Controllable Human-centric Keyframe Interpolation with Generative Prior

Zujin Guo[1] **Size Wu**[1] **Zhongang Cai**[2] **Wei Li**[1*] **Chen Change Loy**[1*]
[1]S-Lab, Nanyang Technological University
[2]SenseTime Research
https://gseancdat.github.io/projects/PoseFuse3D_KI

Abstract

Existing interpolation methods use pre-trained video diffusion priors to generate intermediate frames between sparsely sampled keyframes. In the absence of 3D geometric guidance, these methods struggle to produce plausible results for complex, articulated human motions and offer limited control over the synthesized dynamics. In this paper, we introduce PoseFuse3D Keyframe Interpolator (PoseFuse3D-KI), a novel framework that integrates 3D human guidance signals into the diffusion process for Controllable Human-centric Keyframe Interpolation (CHKI). To provide rich spatial and structural cues for interpolation, our PoseFuse3D, a 3D-informed control model, features a novel SMPL-X encoder that transforms 3D geometry and shape into the 2D latent conditioning space, alongside a fusion network that integrates these 3D cues with 2D pose embeddings. For evaluation, we build CHKI-Video, a new dataset annotated with both 2D poses and 3D SMPL-X parameters. We show that PoseFuse3D-KI consistently outperforms state-of-the-art baselines on CHKI-Video, achieving a 9% improvement in PSNR and a 38% reduction in LPIPS. Comprehensive ablations demonstrate that our PoseFuse3D model improves interpolation fidelity.

1 Introduction

Frame interpolation aims to generate new frames between two consecutive video frames to improve temporal smoothness. Traditional interpolation methods [12, 46, 17] assume small, simple motion over short time spans. These methods are challenged when the input frames are widely separated, known as keyframe interpolation or generative inbetweening [56, 47, 41, 11], where the motion between them becomes complex and ambiguous. This challenge is magnified in human-centric videos, where articulated body movements encompass diverse poses and shapes. With human subjects being prevalent in today's video content, there is a growing need for interpolation methods to handle large temporal gaps and intricate human motion while offering plausible results.

Human-centric keyframe interpolation remains challenging for current methods. Recent approaches [11, 47, 41] leverage generative priors from image-to-video (I2V) models to bridge the temporal gap. These methods condition the interpolation solely on the input keyframes without intermediate guidance. Consequently, they often struggle to resolve motion ambiguities and accurately capture the complex articulated dynamics of human motion. For instance, when keyframes involve large occlusions or non-rigid joint movements, these methods often produce implausible or distorted interpolations (Figure 1(a)) due to insufficient intermediate guidance. FCVG [56] has explored interpolation keyframes with 2D skeletons as control signals for human subjects. However, 2D lines cannot convey full body shape and geometry, leading to unrealistic results (see Figure 1(b)). These

* Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

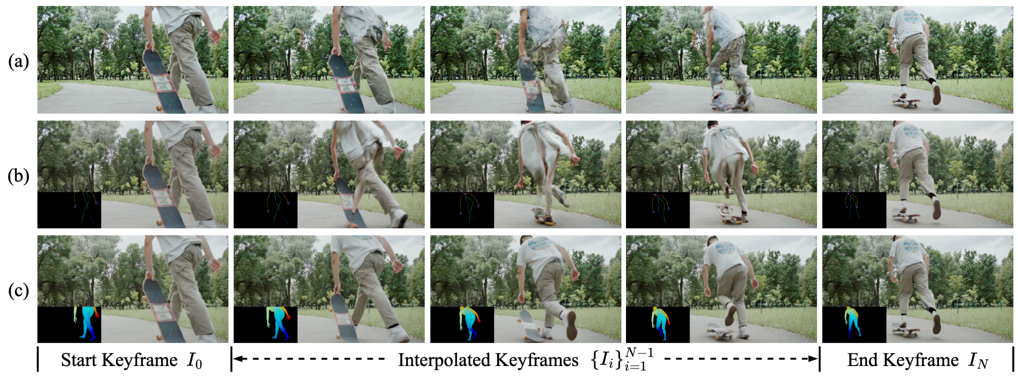


Figure 1: **Keyframe Interpolation with Different Strategies.** (a) Interpolation using I2V models without intermediate guidance often yields implausible or distorted frames, especially under large motion or occlusion. (b) Skeleton-guided interpolation offers structural cues but lacks geometric detail, resulting in unrealistic body shape and appearance. (c) **Our PoseFuse3D-KI** employs dense human-centric guidance, enabling temporally coherent and visually plausible interpolations.

methods lack fine-grained control over the interpolation process, limiting their ability to produce flexible, high-fidelity human-centric interpolations.

In this study, we investigate the integration of 3D human conditions into the human-centric keyframe interpolation pipeline. Drawing inspiration from recent advances in human animation [55, 54], we propose to integrate 2D human poses [48] with 3D SMPL-X models [30] as intermediate control signals. These signals provide precise guidance for complex articulated motions: 2D poses offer concise representations of human joint poses, while 3D models capture rich spatial geometry. However, effectively processing these control signals poses challenges. First, common practice renders 3D human models into 2D proxies (e.g., colored surface, normals, depth maps) before encoding, leading to substantial loss of spatial information in occluded regions. Therefore, we need to develop a dedicated encoder that preserves occluded 3D details when converting models into control signals. Second, fusing signals with different information content and granularity is nontrivial. This necessitates designing appropriate neural architectures that can accurately extract 3D cues and harmonize them with 2D poses into a unified, informative control input.

To this end, we introduce PoseFuse3D Keyframe Interpolator, termed as *PoseFuse3D-KI*, a novel pose-control framework for controllable human-centric keyframe interpolation (Figure 1(c)). Our framework is unique in its 3D-informed control model PoseFuse3D, which comprises three jointly trained modules. The first derives control features from visualized conditions; the second encodes and aggregates 3D body geometry into 2D image conditioning features; and the third combines outputs from the first two modules into a unified control signal for interpolation. In contrast to Champ [55], which fuses 3D features from rendered visualizations, our encoder processes features directly in 3D and integrates projected features through feature aggregation.

To evaluate the proposed PoseFuse3D-KI, we have also built a high-quality video dataset for Controllable Human-centric Keyframe Interpolation (CHKI). Most existing interpolation datasets, such as SportsSlomo [6], target small temporal gaps, lack annotations for 2D poses or 3D human models, and offer limited human-centric motion diversity. Therefore, we introduce *CHKI-Video*, a new dataset for a systematic evaluation of CHKI algorithms. CHKI-Video comprises 2,614 high-quality video clips of over 180K frames sourced from SportsSlomo [6] and Pexels [1] website that hosts high-quality stock videos. Each frame is carefully annotated with bounding boxes, segmentation masks, 2D human poses, and SMPL-X parameters, using state-of-the-art tools [4, 20, 25, 34, 43] supplemented by manual verification. From this collection, we derive a benchmark specifically for the CHKI task. We hope this benchmark will help improve controlled keyframe interpolation techniques with its high-quality videos and diverse examples.

Our contributions are threefold: (i) We present PoseFuse3D-KI, an effective interpolation framework for human-centric keyframe interpolation, characterized by a novel pose-control model, PoseFuse3D. It effectively extracts control signals from 3D SMPL-X and fuses 2D signals, allowing precise and informative control. (ii) For evaluation, we construct CHKI-Video, a benchmark dataset with comprehensive human-centric annotations, which are absent in existing interpolation benchmarks. (iii) Through extensive experiments, we demonstrate that PoseFuse3D-KI delivers state-of-the-art performance on our CHKI-Video benchmark with an improvement of 1.85 dB in PSNR and a reduction of 0.0796 in LPIPS.

2 Related Work

Frame Interpolation. Traditional frame interpolation methods are primarily designed for temporally adjacent frames and rely on either direct synthesis using convolutional networks [8, 19], or motion representations such as dynamic kernels [28, 29, 32, 7, 9, 22] and optical flows [16, 23, 21, 37, 14, 2, 12, 17, 44, 50]. Recent advances [11, 47, 41] extend this task to more challenging keyframe interpolation scenarios by leveraging the generative priors of image-to-video diffusion models [3, 39, 35, 13]. These methods combine temporal forward and reverse denoising predictions in a unified process to enable interpolation. However, such approaches struggle when faced with complex articulated motions or ambiguous transitions from human keyframes. To alleviate motion ambiguity in interpolation, FCVG [56] introduces 2D matched lines as control signals. However, this control signal lacks the 3D geometric context required for plausible human-centric interpolation, resulting in unrealistic body shape and appearance.

