PoseAnything: Universal Pose-guided Video Generation with Part-aware Temporal Coherence

Anonymous CVPR submission

Paper ID 2407

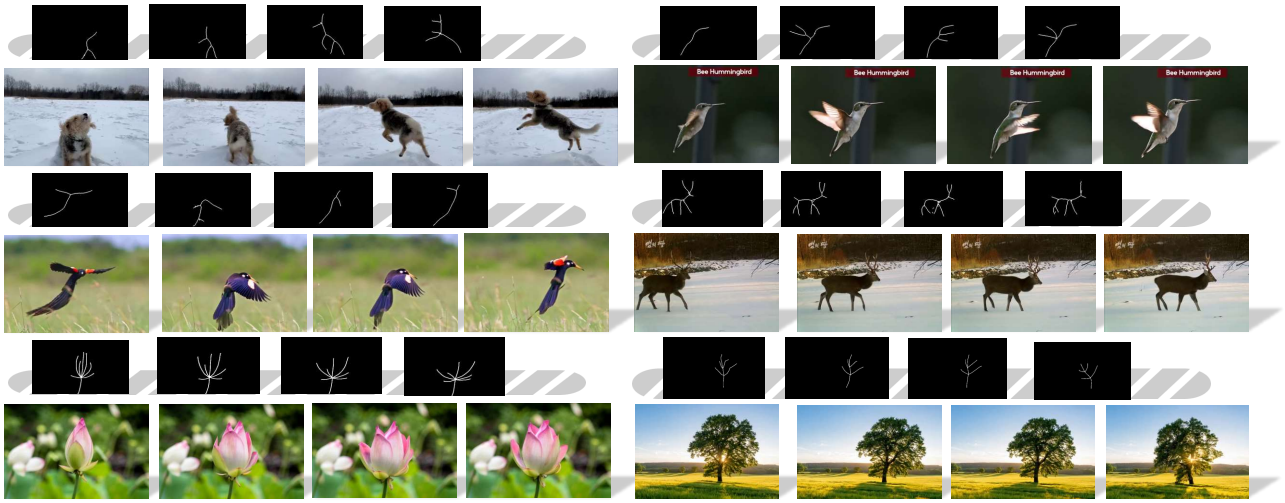

Figure 1. Animations produced by **PoseAnything**, which extends beyond human to non-human characters with various categories.

Abstract

*Pose-guided video generation refers to controlling the motion of subjects in generated video through a sequence of poses. It enables precise control over subject motion and has important applications in animation. However, current pose-guided video generation methods are limited to accepting only human poses as input, thus generalizing poorly to pose of other subjects. To address this issue, we propose **PoseAnything**, the first universal pose-guided video generation framework capable of handling both human and non-human characters, supporting arbitrary skeletal inputs. To enhance consistency preservation during motion, we introduce Part-aware Temporal Coherence Module, which divides the subject into different parts, establishes part correspondences, and computes cross-attention between corresponding parts across frames to achieve fine-grained part-level consistency. Additionally, we propose Subject and Camera Motion Decoupled CFG, a novel guidance strategy that, for the first time, enables independent camera movement control in pose-guided video generation, by separately injecting subject and camera motion control information into the positive and negative anchors of CFG. Furthermore, we present **XPose**, a high-quality public dataset containing 50,000 non-human pose-video pairs, along with an automated pipeline for annotation and filtering. Extensive experiments demonstrate that Pose-Anything significantly outperforms state-of-the-art methods in both effectiveness and generalization.*

1. Introduction

Pose-guided video generation refers to controlling the motion of subjects in video generation through a sequence of poses. By leveraging explicit pose information, it overcomes the limitations of traditional video generation methods, which often struggle to accurately and flexibly manipulate character poses and movements. It demonstrates substantial potential for a wide range of applications, includ-

ing entertainment video production, personalized animation and performance-driven avatar animation.

The rapid advancement of diffusion models has led to numerous methods for pose-guided video generation based on this architecture. For example, Disco [19] modifies Stable Diffusion and incorporates background features via ControlNet, but struggles to preserve detailed character features and suffers from inter-frame jitter. AnimateAnyone [4] introduces ReferenceNe to better retain character appearance and improve pose control and temporal coherence, yet generating natural and continuous movements remains challenging. Although extensive research has been conducted in pose-guided video generation, they only focus on human-pose driven video generation. Recent work likeAnimate-X [14] explores pose-guided video synthesis for non-human subjects by adapting human skeletons, but cannot accommodate diverse non-human skeletal structures. Currently, no method supports skeleton-driven video generation for arbitrary skeleton types.

Beyond pose-guided video generation, other studies have adopted other conditions to control the motion in the generated video. These include trajectory-based controllable generation methods (e.g., TORA [28], SG-I2V [9], ATI [17] and LeviTor [18]), which excel at guiding object position changes (*e.g.,* overall translation, scaling), but lack the granularity to precisely control pose variations and part-level movements. Sketch-based controllable generation methods (e.g., SketchVideo [17]), on the other hand, often require labor-intensive input and can struggle to maintain temporal consistency across frames, highlighting the need for a more versatile and precise control mechanism.

