompts, we use text-conditioned image-to-video model [74] to generate motion-rich videos. We use an MLLM [12] to automatically assess videos and filter out low-quality ones based on a predefined score threshold. We also remove those with minimal or excessive motion measured by flow magnitude [57].

Novel View Priors Distillation. To enhance the reconstruction generalization, we use multi-view video model [61, 70] to obtain geometric priors by generating object novel views.

3.2. Modeling Diverse Motions in Latent Space

To jointly model diverse motion patterns within a single generative model, we must have a proper motion representation to effectively model the underlying distributions. Following this, we first model each motion sequence with explicit, compact, and structured key point trajectories. Then we embed motions into the latent space and train a latent-conditioned motion decoder to jointly learn diverse motion distributions. We further attach canonical Gaussians [20] to capture geometry and fuse them for 4D optimization.

3.2.1. Key Point Trajectories as Motion Representation

Although the geometry and appearance of dynamic objects are often complex and include high-frequency details, the underlying motion that drives these geometries is usually compact (low-rank) and smooth. Inspired by prior works on the Motion Factorization [15, 23, 26, 51, 64, 82], we propose to factorize each motion into sparse neural trajectories of key points $\mathcal{P} = \left\{ \left(p_k \in \mathbb{R}^3, r_k \in \mathbb{R}^+ \right) \right\}_{k=1}^{N_k}$ in canonical space, which act as a low-rank motion basis. Specifically, each key point k is parametrized by a canonical position p_k and a global control radius r_k, which parameterizes a radial basis function (RBF) that describes its influence weight $w_{jk} \in \mathbb{R}^+$ on the nearby point p_j:

$$w_{jk} = \text{Normalize}\left(\exp\left(-\left\| p_j - p_k \right\|_2^2 / 2 r_k \right) \right). \quad (1)$$

With the low-rank motion basis $\mathcal{P}$, then each motion sequence can be parametrized as neural curves formed by the key points' 6DoF transformations $\mathcal{E} = \{\mathcal{E}_k^t\}_{t=1}^T$, $k \in \{1, \cdots, N_k\}$, where each pose $\mathcal{E}_k^t \in \mathbf{SE}(3)$ at a timestamp t consists of a 3DoF rigid translation $\mathbf{T}_k^t \in \mathbb{R}^3$ and 3DoF rotation quaternion $\mathbf{R}_k^t \in \mathbf{SO}(3)$ of each key point k. Formally, each key point based neural curve $\mathbf{c}_k$ is defined as:

$$\mathbf{c}_k = \left(\left[\mathcal{E}_k^1, \mathcal{E}_k^2, \mathcal{E}_k^3, \ldots, \mathcal{E}_k^T \right], r_k \right), \quad (2)$$

To maintain the correct topology correlation, we construct the motion graph that connects key points based on their neural trajectories. We define the graph edge Ω_k as:

$$\begin{aligned} \Omega_k &= \mathrm{KNN}_{j\in\{1,\dots,N_k\}}\left[d(p_j^{traj}, p_k^{traj})/T\right], \\ p_k^{traj} &= \mathrm{concat}(p_k^1, p_k^2, \cdots, p_k^T) \end{aligned} \tag{3}$$

where $d(\cdot,\cdot)$ is the Euclidean distance function, T defines the motion sequence length and KNN denotes the K-nearest neighbors under the distance between two trajectories that capture the global topological changes across time.

3.2.2. Joint Learning of Diverse Motion Distributions

To model the time-varying deformation, previous 4D reconstruction and generation works train a specific deformation network to overfit a single motion, which is time-consuming and not generalizable to diverse motion joint modeling. Most importantly, it doesn't leverage the common motion patterns of the target object. Therefore, to jointly learn diverse motion distributions, we embed a wide variety of motions of the same object into a low-dimensional latent space with latent code $\boldsymbol{z}$ and train a shared deformation network.

