
Controllable Human-centric Keyframe Interpolation with Generative Prior: Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

1 Appendix

In this section, as referenced in the main text, we first provide detailed descriptions of our method and dataset in Sec. 1.1, Sec. 1.2, and Sec. 1.3. We then present additional experiments on our approach, including in-the-wild interpolation results (Sec. 1.4), extended benchmarking on the FCVG [13] test set (Sec. 1.5), and ablation studies on the control signals (Sec. 1.6). as well as the SMPL encoder (Sec. 1.7). Additional qualitative results on CHKI-Video are provided in Sec. 1.8. Finally, we discuss the limitations and broader impact of our method in Sec. 1.9 and Sec. 1.10.

1.1 Wan2.1 for Keyframe Interpolation

We adapt the Wan2.1 [8] 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 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.

1.2 CHKI-Video: Detailed Construction Stages.

Stage 1: Dataset Collection. We begin by collecting video clips from SportSlomo [2], 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, 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 [10] 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 [5] with SAM2 [7] to achieve reliable human detection. For challenging sports scenes, we prioritize

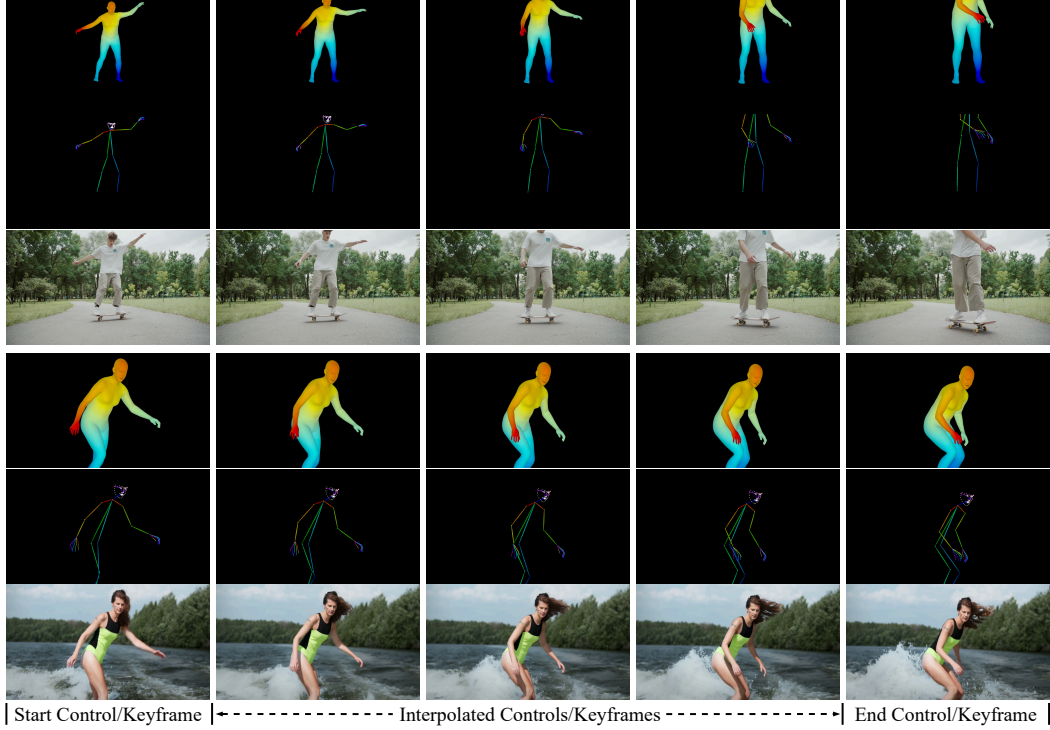


Figure 1: **Qualitative Results of In-the-wild Control and Keyframe Interpolation.**

36 videos with prominent foreground humans and relatively static or blurred backgrounds, striking a
 37 balance between annotation complexity and scenario diversity. Additionally, we exclude videos
 38 containing more than three people or fewer than 20 frames to maintain clean motion patterns and
 39 ensure sufficient temporal coverage. All sports videos are manually reviewed to verify compliance
 40 with these criteria and to confirm the accuracy of the human detections.

