This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centric Video Generation

Ruineng Li[*,1], Daitao Xing[*,1], Huiming Sun[1], Yuanzhou Ha, Jinglin Shen[1], Chiuman Ho[1]
[1]OPPO US AI Center
{ruineng.li2, daitao.xing1, huiming.sun2, jinglin.shen, chiuman}@oppo.com, hayuanzhou0313@gmail.com

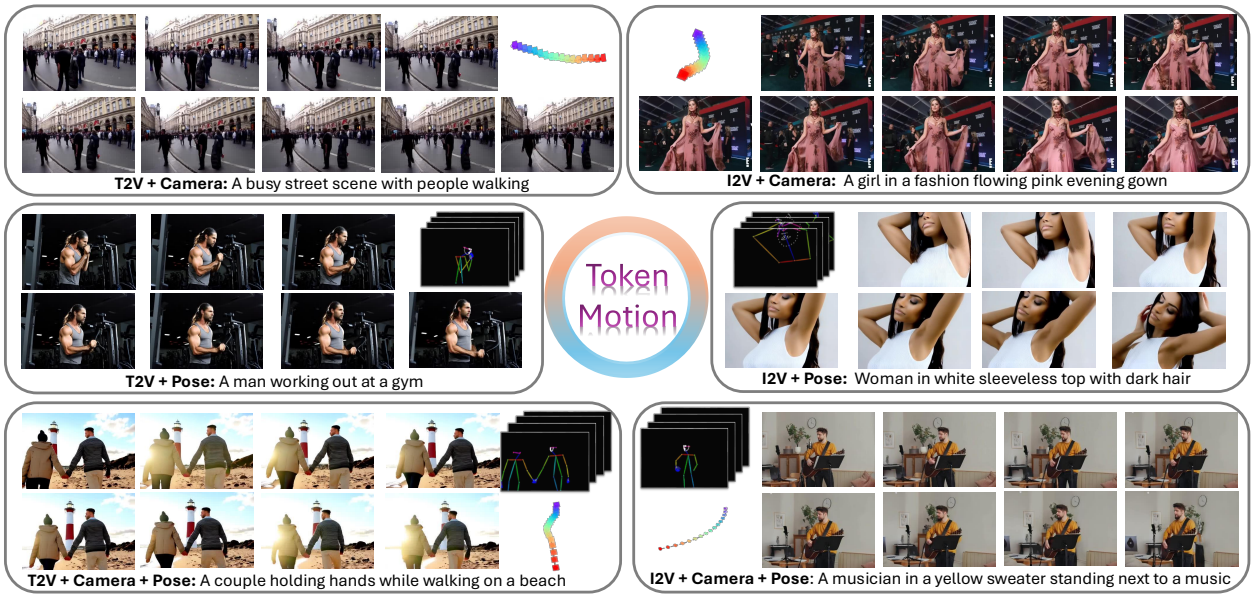


Figure 1. *TokenMotion* is a transformer-based video generation framework that enables simultaneous control of camera trajectories and human kinematic patterns. The framework demonstrates versatility across both text-to-video and image-to-video generation paradigms, while supporting flexible control configurations. *Text prompts are abbreviated for conciseness.

Abstract

Human-centric motion control in video generation remains a critical challenge, particularly when jointly controlling camera movements and human poses in scenarios like the iconic Grammy Glambot moment. While recent video diffusion models have made significant progress, existing approaches struggle with limited motion representations and inadequate integration of camera and human motion controls. In this work, we present TokenMotion, the first DiT-based video diffusion framework that enables fine-grained control over camera motion, human motion, and their joint interaction. We represent camera trajectories and human poses as spatio-temporal tokens to enable local control granularity. Our approach introduces a unified modeling framework utilizing a decouple-and-fuse strategy, bridged by a human-aware dynamic mask that effectively handles the spatially-and-temporally varying nature of combined motion signals. Through extensive experiments, we demonstrate TokenMotion's effectiveness across both text-to-video and image-to-video paradigms, consistently outperforming current state-of-the-art methods in human-centric motion control tasks. Our work represents a significant advancement in controllable video generation, with particular relevance for creative production applications.

[*]Equal contribution.

1. Introduction

Video diffusion models have achieved significant progress in text-to-video [6, 11, 21, 36] and image-to-video [2, 7, 42] generation, with recent works incorporating external control signals such as depth maps [12, 35], subject appearance[39, 53] and motion [15, 38, 41, 49] for better generation controllability. Among these, human-centric motion control emerges as a crucial challenge in video generation. Consider the iconic Grammy Glambot moment, which requires the joint control of both dramatic camera movements and human pose changes - exactly the type of control we aim to achieve. The importance of this capability stems from the explosive growth of AI-generated content and its great potential in creative production, from films to viral short-form videos. However, this critical aspect remains largely unexplored, with existing works providing only limited solutions.

