Articulated Kinematics Distillation from Video Diffusion Models

https://research.nvidia.com/labs/dir/akd/

Xuan Li[1,2,*] Qianli Ma[2] Tsung-Yi Lin[2] Yongxin Chen[2]
Chenfanfu Jiang[1] Ming-Yu Liu[2] Donglai Xiang[2]
[1] UCLA, [2] NVIDIA

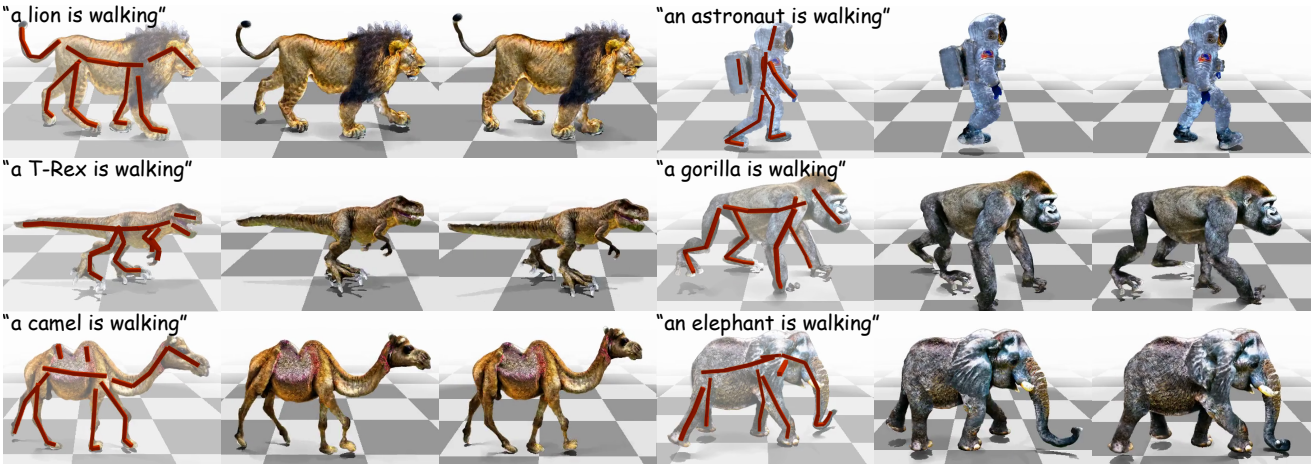

Figure 1. By incorporating articulation into static assets, AKD synthesizes realistic motions distilled from large video diffusion models.

Abstract

We present Articulated Kinematics Distillation (AKD), a framework for generating high-fidelity character animations by merging the strengths of skeleton-based animation and modern generative models. AKD uses a skeleton-based representation for rigged 3D assets, drastically reducing the Degrees of Freedom (DoFs) by focusing on joint-level control, which allows for efficient, consistent motion synthesis. Through Score Distillation Sampling (SDS) with pre-trained video diffusion models, AKD distills complex, articulated motions while maintaining structural integrity, overcoming challenges faced by 4D neural deformation fields in preserving shape consistency. This approach is naturally compatible with physics-based simulation, ensuring physically plausible interactions. Experiments show that AKD achieves superior 3D consistency and motion quality compared with existing works on text-to-4D generation.

1. Introduction

In traditional 3D graphics, a skeleton-based character animation pipeline involves steps like shape modeling, rigging, motion capture, motion retargeting, and editing. As a mature technology, such pipelines can achieve high realism and good controllability over the motion, but they typically require extensive manual work from digital artists, making the process time-consuming and thus hardly scalable. Recent advances in video generation models [6, 59] offer a promising avenue for streamlining the animation authoring process: with a text-to-video model, generating a sequence of character animation only requires a text prompt. However, existing video generation models still struggle to generate high-fidelity dynamics for real-world objects because of a lack of 3D information. Common issues include failing to preserve the 3D structure consistency (e.g. number of limbs of a character) during animation, or producing physically implausible articulated motion, such as foot-skating and ground penetration.

Recent works on text-to-4D generation [1, 2] leverage these video generation models to distill the learned dynamic motion into consistent 4D sequences. These frameworks commonly rely on neural deformation fields which predict displacements at each location in a pre-defined 3D volume to deform a 3D shape. Animation is thus a temporal sequence of such deformed shapes. While flexible, this approach introduces a large number of Degrees of Freedom

* Work done during an internship at NVIDIA.