41 **Stage 3: Human-centric Annotation.** We perform frame-wise human-centric annotations for all
 42 video clips based on the detections in the previous stage. First, we use Sapiens [4] to estimate
 43 whole-body keypoints. To ensure the dataset remains strictly human-centric, we perform whole-body
 44 detection based on these keypoints. Specifically, we extract the keypoints into DWPose to better
 45 define the human figure. We merge all head keypoints into a single point, as significant motion rarely
 46 occurs in that region. A whole-body detection is considered valid if it contains fewer than three
 47 invalid keypoints, using a keypoint score threshold of 0.3. We further filter video clips to retain only
 48 those with more than 20 consecutive valid frames. Finally, we apply SMPLer-X [1], which provides
 49 high re-projection accuracy, to fit detailed SMPL-X models to each frame and generate reliable 3D
 50 body parameters.

51 1.3 Implementation Details

52 We fine-tune the entire PoseFuse3D-KI framework in an end-to-end manner using the AdamW
 53 optimizer with a learning rate of 8×10^{-5} . The fine-tuning is applied to our 3D-informed control
 54 model, PoseFuse3D, with additional LoRA adaptation on the input patch embeddings, as well as the
 55 value and output projections of the VDM’s attention modules. Both the LoRA rank and LoRA alpha
 56 are set to 32. For implementation, we leverage Fully Sharded Data Parallel (FSDP) across 4 NVIDIA
 57 A100 80GB GPUs.

58 1.4 In-the-wild Interpolation

59 Our PoseFuse3D-KI framework can be readily applied to interpolate in-the-wild human-centric
 60 keyframes. In this subsection, we present a simple pipeline that uses linear interpolation for human
 61 body joints. Given a human-centric keyframe pair I_0, I_N , we first employ a 3D human model
 62 estimator, such as SMPLer-X [1], to fit SMPL-X models [6] for each keyframe input. Leveraging

Table 1: **Benchmark Results on FCVG-Test-HC.**

Methods	Metrics						
	PSNR \uparrow	PSNR $_{\text{bbox}}\uparrow$	PSNR $_{\text{mask}}\uparrow$	LPIPS \downarrow	LPIPS $_{\text{bbox}}\downarrow$	LPIPS $_{\text{mask}}\downarrow$	HA \uparrow
GIMM-VFI [3]	23.61	16.48	15.51	0.1324	0.0759	0.0587	0.9459
GI [9]	17.21	10.43	9.59	0.2701	0.1422	0.1045	0.9438
Wan2.1-KI (Ours)	21.50	14.26	13.40	0.1553	0.0915	0.0704	0.9312
FCVG [13]	22.49	16.26	15.92	0.1738	0.0734	0.0493	0.9241
PoseFuse3D-KI (Ours)	24.84	18.20	17.49	0.0915	0.0460	0.0340	0.9245

