

# PoseAnything: General Pose-guided Video Generation with Part-aware Temporal Coherence (Supplementary Material)

Ruiyan Wang\* Teng Hu\* Kaihui Huang Zihan Su  
Ran Yi† Lizhuang Ma  
Shanghai Jiao Tong University

Project page: <https://ryan-w2024.github.io/project/PoseAnything/>

## 1. Overview

In this supplementary material, more details about the proposed PoseAnything method and more experimental results are provided, including:

- More Details about our dataset XPose (sec 2)
- Details about attention weights based part matching in part-aware temporal coherence module (sec 3)
- details about motion decoupled CFG (sec 4)
- Ablation on Condition Injection Strategies (sec 5)
- Ablation on Sparse Pose Condition Injection (sec 6)
- Analysis of effect of CFG on pose-guided video generation of PoseAnything(sec 7)
- More generation results of our PoseAnything (sec 8)
- More comparison results (sec 9)
- Qualitative Ablation Results on PTCM (sec 10)
- More results on Generalization (sec 11)

More videos and comparison results are available on the provided **project page** and **demo video**. Please refer to them for further inspection.

## 2. Details about XPose

To effectively tackle the task of universal pose-guided video generation, which needs both robust generalization capabilities and precise pose-to-subject mapping, we curated a high-quality, non-human pose dataset termed **XPose**. As described in the main paper, the XPose dataset was generated from Koala [11] and UltraVideo [14] via a carefully designed pipeline coupled with a filtering algorithm to ensure its fidelity and diversity. In this section, we present a subset of XPose, offering a clear overview of its structural characteristics, complexity, and overall quality. To provide a more comprehensive and systematic demonstration, the dataset is stratified into three complexity levels according to the number of skeleton segments present: *Simple* (1–3 segments), *Medium* (4–6 segments), and *Complex* (7–10 seg-

ments), which corresponds to Fig 6 7 8, respectively.

This visualization highlights two critical features of XPose: (1) **Diversity**: XPose exhibits a high degree of diversity, encompassing a wide range of backgrounds and dynamic motions. The dataset includes scenes set in diverse environments (e.g., underwater, sky, wilderness, home interiors) and features varied movements (e.g., twisting heads, running, swimming). (2) **High Fidelity** of Skeletons: The visualization confirms the high fidelity of the extracted skeletons, ensuring accurate pose representation. The high diversity in subject poses and high quality of the extracted skeletons is crucial ensures that the model learns accurate, noise-resistant pose-to-subject mapping, achieving the necessary robustness and high-fidelity generation required for the universal pose-guided video generation task. It provides strong support not only for our current universal generation method but also for future research in related work.

## 3. Details about Part-aware Temporal Coherence Module

As detailed in the main paper, we designed a part-aware temporal coherence module for ensuring fine-grained inter-frame consistency. It is accomplished by decomposing the subject into distinct parts, establishing correspondences between identical parts across different frames, and enforcing these correspondences through cross-attention between matched part pairs. In the part-aware coherence module, the mechanism of part-matching is realized through the attention weights within the attention layers of the DiT [6] blocks according to the formula:

$$s_{ij'} \sim s_{0j} \iff j' = \arg \max_t \text{attn\_weight}[m_{0j}][m_{it}]. \quad (1)$$

Specifically, the underlying principle dictates that the attention weights between corresponding parts across different frames should be significantly higher than those between non-corresponding parts. However, the diffusion process inherently involves multiple diffusion timesteps ( $t$ ), and

---

\*Equal Contribution

†Corresponding author.

the model’s primary denoise focus and priorities shift significantly as the signal-to-noise ratio changes across these timesteps. For instance, at higher noise levels, the model tends to prioritize global structural features, whereas at lower noise levels, the focus shifts towards denoising and refining fine-grained details. Additionally, each forward pass traverses approximately 30 DiT blocks, with each block attending to different domains of information.