(DoFs), making optimization challenging and often resulting in suboptimal quality. Is it possible to have the best of both worlds, where generative models provide extensive knowledge of diverse motions from internet-scale data, while skeleton-based 3D animation allows low-DoF control, permanence of articulated structures, and even physical grounding via simulation?

To answer this question, we introduce Articulated Kinematics Distillation (AKD), a motion synthesis system that bridges traditional character animation pipelines and generative motion synthesis. Given a rigged 3D asset, we distill articulated motion sequences from a pre-trained video diffusion model using Score Distillation Sampling (SDS). The skeleton-based representation simplifies the distillation process by limiting the number of DoFs to that of a few joints, in contrast to all query points in space-time as in the text-to-4D works [2]. It also offers an effective regularization of the deformation space, enabling the distillation to concentrate on overall motion styles without worrying about maintaining reasonable local structures. More importantly, the skeleton-based formulation is naturally compatible with physics-based simulation, allowing the generated motion to be grounded by physics-based motion tracking to ensure physical plausibility.

Experiments verify the effectiveness of our design: compared to previous text-to-4D methods, our framework produces results with better 3D shape consistency and more expressive motions. We summarize our contributions:

- We introduce a novel text-driven motion synthesis framework for static 3D assets, combining articulated rigging systems and large video diffusion models.
- We demonstrate that incorporating non-uniform ground renderings enhances the video model's adherence to basic physics between the character and the ground.
- Extensive experiments show that our generated motions exhibit higher quality than the state-of-art methods that can synthesize long-trajectory motions.
- Our generated motion can be used in physics-based motion tracking with differentiable physics to further boost its physical realism.

2. Related Work

Deformable Gaussian Splatting In recent years, different types of 3D representations have been introduced to facilitate the reconstruction and generation of 3D/4D scenes, such as neural fields [32], iNGP [34], and 3D Gaussian Splatting [21]. Among these representations, 3D Gaussian Splatting is particularly suitable for representing dynamic scenes due to its explicit nature [52] compared to the NeRF based on neural implicit fields [35, 38], whose deformations are achieved by bending rendering rays [9, 37]. This advantage of 3DGS has sparked a lot of works on 4D reconstruction and modeling from multi-view input, including general scenes [62], facial avatars [10, 54], and full-body avatars [15, 22, 27, 33, 43, 67], where 3D Gaussian kernels are bound to an articulated human model SMPL [28] through learned skinning weight. We adopt GS as our 3D shape representation, which naturally allows SDS gradients to smoothly propagate through the articulated deformation and rendering pipeline.

Articulated Motion Reconstruction A closely related topic to our work is the reconstruction of articulated motions of deformable objects in under-constrained settings, especially from a monocular view. As a special case, the reconstruction of human body poses [11, 20, 51] benefits strongly from the availability of dedicated deformable models such as SMPL [28] and the abundance of 2D/3D pose data [25, 31]. In contrast, the reconstruction of general objects, such as animals and humans in loose clothing, has remained more challenging due to the inherent 3D ambiguity from a monocular view and a lack of reliable priors for their articulation. One line of works following BANMO [23, 46, 56–58] solve for a static 3D template and articulated motion simultaneously from a monocular video by leveraging various image measurements such as segmentation, optical flow, and DensePose. Several works from another line [16, 24, 49, 60, 61] train a neural network that predicts the shape template of animals and their body deformation conditioned on a single input image in a weakly-supervised manner. Ponymation [45] learns a motion VAE for horse motions from a collection of horse videos. There are also works [17, 39, 47, 63] that directly train generative models on articulated motions. Instead of extracting articulation from a particular input, our method focuses on generating novel skeleton motion utilizing video diffusion priors that have learned general visual knowledge.

Generating 4D and Physics-Based Dynamics Generating 4D content is inherently challenging as it demands high consistency not only along the temporal axis to maintain motion, but also across multiple viewpoints to ensure spatial and structural accuracy in the generated content. Some works attempt to directly build priors models in the 4D domain, including diffusion models [19, 53] and reconstruction model [41], where the limited availability of 4D data pose a challenge. Therefore, a large body of works [1, 18, 26, 40, 44, 48, 64, 66] distill 4D motion from a combination of generative models that operate in lower dimensions, including images, videos, and multi-view images (3D), which is a difficult problem to due to spasity of supervision, noisy gradient from SDS, and high-degree of freedom in the optimization variables. Some works aim to improve the controllability of the 4D generation by introducing conditioning of trajectory [50] or sparse control points [50]. None of the works above allow the generated contents to be grounded in a physics-based manner. The exceptions