To address these challenges, we propose **PoseAnything**, the first unified framework that supports pose-guided video generation for both human and non-human characters, accommodating **universal skeletal inputs**. To address the limitations of current methods in maintaining appearance consistency during motion, we introduce the **Part-aware Temporal Coherence Module**. This module ensures fine-grained, part-level consistency by first partitioning the subject into distinct parts, establishing correspondences between them across frames, and then computing cross-attention exclusively among these matched parts, thus refining the control granularity to the part level and guaranteeing superior temporal coherence. Furthermore, we propose the **Subject and Camera Motion Decoupled CFG**, a novel guidance strategy that, for the first time, enables controllable camera movement in pose-guided video generation. By injecting subject and camera motion control information into the positive and negative anchors of CFG respectively, it effectively decouples the two processes, resolving the mutual interference that occurs when both are injected in a coupled manner. Extensive quantitative and qualitative experiments validate the effectiveness of our model, excelling in preserving appearance consistency while allowing flexible control over both subject and camera movement. In addition, to support the universal pose-guided video generation task, we release **XPose**, the first public high-quality non-human pose dataset, consisting of 50,000 non-human pose-video pairs. We design a pose extraction pipeline and a selection algorithm to extract precise and temporally continuous pose sequences from videos, providing a strong foundation for future research in related field.

The main contributions of our work are three-fold:

- We propose **PoseAnything**, the first unified framework that supports pose-guided video generation for both human and non-human characters, accommodating arbitrary skeletal inputs. In addition, we construct XPose, a high-quality public dataset comprising 50,000 non-human pose-video pairs, laying a solid foundation for future research in this field.
- We design a **Part-aware Temporal Coherence Module** to address the challenge of maintaining subject consistency during motion. The module ensures fine-grained, part-level consistency by part segmentation, establishing part correspondences, and part-aware cross-attention, thereby refining control granularity to a finer level.
- We propose **Subject and Camera Motion Decoupled CFG**, enabling camera motion control in pose-guided video generation for the first time. It effectively decouples the subject and camera motion by separately injecting their control conditions into the positive and negative anchors of CFG, eliminating their mutual interference.

2. Related Works

2.1. Diffusion Models for Video Generation

In recent years, diffusion models have achieved rapid development. Early methods, such as Stable Video diffusion [1], primarily used U-Net-based architectures with 3D convolutional layers for temporal modeling. Following the release of Sora [8], DiT-based approaches [10]) have increasingly replaced UNet-based methods in the field. Other DiT-based models like HunyuanVideo [6] Wan [16] MovieGen [11] utilize a 3D causal VAE [26] to handle the encoding and decoding of raw video data. Since these models acquire knowledge of inter-frame consistency and temporal continuity during pretraining, their application to character animation tasks enhances realism in generated characters and improves temporal coherence. Our Pose-Anything framework is built upon the open-source model Wan2.2-TI2V-5B [16], leveraging its robust pretrained capabilities to ensure high-quality visual generation from the outset.

2.2. Pose-guided Video Generation

Recent advances in human pose extraction, such as DWPose [24] and DensePose [3], enable human pose-guided

CVPR #2407

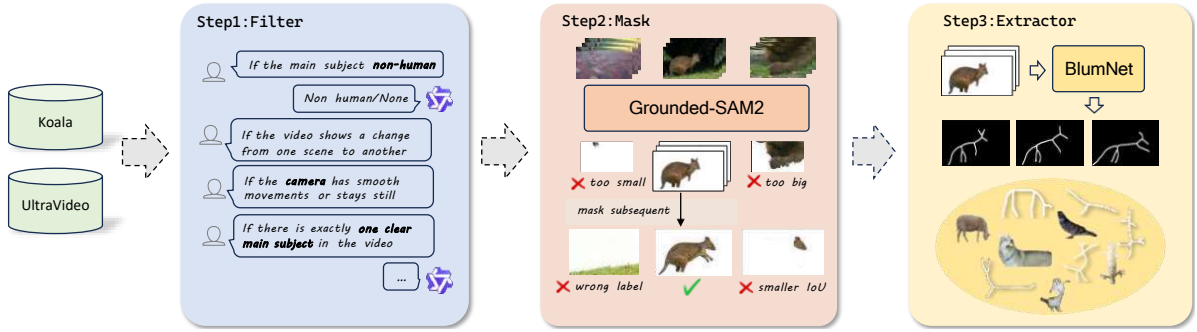

Figure 2. Construction pipeline of **XPose** dataset. This process consists of three stages: (1) selecting videos with **non-human** subjects from the Koala and UltraVideo datasets; (2) employing Grounded-SAM2 to segment the subjects in the videos and applying a **filtering algorithm** to select high-quality mask sequences; and (3) **extracting** poses of the masked subjects using BlumNet.

video generation. Early works like Disco [19] focus on dance actions but were limited in terms of generalization, appearance consistency, and inter-frame continuity. Following this, AnimateAnyone [4] incorporates the ReferenceNet architecture to integrate character appearance features, thereby achieving remarkable performance in consistency preservation. Additionally, temporal layers are employed to facilitate temporal modeling.

Although many studies have been conducted in pose-guided video generation, they focus on human-centric video synthesis. Some works like Animate-X [14] have attempted to generate videos with non-human characters. However, their driving factors are still limited to human skeletons and movements. To the best of our knowledge, our PoseAnything is the first controllable video generation model to enable pose-driven video synthesis for arbitrary subjects.