To accurately achieve the desired part-level matching through attention weights, we conducted a systematic visualization of attention weights across various timesteps and different depths (DiT block indices) within the network. An illustrative example of these visualizations is presented in Fig. 1. In this example, to obtain the visualization results of the attention weights, we designated the region corresponding to the sea turtle’s flipper (highlighted in yellow) in the first frame to compute the query vector ( $q$ ), while the tokens from the corresponding regions in subsequent frames were used to compute the key and value vectors ( $k, v$ ). For clarity of presentation, we only selected the token corresponding to  $frame_i$  to compute  $k$  and  $v$ . The resulting attention weights were averaged across the head dimension and then reshaped into an  $h \times w$  spatial map.

The visualizations consistently demonstrate that the flipper region in subsequent frames appears significantly brighter, indicating that the token corresponding to the flipper in the first frame attends more strongly to the flipper region in subsequent frames. This observation is in accordance with our part-matching principle. Furthermore, this attention intensity difference is more pronounced at higher diffusion timesteps. Comparing attention weights across different block indices reveals that deeper blocks exhibit a more uniform attention distribution, whereas shallower blocks display more dispersed attention. Based on these findings, we select the attention weights from block whose id is 27 at timesteps greater than 975 as the basis for our part-matching mechanism.

#### 4. Details about Motion Decoupled CFG

In the main paper, we propose the subject-camera decoupled motion control CFG. By injecting the control conditions for subject and background motion respectively into the positive and negative anchors of CFG, we effectively prevent mutual interference between these two motion components. In this section, we provide a detailed explanation of the underlying principles behind this approach.

The proposed motion decoupling approach can be rigorously interpreted through the lens of optical flow and latent vector composition. In video generation, optical flow  $\mathbf{F}_t$  characterizes the pixel-wise displacement between consecutive frames  $I_t$  and  $I_{t+1}$ , encapsulating both subject and

background dynamics:

$$\mathbf{F}_t = \mathbf{F}_{\text{subject},t} + \mathbf{F}_{\text{camera},t} \quad (2)$$

where  $\mathbf{F}_{\text{subject},t}$  denotes subject-induced motion, and  $\mathbf{F}_{\text{camera},t}$  represents camera-induced (background) motion.

By explicitly formulating the guidance vectors for subject and camera motion within the Classifier-Free Guidance (CFG) framework, we effectively decompose the overall optical flow field into independent, controllable components in the latent space. Let  $\Delta\epsilon_{\text{subject}}$  and  $\Delta\epsilon_{\text{camera}}$  be the guidance vectors corresponding to subject and camera motion, respectively. The overall latent noise guidance is then expressed as:

$$\tilde{\epsilon} = \hat{\epsilon}_{\theta}(\mathbf{z}_t, \emptyset) + s_s \cdot \Delta\epsilon_{\text{subject}} - s_c \cdot \Delta\epsilon_{\text{camera}} \quad (3)$$

where  $s_s$  and  $s_c$  are scalar weights controlling the strength of subject and camera guidance. Specifically, the subject motion guidance vector  $\Delta\epsilon_{\text{subject}}$  enforces the desired subject trajectory as defined in the pose sequence, steering the latent noise prediction toward the target action. This can be formulated as:

$$\Delta\epsilon_{\text{subject}} = \hat{\epsilon}_{\theta}(\mathbf{z}_t, \mathbf{c}_{\text{subject}}) - \hat{\epsilon}_{\theta}(\mathbf{z}_t, \emptyset) \quad (4)$$

where  $\mathbf{c}_{\text{subject}}$  denotes the subject pose condition.

Meanwhile, camera motion is mathematically equivalent to imposing a spatially uniform optical flow field  $\mathbf{F}_{\text{camera},t}$  across the background, as a camera pan or tilt requires all background pixels to move coherently in the opposite direction of the camera’s intended movement. In the latent space, this is encoded by:

$$\Delta\epsilon_{\text{camera}} = \hat{\epsilon}_{\theta}(\mathbf{z}_t, \mathbf{c}_{\text{camera}}) - \hat{\epsilon}_{\theta}(\mathbf{z}_t, \emptyset) \quad (5)$$

where  $\mathbf{c}_{\text{camera}}$  specifies the camera motion condition.

