# ReVision: High-Quality Video Generation with Explicit 3D Physics Modeling for Complex Motion and Interaction

**Anonymous authors**
**Paper under double-blind review**

## Abstract

In recent years, video generation has seen significant advancements. However, challenges still persist in generating complex motions and interactions. To address these challenges, we introduce ReVision, a plug-and-play framework that explicitly integrates parameterized 3D physical knowledge into a pretrained conditional video generation model, significantly enhancing its ability to generate high-quality videos with complex motion and interactions. Specifically, ReVision consists of three stages. First, a video diffusion model is used to generate a coarse video. Next, we extract a set of 2D and 3D features from the coarse video to construct a 3D object-centric representation, which is then refined by our proposed parameterized physical prior model to produce an accurate 3D motion sequence. Finally, this refined motion sequence is fed back into the same video diffusion model as additional conditioning, enabling the generation of motion-consistent videos, even in scenarios involving complex actions and interactions. We validate the effectiveness of our approach on Stable Video Diffusion, where ReVision significantly improves motion fidelity and coherence. Remarkably, with only 1.5B parameters, it even outperforms a state-of-the-art video generation model with over 13B parameters on complex video generation by a substantial margin. Our results suggest that, by incorporating 3D physical knowledge, even a relatively small video diffusion model can generate complex motions and interactions with greater realism and controllability, offering a promising solution for physically plausible video generation.

## 1 Introduction

Video diffusion models have achieved remarkable success in producing high-quality, temporally coherent videos (Blattmann et al., 2023; Brooks et al., 2024; Polyak et al., 2024; Kong et al., 2024). It has been driven by advances in model architectures (Peebles & Xie, 2023), increases in model complexity, reaching tens of billion parameters (Polyak et al., 2024), and the availability of large-scale high-quality datasets (Chen et al., 2024). However, current models still struggle to generate videos that adhere to realistic physical principles, making it difficult to consistently achieve fine-grained motion control, complex movements, and coherent object interactions. Despite extensive efforts to improve performance through larger models and higher-quality datasets, a recent study (Kang et al., 2024) indicates that scaling model size and data alone is insufficient to fully capture the complexities of the physical world.

On the other hand, human image animation models (Hu, 2024; Tan et al., 2024; Xu et al., 2024) offer valuable insights for addressing persistent challenges in video generation. Despite using smaller models and less data, these methods achieve consistent and precise video outputs with complex motions by following predefined 2D keypoint trajectories. This success suggests that incorporating a well-defined motion prior can substantially reduce the learning complexity of video generative models, enabling them to generate coherent and lifelike motion. However, in general video generation tasks, such strong guiding signals are typically unavailable, limiting the direct applicability of these animation techniques to broader video generation scenarios. This raises a critical question: *Can we develop a video generation framework that leverages the implicit motion information embedded in the generated videos as guidance to further enhance video quality?*

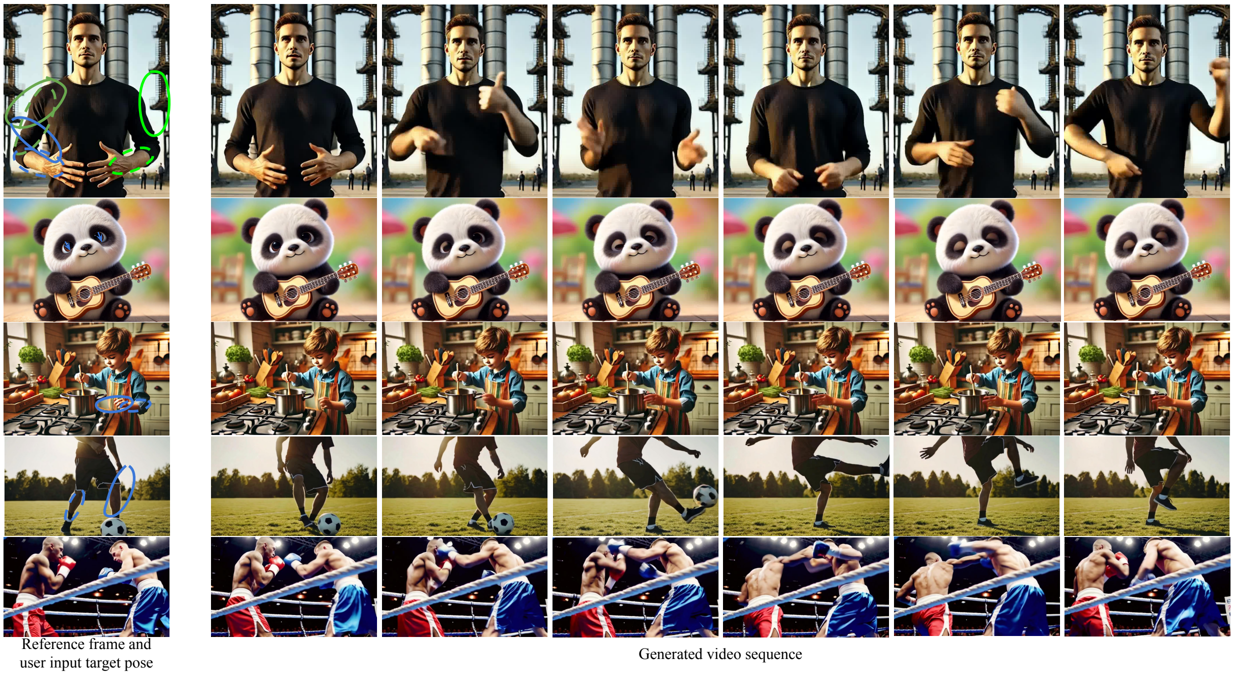

Figure 1: By explicitly leveraging a parameterized 3D physical model, ReVision enhances pre-trained video generation models (*e.g.*, Stable Video Diffusion) to produce high-quality videos with complex motion (row 1), enabling precise motion control (rows 2, 3) and accurate interactions (rows 4, 5). During inference, an *optional* target pose can be specified via a rough sketch (rows 1, 3, 4, colored circles for different parts, dashed lines for the original pose, solid lines for the target pose) or a simple drag operation (blue arrows in row 2) indicating the final position.

In this paper, we propose a simple, general, and plug-and-play video generation framework that incorporates physical knowledge into a conditional video generation model via a parameterized 3D representation, allowing the generation of videos with complex motions and interactions involving *humans*, *animals*, and *general objects*. The core of **ReVision** is to **Re**generate **Vi**deos with explicit 3D physic**s** representat**ion**s, following an *Extract–Optimize–Reinforce* pipeline. Specifically, to effectively leverage 3D knowledge without heavy retraining of the diffusion model while preserving its original ability to generate high-quality visual appearance, we design the pipeline in three stages.

In the first stage, we employ a video diffusion model, *e.g.*, SVD (Blattmann et al., 2023), to generate a coarse video conditioned on the given input. In the second stage, we utilize parametric 3D models (*i.e.*, SMPL-X (Pavlakos et al., 2019) for humans, SMAL (Rueegg et al., 2023; Zuffi et al., 2024) for animals, and 2D binary mask (Yu et al., 2023) with estimated depth (Yang et al., 2024b) for general objects) to *extract* 3D shape and motion features from the coarse video. These 3D representations are subsequently *optimized* by the proposed **P**arameterized **P**hysical **P**rior **M**odel (**PPPM**), producing a more accurate and natural 3D motion sequence. In the third stage, the refined 3D motion sequences are incorporated as additional conditioning inputs to *reinforce* the diffusion model, enabling it to regenerate the video with improved coherence and realism.

Extensive qualitative results and human preference studies confirm that our model excels at generating complex motions and interactions. We first apply ReVision on Stable Video Diffusion (SVD) (Blattmann et al., 2023), substantially improving its ability to generate realistic and intricate motions. We further compare ReVision-SVD with HunyuanVideo (Kong et al., 2024), a state-of-the-art open-source video generation model with 13B parameters, and demonstrate superior motion quality. Finally, on the particularly challenging dance generation task, our model outperforms state-of-the-art human image animation methods that rely on ground-truth pose sequences, surpassing them across all evaluation metrics.