Formally, for the motion indexed by $m \in \{1, \cdots, N_m\}$, we employ a latent-conditioned motion decoder $\mathcal{D}_c$ to condition on its latent code $\boldsymbol{z}_m$ and query canonical location p_k^t of each key point k with timestep t. The motion decoder $\mathcal{D}_c$ then outputs key point motion-specific 6DoF transformation $(\mathcal{E}_k^t)_m = (\mathbf{R}_k^t \mid \mathbf{T}_k^t)_m \in \mathbf{SE}(3)$:

$$(\mathcal{E}_k^t)_m = \mathcal{D}_c\left(\boldsymbol{z}_m, p_k^t, t\right), \tag{4}$$

In motion latent space, we assume the prior distribution over each latent vector $\mathrm{p}\left(\boldsymbol{z}_m\right)$ to be a non-zero-mean multivariate Gaussian as $\boldsymbol{z}_m \sim \mathcal{N}\left(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m\right)$ with *learnable* parameters $(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$. Then motion-associated vector $\boldsymbol{z}_m$ can be sampled using VAE reparameterization trick [21]:

$$\boldsymbol{z}_m = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}) \tag{5}$$

The learnable distribution parameters $(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$ are learned directly through standard back-propagation.

3.2.3. Geometric Modeling with Canonical 3DGS

The compact neural trajectories and latent space have effectively modeled the various underlying motions across time. We further employ a set of canonical 3D Gaussians [20] to capture the object geometry and appearance. Formally, an object is parameterized as Gaussians in canonical space:

$$\mathcal{G} = \{(\mu_i, R_i, s_i, o_i, c_i)\}_{i=1}^{N_g} \tag{6}$$

where $\mu_i, R_i, s_i, o_i, c_i$ are the center, rotation, scale, opacity and spherical harmonics (SH) of the Gaussian primitive $i \in \{1, \cdots, N_g\} (N_k \ll N_g)$. To get the geometry at the timestep t, canonical Gaussians will be deformed by the nearest canonical key point p_k and its graph edges Ω_k using widely-used linear blend skinning (LBS) [51] with weighting factor $w \in \mathbb{R}^+$ calculated in Eq. (1).

$$\begin{aligned} \mathcal{G}(t) &= \left\{\left(\mu_i^t, R_i^t, s_i, o_i, c_i\right)\right\}_{i=1}^{N_g} \\ (\mu_i^t, R_i^t) &= \mathrm{LBS}\left(\left\{w_{ij}, \mathbf{R}_j^t, \mathbf{T}_j^t\right\}_{j\in\Omega_k}\right)(\mu_i, R_i) \end{aligned} \tag{7}$$

The deformed 3D Gaussians $\mathcal{G}(t)$ are fused and then rendered via a splatting-based differentiable rasterization [20].

3.3. Motion-Oriented Optimization

The total learnable parameters include diverse latent codes $\boldsymbol{z}$, canonical key points $\mathcal{P}$, canonical 3D Gaussians $\mathcal{G}$ and a motion decoder $\mathcal{D}_c$. For training efficiency and stability, we adopt a coarse-to-fine motion pre-training schedule.

Rendering-Based Photometric Optimization. We randomly sample the latent code $\boldsymbol{z}_m$ and infer the deformed $\mathcal{G}(t)$ for target motion m at timestamp t. Then, we render frames at training viewpoints and compare them with the video supervision using RGB loss $\mathcal{L}_{\mathrm{rgb}}$, mask loss $\mathcal{L}_{\mathrm{mask}}$ and perceptual LPIPS loss $\mathcal{L}_{\mathrm{lpips}}$. To encourage smooth 3D surfaces, we also involve edge-aware depth smoothness loss $\mathcal{L}_{\mathrm{depth}}$ and bilateral normal smoothness loss $\mathcal{L}_{\mathrm{normal}}$ [61].

$$\mathcal{L}_{\mathrm{photo}} = \mathcal{L}_{\mathrm{rgb}} + \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{lpips}} + \mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{normal}} \tag{8}$$

ARAP-Based Geometric Optimization. To regularize the 3D motion of unconstrained key points, we optimize an *As-Rigid-As-Possible* (ARAP) [49] loss $\mathcal{L}_{\mathrm{arap}}$ to encourage the pairs of nearby key points (p_j within the graph edges Ω_k of p_k) to be *locally* rigid over time. Formally, given two timestamps separated by interval Δt, we define $\mathcal{L}_{\mathrm{arap}}$ as:

$$\mathcal{L}_{\mathrm{arap}} = \sum_{k=1}^{N_k}\sum_{t=1}^{T}\sum_{j\in\Omega_k} w_{jk}\left\|d\left(p_j^t, p_k^t\right) - d\left(p_j^{t+\Delta t}, p_k^{t+\Delta t}\right)\right\|_1 \tag{9}$$

where w_{jk} is an RBF skinning weight calculated in Eq. (1).