Pose-Guided Human Animation. Recent advances in pose-guided human animation [54, 55, 15, 33, 51, 45, 40] harness the power of diffusion models and have achieved remarkable success in generating videos from a single reference image. These methods offer flexible and precise control by incorporating enriched conditioning signals and increasingly sophisticated control mechanisms. 2D human poses, such as OpenPose [5] and DWPose [48], are widely used in existing works [55, 54, 33, 51], but they are limited in capturing fine-grained geometry and motion dynamics. To address this, recent works [55, 54] integrate 3D human parametric models [27, 30], which offer realistic body representation through blend shapes and skinning, resulting in more accurate and expressive human animations. Specifically, Champ [55] renders 3D human models into 2D proxies (e.g., normals, depth, and semantic maps) and combines them with 2D pose visualizations as control input. It operates directly on these visualizations and unifies them using a simple summation operation. To incorporate control signals into diffusion models, many approaches [40, 45, 51, 54] adapt ControlNet [51] for their customized control networks. Some methods [55, 15] introduce task-specific pose guiders but often require retraining the majority of denoiser parameters. ControlNeXt [33] improves control efficiency by encoding conditions with a lightweight convolutional network and injecting them via cross-normalization after the first denoising block, tuning only a minimal subset of parameters. This efficient mechanism enables robust conditioning over large-scale pre-trained video generators [39, 3].

In this paper, we demonstrate the advantage of combining 2D poses (DWPose [48]) with 3D human models (SMPL-X [30]) through comprehensive experiments. We present a novel pose control model that extracts unified, 3D-informed control features to provide precise guidance. Instead of relying on rendered normals and depth maps, the model directly extracts explicit 3D information from human models in 3D space, preserving richer control signals. It adopts a ControlNeXt-inspired strategy to control video diffusion models for keyframe interpolation.

3 Method

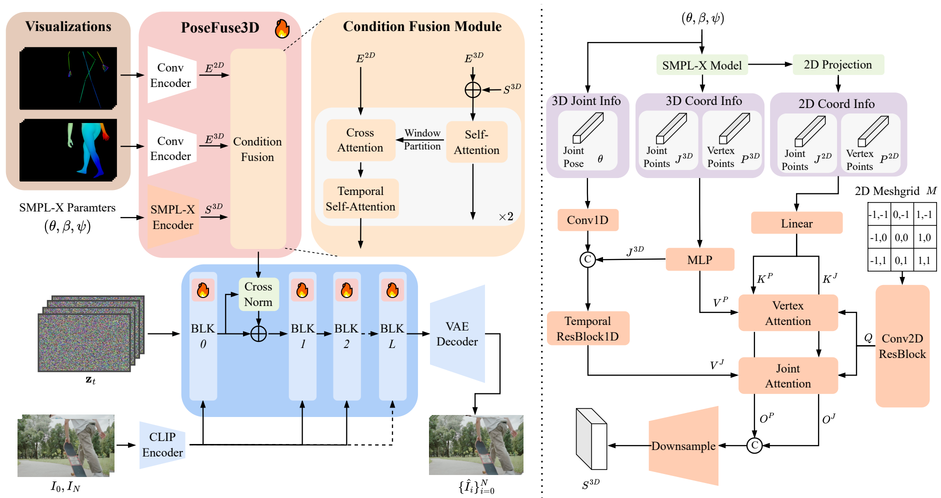
Given a human-centric keyframe-pair $I_0, I_N \in \mathbb{R}^{H\times W\times 3}$ with timesteps $\{0, N\}$, Controllable Human-centric Keyframe Interpolation (CHKI) is formulated as:

$$\{\hat{I}_i\}_{i=1}^{N-1} = \mathcal{G}(I_0, I_N, \{\mathbf{C}\}_{i=0}^{N}), \tag{1}$$

where $\mathcal{G}$ denotes an interpolation model guided by control signals $\{\mathbf{C}_i\}$. In this work, we aim to address CHKI by proposing an effective controllable interpolation framework, **PoseFuse3D-KI**. This framework integrates 3D-aware human-centric signals into a pre-trained Video Diffusion Model (VDM) through our 3D-informed control model PoseFuse3D, as illustrated in Figure 2(a).

3.1 PoseFuse3D

PoseFuse3D is a 3D-informed control model that provides 3D human structure and geometry guidance for plausible human interpolation. This 3D-informed guidance is injected into the base diffusion



(a) PoseFuse3D-KI Framework (b) SMPL-X Encoder

Figure 2: **Model Architecture.** Our PoseFuse3D-KI framework, as shown in (a), comprises a video diffusion model (VDM) and a novel control model, PoseFuse3D. The PoseFuse3D model extracts rich features from both 3D and 2D control signals and fuses them into a unified representation to guide the VDM. The key component of PoseFuse3D is the SMPL-X encoder as illustrated in (b), which provides *explicit 3D signal features*. Specifically, the SMPL-X encoder first extracts 3D information from the SMPL-X model with 2D correspondences via projection. The 3D and 2D information is then encoded in parallel. With features of 2D correspondences, 3D information is aggregated onto the 2D image plane using attention mechanisms. The aggregated features are subsequently processed to produce the final feature S^{3D}.

model after the first denoising block via cross-normalization [33]. Internally, PoseFuse3D comprises three jointly trained components: a visual encoding module that derives control features from 3D SMPL-X [30] renderings and 2D DWPose [48] visualizations; a SMPL-X encoder that directly embeds 3D humans and aggregates them into image conditioning maps; and a fusion module that integrates encoding streams into a unified control tensor to guide interpolation in the base VDM.

Visual Encoding. This visual encoding module extracts conditional features from visualized control images, maintaining natural pixel-level alignment with the controlled video latents. To enhance the control signals, we incorporate visualizations from both DWPose and SMPL-X. While SMPL-X renderings provide rich human surface details, their keypoint information is indirect, mixed with other vertices and mesh faces. Therefore, we add DWPose visualizations to emphasize the skeletal keypoint layout for robust pose understanding. This combination was also demonstrated to be effective in prior work [55]. Specifically, our visual encoding module employs two parallel convolutional encoders to capture comprehensive pose information. One encoder processes DWPose visualizations to capture compact pose information, while the other handles SMPL-X renderings that retain 3D cues such as occlusion boundaries and projected shapes. Notably, to enrich semantic detail, we use SMPL-X Colored Surface [54] during rendering, which assigns a unique color to each vertex. The resulting feature maps, E^{2D} and E^{3D}, are then passed to the fusion module for unified conditioning.

SMPL-X Encoder. Although 3D model renderings offer aligned image maps for conditioning in VDMs, the rendering operation discards parts of 3D information, particularly in *occluded regions*. This results in implausible interpolation for keyframes of large human motion. For example, in the first case of Figure 3, all variants relying solely on rendered controls fail to interpolate the occluded arms with correct spatial position. To enhance controllability with direct 3D information, we introduce a SMPL-X encoder (Figure 2(b)) that processes the SMPL-X model in 3D space and transforms it into an image conditioning feature S^{3D}. Specifically, a SMPL-X model [30] is parameterized with (θ, β, ψ), corresponding to pose, shape, and expression. We obtain structural information by forwarding these parameters into the SMPL-X model to generate vertex and joint coordinates P^{3D}, J^{3D} in 3D space and obtain their corresponding 2D coordinates P^{2D}, J^{2D} through projection:

$$P^{3D}, J^{3D} = \mathrm{SMPLX}(\theta, \beta, \psi), \tag{2}$$

$$P^{2D}, J^{2D} = \mathrm{Projection}(P^{3D}, J^{3D}), \tag{3}$$

where the joint coordinates J^{3D}, J^{2D} correspond to the pose parameter θ that indicates joint rotations. Notably, the projection step establishes a correspondence between the 3D space and the 2D image plane, making it possible to retain 3D spatial structure while producing image conditioning maps. Next, the raw 3D coordinates are processed using an MLP to produce point-wise vertex features V^P and joint features. These joint features are refined through a temporal residual block that fuses them with pose information into expressive joint-level representations V^J. To aggregate these 3D features into 2D image control maps for joints and vertices, we employ separate attention mechanisms:

$$O^J = \mathrm{JointAttn}(Q, K^J, V^J), \tag{4}$$

$$O^P = \mathrm{VertexAttn}(Q, K^P, V^P), \tag{5}$$

where $Q \in \mathbb{R}^{B \times HW \times d}$ is the flattened d-dimension feature of a standard 2D meshgrid extracted by a convolutional encoder. Finally, the outputs O^J and O^P are concatenated and passed through a downsampling block to produce the final SMPL-X control representation S^{3D}, which serves as an informative and compact image embedding of the underlying 3D human structure.