The final guided motion prediction  $\tilde{\epsilon}$  thus emerges as a superposition of these vectors (Eq. 2), analogous to the principle of vector addition in classical mechanics. Each motion component, subject-specific ( $V_s$ ) and background/camera-induced ( $V_{bg}$ ), contributes linearly to the aggregate optical flow. Notably, since the background control condition is injected into the negative anchor of the CFG, the resulting latent guidance vector for background motion  $V_{bg}$ , in the final merged result, is synthesized in a direction opposite to that indicated by the provided control condition. Mathematically, this relationship is captured by the final guided optical flow equation:

$$\mathbf{F}_t = s_s V_s - s_c V_{bg} \quad (6)$$

where  $s_s$  and  $s_c$  are the respective guidance strengths for subject and background,  $V_s$  denotes subject motion, and  $V_{bg}$  is the guidance vector derived from background control information. This principled design guarantees that the model

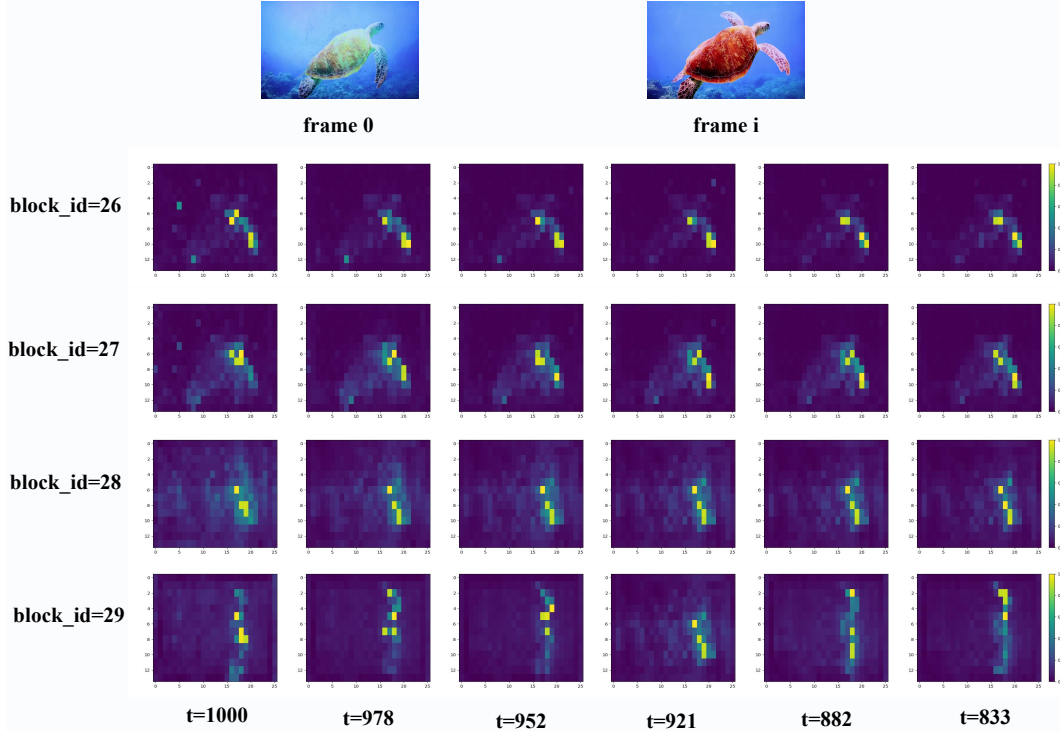


Figure 1. Visualization of attention weights

correctly interprets and synthesizes both subject and background motion trajectories, thereby enabling disentangled and physically coherent video generation under joint control conditions.

This compositionality enables precise and independent control over both foreground and background movements, allowing for the joint or separate manipulation of subject actions and camera transitions via adjustment of the respective guidance strengths  $s_s$  and  $s_c$  within the CFG framework. Consequently, our approach facilitates flexible, disentangled video synthesis, where complex scene dynamics can be intuitively controlled by the user.