Such limitations in controlling video generation with human-centric motion conditions fall into two aspects. First, most current works in motion control focus only on either object motion control [34, 49] or camera control [1, 15, 43], lacking the capability of joint control. Second, while few works [38, 39, 46] have attempted to explore jointly controlling camera motion and object motion, they remain limited for human-centric motion control, due to over-simplified motion representations and ad-hoc motion integration: MotionCtrl [38] uses object-level keypoints, while Direct-A-Video[46] and MotionBooth[39] uses bounding boxes for object trajectories. These coarse representations fail to capture local movements, especially subtle pose changes. Furthermore, these works all directly integrate two motion signals without handling their potential interactions, and report the presence of motion conflicts in complex joint-control scenarios, indicating this challenge remains unsolved for human-centric motion control. While HumanVid [37] shares a similar goal, it focuses on dataset construction and provides only a simple baseline for dataset validation, leaving the technical challenges unaddressed.

A key challenge exists for effectively integrating camera and human motion controls. Intuitively, given any pair of video frames, each pixel's motion arises from different combination of camera and human movements. This spatial-and-temporally-varying nature of motion composition necessitate unified modeling of both motion signals in the same representation space, along with explicit designs to capture their interactions. However, existing approaches lack such dedicated designs. Moreover, current joint-control methods adopt UNet backbones, which separate object and camera motions into spatial and temporal modeling, respectively, making it suboptimal for joint encoding. Despite DiT-based [27] alternatives seem a promising choice, joint-control DiT-based models have not yet been proposed. Besides, recent camera control works [1, 8] find that previous motion encoding techniques require tailored adaptation for effectively working on DiT-based models, a challenge likely amplified for joint motion control.

In this work, we present *TokenMotion* to address all these challenges, which is the first DiT-based video diffusion model that enables human-centric motion control in video generation. TokenMotion achieves fine-grained control over all three fundamental scenarios of human-centric motion control: camera-only, human-motion-only, and their joint control. At the core of our approach is a novel unified modeling framework that handles camera and human motion through a decouple-and-fuse strategy, bridged by a human-aware dynamic mask for effectively integrating two motion signals. Furthermore, to address the challenge of degraded control accuracy in DiT-based frameworks, we introduce motion patchification that encodes motion as fixed-length sequences. We also demonstrate the generalization capabilities and superiority of our approach across both paradigms of text-to-video and image-to-video, consistently outperforming current state-of-the-art methods. We claim three main contributions summarized as follows:

- TokenMotion, the first DiT-based video diffusion framework that enables fine-grained control over camera motion, human motion, and their joint interaction, serving a critical problem in AI-generated video content.
- A unified modeling framework with a decouple-and-fuse strategy, bridged by a human-aware dynamic mask that effectively handles the spatially-and-temporally-varying nature of combined camera and human motion signals.
- We demonstrate the generalization capabilities of our approach across both text-to-video and image-to-video paradigms through extensive experiments, consistently outperforming current state-of-the-art methods in human-centric motion control.

2. Related Works

2.1. Video Diffusion Models

The domain of video generation [2, 3, 10, 19, 21, 36] is undergoing remarkable evolution driven by the advanced generative power of diffusion models [18, 28, 29]. Current approaches primarily utilize UNet-based diffusion backbones to generate videos, including works in text-to-video (T2V) generation [7, 13, 21, 36], and image-to-video generation (I2V) [2, 42, 52]. To further improve the scalability and the generation capabilities of diffusion models, more recent works are shifting their attention towards transformer-based backbones [4, 14, 23, 27, 48]. One of the most recent works, CogVideoX [48] further advances by directly operate in the 3D latent space through the encoding of a 3D Variational Autoencoder (VAE), and a progressive training technique to generate high-quality videos. Besides, another significant line of research focuses on enhancing the generation controllability of video diffusion models through control signals