Latent Distribution Regularization. We regularize the latent distribution by minimizing the Kullback–Leibler (KL) divergence $\mathcal{L}_{\mathrm{KL}}$ between the learned motion distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$ and a standard Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{m}^{N_m} -\frac{1}{2}\left(\log\boldsymbol{\sigma}_m - \boldsymbol{\sigma}_m - \boldsymbol{\mu}_m^2 + 1\right) \tag{10}$$

Coarse-to-Fine Motion Pre-training Schedule. To ensure robust and efficient training, we disentangle the 3D motion and geometry, and adopt a two-stage coarse-to-fine training schedule. In the *first* stage, we obtain a *coarse* motion basis and latent space by pre-training the latent codes $\boldsymbol{z}$, canonical key points $\mathcal{P}$, and motion decoder $\mathcal{D}_c$. Specifically, we initialize N_k key points as 3D Gaussians within a

Figure 3. **Qualitative Results.** During inference, DIMO can ***instantly*** generate ***diverse*** 3D motions and photorealistic 4D contents in ***a single forward pass*** by sampling from latent space. We render three motions for each case under two views at two novel timestamps.

sphere. During optimization, densification and pruning are performed following 3DGS [20]. To avoid the local motion minima, every certain iteration we downsample the key points to N_k using FPS [42] as an annealing process. After this stage, we obtain a desirable key point distribution in canonical space, which acts as the motion basis for subsequent deformation. More importantly, the model now has learned a reasonable latent space for all motion patterns, on top of which joint optimization of geometry and 3D motion is much more efficient and robust.

In the *second* stage, we further incorporate the canonical 3D Gaussians $\mathcal{G}$ to capture the object geometry and jointly optimize all parameters for *fine-grained* motions. To inherit the canonical shape (distribution) modeled in the first stage, we randomly initialize n_g canonical Gaussians within a small sphere around each canonical key point with radius r_k following [43, 56, 69]. To improve training stability, we recycle the key point trajectories $\mathcal{E}$ from the first stage to guide the prediction $\hat{\mathcal{E}}$ in the second stage using a Chamfer Distance regularization defined as $\mathcal{L}_{\text{chamfer}} = \text{CD}\left(\hat{\mathcal{E}}_t, \mathcal{E}_t\right)$. For training efficiency and stability, we gradually increase the rendering resolution from 128×128 to 512×512. At each iteration, we randomly sample 4 motions × 3 views × 2 frames within a batch, providing multi-motion, multi-view, and multi-frame constraints to guide gradient optimization.

Language-Guided Motion Generation. Since the videos are generated from text prompts, the learned motion latent space should also be compatible with language embeddings. To establish the relationship between natural language and