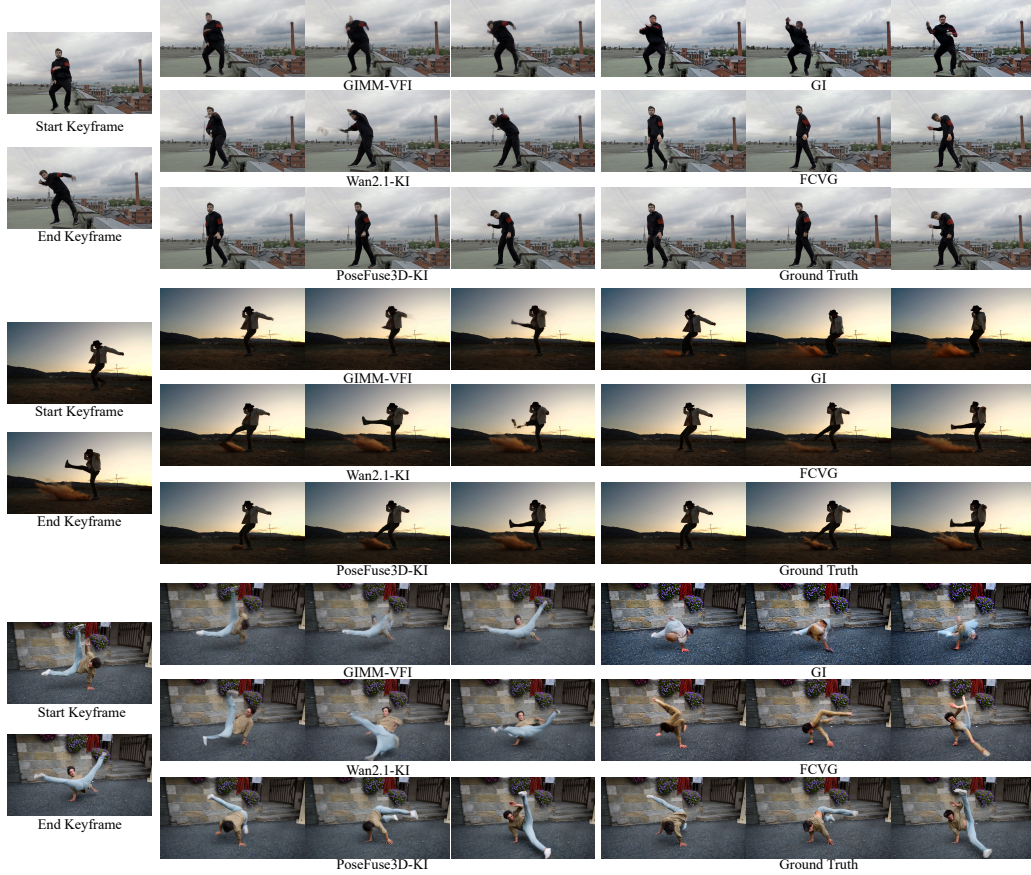


Figure 2: **Qualitative Comparisons on FCVG-Test-HC.**

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 1 shows the results of the interpolated SMPL-X models and the corresponding video frames. As part of future work, the intermediate SMPL-X models could be generated more flexibly using text-to-motion models [12], which synthesize in-between motion from textual descriptions.

1.5 Evaluation on Additional Benchmark

Setup. We conduct an additional evaluation on the test set from FCVG [13]. 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. 1.2. 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

Table 2: Ablation on the Visual Encoding.

Model Variant	Evaluation Metrics					
	PSNR \uparrow	PSNR $_{\text{bbox}}\uparrow$	PSNR $_{\text{mask}}\uparrow$	LPIPS \downarrow	LPIPS $_{\text{bbox}}\downarrow$	LPIPS $_{\text{mask}}\downarrow$
Non-Vis	19.63	11.25	9.92	0.2097	0.1232	0.0889
Non-2D	21.71	13.88	12.51	0.1438	0.0738	0.0531
Full	22.14	14.53	13.24	0.1330	0.0653	0.0464

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 1. Our PoseFuse3D-KI outperforms other methods for human-centric keyframe interpolation. Compared with the previous state-of-the-art method FCVG [13], our method achieves a 10% improvement in PSNR $_{\text{mask}}$ and a 31% reduction in 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 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 [3] 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 2. 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.

1.6 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 [11]. 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 2, this modification results in a significant performance degradation, with a 4.32 dB drop in PSNR $_{\text{mask}}$ and a 0.0425 increase in 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 PSNR $_{\text{bbox}}$ and a 13% increase in LPIPS $_{\text{bbox}}$, demonstrating the significance of encoding 2D visualizations in the visual encoding module.

1.7 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

Table 3: Ablation on SMPL-X Encoder.