2.3. Universal Animation

Various methods have been proposed to inject control information for animation. Sketch-based approaches like VideoComposer [20] and SketchVideo [7], use pose sequences for motion guidance but are challenging to operate in practice. Beginning with DragNUWA [25], some studies have explored motion control based on trajectory conditions, such as TORA [28], SG-I2V [9], ATI [17], and LeviTor [18]. Nevertheless, these approaches struggle to capture fine-grained variations in subject pose. In contrast, our universal pose-guided video generation model, PoseAnything, not only enables flexible control over object positions but also provides precise manipulation of diverse poses.

3. XPose: A Universal Pose Dataset

Pose-driven video generation for arbitrary characters requires that includes a wide variety of subjects, however, such data are absent from existing public datasets. To address this gap, we release **XPose**, a high-quality public dataset comprising 50,000 **non-human** pose-video pairs. As the first dataset focused on non-human poses, XPose provides crucial data for pose-guided video generation with diverse entities, paving the way for future research and applications with a wider range of characters. As shown in Fig. 2, our construction pipeline consists of three stages: video filtering, subject masking, and pose extraction.

Stage 1: Video Filtering. To reduce noise in extracted skeletons, XPose focuses on videos featuring single non-human object. To filter out videos that fail to satisfy the criteria, we employ Qwen-2.5-VL-7B-Instruct [15] to filter samples from Koala [13] and UltraVideo [23] datasets.

Stage 2: Subject Masking. We utilize Grounded-SAM-2 [12]to generate segmentation masks for the primary object in each video. At this stage, we design an algorithm to filter out invalid skeletons ensuring the consistency of the extracted subject across frames. First, to ensure the completeness of the subject and valid motion information, we discard videos in which the mask region is excessively large or small. Specifically, we calculate the area S of the masked subject and retain videos where the ratio of S to the entire image $\frac{S_t}{H \times W}$ falls within the interval $(0.2, 0.8)$. Second, as the Grounded-SAM-2 model produces multiple object masks, we select the largest mask in the first frame and designate its corresponding subject as the target :

$$M_1^* = \arg \max_{M \in \mathcal{M}_1} \text{Area}(M). \quad (1)$$

For subsequent frames, we select masks with the same label as the first frame. If multiple candidates exist, we choose the one with the highest intersection-over-union (IoU) with the last selected mask to maintain temporal consistency:

$$M_t^* = \arg \max_{M \in \mathcal{M}_t, \text{label}(M)=\text{label}(M_1^*)} \text{IoU}(M, M_{t-1}^*). \quad (2)$$

If no valid mask is found, the frame is skipped.

Stage 3: Pose Extraction. Finally, we apply BlumNet [27] to extract skeletons from the masked images. If the number of frames with successfully extracted skeletons T_{skel} in a video is less than 80% of the total frames T, *i.e.*, $\frac{T_{skel}}{T} < 0.8$, the video is discarded. This data extraction

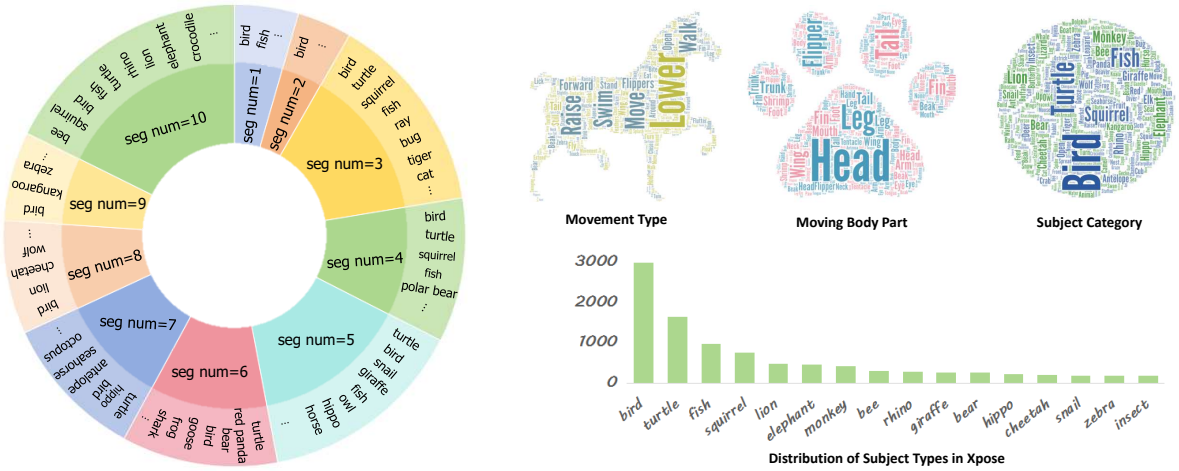

Figure 3. Comprehensive statistics of XPose in several aspects. The left shows the distribution of poses with different **numbers of segments** in XPose, as well as the corresponding **subject categories** for each segment count. The right illustrates the dataset distribution across three dimensions: **motion type**, **motion body part**, and **subject category**. As shown in the figures, our dataset exhibits good diversity.

pipeline and filtering algorithm effectively ensure the accuracy of the extracted skeleton information and the temporal consistency across frames, thereby laying a solid foundation for the construction of high-quality datasets.

Dataset Analysis. Fig. 3 presents a comprehensive statistical overview of XPose. The left panel shows the distribution of skeleton segment counts and corresponding subject categories. The word cloud illustrates the diversity and frequency of subject types, motion types, and motion parts, while the bar chart presents the distribution of subject categories. Together, these analyses demonstrate the richness and diversity of the XPose dataset.