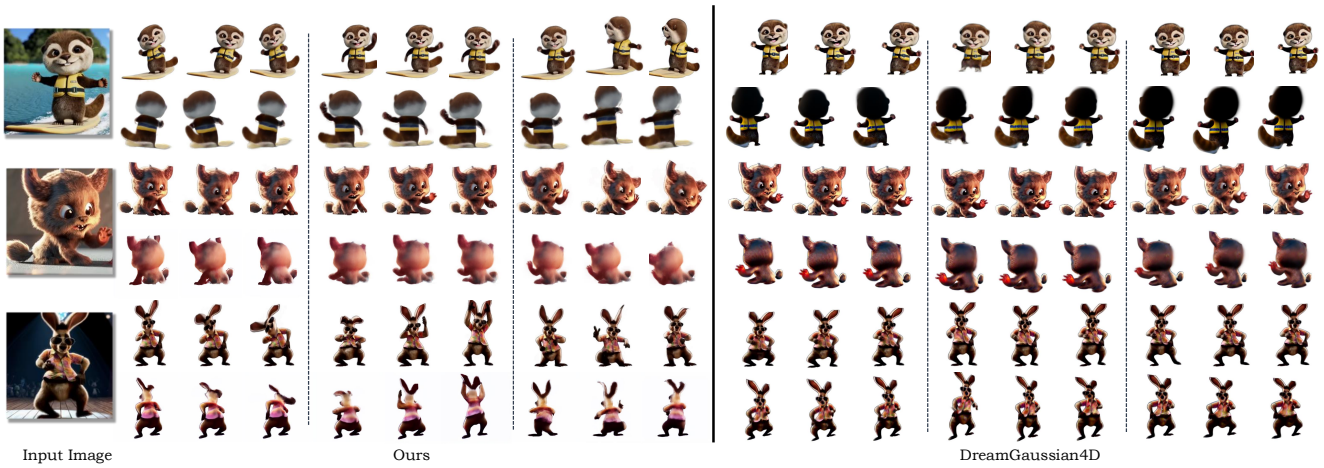

Figure 4. **Visual Comparison on 3D Motion Generation.** DIMO can generate diverse and high-fidelity 3D motions, whereas the baseline fails to produce noticeable motions (DG4D-generated kangaroo is a thin slice so the back side appears as a mirror image of the front).

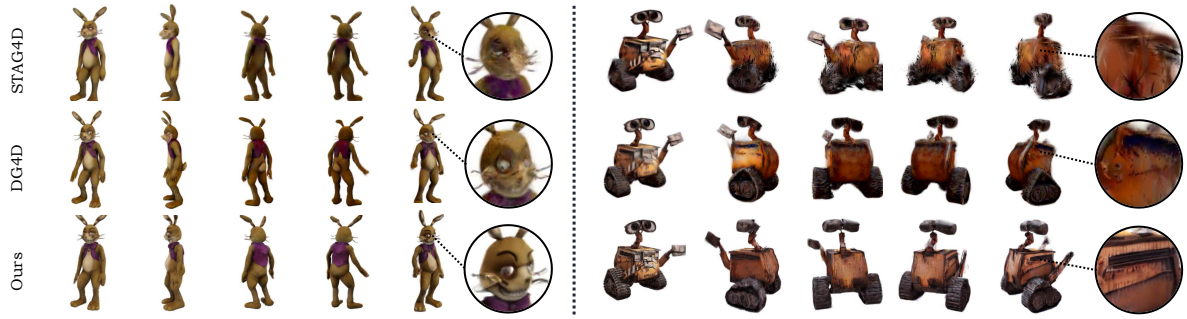

Figure 5. **Visual Comparison on Image-to-4D.** DIMO can generate high-quality 4D contents for both synthetic and in-the-wild objects.

motion space, we first encode the text prompts into text embeddings with pre-trained BERT [10] encoder and then train a lightweight MLP to project the text embedding into the motion latent code. After this, users could directly provide novel text prompts and our model can generate the corresponding motion and 4D content in a feed-forward manner, as indicated in Fig. 7. Note that to allow people to use simple prompts during inference, we use ChatGPT [1] to summarize the detailed video caption we use for video generation into a simple motion description such as "lift the right hand" before feeding them into BERT.

4. Experiments

We conducted experiments in standard settings and benchmarks to demonstrate the effectiveness of DIMO in generating diverse and high-fidelity 3D motions and 4D contents. Furthermore, we highlight our applications in motion interpolation, language-guided motion generation, and test motion reconstruction by learning a motion latent space.

4.1. Experimental Setup

Dataset and Object Diversity. We use Objaverse [9], Animate124 [83], Consistent4D [18], DAVIS [38], SORA [7]-generated and self-collected datasets for experiments. For object diversity, we select a wide range of representative *synthetic* and *real-world* objects including humans, robots, bipods, quadrupeds, birds, plants, fluid and deformable objects with a total of 58 species, each with over 50 motions.