In summary, we make the following contributions:

- We show that optimizing physical knowledge of generative models enhances their ability to generate complex motions and interactions, suggesting a promising direction for improving video generation.
- We introduce ReVision, a three-stage pipeline that significantly improves the motion quality of pre-trained video generation models by explicitly optimizing parameterized 3D physical knowledge extracted from generated videos.
- We propose PPPM, a lightweight and robust parameterized physical prior model that effectively refines motion information in generated videos.

## 2 Related Work

**Video Generation.** With the success of diffusion models in image generation (Rombach et al., 2022; Esser et al., 2024; Liu et al., 2024a; Betker et al., 2023), driven by advancements in both generative modeling strategies (Ho et al., 2020; Song et al., 2020; Lipman et al., 2022; Liu et al., 2022; 2024b) and model architectures (Bao et al., 2023; Peebles & Xie, 2023; Liu et al., 2024c; Ma et al., 2024), video generation (Ho et al., 2022; Singer et al., 2022; Wang et al., 2024c; Yang et al., 2023; Zhang et al., 2024a; Zhou et al., 2022; Bar-Tal et al., 2024; Polyak et al., 2024; Brooks et al., 2024) has recently attracted significant attention. Parallel to text-to-video (T2V) generation, image-to-video (I2V) methods (Babaeizadeh et al., 2017; Li et al., 2018; Xiong et al., 2018; Pan et al., 2019; Zhang et al., 2020) generate videos from a single starting frame. However, existing methods still struggle to handle complex motions and interactions, and often fail to maintain physical plausibility. To overcome these challenges, recent approaches incorporate additional conditions to enhance motion control in video generation. Common conditional inputs include text descriptions (Hu et al., 2022; Girdhar et al., 2023; Chen et al., 2023; Ren et al., 2024b; Zeng et al., 2024), which can further guide motion modeling. For example, MAGE (Hu et al., 2022) introduces a spatially aligned motion anchor to blend motion cues from text, and SEINE (Chen et al., 2023) uses a random-mask video diffusion model to create transitions guided by textual descriptions. Another popular condition is optical flow, where models (Mahapatra & Kulkarni, 2022; Ni et al., 2023; Shi et al., 2024) estimate rough flow from user-provided arrows or text to guide complex motion generation. In contrast, ReVision leverages implicit motion features already embedded in the generated video through 3D parameterized object representations. This allows it to directly extract, optimize, and reinforce accurate and reliable motion features from the generated video itself, resulting in precise motion sequences that enhance coherence and fidelity.

**Human Image Animation.** Human image animation focuses on transferring motion from a source human to a target human by *using ground-truth posture sequences*, which can be represented as flow (Wang et al., 2004), keypoints (Hu, 2024; Tan et al., 2024), or human part masks (Xu et al., 2024). Extensive efforts have gone into extracting improved motion features. For example, MagicAnimate (Xu et al., 2024) leverages an off-the-shelf ControlNet (Zhang et al., 2023a) to obtain motion conditions, Hu *et al.* (Hu, 2024) introduce a Pose Guider network to align pose images with noise latents, and Animate-X (Tan et al., 2024) utilizes both implicit and explicit pose indicators to generate motion and pose features. Such strong guidance enables high-quality video generation in human image animation, as each posture sequence directly dictates the synthesis of corresponding frames. However, *in general video generation, the ground-truth dense guidance is typically unavailable*, and there is usually no reference video for extracting a motion sequence. To overcome this limitation, ReVision introduces a three-stage process: it first extracts an implicit, rough motion sequence from the generated video, then refines it using the proposed PPPM, and finally leverages the refined motion to guide video regeneration. This approach provides effective guidance for video generation, significantly improving video quality.

## 3 Preliminary

**Latent Diffusion Model.** Diffusion models (Ho et al., 2020) generate data through a denoising process that learns a probabilistic transformation. Latent diffusion models (Rombach et al., 2022) move this process from pixel space to the latent space of a Variational Autoencoder (Kingma & Welling, 2014). Specifically, we consider the latent representation $z_0$ of the input data. In the forward diffusion process, Gaussian noise

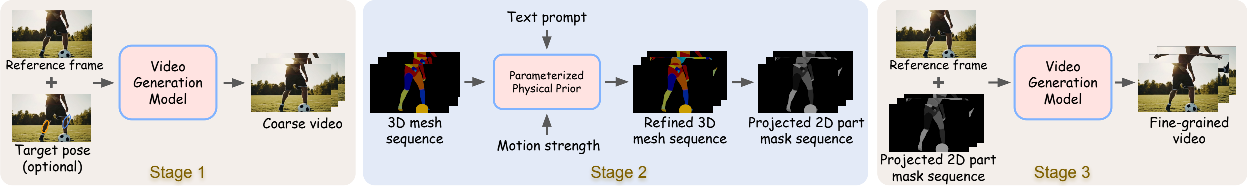

Figure 2: **Method overview.** Given the video generation model, ReVision operates in three stages. Stage 1: A coarse video is generated based on the provided conditions (*e.g.*, target pose, marked in blue, indicating the rough position of the yellow part in the last frame). Stage 2: 3D features from the generated coarse video are extracted and optimized using the proposed PPPM. Stage 3: The optimized 3D sequences are used to regenerate the video with enhanced motion consistency. Best viewed when zoomed in.

is incrementally added to $z_0$:

$$q(z_t|z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\gamma_t} z_{t-1}, \gamma_t \mathbf{I}\right), \tag{1}$$

where $z_t$ represents the noisy latent representation at time step $t$, and $\gamma_t$ is a predefined noise schedule with $t \in (0, 1)$. As $t$ increases, the cumulative noise applied to $z_0$ intensifies, gradually transforming $z_t$ closer to pure Gaussian noise. We express the transformation from $z_0$ to $z_t$ directly as: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\gamma_i)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The latent diffusion model, parameterized by $\Theta$, learns to reverse this noising process by taking $z_t$ as input and reconstructing the clean data with the objective: $\mathcal{L} = \|\epsilon - \epsilon_\Theta(z_t, t, c)\|_2^2$, where $c$ is the condition to guide the denoising process. Once the latent space is reconstructed, it is decoded via the VAE decoder.

**Video Latent Diffusion Model.** We use SVD (Blattmann et al., 2023) as our base video diffusion model, which extends Stable Diffusion 2.1 (Rombach et al., 2022) to video. The main architectural difference from image diffusion models is that SVD incorporates a temporal UNet (Ronneberger et al., 2015) by adding temporal convolution and (cross-) attention (Vaswani, 2017) layers after each corresponding spatial layer.

**3D Human and Animal Mesh Recovery.** We utilize the SMPL-X (Pavlakos et al., 2019) and SMAL (Zuffi et al., 2017) parametric models to represent humans and animals, respectively. These models parameterize 3D meshes with pose parameters $\theta$ and shape parameters $\beta$. Additionally, SMPL-X model includes expression parameters $\psi$ to capture facial expressions through blend shapes. Given these parameters, SMPL-X model is a differentiable function that outputs a posed 3D human mesh $\mathcal{M}_{SMPL-X}(\theta_h, \beta_h, \psi_h) \in \mathbb{R}^{10475\times 3}$, where pose $\theta_h \in \mathbb{R}^{165}$, shape $\beta_h \in \mathbb{R}^{10}$, and expression $\psi_h \in \mathbb{R}^{10}$. Similarly, SMAL model represents a posed 3D animal mesh with $\mathcal{M}_{SMAL}(\theta_a, \beta_a) \in \mathbb{R}^{3889\times 3}$, where pose $\theta_a \in \mathbb{R}^{105}$ and shape $\beta_a \in \mathbb{R}^{41}$. In this project, we recover 3D meshes by fitting SMPL-X and SMAL to our data and to the generated videos. This approach provides accurate part labeling and incorporates spatial priors of human and animal bodies, improving upon direct 2D shape predictions. Additionally, it enables motion strength computation by measuring movement speed in 3D space, offering higher accuracy than 2D pixel-based predictions.