Condition Fusion. The condition fusion module combines control features from both 2D and 3D signals into a unified representation to guide keyframe interpolation. For robust feature representation, we introduce a coarse-to-fine fusion strategy that progressively integrates rich geometric information from the 3D features into the compact 2D pose features. Specifically, we adopt two attention-based fusion blocks to perform this integration, where each block contains three attention layers for progressive refinement. The first layer is a self-attention module that processes the 3D features by operating on the sum of rendering encoding E^{3D} and SMPL-X features S^{3D}. The second layer performs cross-attention, aligning the 3D features with the 2D encoding E^{2D} through a spatially localized interaction scheme. Notably, we adapt the shifted window-partition strategy [26, 50] to restrict attention computation to adjacent regions, enhancing local alignment. The third layer applies temporal self-attention to capture temporal dynamic correlations in the fused representation. We use the second fusion block output as the final control signal, which is injected into the base interpolation engine to provide fine-grained, structure-aware guidance during synthesis.

3.2 VDMs as Base Interpolation Engine

To supply generative priors for human-centric keyframe interpolation, we adapt pre-trained Video Diffusion Models (VDMs) under the latent diffusion framework [35]. VDMs perform the diffusion process in a VAE-encoded latent space, conditioning video synthesis on input frames. Our main experiments investigate Wan2.1 [39] as the base model for keyframe interpolation.

Wan2.1 consists of a scaled-up DiT-based denoiser [31] and a causal 3D-VAE that performs spatiotemporal compression. It employs the flow matching strategy [24, 10] for the diffusion process. It formulates the forward diffusion process as a linear interpolation between the clean video latent $\mathbf{z}$ and the noise $\boldsymbol{\epsilon}$, which adds the noise by: $\mathbf{z}_t = (1-t)\mathbf{z} + t\boldsymbol{\epsilon}$. In the backward process, the denoiser f_θ iteratively refines $\mathbf{z}_t$ conditioned on the first frame I_0. The training objective is formulated as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z},t,I_0,\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}[||f_\theta(\mathbf{z}_t, I_0, t) - \mathbf{y}||_2^2], \tag{6}$$

where the target objective $\mathbf{y}$ is $\frac{d\mathbf{z}_t}{dt} = \boldsymbol{\epsilon} - \mathbf{z}$. Since the latent space is unevenly compressed in time, noise fusion strategies [11, 56] that combine temporally forward and reverse denoising paths are ineffective. Instead, we adapt Wan2.1 for keyframe interpolation in one unified denoising process. Specifically, we condition Wan2.1 on both endpoint frames I_0, I_N, and apply LoRA tuning to its input patch embeddings. Additional details are provided in the supplementary material.

4 The CHKI-Video Dataset

Existing interpolation datasets [46, 8, 6, 36, 38], which are designed for interpolating temporally adjacent frames, are not suitable for the CHKI task. To address this gap, we built the CHKI-Video dataset in three consecutive stages. More details can be found in the supplementary material.

Stage 1: Dataset Collection. To create a dataset with challenging and diverse human motion, we curate video clips from SportSlomo [6] and Pexels [1]. While SportsSlomo provides challenging human-centric videos, its exclusive focus on sports limits the diversity of activities. To enhance diversity versus the reality [49], we compile a list of keywords, spanning everyday activities to high-intensity actions. We use these keywords to retrieve videos from Pexels, complementing the sports videos. The curated videos are then downsampled to eliminate frame redundancy unnecessary for keyframe interpolation.

Stage 2: Pre-annotation Processing. We first perform general filtering for the low-quality videos according to brightness changes and assessed scores [43]. Then, we use Grounding-DINO [25] and SAM2 [34] to detect, segment, and track human instances in each video. We discard any video of more than three people or that is shorter than 20 consecutive frames to ensure a sufficient temporal span for keyframe interpolation. After automated processing, we manually review and filter detections in complex sports scenarios.

Stage 3: Human-centric Annotation. Building on the accurate human detections obtained in Stage 2, we annotate each clip for precise human-centric information. First, we employ Sapiens [20] to extract 2D human keypoints and perform whole-body detection to filter out clips with incomplete human figures. This ensures our dataset remains strictly human-centric. Finally, we apply SMPLer-X [4], leveraging its high re-projection accuracy to fit detailed SMPL-X models for human images and produce reliable 3D body parameters for each frame.

As a result, CHKI-Video comprises 2,614 video clips of over 180K frames, carefully annotated with human bounding boxes, masks, 2D keypoints and 3D parametric models. To prepare the train and test split, we follow the original division for SportsSlomo [6] videos and distribute the Pexels videos according to their keyword frequencies to maintain balanced coverage of all motion categories.

5 Experiments

We present quantitative and qualitative results in Sec. 5.1 to validate the effectiveness of our 3D control strategy in PoseFuse3D. We compare our interpolation performance against state-of-the-art methods in Sec. 5.2 and analyze the scalability across temporal gaps in Sec. 5.3. We further assess the in-the-wild interpolation capability in Sec. 5.4, where ground-truth control signals are not available. Finally, Sec. 5.5 provides a detailed ablation study to justify our model design.

Implementation Details. We fine-tune PoseFuse3D-KI on the CHKI-Video training split for 70k iterations. Specifically, we fine-tune our 3D-informed control model PoseFuse3D, and employ LoRA on the input patch embeddings, as well as the value and output projections of the VDM's attention modules. During training, we randomly sample 25 consecutive frames from video clips and process them to a resolution of 512×320. To maximize human-centric content, we crop each clip around its largest annotated human bounding box. Given the ratio from the target input size, we first crop the maximum scale of the image with the largest box as the center. The cropped image is then resized to match the target resolution. More implementation details are provided in the supplementary material.

Evaluation Protocols. Unless otherwise noted, all methods are evaluated on the CHKI-Video test set using **ground-truth annotations** for controllable interpolation. With motion ambiguity limited by the ground truth controls, we adopt the standard interpolation metrics of PSNR and LPIPS computed on the whole images for evaluation. To quantify performance specifically on human regions, we further leverage the annotated human boxes and masks, yielding $\text{PSNR}_{\text{bbox}}$, $\text{LPIPS}_{\text{bbox}}$, $\text{PSNR}_{\text{mask}}$, and $\text{LPIPS}_{\text{mask}}$ metrics. Notably, we apply binary dilation to expand and smooth the masks before computing the metrics, preventing potential artifacts along mask boundaries.

5.1 3D Control Strategy in PoseFuse3D

Setup. We evaluate the effectiveness of the 3D-informed control design in PoseFuse3D via comparisons among 3D control strategies, including 'VE', 'VE+DN' and 'VE+SE'. Specifically, VE refers to the visual encoding on the SMPL-X colored surface rendering, removing the SMPL-X encoder from the full design of our control model. VE+DN extends VE by incorporating depth and normal renderings, using two additional encoders identical to the one used in VE. VE+SE represents our proposed strategy in PoseFuse3D that directly encodes 3D information via the SMPL-X encoder

Table 1: **Comparisons of 3D Control Strategies.** This table presents quantitative results of different 3D control strategies. The PSNR and LPIPS metrics are calculated for the whole image, as well as for the human-centric parts via annotated boxes or masks. Best results are highlighted with **boldface**.