Implementation Details. We leverage Llama3 [59] and GPT4o [16] to generate $N_m \geq 50$ diverse motion prompts for each object and adopt CogVideoX5B-I2V [74] to generate motion-rich videos. During training, we use 8 views $\times$ 21 frames for each motion sequence and set the virtual camera FOV to 33.9° with a fixed radius of $2m$. We use $N_k = 512$ key points as the motion basis shared by all motion patterns. For the motion latent space, we employ the Gaussian distribution parameters $(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$ as learnable latent variables for each motion m, with a latent dimension of 32. The motion decoder $\mathcal{D}_c$ is implemented as an 8-layer MLP with a skip connection at the 4th layer. The MLP receives motion latent codes, the sinusoidal positional embeddings [34] of time, and key points' canonical positions as input, predicting key points' 6DoF transformations. We train a separate generative model for each object. For the 50-motion joint training setting, we pretrain the motions for 2.8k steps in the first stage, which takes roughly 40 minutes on a single 40GB A100 GPU, and for another 8k steps in the second stage for joint optimization of geometry and motion with an additional 3 hours. The rendering speed is 250 FPS at 512×512 resolution with ∼80k canonical 3DGS.

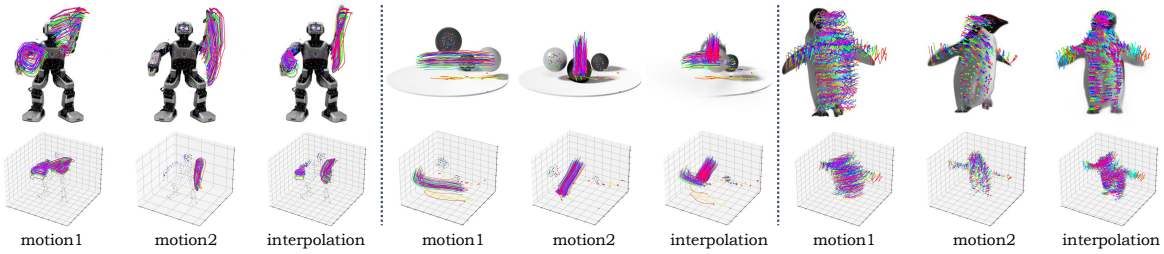

Figure 6. **3D Motion Interpolation.** We generate new motion by linearly interpolating between two motions sampled from latent space.

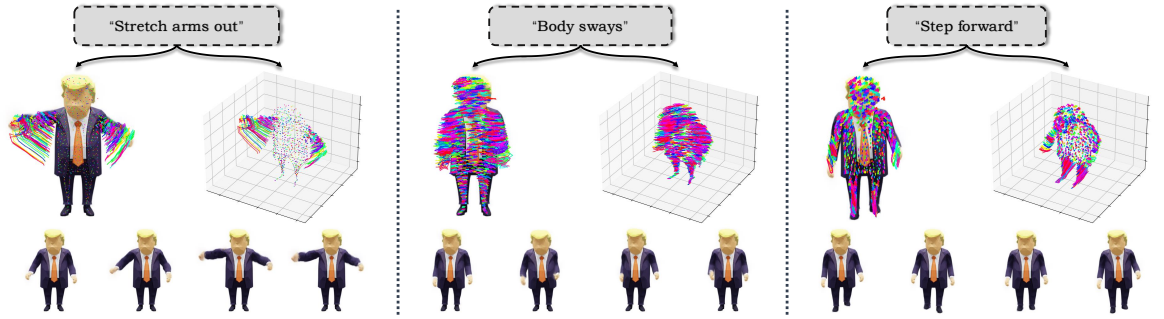

Figure 7. **Language-Guided Motion Generation.** We project language into latent code and enable feed-forward 3D motion generation.