## 5. Ablation on Condition Injection Strategies

In the main paper, we propose three distinct strategies for incorporating pose condition into our framework: (1) **Concatenation by Channel** as [2] [9], (2) **Concatenation by Width** used in [1], and (3) **Multi-layer Perceptron (MLP)-based fusion**, in which the pose latent  $Z_p$  is first processed by an MLP to align its shape with the reference latent  $Z_0$ , and subsequently added to  $Z_0$ . The architectural details of the three injection mechanisms are illustrated in Fig. 2. To systematically compare the effectiveness of the proposed injection methods, we performed a dedicated ablation study, evaluating each approach both quantitatively and qualitatively on the TikTok dataset [4]. For this comparison, models incorporating only the respective injection strate-

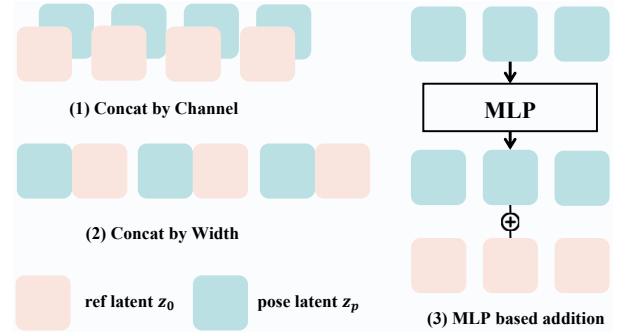


Figure 2. Different Condition Injection Strategies

gies—excluding the Part-aware Temporal Coherence Module (PTCM)—were trained for 3,000 iterations. All models were trained using consistent hyper parameters: a batch size of 32 and a learning rate of  $2e - 5$ .

**Quantitative results.** The following five metrics are employed to quantitatively compare the effectiveness of the three skeleton information injection strategies. For image-based metrics, the metric is first computed for each frame of a video, and the average across all frames is then taken as the final metric value for the video.

- **PSNR [3]:** The Peak Signal-to-Noise Ratio (PSNR) is one of the most prevalent and extensively utilized metrics for assessing image quality. A higher PSNR value indicates a superior quality of image reconstruction.

- SSIM [13]: Structure Similarity Index Measure is derived from three aspects of image similarity: luminance, contrast and structure, based on the idea that the pixels have strong inter-dependencies especially when they are spatially close. The higher the SSIM score is, the more similar the two images are.
- L1: The L1 metric, quantifies the average absolute difference between the predicted and reference images. A lower L1 value signifies a closer resemblance between the reconstructed image and the ground truth.
- LPIPS [15]: The Learned Perceptual Image Patch Similarity (LPIPS) metric measures perceptual similarity between images using deep neural network features. Lower LPIPS scores indicate higher perceptual similarity, reflecting how closely the generated image aligns with human visual perception.
- FVD [8] : The Fréchet Video Distance (FVD) is a widely adopted metric for evaluating the quality of generated videos. It compares the distributions of real and generated videos in a feature space, with lower FVD values indicating greater superior video generation quality.

The experimental results clearly demonstrate that the **Concatenation by Channel** approach consistently achieves superior performance across all evaluated metrics. Specifically, this method yields the highest perceptual quality, as evidenced by a PSNR of **31.50** and an SSIM of **0.8362**. In addition, it attains a substantially lower reconstruction error, with an **L1** value of  $2.79 \times 10^{-5}$ , corresponding to a 29.7% reduction compared to the MLP method ( $3.97 \times 10^{-5}$ ). The perceptual similarity, measured by LPIPS, is also highest for the Concatenation by Channel method, achieving the lowest score of **0.224**. Most notably, with respect to temporal coherence, this method significantly outperforms all other strategies, achieving an FVD of **133.95**, a 52.8% reduction relative to the next best method, Concat by MLP, which yields an FVD of 283.79. These results collectively highlight the effectiveness of the Concatenation by Channel strategy for skeleton information injection.