Method	Backbone	3D Control Strategy	Evaluation Metrics					
			PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓
FCVG	SVD	N.A.	20.42	18.05	16.91	0.2100	0.0899	0.0606
PoseFuse3D	SVD	VE-SVD	20.96	18.56	17.47	0.1975	0.0835	0.0553
PoseFuse3D	Wan2.1-I2V	VE	21.91	19.13	17.87	0.1400	0.0682	0.0484
PoseFuse3D	Wan2.1-I2V	VE+DN	22.07	19.12	17.80	0.1363	0.0667	0.0473
PoseFuse3D	Wan2.1-I2V	VE+SE (Ours)	**22.14**	**19.30**	**18.01**	**0.1330**	**0.0653**	**0.0464**

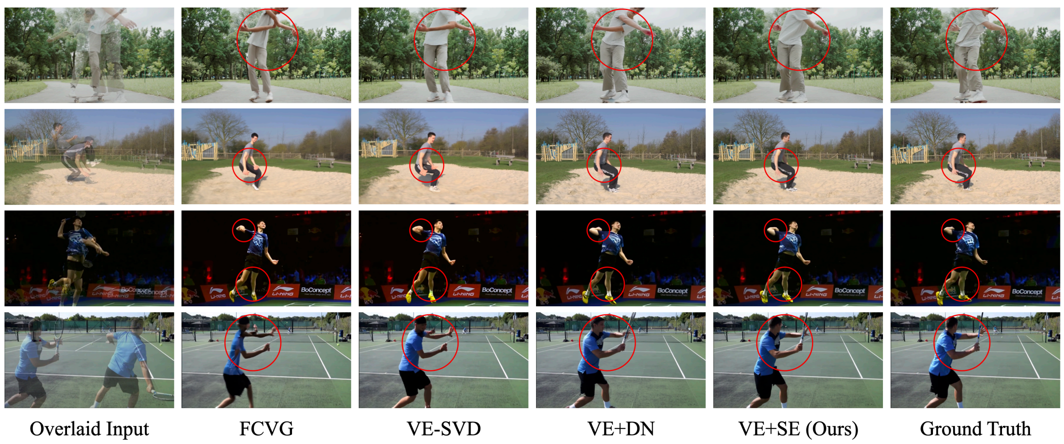


Figure 3: **Qualitative Results of Different 3D Control Strategies.** We use red circles to highlight regions where the 3D controls and our strategy significantly improve the interpolation quality.

and integrates it with VE. Experiments for these strategies use Wan2.1-I2V [39] as the interpolation backbone. For efficiency, they are trained for 40K iterations.

To assess the necessity of 3D information, we compare against FCVG [56], which is conditioned solely on 2D signals. For fair comparison, we create a variant of VE by replacing the backbone with SVD [3]. We adapt SVD for interpolation with the temporal forward and reverse denoising path fusion strategy used in FCVG. This variant is referred to as VE-SVD for ease of analysis.

Results. We provide quantitative comparisons of 3D control strategies in Table 1. In comparisons between FCVG and VE-SVD, we find that adding 3D control improves the interpolation performance. VE-SVD outperforms FCVG across metrics with a more than **0.50 dB** increase on all the PSNRs, indicating the improvements on both whole-image and human-centric levels. Moreover, our study highlights the importance of explicit 3D information. VE+DN and VE+SE, which incorporate depth and normal maps or direct SMPL-X information, outperform the simpler strategy VE. VE+DN and VE+SE show clear improvements in perceptual quality as reflected by the LPIPS metrics. Notably, our VE+SE strategy, which directly encodes information in 3D space, delivers the best performance, achieving the lowest $\text{LPIPS}_{\text{bbox}}$ of **0.0653** and the highest PSNR of **22.14 dB**.

Visualizations. Besides our quantitative analysis, we perform a qualitative comparison of the 3D control strategies, as shown in Figure 3. We find that incorporating 3D controls better preserves human shape during interpolation. Take the Tennis case in Figure 3 for example, methods with 3D control strategies interpolate the player's body close to the ground truth, whereas FCVG exhibits noticeable distortion. Moreover, our VE+SE, which directly encodes 3D information from SMPL-X, proves effective in handling occluded human motion. In both the Skateboarding (1^{st} row) and Jumping (2^{nd} row) cases in Figure 3, we can observe that our VE+SE strategy produces plausible results for the occluded arms, demonstrating its advantage in complex scenarios.

Table 2: **Comparisons with State-of-the-art Interpolation Methods.**

Methods	Metrics						
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓	HA↑
GIMM-VFI [12]	20.29	16.36	14.93	0.1954	0.1187	0.0860	0.9146
GI [41]	15.81	13.04	12.03	0.3364	0.1672	0.1146	0.8954
Wan2.1-KI (Ours)	19.53	15.96	14.62	0.2081	0.1208	0.0868	0.9180
FCVG [56]	20.42	18.05	16.91	0.2100	0.0899	0.0606	0.9187
PoseFuse3D-KI (Ours)	**22.27**	**19.49**	**18.24**	**0.1304**	**0.0636**	**0.0450**	**0.9189**

Table 3: **Results across Temporal Gaps.**

Temporal Gap	Metrics					
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓
24-frame	23.86	21.64	20.70	0.1144	0.0508	0.0332
48-frame	21.46	19.48	18.57	0.1566	0.0644	0.0432
64-frame	20.37	18.34	17.45	0.1882	0.0755	0.0514
96-frame	19.44	17.50	16.66	0.2208	0.0836	0.0577

5.2 Benchmark Results

Setup. We compare PoseFuse3D-KI against several advanced interpolation methods on our CHKI-Video dataset. The main comparison is with FCVG [56], which also enables intermediate control during interpolation. For broader coverage, we also include the keyframe interpolation method GI [41] and the traditional video frame interpolation method GIMM-VFI [12]. We also include Wan2.1-KI, an adaptation of Wan2.1 [39] for keyframe interpolation, following the strategy in Sec. 3.2. Since not all baselines support ground truth controls, we additionally compute the Human Anatomy (HA) [53] score to assess the quality of human synthesis and support fair comparison.

Results. We report quantitative results of PoseFuse3D-KI on the CHKI-Video benchmark in Table 2. Our method delivers state-of-the-art performance on human-centric keyframe interpolation. On whole-image metrics, it boosts PSNR by **1.85 dB** and lowers LPIPS by **0.0796** in comparison with the state-of-the-art method FCVG [56]. Crucially, it also outperforms on human-centric metrics compared with other methods, achieving a $\text{PSNR}_{\text{bbox}}$ of **19.49 dB**, an $\text{LPIPS}_{\text{mask}}$ of **0.045**, and an HA score of **0.9189**. This indicates that our method produces plausible, high-fidelity human interpolations that closely follow the ground-truth dynamics, demonstrating the effectiveness of our method.

Visualizations. For qualitative evaluation, we qualitatively compare PoseFuse3D-KI with other advanced methods, as illustrated in Figure 4. Our approach delivers more accurate human interpolations, faithfully following real-world motions and preserving body shape. For example, in the 2^{nd} 'Fencing' case and 4^{th} 'Stunt Bike' case, only PoseFuse3D-KI correctly interpolates leg and arm movements while maintaining consistent shapes. Moreover, our method naturally handles the occluded human motion well with 3D-informed control. We observe that our method interpolates correct spatial positions for the occluded legs (1^{st} case) and arms (3^{nd} case), achieving significant improvements over FCVG. Furthermore, although the control-free keyframe interpolation methods GI and Wan2.1-KI occasionally generate undistorted humans, they often generate implausible motion that violates real-world dynamics, as observed in 1^{st}, 2^{nd}, and 4^{th} cases.

5.3 Scalability Across Temporal Gaps

Setup. We evaluate the scalability of PoseFuse3D-KI in interpolating frames across different temporal gaps. The temporal gap is defined as the number of frames between the input keyframes, ranging from 24 to 96 frames. Evaluations are conducted on the CHKI-Video test set using PSNR and LPIPS as quantitative metrics.

Results. We provide quantitative results of PoseFuse3D-KI across different temporal gaps in Table 3. As a common trend observed in most interpolation methods, the interpolation performance gradually declines as the interval between keyframes increases. Nevertheless, PoseFuse3D-KI maintains strong performance over a wide range of temporal settings. Notably, it achieves a PSNR of **23.86 dB** and an LPIPS of **0.1144** at 24 frames, and a PSNR of **21.46 dB** and an LPIPS of **0.1566** at 48 frames. Even at a large gap of 96 frames, where the task becomes substantially more challenging,

Figure 4: **Qualitative Comparisons with State-of-The-Art Methods.**

PoseFuse3D-KI remains competitive, achieving a PSNR of **19.44 dB** and an LPIPS of **0.2208**. These results demonstrate the robustness of PoseFuse3D-KI under long-range controllable interpolation.

5.4 In-the-wild Interpolation Results

Our PoseFuse3D-KI framework can be readily applied to interpolate in-the-wild human-centric keyframes **without ground-truth control signals**.

Setup. We assess the in-the-wild interpolation capability of PoseFuse3D-KI on the CHKI-Video dataset by discarding the ground-truth annotations. The comparison is conducted against the state-of-the-art method FCVG [56]. Since ground-truth controls are unavailable in this setting, we employ a simple strategy that linearly interpolates human body joints to generate intermediate control signals. Details of this pipeline are provided in the supplementary material. For performance measurement, we compute the same metrics as described in Sec. 5.2.