4. Method

Pose-guided video generation takes a sequence of poses as input, alongside a reference image and a textual prompt, and aims to generate a video whose subject movement faithfully aligns with the specified pose sequence. In contrast to previous studies, our proposed **PoseAnything** is capable of accommodating **universal skeletal inputs**, including both non-human and human poses, which is the first work to accomplish this task. Moreover, we introduce **Part-aware Temporal Coherence Module**, a fine-grained mechanism for controlling appearance consistency across frames. This method involves partitioning the subject into multiple parts, establishing part correspondence and computing cross attention between matched parts across frames, thereby facilitating enhanced finer-grained part-level consistency control. Furthermore, we develop **Subject and Camera Motion Decoupled CFG**, a CFG-based decoupled control method for subjects and camera movements. It separately injects subject pose control conditions and camera motion control conditions into the positive and negative anchors, respectively, effectively mitigating mutual interference between the two types of motion conditions.

The overall framework of PoseAnything is illustrated in Fig. 4, based on Wan2.2-TI2V-5B [16]. We fuse the latent representation of the reference image and pose by concatenating them along the channel dimension as the input of DiTBlock. The Part-aware Temporal Coherence Module is incorporated after the original cross-attention layer within each DiTBlock, with the aim of enhancing appearance consistency at a finer-grained level (detailed in Sec. 4.2).

4.1. Analysis of Condition Injection Strategies

We utilize Wan2.2-TI2V-5B [16] as our base model for image to video generation. As shown in Fig. 4, the original model takes a reference image I_r and employs the pretrained Wan2.2VAE to encode it into a latent representation Z_i. Z_i is concatenated with noise latent ϵ along the temporal dimension and patchified to form the input Z_0 of the DiTBlock. For pose-guided generation, an additional pose sequence P is taken as input. For easy integration, we encode the pose sequence P into pose latent presentations Z_p, also using the pretrained Wan2.2 VAE. To effectively incorporate skeletal information, we compare three different injection strategies: concatenation by channels, multi-layer perceptron (MLP), and concatenation by width.

Strategy 1: Concatenation by channel. Given the initial latent Z_i and the pose latent Z_p, we first concatenate Z_i with a noise map ϵ to get $Z_0 = [Z_i, \epsilon]$. Next, Z_0 and Z_p are concatenated along the channel axis to obtain the aggregated latent Z_{agr}. In the patchify module, we increase the number of input channels for the convolutional layers to accommodate the additional skeleton dimensions, while maintaining the channel number of input of DiT block con-

CVPR
#2407

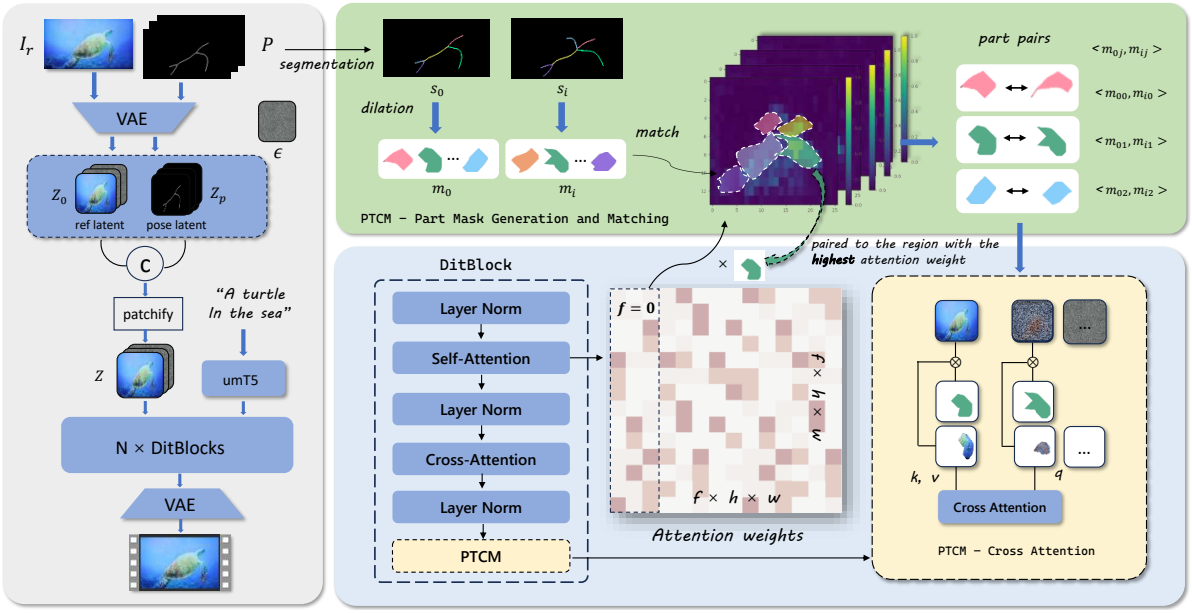

Figure 4. Overview of our **PoseAnything**. Given a reference image I_r and a pose sequence P, we first encode P into pose latent Z_p, and then concatenate it with the latent Z_0 of I_r along the channel dimension. Additionally, we propose **Part-aware Temporal Coherence Module** for fine-grained appearance consistency control: 1) We segment the pose into separate segments s_{ij} and dilate each segment to obtain the subject part masks m_{ij}; 2) We then use attention patterns to match the same parts across different frames; 3) For each pair $< m_{0j}, m_{ij} >$, we introduce a part-aware cross-attention module in the DiTBlock to compute cross-attention between matched parts. By performing consistency control at the part level, the part-aware coherence module achieves enhanced subject appearance consistency in a fine-grained manner.