**2.5D Parameterized Object Representation.** Unlike humans and animals, there is no straightforward way to parameterize general objects in 3D space. Here, we represent objects in 2.5D by combining 2D bounding boxes (Varghese & Sambath, 2024), segmentation masks (Peng et al., 2020; Wu et al., 2024), and estimated depth (Yang et al., 2024b).

## 4 Method

ReVision requires extending a pre-trained video diffusion model to accept additional motion conditions as input. In Sec. 4.1, we describe how to adapt SVD into a motion-conditioned video generation model with minimal modifications. In Sec. 4.2, we introduce ReVision, a three-stage video generation pipeline built upon the extended SVD, incorporating a Parameterized Physical Prior Model (PPPM) to provide accurate motion sequences as conditioning inputs. An illustration is provided in Fig. 2.

### 4.1 Motion-Conditioned Video Generation

Since SVD does not natively support motion-conditioned video generation, we extend its design to enable this capability, with a focus on simplicity to preserve its original generation quality and minimize deviations from user-provided inputs. Concretely, we begin with a pre-trained SVD and fine-tune it within a carefully structured strategy. We concatenate two additional conditioning channels to the original condition: one for a part-level segmentation mask derived from the 3D motion sequence, and another for a confidence map indicating the reliability of the part mask, as illustrated in Fig. 3. We also design a fine-tuning pipeline that integrates three scenarios with varying levels of part mask guidance, allowing the model to flexibly handle diverse inputs. We detail those three scenarios below.

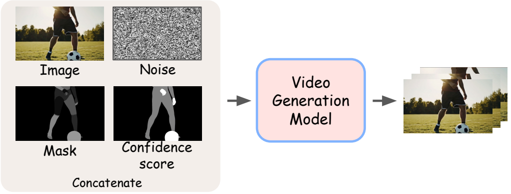

Figure 3: **Motion-conditioned video generation.** We enable motion-conditioned generation by introducing two extra conditioning channels: (1) part segmentation mask derived from the 3D motion sequence, and (2) its corresponding confidence map.

First, when the full motion sequence is provided (40% of training examples), the part-level mask is generated by merging all 2D part segmentation masks projected from 3D parametric mesh models. Since the motion sequence provides dense and precise control over video generation, we assign a confidence score of 1 to the corresponding confidence map. Our experiments confirm that these 3D-projected masks are more robust than existing part segmentation models.

Second, when only the target pose is provided (30% of training examples), we convert the projected part segmentation masks into polygons. This aligns with users' inference input, where they provide simple sketches (*e.g.*, circles or ovals) to indicate the final positions of specific targets or parts (*e.g.*, a hand or arm). These user-friendly sketches are then converted into polygonal masks, similar to the part segmentation mask polygons used during training. Since polygon conversion introduces unavoidable errors, we assign a confidence score of 0.5 in this case.

Last, to preserve SVD's ability to generate videos without motion conditioning, the remaining 30% of training examples provide an empty part mask, with a corresponding confidence score of 0.

Note that all three settings use the same model architecture, with minimal modifications limited to the first convolutional block of SVD. This design enables fine-tuning only the initial convolutional block and the temporal layers, avoiding the need to train SVD from scratch. As a result, the extended SVD can generate videos conditioned on various types of motion inputs, while still retaining its ability to generate videos from just text and the first frame.

### 4.2 Proposed Method: ReVision

**Overview.** As shown in Fig. 2, ReVision consists of three stages. In stage one (S1), we generate a coarse video based on the provided conditions. In stage two (S2), we extract both 2D and 3D features from the coarse video and refine the 3D motion sequences through the proposed PPPM. In stage three (S3), we use the refined 3D motion sequences as strong conditioning, guiding the video generation model to regenerate the video, resulting in significantly higher-quality output even for complex motions and interactions.

**S1: Coarse Video Generation.** Given the first frame and an *optional* user-specified target motion in the final frame, we use the fine-tuned SVD model to generate the video. Since the generation relies only on the target motion in the final frame or an empty motion, rather than a complete motion sequence, the resulting video often exhibits poor motion quality, leaving room for refinement. Therefore, we refer to this stage as coarse video generation.

Although we utilize only 2D and 3D motion features from the coarse video generated in Stage 1, this phase remains foundational, as it is critical for capturing rich motion patterns, intricate object interactions, and authentic camera movements in complex real-world settings. Video generation is inherently a multifaceted task. Beyond producing realistic object appearances, it also requires an understanding of dynamic motion, scene context, coherent camera trajectories, and the diverse interplay of these elements. While training a motion generation model directly for this task is theoretically feasible, it proves challenging in practice.

Current state-of-the-art models (Guo et al., 2024; Zhang et al., 2024b; 2023b) are constrained by the lack of large datasets, limiting their capacity to model only simplistic motions, such as human-like activities like running or dancing. Consequently, they struggle to generalize to more complex motions, diverse object interactions, and fail to generate motions that align with realistic camera dynamics and scene context.

Stage 1 mitigates these limitations by directly generating videos and extracting detailed motion patterns, camera trajectories, scene transitions, and meaningful interactions. By leveraging extensive motion priors learned from billions of videos, it constructs a comprehensive sketch of motion and scene structure. This sketch serves as a critical foundation, providing the diversity and realism necessary to produce lifelike, engaging videos. The subsequent stages build on this: Stage 2 refines the motion further, while Stage 3 focuses on generating the video's visual appearance based on the refined motion feature.

Notice that, because only rough motion features without detailed visual information are needed from the coarse video, **we can significantly reduce computational overhead**. For instance, our experiments show that the compute time for Stage 1 can be reduced from 36 seconds to 8 seconds by generating the coarse video at a lower resolution (1/4 of the original), with fewer frames (1/2 of the original), and fewer denoising steps (32 vs. 50), while still preserving comparable final video quality. This optimization makes the overall generation process more efficient and cost-effective (See Tab.2).

**S2: Object-Centric 3D Optimization.** After generating coarse videos, we parameterize the 3D information in the scenes for further optimization. For *humans* and *animals*, we employ well-established 3D parametric mesh models (Loper et al., 2015; Rueegg et al., 2023; Zuffi et al., 2024). For *general objects*, where no unified 3D representation or well-established modeling approach exists, we construct a parameterized representation by combining 2D bounding boxes (Varghese & Sambath, 2024), segmentation masks (Yu et al., 2023; Peng et al., 2020), and estimated depth (Yang et al., 2024b). Specifically, given the detected bounding box and segmentation mask, we extract a contour from the mask and approximate it with 16 vertices. We then combine these with 4 bounding box corners and the box center, yielding a total of 21 key 2D points. These points are lifted into 3D space using the estimated depth, resulting in a compact point-based representation for each object, denoted as $p_o \in \mathbb{R}^{21\times3}$.

However, due to the poor motion quality and inconsistencies in the coarse video generated in S1, the 3D parameters extracted also suffer from instability and inconsistencies. To address this, we propose a Parameterized Physical Prior Model (PPPM) to optimize the 3D motion sequence, based on the text information and the motion strength.

PPPM first extracts text embeddings from the text description using a pre-trained CLIP encoder (Radford et al., 2021). Motion strength is computed from the differences in parametric 3D model parameters between adjacent frames, providing a measure of motion speed. Since the 3D motion sequences are already parameterized as 3D vectors, PPPM employs a series of transformer blocks to iteratively refine the motion sequence based on these conditioning inputs. Within each block, motion features undergo self-attention, followed by cross-attention with the conditioning inputs (text embeddings and motion strength) and a feed-forward network to generate the final output. Finally, the optimized 3D parameterized motion sequences are converted into 3D mesh sequences and projected into 2D as part segmentation masks and confidence maps, providing more accurate motion guidance. Architectural details are provided in Sec A of the Supp, and the effectiveness of PPPM is demonstrated in Sec. 5.3.

To effectively train PPPM, we introduce small perturbations to ground-truth motion sequences during training. Three types of perturbations are randomly applied: (1) adding small random noise to the motion sequence, (2) shuffling the internal order of the sequence, and (3) dropping a small segment while repeating the remaining segments to maintain the original length. Through this process, PPPM learns to denoise perturbations, improving its ability to recover smooth and robust motion sequences.