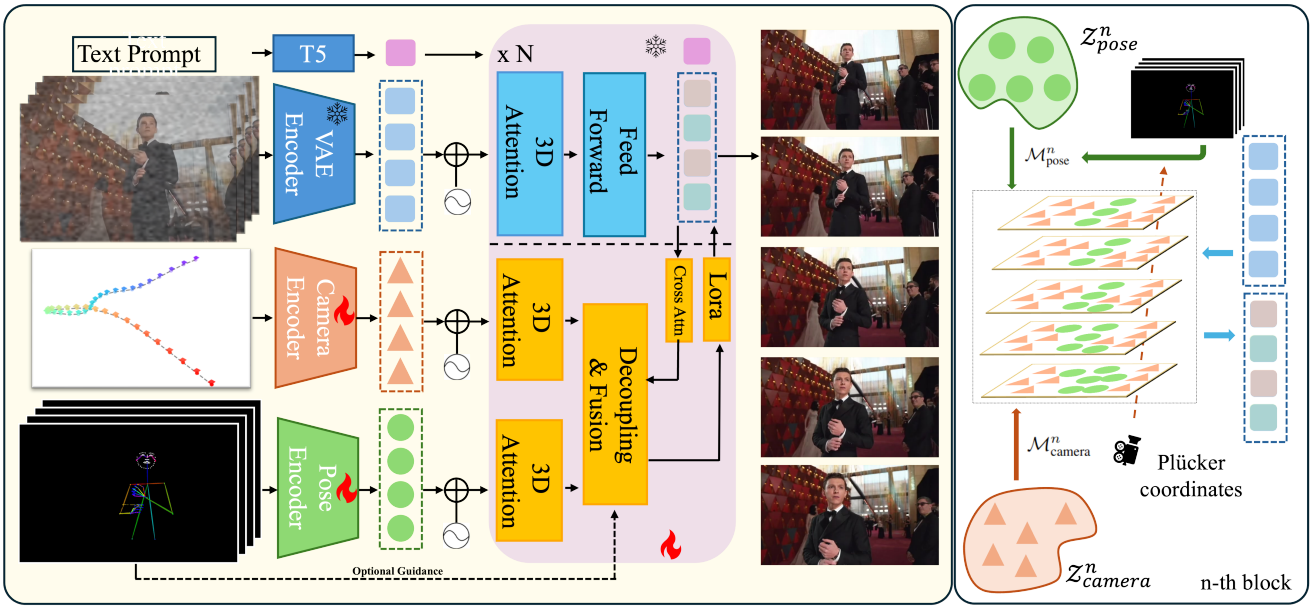


Figure 2. **Architectural Overview.** TokenMotion presents a novel video generation framework that combines a transformer-based video diffusion model with content-aware motion guidance. The architecture employs dual motion encoders that extract spatio-temporal motion tokens. These motion features are then processed through a specialized decoupling and fusion module, which dynamically modulates the strength of motion guidance based on content characteristics, enabling fine-grained control over temporal consistency.

such as dense map [9, 45], subject appearance [5, 26, 40] and motion [15, 20, 34, 38, 49].

2.2. Motion Control in Video Generation

Recent works in motion control introduce different types of motion signals to gain more accurate controllability in video generation, especially for object motion [20, 24, 34, 41, 44, 46, 49, 56] and camera motion [1, 8, 15, 43, 54]. CameraCtrl [15] pioneered in using Plücker Embedding [31] instead of start-and-end camera shifts [38, 39], enabling fine-grained control over intermediate camera poses throughout the trajectories. MagicAnimate [44] and AnimateAnyone [20] utilizes skeletons to enable subtle pose changes beyond object-level motion. However, most works only focus on either type of motions, limiting the scope of generation controllability. Some recent studies take first steps toward jointly controlling object motion and camera motion [38, 39, 46]. MotionCtrl [38] and Direct-A-Video [46] adopt a data-driven decoupling approach, which use different data to learn different motion concepts. However, both methods claim to show motion conflicts. Related to these works is ImageConductor[25], which decouples camera motion and object motion for better individual control due to their entanglement in training data, but doesn't explore their joint operation.

Another line of research proposes that the limited motion understanding these current models show originates from the inferiority of their UNet-based backbones [2, 7, 13], which are claimed to have weaker scalability and limited generative power compared to DiT-based counterparts [4, 27]. Recently, VD3D [1] introduced the first camera-controlled DiT-based models. Currently, joint-control DiT-based models have not yet been proposed.

3. Methodology

3.1. Preliminary

Recent developments in transformer-based diffusion frameworks [23, 48] have markedly advanced the field of video generation, particularly through the synergistic integration of 3D Causal Variational Autoencoders (VAE) with sophisticated spatiotemporal (3D) attention mechanisms. Specifically, the 3D Causal VAE converts a video clip of shape $T \times H \times W \times C$ into sequences of visual tokens z_{visual} of length $\frac{T}{q} \cdot \frac{H}{p} \cdot \frac{W}{p}$, where H, W, and T denote the height, width, and length of video, and p and q are spatial and temporal compression ratios, respectively. z_{visual} is flattened into 1-dimensional vectors and fed into N stacked 3D fully attention blocks, together with positional encoding, text prompt encoding, and timestamps for the denoising process.

The video diffusion framework learns a denoiser model $D_{\boldsymbol{\theta}}$ which predict the clean visual tokens from noise inputs $\tilde{z} = z + \sigma\varepsilon, \sigma \in \mathbb{R}_+, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ following the guidance including text prompt c and additional structure control signals s (i.e. camera trajectories and human poses). Consequently, the denoiser model $D_{\boldsymbol{\theta}}$ is parametrized as $D_{\boldsymbol{\theta}}(z_{\text{visual}}; \sigma) = c_{\text{skip}}(\sigma) z_{\text{visual}} +$



Figure 3. Visualization of joint-control video generation results from Direct-A-Video [46], MotionCtrl [38], MotionBooth [39] and our TokenMotion-T. Above cases shows that our TokenMotion method succeeds in jointly handling controls of both human motion and camera motion, while being consistently aligned with the input prompts at the same time.