Results. We report quantitative results of PoseFuse3D-KI under the in-the-wild interpolation setting in Table 4. Our method achieves state-of-the-art performance and delivers significant improvements over FCVG [56]. On whole-image metrics, it boosts PSNR by **4.5%** and reduces LPIPS by **22.1%**

Table 4: **Comparisons with FCVG on In-the-wild Interpolation.**

Methods	Metrics						
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓	HA↑
FCVG [56]	18.47	15.31	14.08	0.2607	0.1321	0.0921	0.9284
PoseFuse3D-KI (Ours)	**19.30**	**15.75**	**14.46**	**0.2031**	**0.1194**	**0.0859**	**0.9289**

Table 5: **Ablation Study on Visual Encoding and Fusion Module.**

Model Variants		Evaluation Metrics					
Visual Encoding	Fusion Module	PSNR	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓
Conv-Enc (2D)	N.A.	20.29	16.94	16.10	0.1990	0.1096	0.0836
Dual Conv-Enc	Sum	20.50	17.20	16.36	0.1953	0.1059	0.0806
Dual Conv-Enc	Non-TSA	20.30	16.96	16.12	0.1956	0.1053	0.0790
Dual Conv-Enc	Non-WP	20.46	17.19	16.37	0.1938	0.1038	0.0775
Dual Conv-Enc	Full	**20.55**	**17.30**	**16.48**	**0.1927**	**0.1031**	**0.0773**

compared with FCVG. More importantly, PoseFuse3D-KI attains the best results on perceptual human-centric metrics, with an $\text{LPIPS}_{\text{mask}}$ of **0.0859** and an HA score of **0.9289**. These results demonstrate that PoseFuse3D-KI effectively handles in-the-wild interpolation.

5.5 Ablation Study

In this section, we ablate the visual encoding and fusion modules of PoseFuse3D in Table 5. For efficiency, we use SVD as the backbone and process video clips into 9 consecutive frames of 256×256. We use 3× temporally downsampled videos from the CHKI-Video test set for evaluation.

3D Visual Encoding. Our visual encoding module includes two convolutional encoders for 2D and 3D control maps, respectively. We denote the variant including this entire module as 'Dual Conv-Enc', and the one using only the 2D encoder as 'Conv-Enc (2D)'. Removing the 3D visual encoding leads to a performance drop of 0.26 dB on both $\text{PSNR}_{\text{bbox}}$ and $\text{PSNR}_{\text{mask}}$, highlighting the importance of 3D visual encoding.

Fusion Module. In PoseFuse3D, condition features are fused through a carefully designed fusion module. To validate its effectiveness, we replace it with a simple summation operation [55], denoted as 'Sum' in Table 5. This change leads to a significant performance drop, particularly in perceptual quality, with an increase of 0.0033 in $\text{LPIPS}_{\text{mask}}$. These results demonstrate the fusion module's contribution to providing informative control for high-quality interpolation.

Window-Partition Strategy. PoseFuse3D employs a cross-attention layer with a shifted window-partition strategy to fuse features across neighboring windows. To validate this design, we remove the window partitioning, denoted as Non-WP. This results in notable drops of 0.11 dB in both $\text{PSNR}_{\text{bbox}}$ and $\text{PSNR}_{\text{mask}}$, indicating that the window-partition strategy enhances controlled interpolation.

Temporal Attention in Fusion Module. To justify the efficacy of the temporal self-attention (TSA) layer in the fusion module, we conduct experiments excluding the TSA layer (*Non-TSA*). This removal causes increases of 0.0022 and 0.0017 in $\text{LPIPS}_{\text{bbox}}$ and $\text{LPIPS}_{\text{mask}}$, demonstrating the crucial role of the temporal self-attention layers in the fusion module.

6 Conclusion

We propose PoseFuse3D-KI, a controllable human-centric keyframe interpolation framework powered by our novel 3D-informed control model, PoseFuse3D. PoseFuse3D embeds rich spatial geometry from 3D human signals together with 2D poses into a unified control feature, enabling the generation of more plausible and realistic intermediate frames. For evaluation, we construct a CHKI-Video dataset with comprehensive human-centric annotations. Extensive experiments on the benchmark demonstrate that PoseFuse3D-KI outperforms previous interpolation methods with a 9% improvement in PSNR and a 38% reduction in LPIPS.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Pexels. URL `https://www.pexels.com/videos/`, 2025.

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, 2019.

[3] 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.

[4] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[6] Jiaben Chen and Huaizu Jiang. Sportsslomo: A new benchmark and baselines for human-centric video frame interpolation. In *CVPR*, 2024.

[7] Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE TPAMI*, 2021.

[8] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *AAAI*, 2020.

[9] Tianyu Ding, Luming Liang, Zhihui Zhu, and Ilya Zharkov. CDFI: Compression-driven network design for frame interpolation. In *CVPR*, 2021.

[10] 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 *ICML*, 2024.

[11] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *ECCV*, 2024.

[12] Zujin Guo, Wei Li, and Chen Change Loy. Generalizable implicit motion modeling for video frame interpolation. In *NeurIPS*, 2024.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[14] Xinyu Hou, Liming Jiang, Rui Shao, and Chen Change Loy. Video infilling with rich motion prior. In *BMVC*, 2023.

[15] Li Hu. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024.

[16] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 2022.

[17] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018.

[18] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. MiraData: A large-scale video dataset with long durations and structured captions. In *NeurIPS*, 2024.

[19] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. FLAVR: Flow-agnostic video representations for fast frame interpolation. In *WACV*, 2023.

[20] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*. Springer, 2024.

[21] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. IFRNet: Intermediate feature refine network for efficient frame interpolation. In *CVPR*, 2022.

[22] Hyeongmin Lee, Taeoh Kim, Tae-Young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. AdaCoF: Adaptive collaboration of flows for video frame interpolation. In *CVPR*, 2020.

[23] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, 2023.

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2022.

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015.

[28] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *CVPR*, 2017.

[29] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017.

[30] 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 *CVPR*, 2019.

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[32] Tomer Peleg, Pablo Szekely, Doron Sabo, and Omry Sendik. Im-net for high resolution video frame interpolation. In *CVPR*, 2019.

[33] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. ControlNeXt: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[36] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *ICCV*, 2021.

[37] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. In *CVPR*, 2021.

[38] Alexandros Stergiou. Lavib: A large-scale video interpolation benchmark. In *NeurIPS*, 2024.

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

[40] 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 *CVPR*, 2024.

[41] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. Generative Inbetweening: Adapting image-to-video models for keyframe interpolation. In *ICLR*, 2025.

[42] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. HumanVid: Demystifying training data for camera-controllable human image animation. In *NeurIPS*, 2024.

[43] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023.

[44] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019.

[45] 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 *CVPR*, 2024.

[46] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019.

[47] Serin Yang, Taesung Kwon, and Jong Chul Ye. ViBiDSampler: Enhancing video interpolation using bidirectional diffusion sampler. In *ICLR*, 2025.

[48] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023.

[49] Mengying Yu and Aixin Sun. Dataset versus reality: Understanding model performance from the perspective of information need. *JASIST*, 2023.

[50] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *CVPR*, 2023.

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[52] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *ICLR*, 2024.

[53] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[54] Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. RealisDance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024.

[55] 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 *ECCV*, 2024.

[56] Tianyi Zhu, Dongwei Ren, Qilong Wang, Xiaohe Wu, and Wangmeng Zuo. Generative inbetweening through frame-wise conditions-driven video generation. In *CVPR*, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made clear claims in the abstract and introduction that accurately reflect our paper's contributions to the studied problem. Please refer to the abstract and introduction for more detailed information.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have discussions on the limitations in the supplementary materials.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our work does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the implementation details in the main text and training hyperparameters in the supplementary materials.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are unable to provide our code upon submission, but releasing the code to the public in the future is our plan.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the general training information and evaluation settings in the main text. The full implementation details and training hyperparameters are presented in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the type and the number of GPUs used in our experiments.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impacts in the supplementary materials.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In our paper, we have cited the related papers and dataset sources.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Upon submission, we do not provide the source code and the dataset. But we will make them public in the future after acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

7 Appendix

In this section, as referenced in the main text, we first provide detailed descriptions of our method and dataset in Sec. 7.1, Sec. 7.2, and Sec. 7.3. We then present additional experiments on our approach, including in-the-wild interpolation results(Sec. 7.4), extended benchmarking on the FCVG [56] test set (Sec. 7.5), user study 7.6, and ablation studies on the control signals (Sec. 7.7). as well as the SMPL encoder (Sec. 7.8). We also provide supplementary visualizations in Sec. 7.9 and analyze model runtime in Sec. 7.10. Finally, we discuss the limitations and broader impact of our method in Sec. 7.11 and Sec. 7.12.