**S3: Fine-grained Video Generation.** In the final stage, we regenerate the video using the same SVD model but with the improved motion sequence as additional conditioning. Unlike the coarse generation in stage one, which uses only the target pose in the last frame or no motion information, we now utilize the full motion sequence as part masks optimized in 3D space. With this stronger conditioning, the final output exhibits significantly improved motion consistency compared to the coarse video, as illustrated in Fig. 8.

# 5 Experimental Results

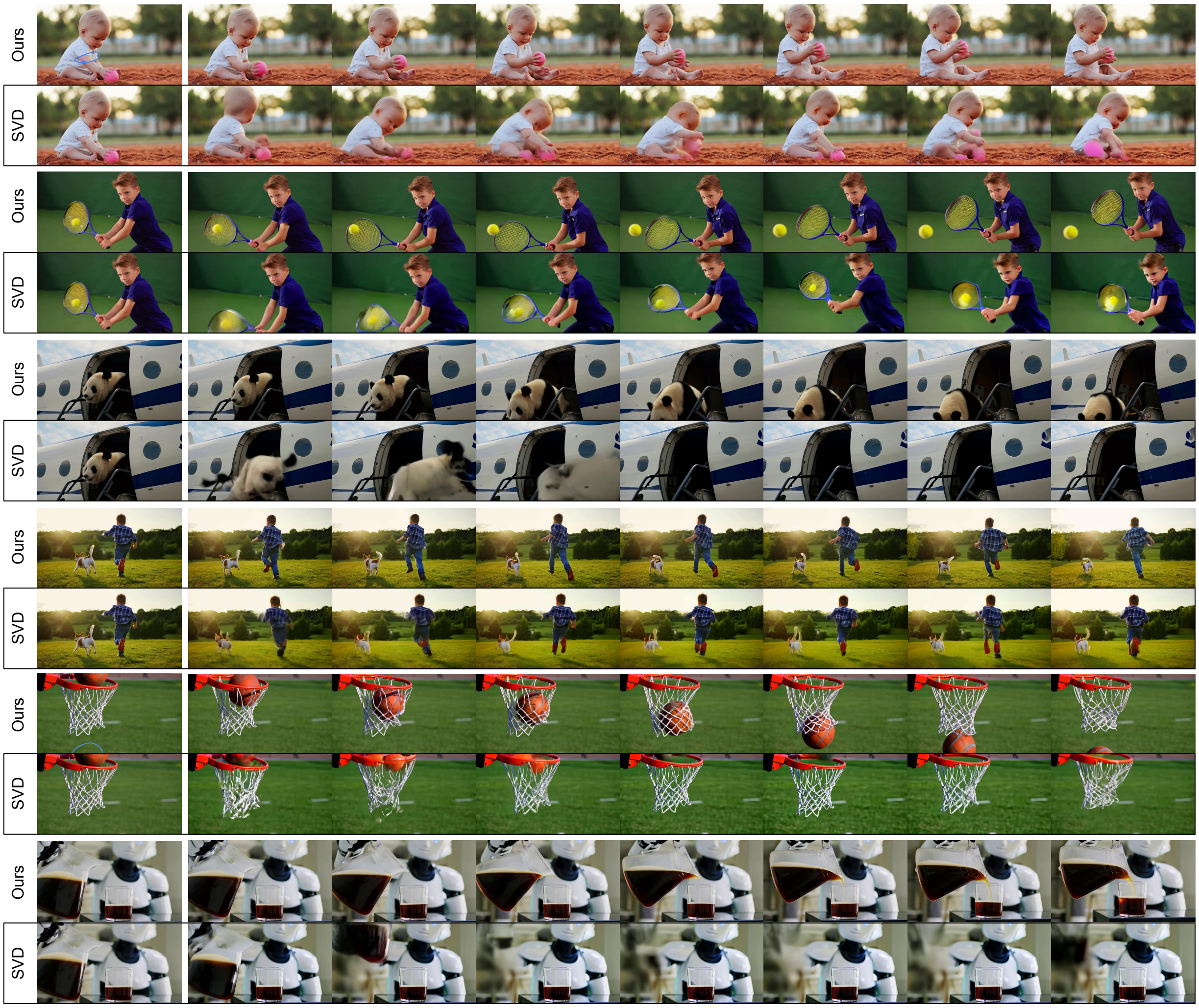

Figure 4: **Qualitative comparison.** ReVision generates high-quality videos with complex motions and interactions of *humans, animals, and general objects*. Zoom in for better details. Please find the side-by-side video comparisons in the supplementary video. Reference frames are in the first column.

We first compare our method with SVD (Blattmann et al., 2023) and HunyuanVideo (Kong et al., 2024) in Sec. 5.1, highlighting how it enhances SVD to support more controllable and complex motion generation while maintaining efficiency, effectively handling occlusions, and enabling long video generation. Next, in Sec.5.2, we compare our model with Human Image Animation methods, demonstrating its ability to generate complex motions. We then evaluate the effectiveness of the proposed Parameterized Physical Prior Model in Sec.5.3. Due to space limitations, additional details and results, including ablations on parametric 3D mesh, text prompt, and motion strength, are provided in Sec. C of the Supp.

## 5.1 Image-to-Video Generation

**Dataset.** Both the motion-conditioned video generation model and the Parameterized Physical Prior Model (PPPM) need to be fine-tuned (trained) on a small yet high-quality video dataset with object-centric annotations. Existing large-scale video datasets (Bain et al., 2021; Chen et al., 2024) mainly provide text-image pairs without detailed object-centric annotations. To address this limitation, we use a suite of off-the-shelf models across various tasks to generate 2D and 3D object-centric annotations. We annotate a total of 20K

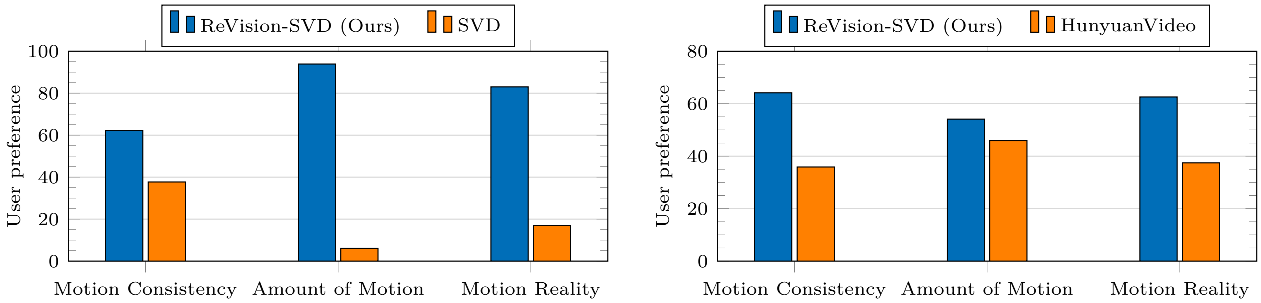

Figure 5: **User preference comparison.** Our model enhances the motion generation capability of the pre-trained SVD. It even surpasses HunyuanVideo, a SOTA model with 13B parameters. These results highlight the effectiveness of our model in generating complex motions and interactions.

Table 1: **Quantitative comparison on VBench++.** We achieve a significantly higher Dynamic Degree while maintaining similar performance across all metrics of consistency, smoothness, and quality.

| | **I2V Subject** | **I2V Background** | **Subject Consistency** | **Background Consistency** | **Motion Smoothness** | **Dynamic Degree** | **Aesthetic Quality** | **Imaging Quality** |
|---|---|---|---|---|---|---|---|---|
| Wan2.1-I2V-14B-720P | 96.95% | 96.44% | 94.86% | 97.07% | 97.90% | 51.38% | **64.75%** | 70.44% |
| Gen-4-I2V | 97.84% | 97.46% | 93.23% | 96.79% | 98.99% | 55.20% | 61.77% | 70.41% |
| HunyuanVideo-I2V | **98.53%** | 97.37% | 95.26% | 96.70% | **99.23%** | 22.20% | 62.55% | 70.1% |
| SVD-XT-1.1 | 97.51% | 97.62% | 95.42% | 96.77% | 98.12% | 43.17% | 60.23% | 70.23% |
| ReVision-SVD (Ours) | 97.94% | **98.06%** | **96.13%** | **97.89%** | 98.88% | **83.15%** | 60.18% | **71.48%** |

videos from the Panda-70M (Chen et al., 2024) dataset. For each video, we provide frame-wise 2D bounding boxes, semantic masks, depth estimation maps, and 3D parametric mesh reconstructions for detected humans and animals. The details are outlined in Sec. B in the Supp.