$c_{\text{out}}(\sigma) F_{\boldsymbol{\theta}}\left(c_{\text{in}}(\sigma) z_{\text{visual}}; c_{\text{noise}}(\sigma)\right)$ and $F_{\boldsymbol{\theta}}$ is network to be trained to optimize the objective :

$$\mathbb{E}_{(z,c,s)}\left[\lambda_\sigma \left\| D_{\boldsymbol{\theta}}\left(\tilde{z}; \sigma, c, s\right) - z \right\|_2^2\right] \tag{1}$$

3.2. Motion Control

3.2.1 Camera Motion Representation

We follow CameraCtrl [15] to adopt Plücker embeddings to represent camera pose conditions. For each video clip, its camera condition comprises a sequence of camera parameters, each consisting an intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, and an extrinsic matrix $\mathbf{P} = [\mathbf{R}; \mathbf{t}] \in \mathbb{R}^{3\times4}$, where $\mathbf{R}$ and $\mathbf{t}$ denote rotation and translation components, respectively. To effectively anchor the raw values of camera poses within the pixel coordinates, we view camera control signals as camera rays originating from the camera center, represented with Plücker coordinates given some $\mathbf{K}$ and $\mathbf{P}$. Mathematically, For a random pixel (u, v) in the f-th frame, its corresponding ray $\mathbf{p}_{u,v,f}$ in the Plücker coordinates can be represented as:

$$\mathbf{p}_{u,v,f} = \frac{(\mathbf{d}_{u,v,f}, \mathbf{t_f} \times \mathbf{d}_{u,v,f})}{||\mathbf{d}_{u,v,f}||}, \mathbf{d}_{u,v,f} = \mathbf{R}_f \mathbf{K}_f^{-1} [u, v, 1]^T + \mathbf{t}_f, \tag{2}$$

This parametrizes camera trajectories into a $\ddot{\mathbf{P}} \in \mathbb{R}^{6\times T\times H\times W}$ representation that maintains dimensional consistency with the video's spatiotemporal structure. For model input latent space transformation, we implement a patchification module that initially compresses spatial dimensions to $\left(\frac{H}{p}\right) \times \left(\frac{W}{p}\right)$ using 2D convolutional layers, followed by 3D Causal convolution layers that compress the temporal dimension to $\left(\frac{L}{q}\right)$, yielding the final camera tokens z_{camera}. The compression ratios q and p are aligned with those of the visual Causal VAE.

3.2.2 Human Motion Representation

To condition generated videos on subtle human dynamics, the representation space must have sufficient expressiveness to encode such kinematic subtleties. In light of this, we consider coarse controls, such as bounding boxes and object-level keypoints in previous works [34, 39], suboptimal choices, and instead adopt a pose-based representation computed by DWPose [47] for encoding human motion signals. This representation effectively captures subtle changes in facial expressions and postural nuances. Its robustness holds for both single-person and multi-person scenarios, making it an ideal choice for complex human-centric motion control. The same patchification module structure as described in Sec. 3.2.1 is then used to encode the human motion conditions.

3.2.3 Decouple-and-Fuse Strategy

Having encoded camera conditions and human-motion conditions separately, we now focus on enabling the model to learn the interactions between the two motions. Intuitively, the model should first project the learned camera conditions globally across the entire latent, then project human-motion conditions only to human-relevant regions, while simultaneously ensuring these localized human poses harmonize with the established camera perspective.