Model Variant	Evaluation Metrics					
	PSNR \uparrow	PSNR $_{\text{bbox}}\uparrow$	PSNR $_{\text{mask}}\uparrow$	LPIPS \downarrow	LPIPS $_{\text{bbox}}\downarrow$	LPIPS $_{\text{mask}}\downarrow$
Non-JA	22.07	14.47	13.18	0.1348	0.0659	0.0466
Non-VA	22.15	14.50	13.22	0.1374	0.0667	0.0470
Full	22.14	14.53	13.24	0.1330	0.0653	0.0464

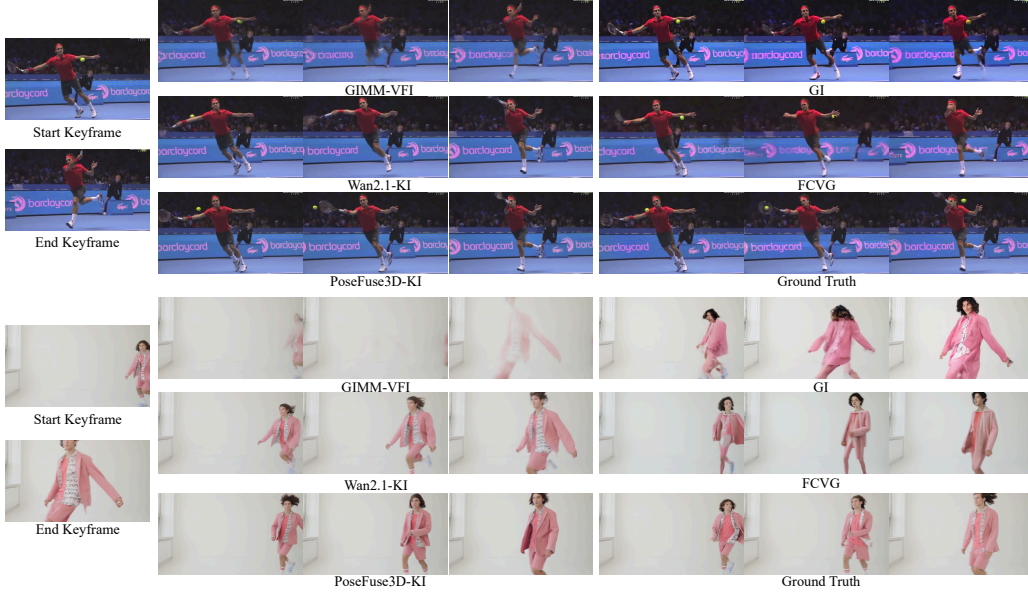


Figure 3: Additional Qualitative Results on CHKI-Video.

as ‘Non-JA’. As shown in Table 3, this leads to a 0.06 dB drop in both PSNR $_{\text{bbox}}$ and 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 3, 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 LPIPS $_{\text{bbox}}$. These results demonstrate the significance of incorporating 3D vertex information for effective control representation from SMPL-X.

1.8 Additional Qualitative Comparisons

We present additional qualitative comparisons with other interpolation methods in Figure 3. Our PoseFuse3D-KI framework consistently produces more plausible interpolations, closely capturing real-world dynamics observed in the ground truth.

1.9 Limitations

There are several known limitations to our method. First of all, 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.

141 **1.10 Broader Impacts**

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

References

- [1] 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.
- [2] Jiaben Chen and Huaizu Jiang. SportssloMo: A new benchmark and baselines for human-centric video frame interpolation. In *CVPR*, 2024.
- [3] Zujin Guo, Wei Li, and Chen Change Loy. Generalizable implicit motion modeling for video frame interpolation. In *NeurIPS*, 2024.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.
- [11] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023.
- [12] Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *ICLR*, 2024.
- [13] Tianyi Zhu, Dongwei Ren, Qilong Wang, Xiaohe Wu, and Wangmeng Zuo. Generative inbetweening through frame-wise conditions-driven video generation. In *CVPR*, 2025.