**Experimental Setup.** We use SVD (Blattmann et al., 2023) as the base video generation model and modify it by introducing two additional channels for conditional generation. We fine-tune SVD on our annotated dataset for 300K iterations with a batch size of 64 and a constant learning rate of $2 \times 10^{-5}$. During training, we randomly sample 16-frame video clips with a stride of 4 at a resolution of $1024 \times 576$. To enable various control, we incorporate different conditioning strategies: 40% of video clips contain accurate part masks for each frame, 30% contain a polygon mask for random parts in the final frame, and the remaining clips have no additional conditioning.

**Benchmark on General Video Generation.** To better evaluate our model on general video generation, we used VBench++ (Huang et al., 2024), which provides a comprehensive, detailed, multi-dimensional assessment of general video generation quality. As our model is designed for image-to-video generation, we primarily compared it against SVD-XT-1.1. Results are provided in Tab. 1. Our ReVision-SVD consistently outperforms SVD on nearly all metrics, particularly in dynamic degree (83.15% vs. 43.17%) and various consistency and smoothness measures, while maintaining similar aesthetic quality.

**User Study.** To better compare our model with the baseline SVD and HunyuanVideo, we conduct user studies to assess user preferences. Specifically, we generate 500 text descriptions of humans and animals engaged in daily activities using GPT-4o (Hurst et al., 2024). For the comparison with SVD, we use GPT-4o to generate five $16:9$ images for each prompt, which are resized to $1024 \times 576$ as input. For the comparison with HunyuanVideo, we first use their released model to generate five videos at a resolution of $1280 \times 720$ for each prompt, then extract the first frame of each video as the input image for our model to generate the corresponding video. No target pose is provided for any model. For each image, we generate one video per model using the same random seed (42), resulting in a total of $5,000$ video pairs. Each video pair is evaluated by three randomly selected users on Amazon MTurk, leading to a total of $15,000$ comparisons. Users are shown two videos side by side, generated by different models, with the order randomized. They are instructed to assess the videos based on Motion Consistency, Amount of Motion, and Motion Realism. The results are reported in Fig. 5. Our model significantly enhances the motion generation capabilities of SVD, producing videos with superior motion quantity, consistency, and realism. Furthermore, it even surpasses

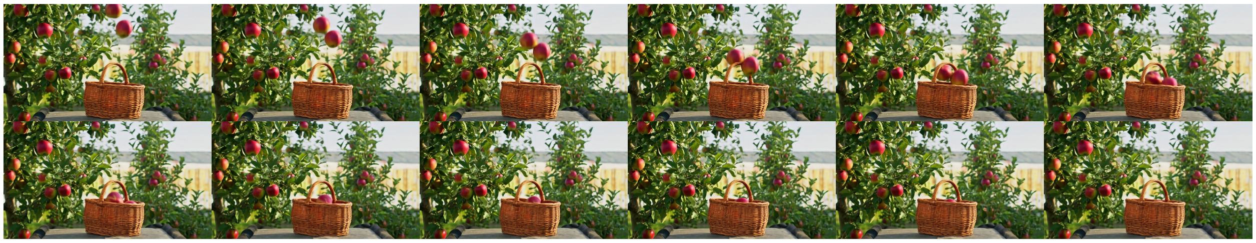

Figure 6: **Handling occlusion.** As illustrated by the two apples falling into the basket, ReVision handles occlusions by lifting and optimizing motion in 3D space, which allows explicit reasoning about object spatial relationships, effectively resolving occlusions that are ambiguous in 2D.

HunyuanVideo, a state-of-the-art video generation model with 13B parameters, in terms of motion quality. These results highlight the effectiveness of our model in generating complex motions and interactions.

**Qualitative Comparison.** We provide samples of generated videos in Fig. 4. Our ReVision produces realistic movements that closely follow user instructions. It also generates high-quality videos that involves complex motions and interactions, such as running with dogs, picking up a ball, and hitting a tennis ball. More visualizations are available in the supplementary videos.

**Inference Speed.** We compare the inference speed of our model against two baselines in Tab. 2. Despite the three-step pipeline, the coarse video generation (S1) takes only 8 seconds after our optimization, and the additional 3D detection and refinement modules (S2) add 5 seconds to the inference time on a single A100. Together, these two stages are significantly faster than the original SVD, which requires 36 seconds. More importantly, with just 1.5B parameters and a runtime of 49 seconds, our model generates high-quality videos with complex motions – comparable to or even surpassing state-of-the-art models like HunyuanVideo (see Fig. 4 and 5), which uses over 13B parameters and requires an average of 411 seconds.

Table 2: **Inference speed.** Average time to generate a 32-frame video. Our ReVision-SVD matches SVD in speed (8.4x faster than HunyuanVideo) while surpassing HunyuanVideo in generating complex motions and interactions.

| | SVD | ReVision-SVD | HunyuanVideo |
|---|---|---|---|
| Time (s) | 36 | 49 (8 + 5 + 36) | 411 |

**Handling Occlusion.** Occlusion becomes a significant challenge when generating videos with multiple objects and large motions. However, by lifting everything into 3D, occlusion is naturally resolved: Since we estimate depth, all objects are fully represented with their spatial positions, allowing us to reason about their relative locations in 3D. And after optimizing the motion in 3D, we project it back to 2D using the depth information, which restores accurate occlusion relationships in the camera coordinate. An example is shown in Fig. 6, where two apples are generated dropping into a basket. Our model effectively captures spatial relationships, producing realistic videos in which the apples fall *into* the basket with appropriate occlusion.

To further evaluate our model's ability to handle occlusion, we conducted an additional user study using the 5,000 video pairs generated for the main experiment. Users on Amazon MTurk were asked to assess the quality of occlusion handling and object interactions for each video pair. If no object interaction was observed, users were instructed to select "no interaction/occlusion." Similarly, each video pair was evaluated independently by three randomly selected users to ensure reliability. The study yielded 15,000 evaluations, including 10,207 valid comparisons and 4,793 responses marked as "no interaction/occlusion". The results of these 10,207 valid comparisons are summarized in Tab. 3, highlighting that our model consistently outperforms both SVD and HunyuanVideo across a diverse range of occlusion scenarios.

Table 3: **User preference comparison for occlusion and interaction handling.**

| | ReVision-SVD (Ours) | SVD |
|---|---|---|
| Preference | 97.63% | 2.37% |
| | **ReVision-SVD (Ours)** | **HunyuanVideo** |
| Preference | 63.99% | 36.01% |

**Long Video Generation.** Another advantage of our model is its ability to generate complex and large-scale motions over long video sequences (Fig. 7). PPPM optimizes motion in a 3D parameterized space, enabling smooth and realistic interpolation and extrapolation to arbitrary lengths. The resulting long 3D

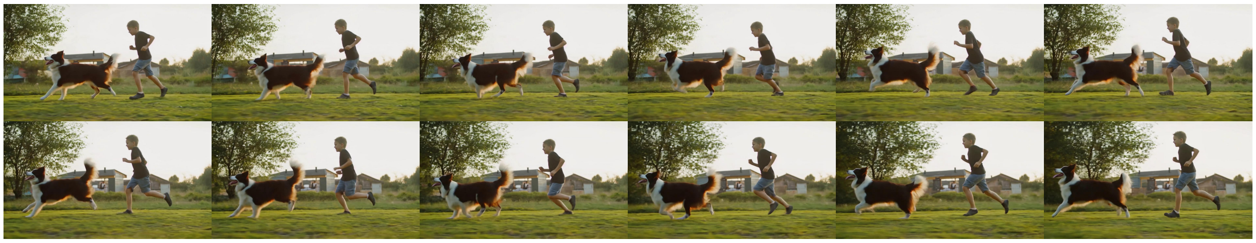

Figure 7: **Long video generation.** Our PPPM extends a 32-frame 3D motion to 128 frames through interpolation (32 → 64), extrapolation (64 → 128), and refinement, enabling complex, large-scale motion generation over long video sequences. See supplementary videos for details.