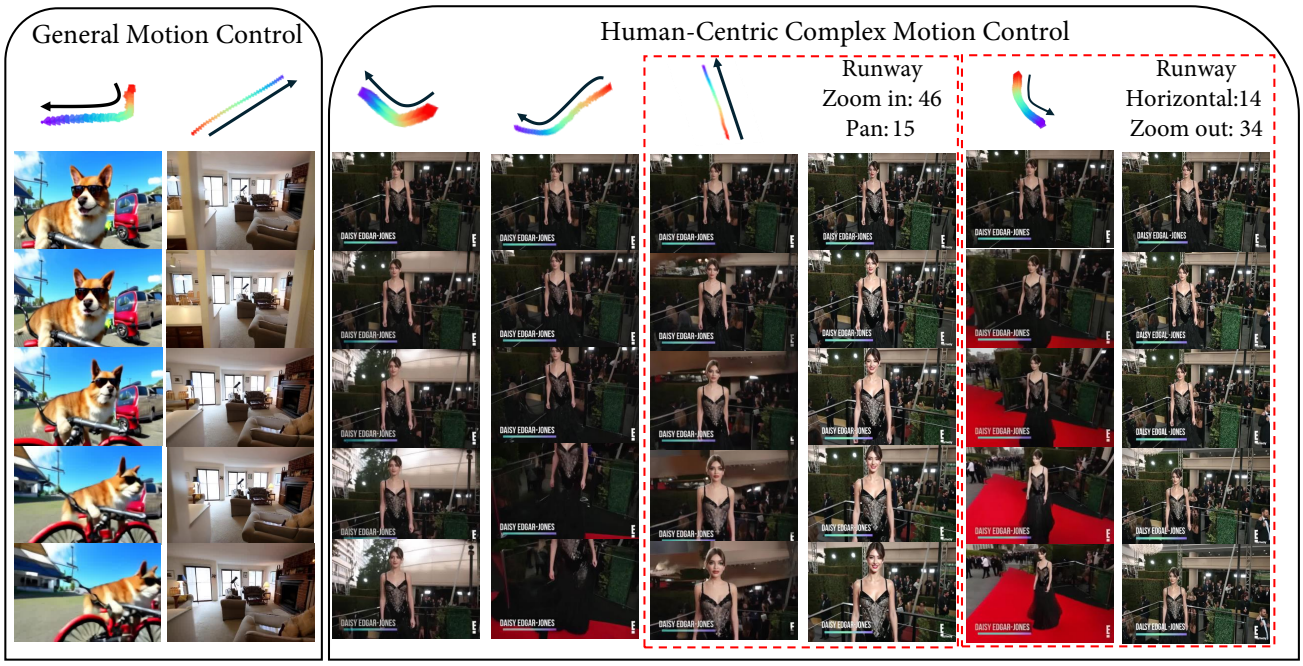


Figure 4. Visualization of camera control over diverse objects, scenarios, and complex camera motion. Our model shows superior visual fidelity and enhanced control flexibility compared to commercial tools (Runway). *Text prompts are omitted to save space.

As illustrated in Fig.2, the projection of two motion conditions begins by processing flattened motion tokens through parallel self-attention blocks, enabling the learning of global dependencies that maintain motion consistency across frames. Then, the localized nature of human-motion conditions compared to the global nature of camera-motion conditions necessitates an implicit decoupling approach that respects their distinct operational domains. We implement this decoupling with a masking strategy, which aims to enable the model to isolate human-motion effects only to relevant regions. We further relax the learning of this constraint by employing a dynamic mask that combines a learnable component with a hard human-pose prior. Specifically, given the camera tokens z^n_{camera} and human pose tokens z^n_{pose} on n-th attention block, we first apply a dimension reduction layer following with a normalization layer to get the learnable parts of the motion masks, $\mathcal{M}^n_{\text{pose}}$ and $\mathcal{M}^n_{\text{camera}}$:

$$\begin{aligned} \mathcal{M}^n_{\text{pose}} &= \text{LN}(\text{Linear}^n_{\text{pose}}(z^n_{\text{pose}})) \\ \mathcal{M}^n_{\text{camera}} &= \text{LN}(\text{Linear}^n_{\text{camera}}(z^n_{\text{camera}})) \end{aligned} \tag{3}$$

We then obtain the human-pose prior $\mathcal{M}^{\text{n, prior}}_{\text{pose}}$ by first resizing the raw human-pose representation to match the latent dimension of the n-th block, then dilate it to delineate human-relevant regions, and transform it into a binary mask. $\mathcal{M}^{\text{n, prior}}_{\text{pose}}$ is then added to the learned human-pose mask $\mathcal{M}^n_{\text{pose}}$. We finally compare the token attention mask token-wise with softmax and fuse the tokens accordingly:

$$z^n_{\text{fused}} = \text{Softmax}\left([\mathcal{M}^n_{\text{pose}} + \mathcal{M}^{\text{prior}}_{\text{pose}}; \mathcal{M}^n_{\text{camera}}]\right)[z^n_{\text{pose}}; z^n_{\text{camera}}] \tag{4}$$

We then encode z^n_{visual}, the visual tokens on the n-th block, with z^n_{fused} via the cross-attention mechanism following a ControlNet [50] block architechture. To effectively enable the model to learn the interaction between two motion conditions, we restrict the cross-attention to only the fused motion condition rather than separate motion conditions, forcing the model to disentangle motions from the fused representation to align with the two conditions for minimizing the distance from the reference videos during training. Specifically, z^n_{visual} are projected into queries Q^n_{visual}, while the fused motion tokens are transformed into corresponding keys and values K^n_{fused} and V^n_{fused} with three learnable linear layers. To maintain optimal balance between motion controllability and video generation fidelity, we incorporate a LoRA (Low-Rank Adaptation) layer prior to updating the visual tokens with the learned motion information. This fusion process can be written as:

$$z^n_{\text{visual}} += \text{Lora}(\text{CrossAttn}(Q^n_{\text{visual}}, K^n_{\text{fused}}, V^n_{\text{fused}})) \tag{5}$$

This architectural design enables fine-grained control while preserving the fundamental quality characteristics of the generated video sequences.

	Method	General Metrics			Camera Metrics			Human-Motion Metrics	
		LPIPS↓	FID↓	FVD↓	KptsErr↓	RotErr↓	TransErr↓	PoseErr↓	DetErr↓
T2V	MotionCtrl [38]	0.62	116.41	1185.83	6.72	0.47	8.66	172.44	14.73
	Direct-A-Video [46]	0.67	132.08	941.05	**3.85**	**0.24**	**5.06**	173.26	10.68
	MotionBooth [39]	0.69	125.62	795.59	5.24	0.36	6.42	165.49	13.19
	CogVideoX*-2B [48]	0.57	88.11	402.34	-	-	-	-	-
	Ours	**0.46**	**82.36**	**361.03**	4.43	0.27	5.34	**45.24**	**2.50**
I2V	ImageConductor [25]	0.43	131.36	878.59	4.32	0.44	5.44	164.28	9.18
	Ours	**0.42**	**74.65**	**332.27**	**3.77**	**0.22**	**2.98**	**83.82**	**3.63**

Table 1. Quantitative comparisons for joint controlling camera and human motion for both T2V and I2V generation. * denotes the original CogVideoX model, whose camera metrics and human-motion metrics are not calculated because no motion control is performed.