sistent with the original Wan model:

$$Z_{agr} = [Z_0, Z_p] \in F \times H \times W \times 2C,$$
$$Z = Conv(Z_{agr}) \in f \times h \times w \times c. \tag{3}$$

Strategy 2: Multi-layer Perceptron. Z_p is converted to the same shape as Z_0 using a MLP. The resulting features are then fused with the initial latent Z_0 by element-wise addition, yielding the DiT model input Z:

$$Z = Z_0 + \text{MLP}(X_p). \tag{4}$$

Strategy 3: Concatenation by width. Z_0 and Z_p are concatenated along the width dimension to form an aggregated latent Z, which is directly used as the input of DiT block:

$$Z = \text{Concat}_{\text{width}}(Z_0, Z_p) \in F \times H \times 2W \times C. \tag{5}$$

Comparison of Injection Strategies. Our experimental results demonstrate that channel-wise conditioning methods exhibit significant advantages in pose-guided video generation (presented in #Suppl). Consequently, we employ this strategy to inject skeletal information into our model, enabling more effective utilization of poses.

4.2. Part-aware Temporal Coherence Module

In pose-guided video generation, existing methods often struggle to maintain the consistency of the object's appearance throughout the motion, especially during large pose changes. Previous works attempt to address this issue by utilizing ControlNet or cross-attention mechanisms to capture the overall appearance of the object in the reference image. However, these approaches still face inconsistencies or distortions in fine details. To tackle this, we propose a **Part-aware Temporal Coherence Module (PTCM)**. We divide the object into multiple smaller parts, utilize attention weight to match corresponding parts across different frames, and perform cross-attention between the matched parts. By decomposing overall appearance consistency into *finer-grained part-level consistency control*, our method achieves superior performance in maintaining temporal coherence. As shown in Fig. 4, the Part-aware Temporal Coherence Module (PTCM) consists of three steps.

Part Mask Generation. We first segment each pose into segments, denoted as s_{ij}, where i denotes the frame index corresponding to the skeleton image, and j denotes the index of the segment of the current pose. To obtain the pixels m_{ij} corresponding to s_{ij}, we dilate each s_{ij} by α:

$$m_{ij} = \text{Dilate}(s_{ij}, \alpha), \tag{6}$$

where the expansion coefficient α is calculated by continuously dilating the skeleton until it can cover the main body

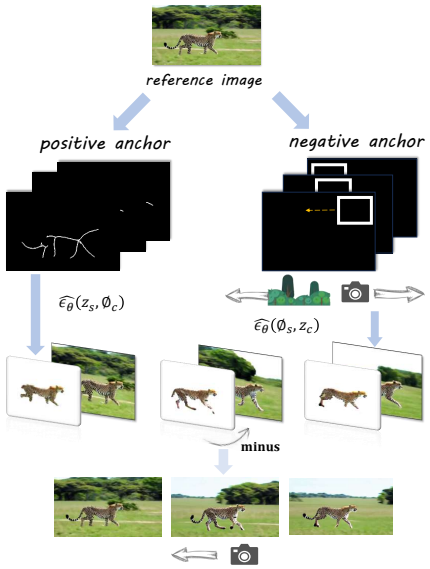

Figure 5. Subject and Camera Decoupled Control Based on CFG.

in the reference image:

$$\alpha_{ij} = \min\left\{\alpha, 100 \;\middle|\; \mathrm{IoU}\left(\mathrm{Dilate}(s_{ij}, \alpha), \mathrm{Body}_{ij}^{\mathrm{ref}}\right) \geq 1\right\}. \tag{7}$$

Part Matching using Attention Patterns. Next, we establish correspondences between parts across frames. Based on the observation that the attention weight between the same parts in different frames is higher than that between different parts, we match each part in the first frame to its counterpart in subsequent frames by:

$$s_{ij'} \sim s_{0j} \iff j' = \arg\max_t \ \text{attn\_weight}[m_{0j}][m_{it}]. \tag{8}$$

In implementation, we first perform several steps of inference to compute the attention weights between the first frame and subsequent frames. Then, using the aforementioned method, we match the masks of the first frame to those of the subsequent frames based on these attention weights, as shown in Fig. 4.

Part-aware Cross Attention. For each pair $< s_{0j}, s_{ij} >$, we calculate cross-attention by calculating K and V using the tokens corresponding to s_{0j} in the first frame, and calculate Q using the tokens corresponding to s_{ij} in subsequent frames:

$$\begin{aligned} &x' = x + \text{Cross-Atten}(Q_j, K_j, V_j), \\ &Q_j = m_{ij} X W_q, K_j = m_{0j} X_0 W_k, V_j = m_{0j} X_0 W_v. \end{aligned} \tag{9}$$

This module is inserted after the final cross-attention layer in the DiT block, as shown in Fig. 4.