Table 4: **Quantitative comparison for dance generation.** 'ReVision (w/ full motion)' follows baselines and takes full motion sequences as condition, while 'ReVision (w/ target pose)' uses the target pose from the final frame.

| | **SSIM ↑** | **PSNR ↑** | **LPIPS ↓** | **FVD ↓** |
|---|---|---|---|---|
| MagicAnimate | 0.714 | 29.16 | 0.239 | 179.07 |
| Animate Anyone | 0.718 | 29.56 | 0.285 | 171.90 |
| Champ | 0.802 | 29.91 | 0.234 | 160.82 |
| VividPose | 0.758 | 29.83 | 0.261 | 152.97 |
| ReVision (image only) | - | - | - | 136.43 |
| ReVision (w/ target pose) | - | - | - | 130.14 |
| ReVision (w/ full motion) | **0.864** | **30.08** | **0.210** | **121.26** |

Table 5: **PPPM acts as a general motion denoiser, improving performance on human motion generation.** Following MoMask, we report R-Precision at Top-1, Top-2, and Top-3. Our PPPM achieves state-of-the-art performance on two widely used benchmarks.

| | **HumanML3D** | | | **KIT-ML** | | |
|---|---|---|---|---|---|---|
| | **R-P@1** | **R-P@2** | **R-P@3** | **R-P@1** | **R-P@2** | **R-P@3** |
| MotionDiffuse | 0.491 | 0.681 | 0.782 | 0.417 | 0.621 | 0.739 |
| ReMoDiffuse | 0.510 | 0.698 | 0.795 | 0.427 | 0.641 | 0.765 |
| MoMask | 0.521 | 0.713 | 0.807 | 0.433 | 0.656 | 0.781 |
| MoMask + PPPM | **0.544** | **0.735** | **0.810** | **0.471** | **0.673** | **0.785** |

motion sequences are then used to generate multiple overlapping video clips, which are stitched together to form extended videos with consistent motion. *The 3D representation plays a key role in maintaining smooth temporal continuity.* While long video generation is not our main focus, using advanced techniques like temporal compression (Bar-Tal et al., 2024), beyond simple overlap, could further improve visual coherence. We leave their integration with our 3D-aware framework as future work.

### 5.2 Complex Motion Generation

**Experimental Setup.** To demonstrate our model's ability to generate videos with complex motion, we compare our approach with state-of-the-art human image animation models on the TikTok Dancing dataset (Jafarian & Park, 2021), using the Disco (Wang et al., 2024b) split. For compatibility with the SVD model architecture, all videos are cropped to 576 × 1024. We fine-tune the original SVD only on the training split for 30K iterations, with a batch size of 8 and a learning rate of $1 \times 10^{-5}$.

**Evaluation Metrics.** We follow baselines and report Peak Signal-to-Noise Ratio (PSNR) (Hore & Ziou, 2010), Structural Similarity Index (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) to measure the visual quality of the generated results. We also report and Fréchet Video Distance (FVD) (Unterthiner et al., 2018) for video fidelity comparision.

**Experimental Results.** We compare ReVision with human image animation methods in Tab. 4, where we achieve state-of-the-art performance across all metrics. Notably, we observe a significant improvement in FVD, highlighting substantial gains in video generation quality. It is important to note that all baselines in this task *rely on ground-truth motion sequences*, which are challenging to obtain in practical scenarios, limiting their applicability. In contrast, our method can generate realistic and high-quality videos *using only the input inference image or inference image with a target pose.*

### 5.3 Parameterized Physical Prior Model (PPPM)

We demonstrate the effectiveness of the proposed Parameterized Physical Prior Model in this section.

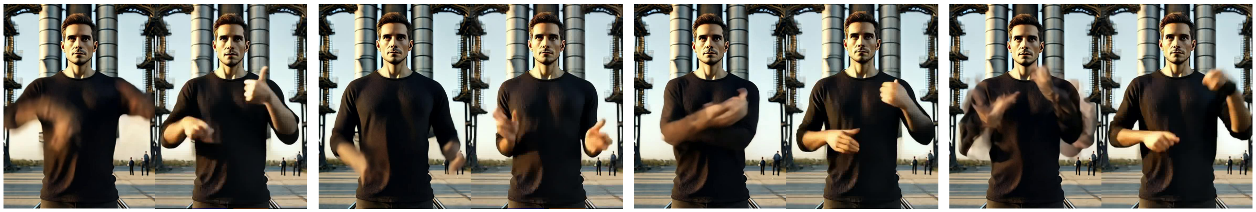

Figure 8: PPPM improves motion and visual quality, generating realistic videos with large motions.

Table 6: **User studies for PPPM.** PPPM improves object and motion consistency, while reducing morphological failure rates.

| | Object Consistency ↑ | Motion Consistency ↑ | Morphological Failure Rate ↓ |
|---|---|---|---|
| w/o PPPM | 12.4 | 4.0 | 83.5 |
| w/ PPPM | 87.6 | 96.0 | 14.3 |

**PPPM Enables High-Quality Video Generation with Complex Motions and Interactions.** To demonstrate the effectiveness of PPPM, we select a complex dance scenario and visualize outputs with and without the proposed PPPM in Fig. 8. We also show quantitative improvements of the generated videos with PPPM in Tab. 6, where 500 video pairs were evaluated by random users on Amazon MTurk. Each pair was rated by three different users, resulting in a total of 1,500 evaluations. The results show that the video generation model alone still struggles to produce high-quality videos with accurate motion. However, *leveraging the object-level priors from our Parameterized Physical Prior Model enables the generation of realistic videos with enhanced motion and visual quality.*

**PPPM Prevents Error Accumulation in Multi-stage Video Generation.** In addition, Fig. 8 shows that even when the generated videos exhibit severely broken motion, our PPPM can still recover (predict) a smooth and coherent motion sequence using the ground-truth first frame and target pose, enabling successful final video generation. *This correction mitigates motion errors and prevents error accumulation, highlighting the robustness of our pipeline.*

**PPPM as a General Motion Denoiser.** We focus on a more specific human motion generation task and show that our model improves the performance of the state-of-the-art method, MoMask (Guo et al., 2024), on standard benchmarks (see Tab. 5). Specifically, we use PPPM to refine the motion sequences generated by MoMask and compare the results with MoMask and other methods on HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016) benchmarks. We adopt the same training dataset as MoMask and apply the perturbations described in Sec. 4.2 to train PPPM. *Serving as a general motion denoiser, our model consistently enhances motion generation quality of the current best models across multiple benchmarks.*

## 6 Conclusion

We introduced ReVision, a three-stage framework for video generation that improves motion consistency by integrating 3D motion cues. ReVision leverages a pretrained video diffusion model to generate coarse videos, refines 3D motion sequences with PPPM, and reconditions the generation process with enhanced motions to improve fine-grained and complex motion generations. Evaluations show that ReVision significantly outperforms existing methods in motion fidelity and coherence.

## References

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

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.

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.

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

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18219–18228, 2022.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

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.

Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12753–12762, 2021.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

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.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 600–615, 2018.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024a.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A framework for cross-modality evolution. *arXiv preprint arXiv:2412.15213*, 2024b.

Qihao Liu, Zhanpeng Zeng, Ju He, Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. *Advances in Neural Information Processing Systems*, 37:133879–133907, 2024c.

Qihao Liu, Yi Zhang, Song Bai, Adam Kortylewski, and Alan Yuille. Direct-3d: Learning direct text-to-3d generation on massive noisy 3d data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6891, 2024d.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2022.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18444–18455, 2023.

Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2019.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8533–8542, 2020.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4 (4):236–252, 2016.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, et al. Movie gen: A cast of media foundation models, 2024. URL `https://arxiv.org/abs/2410.13720`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL `https://arxiv.org/abs/2408.00714`.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024a.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, 2015.