4. Experiments

4.1. Implementation Details

We implement two variants of our TokenMotion framework: TokenMotion-T for text-to-video (T2V) and TokenMotion-I for image-to-video (I2V) generation. Both variants leverage the open-source CogVideoX-2B [48] as the backbone. Specifically, since CogVideoX-2B only provides the official T2V checkpoint, we implement TokenMotion-I by fine-tuning this checkpoint for 100,000 steps, incorporating input images as an additional conditioning signal. For both models, we train them on 2×8 A100(80G) GPUs for 5 days using the Adam optimizer [22], with a learning rate of 1e-5 and a batch size of 2. During fine-tuning, we only train the two encoders of camera motion and human motion, along with the decouple-and-fuse module, while all the remaining parameters being frozen. For training data, we construct a mixed dataset consisting of randomly-sampled 49-frame video clips in a resolution of 256×384, which is a combination of HumanVid [37] and RealEstate10K [55]. Specifically, HumanVid contains videos with both control signals of human motion and camera motion, while RealEstate10K contains videos with only camera motion signals. To accommodate these differences, we input DWPose-extracted poses for HumanVid samples and blank human-motion cues for RealEstate10K samples during training, respectively. The inference is performed with 50-step denoising using DDIM Sampler [32], with a classifier-free guidance [17] scale of 7.5. More details of preprocessing steps and model configurations are provided in the supplementary.

4.2. Joint Camera and Human Motion Control

Evaluation Dataset. We carefully curated an evaluation set containing 500 real-world videos from HumanVid [37] which comprises both camera and human motions. The curation begins with 1000 randomly-sampled videos, from which we removed the 500 videos exhibiting minimal camera motions. This selection complements the focus of HumanVid's real-world set, which prioritizes human motion magnitude with some videos having limited camera dynamics.

Comparison Methods. We compare both variants of TokenMotion against state-of-the-art joint-control approaches. For T2V generation, we compare TokenMotion-T with MotionCtrl [38], Direct-A-Video [46] and MotionBooth [39]. For I2V generation, as no existing methods support joint-control, we compare TokenMotion-I with ImageConductor [25], the closest method, which offers only separate control of the two motions. We implement the joint-control in a cascaded manner for ImageConductor.

Evaluation Metrics. We evaluate joint-control of camera motion and human motion from three aspects: (1) General visual quality, for which we use LPIPS [51], FID[16] and FVD[33]. (2) Camera movement alignment, for which we follow CameraCtrl [15] and CamCO [43] to extract camera trajectories using COLMAP [30], which are then leveraged to compute the keypoints error (KptsErr), rotation error (RotError) and transition error (TransErr). (3) Human-motion alignment, for which we follow AnimateAnyone [13] to use DWPose [47] for human pose extraction, and calculate the accuracy for generated motion. This measurement includes the pose-landmark detection failure (DetErr) rate and the Euclidean distances between detectable landmarks of ground-truth and generated poses (PoseErr).

Quantitative Results. Results shown in Tab.1 demonstrate that our approach surpass all baselines in both general visual quality and human-motion alignment in both scenarios of T2V and I2V generation. Specifically, for human-motion control in T2V generation, our TokenMotion-T achieves a pose error of 83.82 and a landmark-detection failure rate of 2.5%, representing significant improvements of PoseErr and DetErr (164 $\rightarrow$ 83; 9.18%$\rightarrow$ 3.63%), which indicates that the model manages to decouple the two motions well. In terms of the general visual quality, our TokenMotion-T also achieves consistent improvements across three metrics, with a significant reduction in FVD from 795.59 to 361.03.

This indicates that the unified modeling of human motion and camera motion is effective in achieving the spatiotemporal consistency and the balance between two motions. Although TokenMotion-T achieves worse performance compared to Direct-A-Video, it still outperforms all other baselines. Specifically, Direct-A-Video achieves almost the worst human-motion control across all baselines, indicating that it fails to achieve balance in joint-control.

For joint-control in I2V generation, our TokenMotion-I consistently outperforms ImageConductor [25] across all metrics, highlighting that controlling human motions and camera motions in a decoupling manner without handling their interactions yields compromised performance, indicating the effectiveness of a decouple-and-fuse strategy proposed in our work.

Qualitative Results. Fig.3 further visualizes the superiority of TokenMotion-T with two challenging joint-control scenarios. We adopt TokenMotion-T for qualitative comparisons and our TokenMotion-T achieves the best performance in following the joint-control signals, while baseline methods fail in different ways: MotionCtrl fails to follow the human motion accurately in the first case while MotionBooth and Direct-A-Video lose more human details in the second case. Some confuse the camera motion with the human motion, such as MotionCtrl [38] in Fig.3 (a) misusing the right-moving camera control as the right-moving of the two people. Some sacrifice the visual quality for motion-control, such as MotionBooth [39] and Direct-A-Video generating [46] unrealistic background scenes in Fig.3 (b). Meanwhile, our TokenMotion can accurately decouple camera motion from human motion, while maintaining the overall visual quality, the same as Tab.1 suggests.