Table 1. Quantitative Comparison of Injection Strategies.

Method	PSNR↑	SSIM↑	L1↓	LPIPS↓	FVD↓
Concat by width	29.16	0.7042	7.25E-05	0.370	415.13
Concat by MLP	30.92	0.7829	3.97E-05	0.278	283.79
<b>Concat by channel</b>	<b>31.50</b>	<b>0.8362</b>	<b>2.79E-05</b>	<b>0.224</b>	<b>133.95</b>

**Qualitative Results.** As shown in Fig 3, the qualitative results indicate that the concat-by-channel approach produces more stable outputs compared to concat-by-width and MLP-based addition methods. Specifically, this is reflected in improved consistency of human motion, appearance uniformity, and overall image quality. Based on the aforementioned qualitative and quantitative comparison results, we adopt the concat-by-channel approach as our skeletal injection method at the coarse granularity level.

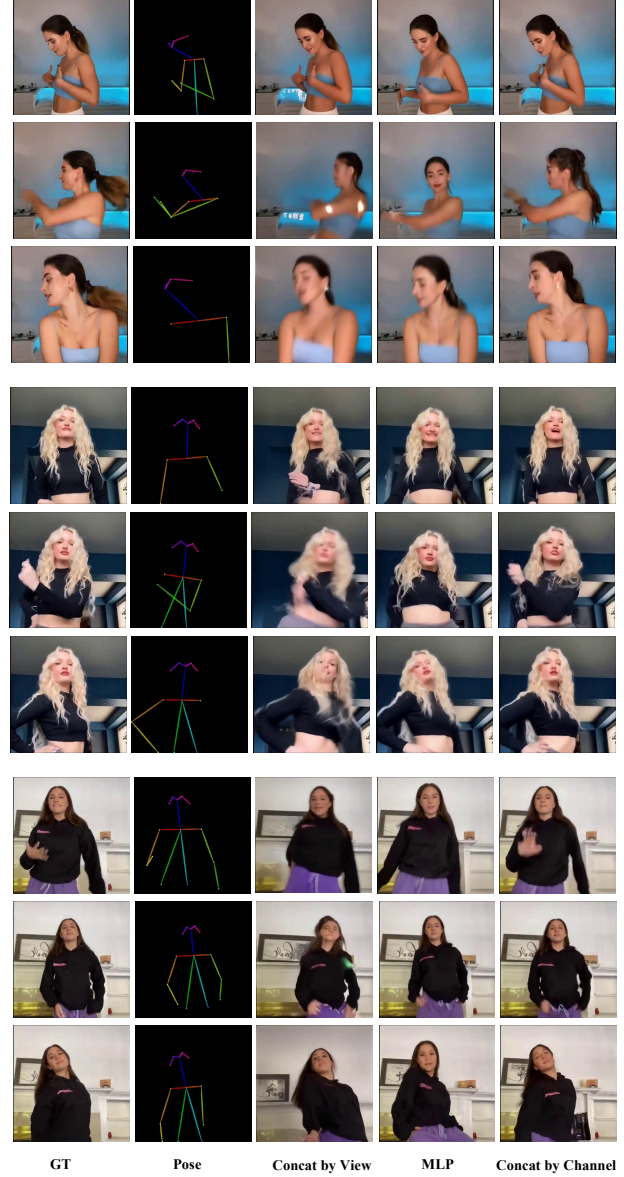


Figure 3. Qualitative Comparison of Different Injection Strategies

## 6. Sparse Pose Condition Injection

Current state-of-the-art human pose-guided video generation methods, including UniAnimate [12], Animate-X [7], and trajectory-control approaches such as ATI [10], Tora [16], and SG-I2V [5], typically mandate dense control injection, requiring cues at every frame or across a substantial portion (i.e.,  $\geq 50\%$ ) of the sequence. This dense dependency contrasts with real-world applications where crucial control information is often concentrated within only a few **key frames**. This inherent disparity motivates us to enhance our model’s capacity to learn effective temporal dynamics and motion propagation from **sparse pose injection**.