Nadine Rueegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *Proceedings IEEE Conferecne on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Remy Sabathier, Niloy Jyoti Mitra, and David Novotny. Animal avatars: Reconstructing animatable 3D animals from casual videos. *ArXiv*, abs/2403.17103, 2024.

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

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.

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.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6. IEEE, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9326–9336, 2024b.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pp. 1–20, 2024c.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373, 2018.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9099–9109, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024b.

Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint arXiv:2306.01872*, 2023.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8850–8860, 2024.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pp. 1–15, 2024a.

Jiangning Zhang, Chao Xu, Liang Liu, Mengmeng Wang, Xia Wu, Yong Liu, and Yunliang Jiang. Dtvnet: Dynamic time-lapse video generation via single still image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 300–315. Springer, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023a.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023b.

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*, 46(6):4115–4128, 2024b.

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*, pp. 586–595, 2018.

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

Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5524–5532, 2017.

Silvia Zuffi, Ylva Mellbin, Ci Li, Markus Hoeschle, Hedvig Kjellström, Senya Polikovsky, Elin Hernlund, and Michael J. Black. VAREN: Very accurate and realistic equine network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

## Appendix

The supplementary material includes the following additional information.

- Sec. A provides the architectural details for PPPM.
- Sec. B provides additional details for our annotated dataset.
- Sec. C provides additional ablation studies omitted from the main paper.
- Sec. D discusses the limitations of our method.
- Sec. E discusses the societal impacts of our method.

We also provide the generated videos used in all figures in the main paper, as well as additional videos demonstrating accurate motion control, in the *supplementary videos*.

## A Architectural details for PPPM

To optimize the 3D motion sequence extracted from the coarse generated video, we propose the Parameterized Physical Prior Model (PPPM). As shown in Fig. 9, PPPM utilizes a transformer architecture with self-attention, cross-attention, and feedforward layers as its backbone. It takes the parameterized motion sequence of the coarse video as input and optimizes it based on the input text prompt and motion strength.

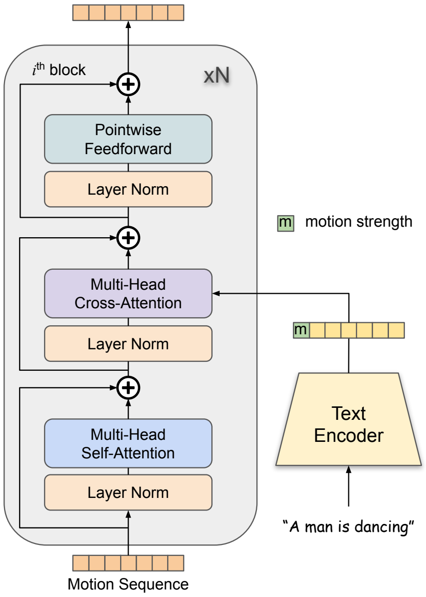

Figure 9: **Architecture of the Parameterized Physical Prior Model**

## B Dataset

Both the motion-conditioned video generation model and the Parameterized Physical Prior Model (PPPM) need to be fine-tuned (trained) on a small yet high-quality video dataset with object-centric annotations. Existing large-scale video datasets (Bain et al., 2021; Chen et al., 2024) mainly provide text-image pairs without detailed object-centric annotations. To address this limitation, we use a suite of off-the-shelf models across various tasks to generate detailed 2D and 3D object-centric annotations. We annotate a total of 20K videos from the Panda-70M (Chen et al., 2024) dataset. For each video, we provide frame-wise 2D bounding boxes, semantic masks, depth estimation maps, and 3D parametric mesh reconstructions for detected humans and animals. The details are outlined below.

**High-Quality Motion Videos Filtering.** To start with, we use LLMs (Yang et al., 2024a) and an open-vocabulary segmentation model (Yu et al., 2023) to curate high-quality motion videos. Specifically, LLM filters videos with evident motion based on their captions. Then, for each selected video, we equally sample 10 frames and apply the segmentation model to identify humans and animals. We evaluate each frame based on the predicted mask size and mask count. Then we retain videos where humans or animals occupy a significant portion of the frame and where the count of humans does not exceed five in each frame.

**Object Detection and Depth Estimation.** Based on the captions of the selected videos, we identify the objects mentioned and detect their bounding boxes (Varghese & Sambath, 2024) and instance masks (Yu et al., 2023). Additionally, we apply Depth Anything V2 (Yang et al., 2024b) to generate the depth maps of each frame.

Table 7: **Quantitative Evaluation Regarding Inference Efficiency.** We reduce Stage 1 compute time from 36 seconds to 8 seconds, while maintaining comparable final video quality.

| | Overall Preference | Visual Quality | Motion Consistency | Amount of Motion |
|---|---|---|---|---|
| ReVision-SVD (full model) | 52.60% | 50.27% | 51.87% | 54.27% |
| ReVision-SVD (efficient model) | 47.40% | 49.73% | 48.13% | 45.73% |

**Human Videos Annotation.** For videos containing humans, we focus on extracting 2D instance segmentation masks, 2D part masks, 2D face keypoints, 3D body pose and shape, and 3D hand pose. We begin by using YOLO-V8 (Varghese & Sambath, 2024) to segment all humans in each frame, providing accurate human masks. Next, we apply a state-of-the-art face keypoint detector, RTMPose (Jiang et al., 2023), to predict facial keypoints for each detected human. Simultaneously, we use 4D-Human (Goel et al., 2023) and HaMeR (Pavlakos et al., 2024) to estimate the 3D body and hand meshes. The resulting SMPL (body mesh) (Loper et al., 2015) and MANO (hand mesh) (Romero et al., 2022) parameters are then fit into a unified SMPL-X (Pavlakos et al., 2019) representation, which contains both human body and hand meshes. We then project the 3D SMPL-X human mesh onto 2D to obtain part masks, as each vertex in the SMPL-X mesh is labeled by body part. Finally, we project the face keypoints and 3D human mesh onto the instance mask, allowing us to compute the overlap between the projected keypoints, projected human mask, and detected 2D human mask. This overlap is quantified using an IoU score, which is used to filter out annotations with high errors. As a result, for each video, we obtain annotations including human instance masks, 2D facial keypoints, 3D SMPL-X meshes for the body and hands, and 2D part-level segmentation masks.

**Animal Videos Annotation.** We start by using Grounded SAM 2 (Kirillov et al., 2023; Ravi et al., 2024; Ren et al., 2024a) to segment animal masks in each frame. Next, we apply a state-of-the-art camera estimation algorithm, VGGSfM (Wang et al., 2024a), to optimize the camera's intrinsic and extrinsic parameters across the video. To ensure a reliable camera estimate, we set thresholds on mean projection errors and mean track lengths, filtering out videos that do not meet these criteria. We then use AnimalAvatar (Sabathier et al., 2024) initialized with Animal3D (Xu et al., 2023) to fit SMAL parameters. Each video is divided into segments of 10 consecutive frames, and AnimalAvatar is applied to each segment independently. This strategy helps mitigate the impact of outliers in camera predictions on the overall optimization quality. To ensure the accuracy of SMAL fitting, we impose thresholds on IoU and PSNR (Hore & Ziou, 2010), filtering out video segments that do not meet our accuracy standards. Once accurate SMAL fittings are obtained, we follow a similar pipeline to extract the desired annotations as used in human cases.

## C Ablation Study

Herein, we conduct additional ablation studies to verify the effectiveness of the proposed designs.

**Quantitative Evaluation Regarding Inference Efficiency.**

Stage 1 can be performed with reduced computational overhead while maintaining similar results. To quantify this, we conducted experiments on 500 video pairs using the same input images but with varying settings: lower resolution (1:4), fewer frames (8 vs. 16), and fewer denoising steps (32 vs. 50). The efficient model reduced Stage 1 compute time from 36 seconds to 8 seconds. We ran a human preference study on Amazon MTurk, comparing the full and efficient versions for each video, with results provided in Tab. 7. The results demonstrate that the efficient model achieves similar performance to the full model in visual quality and motion consistency, with only a slight drop in the amount of motion, which is likely due to the reduced number of frames. This confirms that Stage 1 can be significantly accelerated with minimal impact on perceived quality.