For TokenMotion-I, Fig. 5 demonstrates that it can perform fine-grained human-motion controls such as body rotation, leg lifting, and eyes blinking simultaneously with complex camera movements such as left-to-right rotation, up-down rotation and zoom-out. Our both TokenMotion variants achieve such advantages beyond these cases, and we refer readers to more generated samples in the supplementary materials for space limitation.

Methods	Kpts-err↓	Rot-err↓	Trans-err↓
CameraCtrl [15]	5.61	**0.27**	6.48
MotionCtrl-cam [38]	4.46	0.33	5.83
Ours	**4.36**	0.28	**5.39**

Table 2. Quantitative evaluation results for camera-only control video generation.

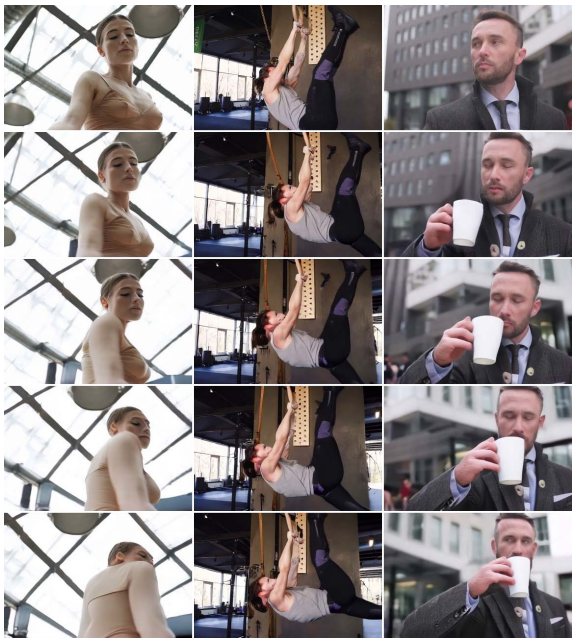


Figure 5. Qualitative results for TokenMotion-I.

4.3. Camera Motion Control

To evaluate the camera motion controllability of our proposed TokenMotion framework, we conducted comprehensive experiments on 1,000 randomly sampled videos from the RealEstate10K [55] test set, following the evaluation protocol established in CameraCtrl [15]. Comparative analyses against state-of-the-art camera-control methods, specifically CameraCtrl [15] and MotionCtrl-camera [38], are presented in Tab. 2, using camera metrics detailed in Section 4.2. Our method achieves superior performance in Trans-err and Kpts-err, while maintaining competitive performance in Rot-err, demonstrating exceptional camera motion control capabilities. Although TokenMotion is jointly trained on both camera and human motion samples, it exhibits robust motion control across diverse objects and scenarios in both image-to-video (I2V) and text-to-video (T2V) generation tasks, as shown in Fig. 4. We further validated its effectiveness in complex camera control scenarios, particularly in human-centric cases. Qualitative results in Fig. 4 illustrate that our model successfully synthesizes temporally coherent video sequences that faithfully adhere to physical camera movements. Furthermore, our method demonstrates performance comparable to the commercial Runway Alpha Gen3's camera control system, showcasing both superior visual quality and enhanced control flexibility.

4.4. Human Motion Control

We show that TokenMotion has strong generalization ability to handle extensive scenarios of human-motion control. As



Figure 6. Human-motion control with complex composition.



(a) Human-motion control with subtle facial movements.



(b) Human-motion control with large body movements.

Figure 7. Control with human motions of different granularity.

shown in Fig. 6, TokenMotion manages to perform compositional human motion control, including single-person object interactions, multi-person interactions and multi-person object interactions stably without artifacts. Additionally, TokenMotion is also able to handle human motions of different granularity. While Fig. 1 has demonstrates that TokenMotion is able to handle full-body movements like trajectories, we additionally present Fig. 7 to show that our model also stably advances in non-full-body movements such as pose changes in Fig. 7b and even-smaller changes in facial expression, such as lip twitching and head swaying, in Fig. 7a.

4.5. Ablation Study

Baseline. To evaluate the impact of the proposed TokenMotion, we conduct experiments on our base model, CogvideoX, whose results are shown in Tab. 3. The base model performs worse than our approach in all three metrics, with a significant performance difference in FVD. This indicates the introduced motion signals largely benefit the spatiotemporal consistency of generated videos.

Token Compression. To demonstrate that the motion signals are effectively encoded, we replace the patchification module in TokenMotion by a ControlNet [50] module, which yields motion representations of larger dimensions, the same as hidden states. As shown in Tab. 3, without patchification, the model produces a worse FVD score, indicating that this model variant has worse motion controllability in generation.