To cultivate this robust generalization capability, we devised a systematic **sparse sampling strategy** during training. Given the standard sequence length of 81 pose images, the dataloader randomly masks the input pose sequence according to a defined probability distribution for the remaining frames: (1) Dense Subset: 21 to 81 frames remain, applied with a probability of 35%, (2) Medium Subset: 11 to 21 frames remain, applied with a probability of 20%, (3) Sparse Subset: 1 to 11 frames remain, applied with a probability of 45%. Following the determination of the number of remaining control frames, we introduce further spatial variance by applying a **random masking scheme** with a 50% probability, and a **uniform masking scheme** with a 50% probability. Through the strategic integration of this stratified sampling approach, we aim to significantly bolster the model’s temporal predictive capacity and improve its resilience and generalization under conditions of highly sparse skeleton input.

To evaluate the model’s generalization ability under sparse pose conditions, we conduct experiments on the XPose-Benchmark described in the main paper. Specifically, we inject skeleton images at different sparsity levels, 100% (full condition), 50%, 20%, 10%, 5%, and 2.5%, and compare the quantitative results.

**Quantitative Results.** Tab. 2 and Fig. 4 summarize the quantitative comparison of model performance under different pose condition injection ratios. The results show that our model consistently maintains high performance across a broad range of injection densities, with only slight degradation observed even under conditions of extreme sparsity. Specifically, when the injection ratio is reduced from full conditioning (100%) to 10%, all major image quality metrics (PSNR, SSIM, L1, LPIPS) experience negligible declines, and the temporal metric FVD is even marginally improved, suggesting strong temporal coherence. Furthermore, under the most challenging scenario (2.5% injection), our model still delivers competitive results: while SSIM and LPIPS show the largest drops, PSNR remains high and FVD stays close to the baseline value. These observations highlight the robustness of our architecture in effectively interpolating and generalizing pose information across unconditioned frames. Such resilience under sparse pose injection conditions indicates the practical value of our method, especially for real-world applications where pose inputs may be infrequent or irregular.

## 7. Effect of CFG in Pose-Guided Generation

Classifier-Free Guidance (CFG) is a crucial technique in conditional diffusion models, designed to enhance the alignment between generated outputs and specified conditioning information (e.g., text, image, or pose). CFG operates by linearly extrapolating the predicted noise ( $\hat{\epsilon}$ ) away from the unconditional estimate ( $\hat{\epsilon}_\theta$ ) toward the conditional

Table 2. Quantitative Result of Sparse Pose Condition Injection

Ratio	PSNR $\uparrow$	SSIM $\uparrow$	L1 $\downarrow$	LPIPS $\downarrow$	FVD $\downarrow$
100%	30.29	0.7114	8.19E-06	0.324	99.97
50%	30.31	0.7145	8.03E-06	0.323	101.1
20%	30.32	0.7097	8.08E-06	0.327	100.88
10%	30.21	0.73334	8.13E-06	0.317	97.02
5%	29.95	0.6934	9.01E-06	0.344	102.06
2.5%	29.82	0.6757	1.01E-05	0.363	99.23

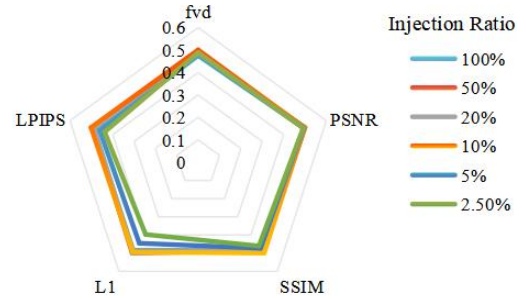


Figure 4. Visualization of Sparse Injection Comparison

estimate ( $\hat{\epsilon}_\theta$ ), controlled by a guidance scale parameter  $s$ :

$$\tilde{\epsilon} = \hat{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) + s \cdot (\hat{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - \hat{\epsilon}_\theta(\mathbf{z}_t, \emptyset)) \quad (7)$$

Within the context of diffusion-based video generation, the CFG scale  $s$  serves as a hyperparameter that governs the degree of fidelity to the motion condition. Increasing  $s$  generally enforces stricter adherence to the input pose sequence, resulting in sharper and more pronounced movements, albeit sometimes at the expense of sample diversity and generation stability.