7.1 Wan2.1 for Keyframe Interpolation

We adapt the Wan2.1 [39] Image-to-Video model (14B, 480P) as the keyframe interpolator Wan2.1-KI along a single temporal forward diffusion path. Specifically, we insert zero-padding between the input keyframes to construct a full-length video sequence. This sequence is then encoded into latents using the VAE encoder of Wan2.1. The resulting latent representation is concatenated with a noisy latent and a latent mask, and passed to the denoising network for prediction. In parallel, the input keyframes are encoded using the image CLIP encoder to produce condition tokens, which guide the denoising process through cross-attention mechanisms. To accommodate changes in both the latent inputs and attention layer inputs, we perform parameter-efficient LoRA fine-tuning on the input embedding layer and on the value and output projection matrices of the attention layers.

7.2 CHKI-Video: Detailed Construction Stages

Our dataset is specifically designed for the Controllable Human-centric Keyframe Interpolation (CHKI) task, emphasizing complex human motions, human-centric annotations, and distant keyframe inputs. To this end, we carefully control the data collection and annotation process, sourcing videos from task-relevant datasets [6] or the internet using specifically curated keywords, rather than relying on large-scale collections [18, 42] that may include irrelevant content.

Stage 1: Dataset Collection. We begin by collecting video clips from SportSlomo [6], which are temporally downsampled to 60 fps due to the large motions typically present in sports scenarios, making them more challenging for keyframe interpolation. To enhance dataset diversity versus reality [49], we additionally crawl high-quality stock videos from the Pexels website. We compile a list of keywords representing fundamental human movements such as 'Walking', 'Kicking', 'Throwing', 'Catching', and 'Climbing', to cover a broad range of human activities. For each keyword, we collect 100 unique videos with resolutions above 720p and durations under 30 seconds. These keywords are grouped into three motion categories: arm motion, leg motion, and general motion, ensuring balanced labels for subsequent train-test splitting. To match the motion characteristics of the SportSlomo videos, we downsample the collected stock videos based on their optical flow scores, ensuring the flow score distributions are aligned.

Stage 2: Pre-annotation Processing. To ensure the quality of the collected videos, we use DOVER [43] to obtain both technical and overall quality scores, and compute brightness change scores between adjacent frames. Videos falling below the bottom 5th percentile in any of these metrics are filtered out. Given the importance of accurate human detection for downstream keypoint and SMPL-X annotation, we design a robust detection pipeline. We combine Grounding-DINO [25] with SAM2 [34] to achieve reliable human detection. For challenging sports scenes, we prioritize videos with prominent foreground humans and relatively static or blurred backgrounds, striking a balance between annotation complexity and scenario diversity. Additionally, we exclude videos containing more than three people or fewer than 20 frames to maintain clean motion patterns and ensure sufficient temporal coverage. All sports videos are manually reviewed to verify compliance with these criteria and to confirm the accuracy of the human detections.

Stage 3: Human-centric Annotation. We perform frame-wise human-centric annotations for all video clips based on the detections in the previous stage. First, we use Sapiens [20] to estimate whole-body keypoints. To ensure the dataset remains strictly human-centric, we perform whole-body detection based on these keypoints. Specifically, we extract the keypoints into DWPose to better define the human figure. We merge all head keypoints into a single point, as significant motion rarely occurs in that region. A whole-body detection is considered valid if it contains fewer than three

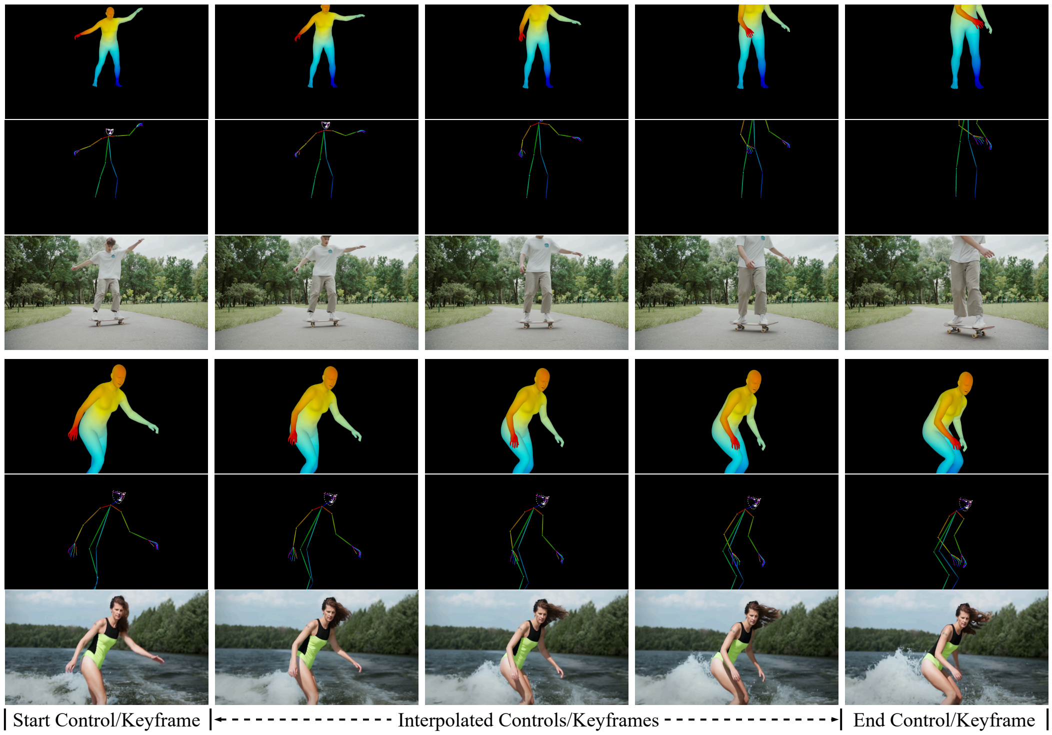


Figure 5: **Qualitative Results of In-the-wild Control and Keyframe Interpolation.**

invalid keypoints, using a keypoint score threshold of 0.3. We further filter video clips to retain only those with more than 20 consecutive valid frames. Finally, we apply SMPLer-X [4], which provides high re-projection accuracy, to fit detailed SMPL-X models to each frame and generate reliable 3D body parameters.

7.3 Implementation Details

We fine-tune the entire PoseFuse3D-KI framework in an end-to-end manner using the AdamW optimizer with a learning rate of 8×10^{-5}. The fine-tuning is applied to our 3D-informed control model, PoseFuse3D, with additional LoRA adaptation on the input patch embeddings, as well as the value and output projections of the VDM's attention modules. Both the LoRA rank and LoRA alpha are set to 32. For implementation, we leverage Fully Sharded Data Parallel (FSDP) across 4 GPUs.

7.4 In-the-wild Interpolation

Our PoseFuse3D-KI framework can be readily applied to interpolate in-the-wild human-centric keyframes. In this subsection, we present a simple pipeline that uses linear interpolation for human body joints. Given a human-centric keyframe pair I_0, I_N, we first employ a 3D human model estimator, such as SMPLer-X [4], to fit SMPL-X models [30] for each keyframe input. Leveraging the strong human body priors from SMPL-X, we linearly interpolate the SMPL-X parameters to generate intermediate 3D human models, which serve as control signals for interpolation. We then extract 2D DWPose keypoints from the 2D projections of the interpolated SMPL-X models. With these steps, all necessary guidance inputs for PoseFuse3D-KI are prepared and can be directly used for keyframe interpolation. Figure 5 shows the results of the interpolated SMPL-X models and the corresponding video frames. While this pipeline offers a straightforward and efficient approach for generating intermediate poses, it is limited in its ability to handle complex human motion. In particular, it struggles to capture realistic dynamics such as acceleration or multi-step actions. To address these limitations, we could explore text-to-motion models [52] as part of future work, which generate intermediate SMPL-X models from textual descriptions, offering a more flexible and semantically rich alternative to linear interpolation.