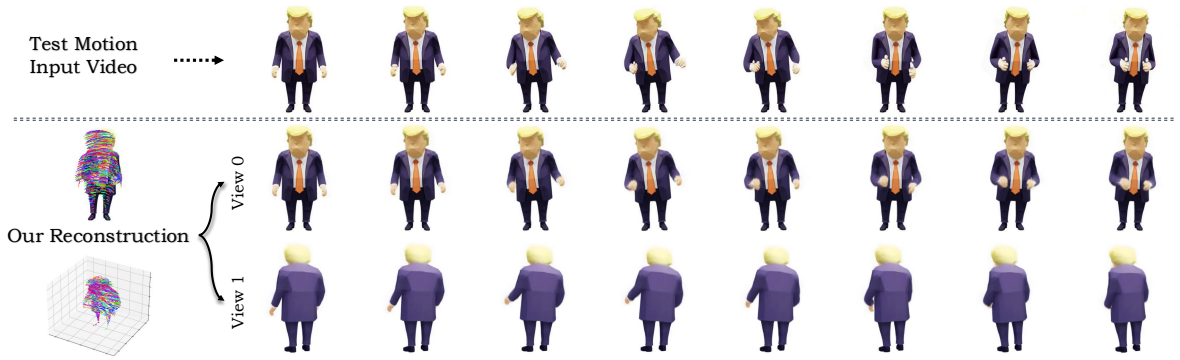

Figure 8. **Test Motion Reconstruction.** The first row is the input video of the test motion, and the rest are our reconstruction results.

4.2. Comparisons

Comparison on 3D Motion Generation. Our method is the first work to generate diverse 3D motions for any general dynamic objects. To evaluate the diversity and quality of our generated 3D motions, we compare our DIMO with DG4D [43] and 4DGen [76]. We repetitively run their video diffusion models 50 times and perform 4D optimization separately. Then we randomly sample a subset for evaluation. Specifically, we select 7 cases including Human, Cat, Bird, Robot, and SORA [7]-generated Kangaroo, Otter, Monster. Each object has five motions. While baselines train each motion sequence separately, our DIMO jointly models all motions in a single generative model. A qualitative comparison of 3D motion generation is shown in Fig. 4. Following [43, 76, 83], we conduct a user study with 35 participants who are asked to evaluate based on four criteria: motion diversity (MD), image alignment (IA), motion quality (MQ), and 3D appearance (3D App.). The results in Tab. 1 indicate that our DIMO has a clear advantage over all baselines across all criteria, achieving better 3D motion diversity and visual fidelity.

Table 1. Quantitative comparison on 3D Motion Generation.

Method	MD↑	IA↑	MQ↑	3D App.↑
DG4D [43]	11.4%	17.2%	8.6%	14.3%
4DGen [76]	8.6%	20.0%	22.8%	25.7%
Ours	**80.0%**	**62.8%**	**68.6%**	**60.0%**

Comparison on Image-to-4D. To quantitatively evaluate our generation results, we compare our method with Animate124 [83], 4DGen [76], STAG4D [77], DG4D [43] and SV4D [70] on Animate124 and Objaverse dataset. Following [43, 76, 83], we quantify image alignment using CLIP-I and temporal coherence using CLIP-F. We conduct the user study to measure the motion diversity. The qualitative com-

parisons are presented in Fig. 5. The quantitative comparison results in Tab. 2 show that our method outperforms the baselines by a significant margin, demonstrating superior performance in visual quality and temporal consistency.

Table 2. Quantitative comparison on Image-to-4D generation.

Method	CLIP-I ↑	CLIP-F ↑	MD ↑	Training Time ↓
Animate124 [83]	0.8596	0.9756	8.1%	4.5h
4DGen [76]	0.9011	0.9854	5.4%	1h
STAG4D [77]	0.9281	0.9868	8.1%	1.5h
DG4D [43]	0.9214	0.9883	10.2%	10min
SV4D [70]	0.9372	0.9904	8.1%	15min
Ours	**0.9505**	**0.9912**	**59.4%**	10min

Comparison on Video-to-4D. We also evaluate DIMO on the widely-used Consistent4D [18] benchmark to measure the 4D generation quality and temporal consistency. As reported in Tab. 3, we achieve superior LPIPS [81], FVD [60], and very competitive CLIP results, showing the faithful consistency and visual details of our generated 4D contents.

Table 3. Quantitative comparison on Video-to-4D generation.