**Parameterized Physical Prior Model.** We briefly discuss the quantitative improvements of the proposed PPPM in the main paper and provide additional experimental details and results here. To further demonstrate the improvements of PPPM, we conducted an additional user study on Amazon MTurk comparing videos generated with and without PPPM. Unlike our previous study, which compared our method with SVD, this evaluation focuses on object consistency and motion consistency. We also report the percentage of videos containing incorrect human or animal structures (*i.e.*, the morphological failure rate). We evaluate 500 video pairs, each rated by three different users, resulting in 1,500 total evaluations. The results are

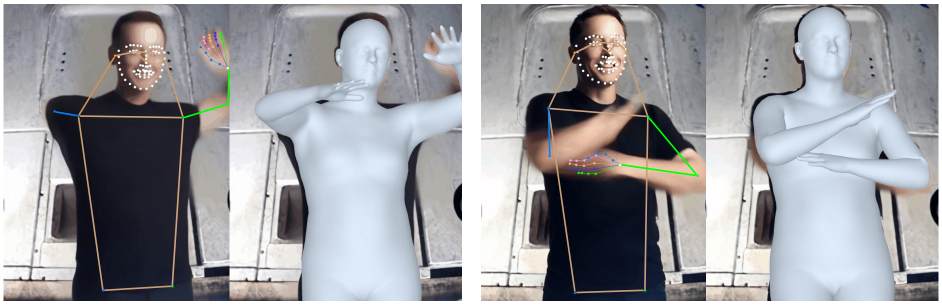

Figure 10: **The parametric 3D mesh serves as an effective object-level prior,** ensuring complete human body structures in the coarse video generated during the first stage. In the left two images, the human keypoint model fails to detect the missing right hand, which is accurately “recovered" by the parametric human mesh model. In the right two images, the human mesh model provides a more accurate prior for both blurred hands.

presented in Tab. 6 of the main paper. By optimizing with a parametric 3D mesh, our approach significantly reduces incorrect human and animal structures, leading to substantial improvements in object and motion consistency.

**Parametric 3D Mesh.** Previous human image animation models mainly rely on 2D pose sequences for each frame to provide motion information. However, this approach is not optimal for general video generation. As shown in Fig. 10, we compare the results of using a parametric human mesh model (Loper et al., 2015) versus a human keypoint model (Sun et al., 2019). Our findings indicate that the human mesh model provides a robust object-level prior, which significantly benefits general video generation. Specifically, current video generation models often misinterpret the structure of humans and animals, occasionally producing unrealistic results, such as a man with three arms, an example of morphological failure. This problem becomes more serious in complex motion generation. However, incorporating human and animal priors from 3D mesh models substantially mitigates these structural inaccuracies, enabling more accurate representations of targets.

**Text Prompt and Motion Strength.** Generating videos from a single image introduces significant ambiguity. To reduce this, we incorporate additional conditioning using a text prompt and a motion strength parameter. Specifically, the text prompt defines the intended motion type, while motion strength controls the speed and complexity of motion within the video. In our experiments, we observe that varying motion strength with the same target pose leads to different motion trajectories. For instance, when moving a hand from point A to B, a video generated with low motion strength results in a direct, simple movement. In contrast, higher motion strength produces a more dynamic and complex trajectory, though it still reaches the same final pose at B. An illustrative example is provided in Fig. 11.

## D Limitations

The proposed method has several remaining limitations. *First*, it relies on parametric 3D mesh models, requiring multiple off-the-shelf models for different object categories, though it adds only 5 seconds to the total inference time. Recent advances in 3D modeling, such as encoding 3D priors of general objects within a single diffusion model (Liu et al., 2024d), are paving the way for more general, efficient models that can be seamlessly integrated into our pipeline for high-quality video generation. *Second*, the model still struggles to generate high-quality details such as fingers and hands. *Finally*, while PPPM can generate realistic motion sequences beyond 32 frames, our current implementation, based on vanilla SVD, is limited by memory constraints (80GB RAM). However, recent methods have demonstrated longer video generation ability using pretrained diffusion models (Chen et al., 2023). Exploring long-video generation with 3D knowledge remains future work.

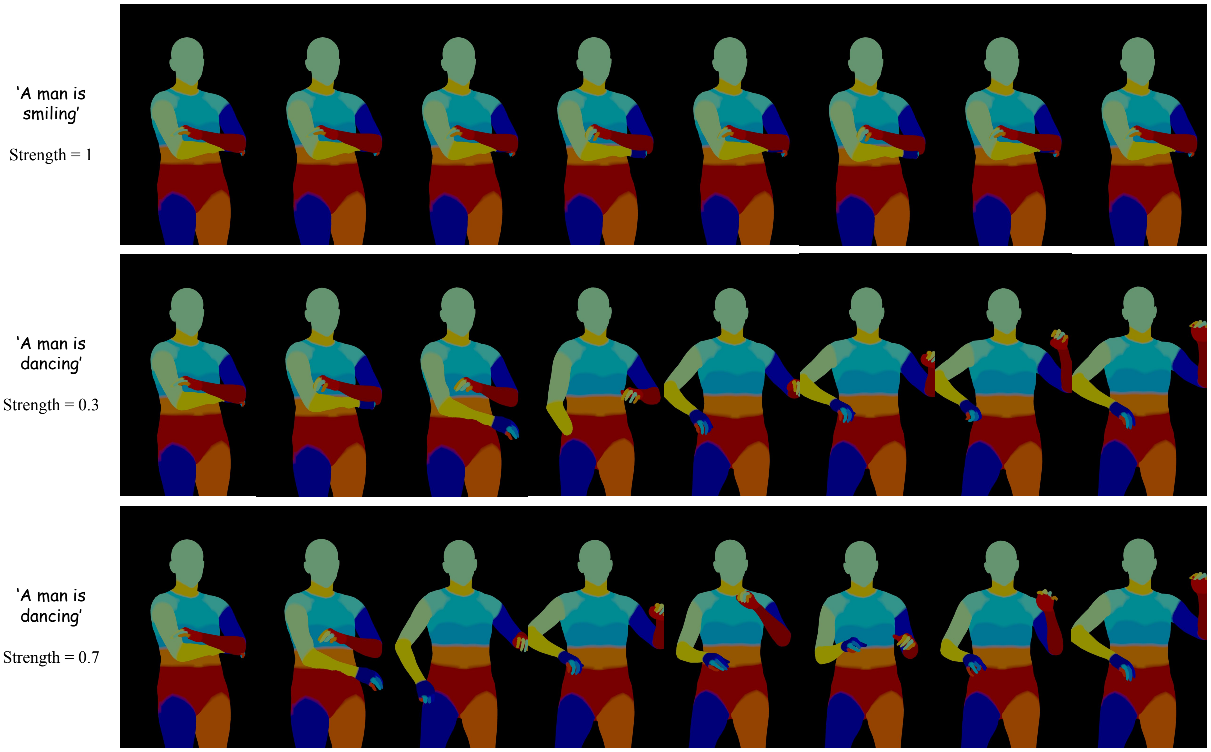

Figure 11: **Text prompt and motion strength.** We show projected motion sequences generated using different text prompts and motion strength parameters. The text prompt helps define the motion style, while the motion strength controls the speed and complexity of the generated motion. For example, a motion strength of 0.3 results in a simple, direct trajectory, whereas a strength of 0.7 produces a more dynamic and complex motion path. The same target pose is used for the second and third rows to highlight the effect of motion strength.

## E Statement of Broader Impact

The proposed ReVision has the potential to facilitate numerous fields through its advanced video generation capabilities. In the realm of creative industries, ReVision can enhance the efficiency and creativity of artists and designers by generating high-fidelity videos. The high-quality generated videos can also contribute to research on synthetic datasets by creating realistic videos, aiding in reducing the annotations required for training vision models. However, with these advancements come ethical considerations, such as the risk of generating deepfakes or other malicious content. It is thus crucial to implement safeguards to minimize potential harms.