Table 6: **Benchmark Results on FCVG-Test-HC.**

Methods	Metrics						
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓	HA↑
GIMM-VFI [12]	23.61	21.25	20.28	0.1324	0.0759	0.0587	**0.9459**
GI [41]	17.21	15.20	14.36	0.2701	0.1422	0.1045	0.9438
Wan2.1-KI (Ours)	21.50	19.03	18.17	0.1553	0.0915	0.0704	0.9312
FCVG [56]	22.49	21.03	20.69	0.1738	0.0734	0.0493	0.9241
PoseFuse3D-KI (Ours)	**24.84**	**22.97**	**22.26**	**0.0915**	**0.0460**	**0.0340**	0.9245

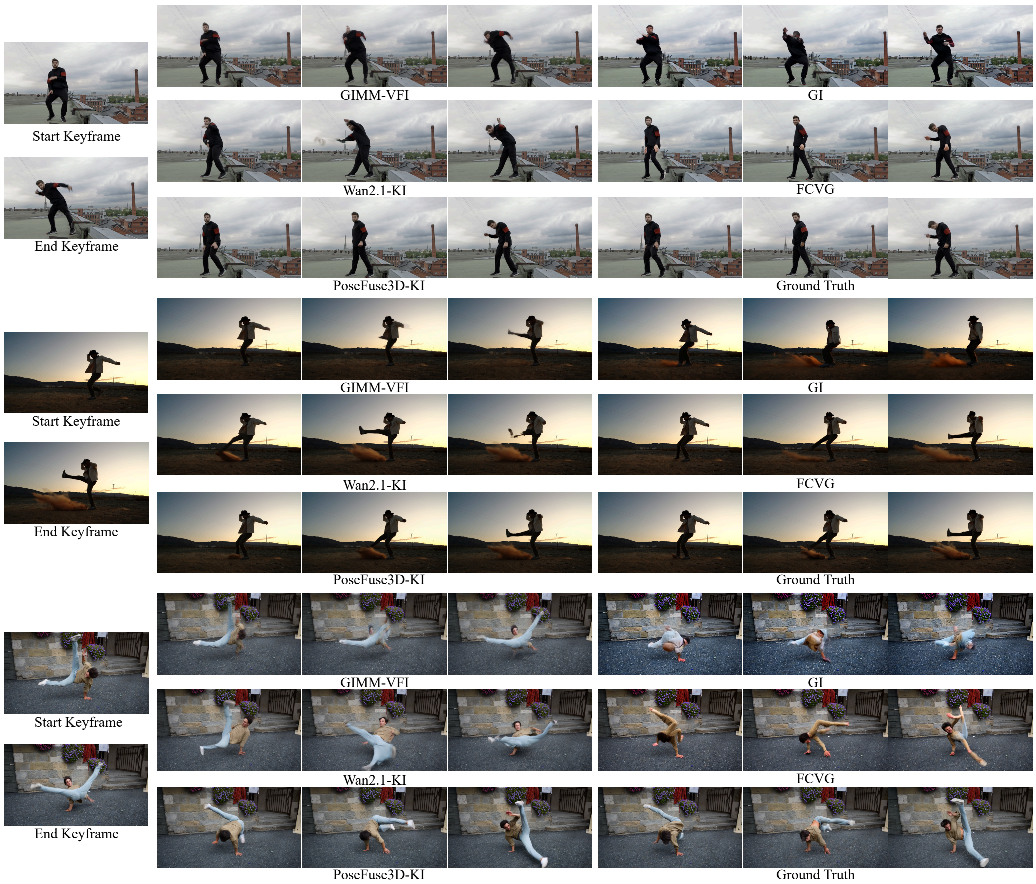


Figure 6: **Qualitative Comparisons on FCVG-Test-HC.**

7.5 Evaluation on Additional Benchmark

Setup. We conduct an additional evaluation on the test set from FCVG [56]. We first perform human detection and extract human-centric video clips from the test set. The extracted videos are then annotated using the same processing pipeline detailed in Sec. 7.2. This results in FCVG-Test-HC, a curated human-centric subset of 54 clips suitable for CHKI benchmarking. The FCVG-Test-HC benchmark is relatively easier than CHKI-Video, primarily consisting of human-centric clips with limited motion rather than more challenging scenarios such as sports and dancing. Other benchmarking settings follow those described in the main paper.

Results. We present quantitative comparisons on the FCVG-Test-HC benchmark in Table 6. Our PoseFuse3D-KI outperforms other methods for human-centric keyframe interpolation. Compared with the previous state-of-the-art method FCVG [56], our method achieves a 7.6% improvement in $\text{PSNR}_{\text{mask}}$ and a 31% reduction in $\text{LPIPS}_{\text{mask}}$. We observe that all methods achieve higher PSNR and lower LPIPS scores on the FCVG-Test-HC benchmark compared to the CHKI-Video dataset, indicating that FCVG-Test-HC is an easier benchmark for interpolation. This aligns with our earlier observation during dataset construction, where FCVG-Test-HC primarily consists of human-centric

Table 7: **User Study.**

Methods	GIMM-VFI [12]	GI [41]	FCVG [56]	PoseFuse3D-KI (Ours)
User Preference (%)	1.75	5.13	6.63	**86.50**

Table 8: **Ablation on the Visual Encoding.**

Model Variant	Evaluation Metrics					
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓
Non-Vis	19.63	16.02	14.69	0.2097	0.1232	0.0889
Non-2D	21.71	18.65	17.28	0.1438	0.0738	0.0531
Full	**22.14**	**19.30**	**18.01**	**0.1330**	**0.0653**	**0.0464**

clips with limited motion. Interestingly, methods with fewer learned priors tend to achieve higher HA scores in this setting. For instance, the traditional interpolation method GIMM-VFI [12] records the highest Human Anatomy (HA) score. This is likely because such methods rely more heavily on the input keyframes. While this reliance leads to motion artifacts under large movement, it better preserves human textures from inputs when the motion between keyframes is small.

Visualizations. We qualitatively compare PoseFuse3D-KI with other advanced methods on the FCVG-Test-HC benchmark, as shown in Figure 6. Consistent with our findings in the Benchmark Results section of the main paper, our method achieves robust human-centric interpolation, accurately follows real-world dynamics, and effectively preserves human body shape. For instance, our method generates plausible interpolations of complex body movements while maintaining the correct leg structure and posture in the last 'Breaking Dance' case.

7.6 User Study

Setup. To further assess the perceptual quality of our controllable human-centric interpolation, we conducted a user study involving 20 interpolation scenarios sampled from both the CHKI-Video test set and FCVG-Test-HC. We compared PoseFuse-KI with three representative state-of-the-art methods: GIMM-VFI [12], GI [41], and FCVG [56]. A total of 40 participants were invited to take part in the study, where each participant was asked to select their preferred interpolation result.

Results. We report the user preference of the study in Table 7. Our method received a strong majority of user preferences (86.5%), consistently outperforming all baselines. These results further demonstrate the perceptual effectiveness of PoseFuse3D-KI.

7.7 Ablation Study on Visual Encoding

The core of our framework is the control module, PoseFuse3D. As detailed in the Method section of the main paper, it includes encoding visualizations from both SMPL-X and DWPose [48]. To evaluate the importance of encoding these visualizations and explore whether 2D visual cues can be omitted, we conduct an ablation study on the visual encoding component of PoseFuse3D.

Necessity of Encoding Visualizations. In PoseFuse3D, we encode visualizations of control signals. Since they preserve natural pixel-level alignment with the video latent, thereby providing direct control signals on the pixel plane. To assess its necessity, we ablate all visual encoding components in PoseFuse3D and rely solely on the SMPL-X encoded information as the control representation. We refer to this variant as 'Non-Vis'. In Table 8, this modification results in a significant performance degradation, with a 4.32 dB drop in $\text{PSNR}_{\text{mask}}$ and a 0.0425 increase in $\text{LPIPS}_{\text{mask}}$, underscoring the critical role of encoding visualizations in achieving high-fidelity results.

Importance of Encoding 2D Visualization. The visual encoding module of PoseFuse3D integrates 2D DWPose visualizations with rendered SMPL-X images. The 2D DWPose visualizations emphasize skeletal keypoint layouts, contributing to robust pose understanding. To assess its importance, we exclude the encoding of 2D visualizations, denoting this variant as 'Non-2D'. This leads to a drop of 0.65 dB in $\text{PSNR}_{\text{bbox}}$ and a 13% increase in $\text{LPIPS}_{\text{bbox}}$, demonstrating the significance of encoding 2D visualizations in the visual encoding module.