Decouple-and-Fuse. To demonstrate the effectiveness of our decouple-and-fuse strategy, we conduct experiments on implementing the joint modeling of camera motion and human motion as direct addition. Results in Tab. 3 shows that without the decouple-and-fuse strategy, the model struggles with maintaining both per-frame quality and spatiotemporal consistency in generated videos, as it achieves the worst among all three metrics. This indicates that dedicated-designed modules are necessary for models to handle the interaction between the two motion signals.

Hybrid-mask Guidance. To demonstrate the effectiveness of introducing explicit pose guidance in mask, we conduct experiments on a model variant which only leverages a learnable mask for decoupling human motion and camera motion. As shown in Tab. 3, while a slight performance drop in FVD is shown when incorporating a hybrid mask, scores of LPIPS and FID are much better with our design.

Methods	LPIPS↓	FID↓	FVD↓
Baseline	0.57	88.11	402.34
w/o decouple-and-fuse	0.66	93.89	890.39
w/o token compression	0.54	88.56	590.62
w/o hybrid mask	0.55	89.63	**358.76**
Ours	**0.46**	**82.36**	361.03

Table 3. Ablation studies about Token Compression, Decouple and fusion, and Hybrid-mask.

4.6. Limitations and Future Works

While TokenMotion can precisely generate videos that follow the joint control signals, the visual quality is limited by the base model. We identify two representative limitations: First, the model shows difficulties in modeling finger movements. Second, the model struggles with facial detail preservation, where facial features can appear blurred or exhibit geometric distortions. Future work could explore adapting TokenMotion's joint-control strategy to larger-scale backbones. For more concrete discussions on limitations, we refer readers to the supplementary materials.

5. Conclusions

This paper presents TokenMotion, the first DiT-based framework for joint motion control, targeting the task of human-centric motion control, which demonstrates superior performance in both text-to-video and image-to-video generation. Our decouple-and-fuse strategy with the dynamic masking mechanism effectively coheres the control signals of camera motion and human motion. Comprehensive evaluations demonstrate that our TokenMotion outperforms state-of-the-art approaches in simultaneously controlling camera motions and human motions.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. *arXiv preprint arXiv.2407.12781*, 2024. 2, 3

[2] 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. *arXiv preprint arXiv.2311.15127*, 2023. 2, 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 22563–22575, 2023. 2

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. https://openai.com/research/video-generation-models-as-world-simulators. 2, 3

[5] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *arXiv preprint arXiv.2407.08674*, 2024. 3

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv.2310.19512*, 2023. 2

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 7310–7320, 2024. 2, 3

[8] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024. 2, 3

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 7312–7322, 2023. 3

[10] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv.2302.03011*, 2023. 2

[11] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 2

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *ECCV 2024*, pages 330–348, 2024. 2

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. 2, 3, 6

[14] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv.2312.06662*, 2023. 2

[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. *arXiv preprint arXiv.2404.02101*, 2024. 2, 3, 4, 6, 7

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 6

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[19] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[20] 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*, pages 8153–8163, 2024. 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 15908–15918, 2023. 2

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 6

[23] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 2, 3

[24] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dragapart: Learning a part-level motion prior for articulated objects. In *ECCV 2024*, pages 165–183, 2024. 3

[25] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image Conductor: Precision Control for Interactive Video Synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 3, 6, 7

[26] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. ReVideo: Remake a Video with Motion and Content Control. *arXiv preprint arXiv:2405.13865*, 2024. 3

[27] 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, 2023. 2, 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685, 2022. 2

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022. 2

[30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[31] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021. 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 6

[33] 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*, 2019. 6

[34] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024. 2, 3, 4

[35] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[36] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv.2309.15103*, 2023. 2

[37] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation. *arXiv preprint arXiv:2407.17438*, 2024. 2, 6

[38] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, page 114, 2024. 2, 3, 4, 6, 7

[39] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv.2406.17758*, 2024. 2, 3, 4, 6, 7

[40] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities. *arXiv preprint arXiv:2408.13239*, art. arXiv:2408.13239, 2024. 3

[41] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion Control for Anything using Entity Representation. *arXiv preprint arXiv:2403.07420*, 2024. 2, 3

[42] Jinbo Xing, Menghan Xia, Yong Zhang, Hao Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV 2024*, pages 399–417, 2024. 2

[43] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-Controllable 3D-Consistent Image-to-Video Generation. *arXiv preprint arXiv.2406.02509*, 2024. 2, 3, 6

[44] 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*, pages 1481–1490, 2024. 3

[45] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender A video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers, SA 2023, Sydney, NSW, Australia, December 12-15, 2023*, pages 95:1–95:11, 2023. 3

[46] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver,*

CO, USA, 27 July 2024- 1 August 2024, page 113, 2024. 2, 3, 4, 6, 7

[47] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 4, 6

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv.2408.06072*, 2024. 2, 3, 6

[49] 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. 2, 3

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5, 8

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[52] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv.2311.04145*, 2023. 2

[53] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290, 2025. 2

[54] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-Controlled Image-to-Video Diffusion Model. *arXiv preprint arXiv:2410.15957*, 2024. 3

[55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 6, 7

[56] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. *arXiv preprint arXiv:2403.14781*, 2024. 3