4.3. Subject and Camera Motion Decoupled CFG

Current pose-guided video generation methods are limited to controlling object motion and do not support camera motion control. Other video generation approaches, such as SG-I2V (based on drag), inject both object and camera motion control conditions simultaneously. Such a coupled injection strategy often results in mutual interference between the two control conditions, hindering the model's ability to comprehensively represent both types of motion information. To address this issue, we ingeniously leverage the positive and negative anchors of classifier-free guidance (CFG) to decouple subject and camera motion control (Fig. 5). This enables a complete separation of the two control conditions and effectively prevents mutual interference.

Decoupled Subject and Camera Motion via CFG. In practice, we find that although our model is trained on subject motion control, it can be generalized to control camera motion as well. However, directly injecting both subject and camera motion control conditions simultaneously leads to mutual interference between the two control signals. To tackle this, we propose decoupled subject and camera motion control via classifier-free guidance: injecting the subject motion control conditions (pose sequence) into the positive anchors of CFG, while injecting the camera motion control conditions into the negative anchors. The underlying principle is illustrated as follows:

$$\begin{aligned} \tilde{\epsilon} &= \hat{\epsilon}_\theta(\emptyset_s, z_c) + s \cdot (\hat{\epsilon}_\theta(z_s, \emptyset_c) - \hat{\epsilon}_\theta(\emptyset_s, z_c)) \\ &= \hat{\epsilon}_\theta(z_s, \emptyset_c) - \hat{\epsilon}_\theta(\emptyset_s, \emptyset_c) - s \cdot [\hat{\epsilon}_\theta(\emptyset_s, z_c) - \hat{\epsilon}_\theta(\emptyset_s, \emptyset_c)] \\ &= (1+s) \cdot \hat{\epsilon}_\theta(\emptyset_s, \emptyset_c) + \hat{\epsilon}_\theta(zt_s, \emptyset_c) + s \cdot \hat{\epsilon}_\theta(\emptyset_s, zt_c), \end{aligned} \tag{10}$$

where Z_c denotes the latents injected with camera motion information, and Z_s denotes the latents injected with subject motion information.

Camera Control via Negative Anchors. Our key idea is to use the negative anchor in CFG to steer the generation away from specific camera states, thereby achieving camera movement control. This requires the control signal injected into the negative anchor to be *opposite* to be the target camera motion. For instance, to generate a leftward camera movement (where the background should move rightward), we generate a skeleton sequence with a rectangle that moves *leftward* and inject it into the negative anchor as shown in Fig. 5. Injecting this "left-moving" negative signal prompts the model to produce a rightward background flow – achieving the desired leftward camera pan. This decoupled CFG design effectively enables precise and independent control over both the subject and the camera motion.

5. Experiments

Implementation Details. We utilize the XPose dataset alongside 15,000 internal human videos as training set. The training process is divided into three stages. In the first stage, we train a baseline model without the part-aware temporal coherence module solely on the human dataset for 3k iterations, with batch size set to 32 and learning rate 5e-5.

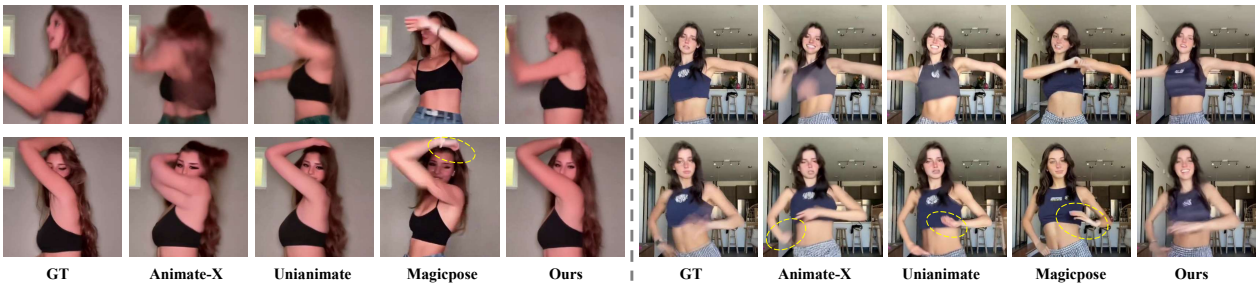

Figure 6. Quantitative comparison between the state-of-the-arts and Ours on TikTok dataset.

Table 1. Quantitative comparisons with the state-of-the-arts on TikTok dataset (Human).

Method	PSNR↑	SSIM↑	L1↓	LPIPS↓	FVD↓
Disco [19]	29.03	0.668	3.78E-04	0.292	292.8
MagicAnimate [22]	29.16	0.714	3.13E-04	0.239	179.07
MagicPose [2]	29.53	0.752	0.81E-04	0.292	-
AnimateAnyone [4]	29.56	0.718	-	0.285	171.9
Champ [29]	29.91	0.802	2.94E-04	0.234	160.82
Unianimate [21]	30.77	0.811	2.66E-04	0.231	148.06
Animate-X [14]	30.78	0.806	2.70E-04	0.232	139.01
PoseAnything	**31.5**	**0.8362**	**2.79E-05**	**0.224**	**133.95**

Table 2. Quantitative comparison between the state-of-the-arts and Ours on XPose-benchmark (Non-human).