Table 9: **Ablation on SMPL-X Encoder.**

Model Variant	Evaluation Metrics					
	PSNR↑	$\text{PSNR}_{\text{bbox}}$↑	$\text{PSNR}_{\text{mask}}$↑	LPIPS↓	$\text{LPIPS}_{\text{bbox}}$↓	$\text{LPIPS}_{\text{mask}}$↓
Non-JA	22.07	19.24	17.95	0.1348	0.0659	0.0466
Non-VA	**22.15**	19.27	17.99	0.1374	0.0667	0.0470
Full	22.14	**19.30**	**18.01**	**0.1330**	**0.0653**	**0.0464**

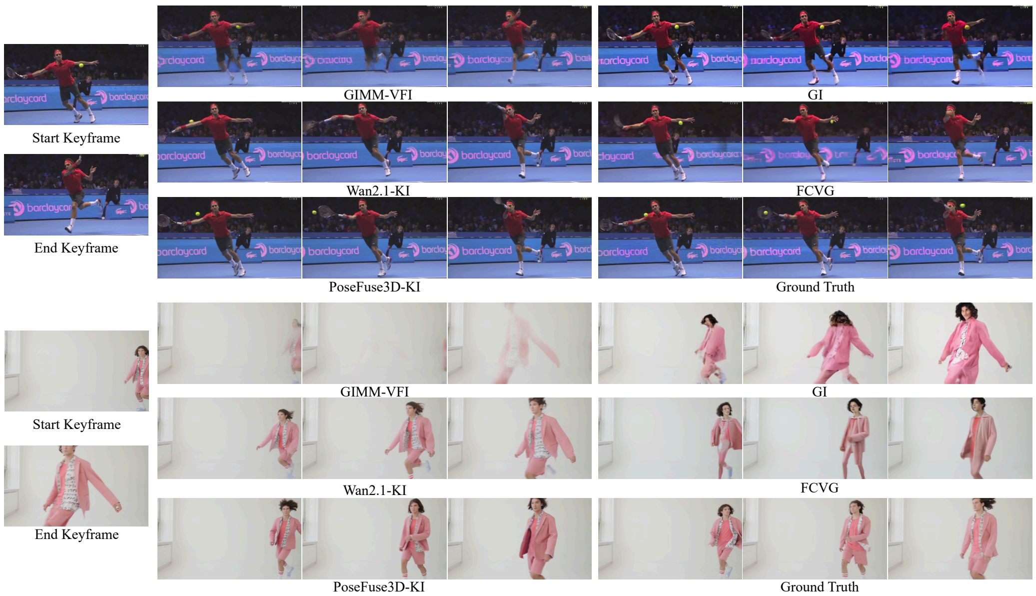


Figure 7: **Additional Qualitative Results on CHKI-Video.**

7.8 Ablation on SMPL-X Encoder

We conduct an additional ablation study on the SMPL-X encoder to justify our design.

Joint Aggregation. The SMPL-X encoder extracts joint motion and position features from 3D space and projects them onto the 2D image plane via an attention mechanism, providing spatial human body motion cues. To assess the impact of this design, we remove the joint aggregation module, denoted as 'Non-JA'. As shown in Table 9, this leads to a 0.06 dB drop in both $\text{PSNR}_{\text{bbox}}$ and $\text{PSNR}_{\text{mask}}$, emphasizing the importance of 3D joint aggregation for accurate body control representation.

Vertex Aggregation. We also apply a separate attention mechanism to aggregate vertex information into the 2D image plane. To examine its necessity, we remove the vertex attention module. We denote this variant as 'Non-VA'. As reported in Table 9, this leads to a noticeable degradation in performance across all LPIPS scores, including a 0.0040 rise in LPIPS and a 0.0014 increase in $\text{LPIPS}_{\text{bbox}}$. These results demonstrate the significance of incorporating 3D vertex information for effective control representation from SMPL-X.

7.9 Supplementary Visualizations

Additional Qualitative Comparisons. We present additional qualitative comparisons with other interpolation methods in Figure 7. Our PoseFuse3D-KI framework consistently produces more plausible interpolations, closely capturing real-world dynamics observed in the ground truth.

Robustness across Corner Cases. PoseFuse3D-KI exhibits robust interpolation performance across diverse corner cases. As shown in Figure 8, we evaluate scenarios such as off-center subjects, camera motion, multiple persons, virtual animation, and robotics videos. Our method consistently produces controllable interpolations with plausible and temporally coherent motion.

Table 10: **Runtime Comparison.**

Methods	GI [41]	FCVG [56]	PoseFuse3D-KI (Ours)
Runtime (s)	975	523	**212**

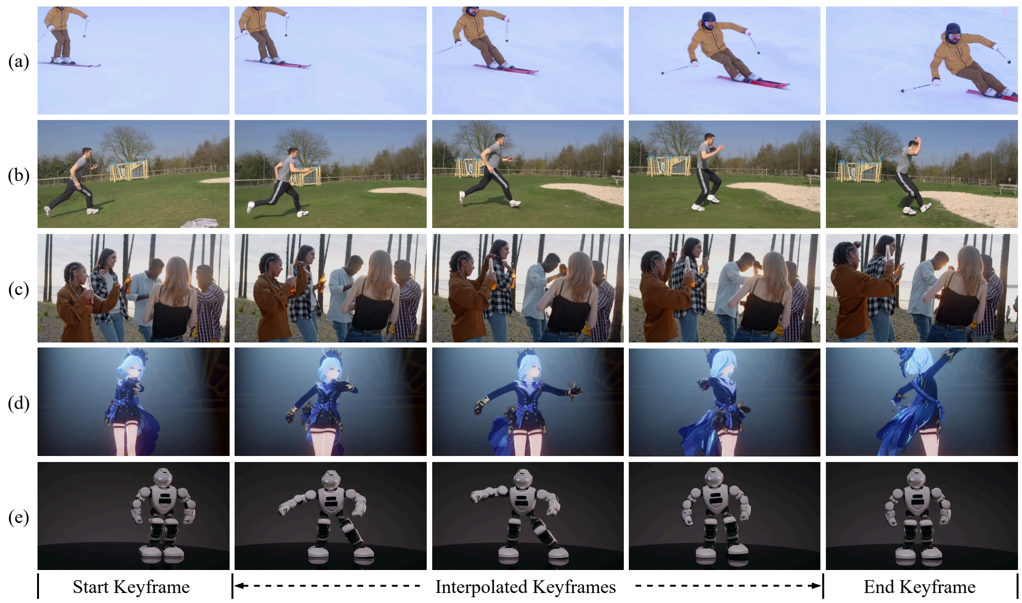


Figure 8: **Qualitative Results on Corner Cases.** (a) Person off-center. (b) Camera motion. (c) Multiple persons. (d) Virtual animation. (e) Robotics video.

7.10 Runtime Comparison

To assess computational efficiency, we compare the runtime of PoseFuse3D-KI with existing diffusion-based baselines. The comparison measures the time required to interpolate 25 frames at a resolution of 1024×576, using the same GPU equipped with 80 GB memory. As summarized in Table 10, PoseFuse3D-KI exhibits the shortest runtime, confirming its efficiency advantage over competing methods.

7.11 Limitations

There are several known limitations to our method. First, PoseFuse3D-KI relies on accurate SMPL-X estimations to generate reliable 3D control signals. Therefore, it inherits the limitations of the 3D human model estimators, where inaccurate predictions can degrade the quality of interpolated results. Additionally, our method, while offering strong control via 3D and 2D fusion, still depends on the base diffusion model's generative priors. As a result, output quality is influenced by the model's learned behavior and inherits its high GPU memory demands. Finally, our method does not explicitly model human-object interactions, which may lead to artifacts or misaligned object motion in scenarios involving close interaction with external objects.

7.12 Broader Impacts

Our proposed method, PoseFuse3D-KI, enables accurate and controllable human-centric keyframe interpolation, with applications in areas such as human animation and video generation. By integrating explicit 3D information from human models and 2D pose cues, our framework supports 3D-informed and semantically meaningful guidance for interpolating realistic human motion across frames. This technique not only enriches creative workflows but also opens new opportunities for research in human motion understanding and video synthesis. While powerful, our method shares common limitations of generative models and may pose risks if misused to produce manipulated or deceptive human videos, highlighting the importance of responsible use and ethical safeguards.