Method	LPIPS ↓	CLIP ↑	FVD ↓
Consistent4D [18]	0.160	0.87	1133.93
STAG4D [77]	0.126	0.91	992.21
DreamGaussian4D [43]	0.160	0.87	-
4DGen [76]	0.160	0.87	-
SV4D [70]	0.118	**0.92**	732.40
Ours	**0.112**	**0.92**	**625.30**

Comparison on Text-to-4D. To evaluate our language-guided motion generation results, we compare with text-to-4D baselines Animate124 [83] and 4D-fy [3] using six Animate124 examples. Specifically, we provide each model with the same 10 prompts generated by GPT4 [1]. It takes 5 hours to generate one Animate124 and 12 hours for one 4D-fy instance, whereas DIMO generates text-guided 3D motions *within seconds* in a single forward pass. We compute the average of the largest 20% optical flows [57] within the instance mask as motion amplitude and conduct a user study to measure the text alignment & motion diversity. As reported in Tab. 4, DIMO can efficiently generate more diverse, noticeable, and high-quality text-guided 3D motions.

Table 4. Quantitative comparison on Text-to-4D generation.

Method	Motion Amplitude ↑	Text Alignment ↑	Motion Diversity ↑
Animate124 [83]	0.428	6.3%	9.3%
4D-fy [3]	0.254	12.5%	15.6%
Ours	**4.319**	**81.3%**	**75.0%**

4.3. Applications

Latent Space Motion Interpolation. To demonstrate the completeness and continuity of our learned motion embedding, we visualize the decoder's output when interpolating between pairs of motions in the latent space, as shown in Fig. 6. The results indicate that the embedded motion latent codes capture meaningful and common motion patterns of the object, which can be effectively linearly interpolated to generate novel, coherent 3D motions.

Language-Guided Motion Generation. Text-to-Motion offers a more user-friendly approach to interactive motion generation. We adopt GPT to generate 300 motion prompts and cluster them into 60 (50 for training and 10 for evaluation) in the BERT embedding space. We project language into the motion latent code and then generate plausible motion sequences in a feed-forward manner, as shown in Fig. 7.

Test Motions Reconstruction. For encoding unseen motions, i.e., those in the held-out test set, DIMO demonstrates strong performance in reconstruction quality and motion alignment, as shown in Fig. 8. Given the multi-view videos from the test set, we optimize the latent code, initialized as a standard Gaussian $\mathcal{N}(0, \boldsymbol{I})$ from scratch, while keeping all other parameters fixed. We only use the reconstruction loss for fine-tuning, achieving convergence within 300 steps.

4.4. Ablation Study

We validate the effectiveness of our model's design choices in Tab. 5. We observe that both motion representation and motion-oriented optimization are critical. Neural key point-based *motion factorization* contributes to effective motion distribution learning and improves the expressiveness of our system. Furthermore, our DIMO leverages the *latent space* to distinguish diverse motion patterns and model the underlying motion distributions. We also verify the effectiveness of *motion pre-training* in achieving robust and precise motion reconstruction. *Multi-motion joint training* within a single generative model forces the network to capture a *shared* 4D geometric structure and learn a *smooth, continuous* motion space. Thus, the generated 3D motions and 4D contents are robust and less sensitive to high-frequency errors or appearance inconsistency of video supervision, compared with single-motion overfitting. More details and qualitative evidence can be found in the supplementary material.

Table 5. Ablation results on different components of DIMO.

Method	LPIPS ↓	CLIP ↑	FVD ↓
Full Model	**0.126**	**0.93**	**587.09**
w/o motion factorization	0.134	0.91	851.83
w/o latent space	0.163	0.87	1077.19
w/o motion pre-training	0.149	0.90	890.26
w/o multi-motion joint training	0.131	0.92	693.49

5. Limitations & Conclusion

Limitations and Future Direction. DIMO relies on video models for object motion prior distillation, indicating that improvements of these models are critical for enhancing our performance. Also, we currently learn language-guided motion generation in two stages by optimizing latent codes and the language projector separately. Jointly learning these two objectives within a single stage will be a key direction.

Conclusion. This paper takes the first step toward diverse 3D motion generation for general objects by learning a motion latent space. We achieve state-of-the-art performance in a wide range of standard settings and benchmarks. We hope this work could help us better understand numerous dynamic objects in our physical world and inspire future research in building a general SMPL-like parametric model.