Method	PSNR↑	SSIM↑	L1↓	LPIPS↓	FVD↓
Tora [28]	30.08	0.6929	9.38E-06	0.3530	103.75
ATI [17]	30.15	0.6810	9.59E-06	0.3706	101.44
SG-I2V [9]	29.86	0.6634	1.28E-05	0.3674	102.97
PoseAnything	**30.29**	**0.7114**	**8.19E-06**	**0.3241**	**99.97**

In the second stage, we further trained the above model with human and non-human mixed data with the same batch size and learning rate. In the third stage, we exclusively train the part-aware temporal coherence module on the mixed dataset while keeping all other modules frozen for 8k iterations, with batch size set to 32 and learning rate 1e-5. All experiments are performed on an NVIDIA H20 80GB GPU.

Evaluation Details. To comprehensively evaluate the model's performance on both human and non-human data, we conduct qualitative and quantitative experiments separately. The generated videos are assessed using five standard metrics: (1) PSNR, (2) SSIM, (3) L1 distance, (4) LPIPS, and (5) FVD.

5.1. Experiment Settings

5.2. Human Pose-Guided Generation

To validate the effectiveness of our method on human data, we conduct experiments on the widely-used benchmark, TikTok [5]. To ensure a fair comparison, we separately train our Pose Anything for 1,500 iterations exclusively on the training set of the TikTok dataset. Both qualitative and quantitative experiments are then conducted on the test split. We compare our model with several state-of-the-art methods, including Disco [19], MagicAnimate [22], Animate Anyone [4], Champ [29], Unianimate [21], Animate-X [14]. **Quantitative results.** The quantitative comparison results between our method and the state-of-the-arts on Tiktok are reported in Tab. 1. PoseAnything achieves the best performance across all metrics. **Qualitative results.** Fig. 6 shows qualitative comparison results of our approach with UniAnimate, MagicPose, and Animate-X. It can be observed that while the results generated by other methods exhibit obvious distortions, PoseAnything demonstrates excellent motion alignment and appearance consistency. *Video comparisons are presented in #Suppl.*

5.3. Non-human Pose-guided Generation

As there is no existing universal pose-guided video generation, we compare our method with controllable generation approaches based on drag-and-control methods, including ATI [17], Tora [28], and SG-I2V [9]. We conduct comparison experiments on 51 videos randomly selected from XPose. For a fair comparison, these videos were not used during training. **Quantitative results** are shown in Table 2, demonstrating that our PoseAnything achieves the best performance in non-human pose-guided generation. **Qualitative results** are presented in Fig. 7, from which we can observe that ATI, SG-I2V, and Tora fail to achieve accurate object pose control. Furthermore, when handling large-magnitude motions, these approaches often result in object deformation and background artifacts. In contrast, our PoseAnything model generates accurate object motions based on skeletal guidance, while simultaneously preserving the integrity and realism of both the object and the background. *Video comparisons are presented in #Suppl.*

5.4. Camera Motion Control

To validate the effectiveness and robustness of our **Subject and Camera Motion Decoupled CFG**, we designed a challenging set of experiments. In this setup, we task the model with handling two distinct and concurrent motion signals: the subject is driven by its own dynamic pose

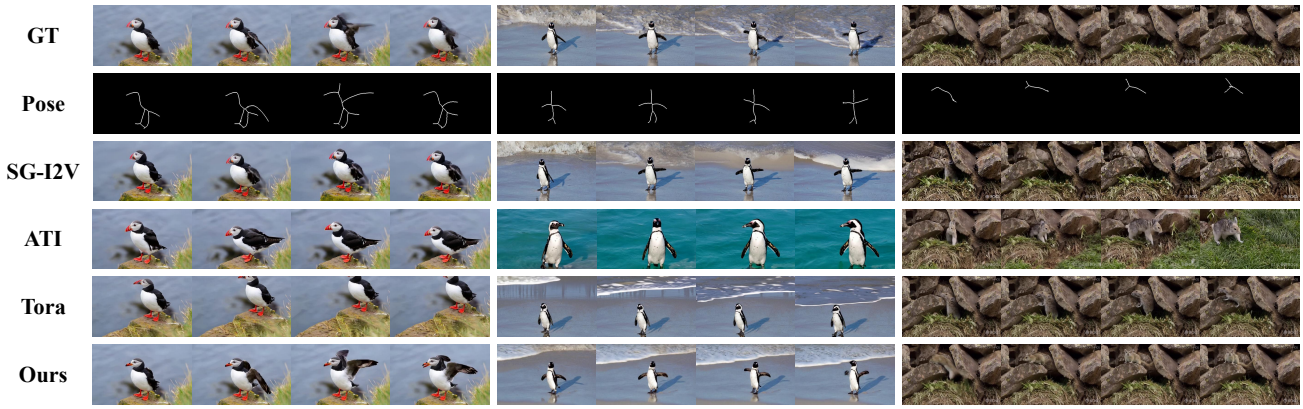

Figure 7. Qualitative comparison with existing state-of-the-art methods on XPose-benchmark.

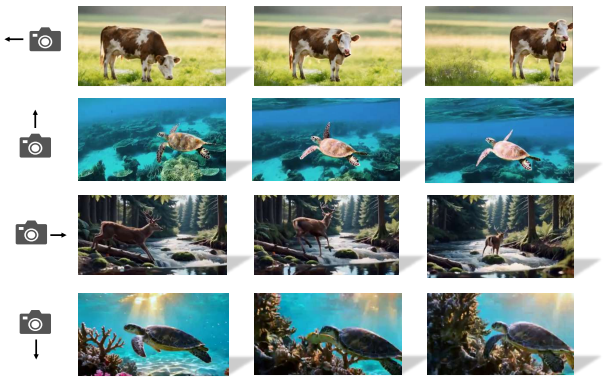

Figure 8. Demonstration of Camera Control Cases.

sequence, while a separate camera motion command (e.g., pan left, tilt up) is simultaneously injected into the negative anchor of the CFG. This scenario directly tests our core claim of preventing mutual interference between subject and camera controls. The qualitative results, presented in Fig. 8, showcase the remarkable success of our approach. The subject accurately performs its intended actions according to the pose guidance, while the camera simultaneously executes the specified movement smoothly and coherently. The ability to maintain high fidelity for both the subject's action and the global camera motion provides strong empirical evidence that our method effectively disentangles the two control signals, achieving the precise and independent manipulation it was designed for.

5.5. Ablation Study

Ablation on Part-aware Temporal Coherence Module. To assess the contribution of the Part-aware Temporal Coherence Module (PTCM), we conducted an ablation study on XPose. Specifically, we compared the baseline model, which employs only concatenation for pose injection, with the full model integrating the PTCM module. Furthermore, we evaluated a configuration in which cross-attention is computed over the entire object region (EC) without part segmentation and matching. Quantitative results are summarized in Tab. 3. Results show that the model without the PTCM module has poorer performance. Furthermore, omitting part segmentation and matching also leads to a degradation in model performance, which effectively validates the contribution of the PTCM module.

Table 3. Quantitative results of ablation study.

Method	PSNR↑	SSIM↑	LPIPS↓	L1↓	FVD↓
Concat	29.85	0.6964	0.3304	9.43E-06	102.30
EC	30.27	0.7107	0.3243	**8.15E-06**	101.50
PTCM	**30.29**	**0.7114**	**0.3241**	8.19E-06	**99.97**

6. Conclusion

In this work, we introduce PoseAnything, the first unified framework supporting arbitrary skeletal inputs for pose-guided video generation. To address the challenge of maintaining consistent object appearance throughout motion sequences, we propose a Part-aware Temporal Coherence Module that enables fine-grained, controllable appearance consistency at the part level. It divides the subject into different parts, establishes part correspondences, and computes cross-attention between corresponding parts across frames to achieve fine-grained part-level consistency. We are also the first to incorporate camera control by decoupling subject and camera motions through separate conditioning branches in classifier-free guidance, which enables a complete separation of the two control conditions and effectively prevents mutual interference. Additionally, we present a novel pipeline and filtering algorithm for extracting skeletal representations from various objects and release a high-quality dataset of 50,000 non-human pose-video pairs. Extensive quantitative and qualitative experiments demonstrate that PoseAnything outperforms the state-of-the-art methods and generalizes well across diverse subjects and poses.

CVPR #2407

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 2

[2] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 7

[3] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7297–7306. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[4] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8153–8163. IEEE, 2024. 2, 3, 7

[5] Yasamin Jafarian and Hyun Soo Park. Self-supervised 3d representation learning of dressed humans from social media videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8969–8983, 2023. 7

[6] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. 2

[7] Feng-Lin Liu, Hongbo Fu, Xintao Wang, Weicai Ye, Pengfei Wan, Di Zhang, and Lin Gao. Sketchvideo: Sketch-based video generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 23379–23390. Computer Vision Foundation / IEEE, 2025. 3

[8] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models. *CoRR*, abs/2402.17177, 2024. 2

[9] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. SG-I2V: self-guided trajectory control in image-to-video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2, 3, 7

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. 2

[11] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam S. Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali K. Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dmitry Vengertsev, Edgar Schönfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *CoRR*, abs/2410.13720, 2024. 2

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 3

[13] Reuben Tan, Ximeng Sun, Ping Hu, Jui-Hsien Wang, Hanieh Deilamsalehy, Bryan A. Plummer, Bryan C. Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13581–13591. IEEE, 2024. 3

[14] 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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2, 3, 7

[15] Qwen Team. Qwen2.5-vl, 2025. 3

[16] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng

Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025. 2, 4

[17] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: any trajectory instruction for controllable video generation. *CoRR*, abs/2505.22944, 2025. 2, 3, 7

[18] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. Levitor: 3d trajectory oriented image-to-video synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 12490–12500. Computer Vision Foundation / IEEE, 2025. 2, 3

[19] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9326–9336. IEEE, 2024. 2, 3, 7

[20] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3

[21] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *CoRR*, abs/2406.01188, 2024. 7

[22] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 1481–1490. IEEE, 2024. 7

[23] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao. Ultravideo: High-quality UHD video dataset with comprehensive captions. *CoRR*, abs/2506.13691, 2025. 3

[24] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 4212–4222. IEEE, 2023. 2

[25] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[26] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: masked generative video transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10459–10469. IEEE, 2023. 2

[27] Yulu Zhang, Liang Sang, Marcin Grzegorzek, John See, and Cong Yang. Blumnet: Graph component detection for object skeleton extraction. pages 5527–5536, 2022. 3

[28] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 2063–2073. Computer Vision Foundation / IEEE, 2025. 2, 3, 7

[29] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LV*, pages 145–162. Springer, 2024. 7