Acknowledgement

We gratefully acknowledge support by the following grants: NSF FRR 2220868, NSF IIS-RI 2212433, ARO MURI W911NF-20-1-0080, and ONR N00014-22-1-2677.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 6, 8

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024. 2

[3] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 1, 2, 8

[4] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 2

[5] 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. 1, 2

[6] 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.

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 1, 2, 3, 6, 7

[8] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 2

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 6

[10] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

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

[12] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 3

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[14] 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

[15] 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. 3

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[17] Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3d: Learning articulated 3d animals by distilling 2d diffusion. In *2024 International Conference on 3D Vision (3DV)*, pages 852–861. IEEE, 2024. 2

[18] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 1, 2, 6, 8

[19] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1965–1974, 2024. 2

[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, 2, 3, 4, 5

[21] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[23] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. *arXiv preprint arXiv:2312.00112*, 2023. 3

[24] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *arXiv preprint arXiv:2405.17414*, 2024. 2

[25] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024. 2

[26] 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. 3

[27] Zhiqi Li, Yiming Chen, and Peidong Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *arXiv preprint arXiv:2410.06756*, 2024. 1, 2

[28] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9752–9762, 2024. 2

[29] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 2

[30] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2

[31] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 1, 2

[32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1, 2, 3

[34] 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, 2, 6

[35] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[36] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024. 1, 2

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[38] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6

[39] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[40] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[42] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[43] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 1, 2, 5, 7, 8

[44] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. L4gm: Large 4d gaussian reconstruction model. *arXiv preprint arXiv:2406.10324*, 2024. 1

[45] 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

[46] Remy Sabathier, Niloy J Mitra, and David Novotny. Animal avatars: Reconstructing animatable 3d animals from casual videos. In *European Conference on Computer Vision*, pages 270–287. Springer, 2024. 2

[47] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[48] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1, 2

[49] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, pages 109–116. Citeseer, 2007. 4

[50] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2

[51] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 3, 4

[52] Jennifer J Sun, Serim Ryou, Roni H Goldshmid, Brandon Weissbourd, John O Dabiri, David J Anderson, Ann Kennedy, Yisong Yue, and Pietro Perona. Self-supervised keypoint discovery in behavioral videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2171–2180, 2022. 2

[53] Jennifer J Sun, Lili Karashchuk, Amil Dravid, Serim Ryou, Sonia Fereidooni, John C Tuthill, Aggelos Katsaggelos, Bingni W Brunton, Georgia Gkioxari, Ann Kennedy, et al. Bkind-3d: self-supervised 3d keypoint discovery from multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9001–9010, 2023. 2

[54] Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In *European Conference on Computer Vision*, pages 100–119. Springer, 2025. 2

[55] Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. Eg4d: Explicit generation of 4d object without score distillation. *arXiv preprint arXiv:2405.18132*, 2024. 1, 2

[56] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 5

[57] 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. 3, 8

[58] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3, 6

[60] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 8

[61] 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. *arXiv preprint arXiv:2403.12008*, 2024. 1, 2, 3, 4

[62] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[63] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 2

[64] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 3

[65] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[66] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 2

[67] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024. 2

[68] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 2

[69] Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. *arXiv preprint arXiv:2404.03736*, 2024. 1, 2, 5

[70] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 1, 2, 3, 7, 8

[71] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2

[72] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16995–17005, 2023. 2

[73] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion ²: Dynamic 3d content generation via score composition of orthogonal diffusion models. *arXiv preprint arXiv:2404.02148*, 2024. 1, 2

[74] 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. 1, 2, 3, 6

[75] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396*, 2025. 1, 2

[76] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 1, 2, 7, 8

[77] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024. 7, 8

[78] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *arXiv preprint arXiv:2405.20674*, 2024. 1, 2

[79] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[80] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8

[82] Boming Zhao, Yuan Li, Ziyu Sun, Lin Zeng, Yujun Shen, Rui Ma, Yinda Zhang, Hujun Bao, and Zhaopeng Cui. Gaussianprediction: Dynamic 3d gaussian prediction for motion extrapolation and free view synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[83] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. 1, 2, 6, 7, 8

[84] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 1, 2

[85] Wenyang Zhou, Zhiyang Dou#, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu#. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, Cham, 2024. 2