To empirically evaluate the performance of PoseAnything to the strength of conditional guidance, we introduced CFG into PoseAnything and conducted a series of experiments. Specifically, as a negative anchor, we injected a pose sequence in which every valid pose was identical to the first frame’s pose. The primary objective was to systematically characterize model performance across a range of CFG scale values, thereby elucidating the trade-off between pose fidelity and overall generation quality. Our experimental protocol encompassed two distinct sparsity settings: (1) Sparse: Pose conditions were injected into 10% of the total generated frames; (2) Dense: Pose conditions were injected into 100% of the generated frames.

**Quantitative results.** As observed from Tab 3, increasing the CFG scale leads to a consistent degradation across all quantitative metrics, with this effect being more pronounced under the dense conditioning setting. Specifically, as the value of *cfg scale* increases, both PSNR and SSIM decrease, while L1, LPIPS, and FVD scores increase, indicating a decline in both reconstruction fidelity and perceptual quality. This trend is especially evident when pose conditions are densely injected (*dense* setting), suggesting that

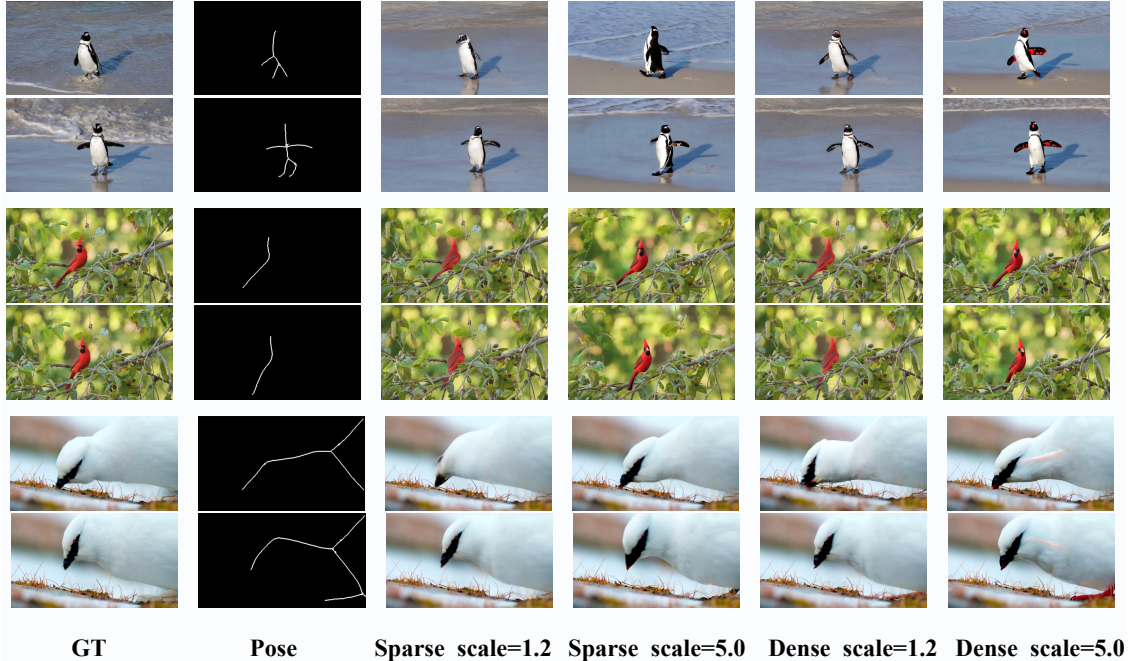


Figure 5. Qualitative Comparison of Different CFG Scale under Sparse/Dense Injection

higher conditioning density amplifies the impact of excessive guidance strength. This phenomenon can be attributed to the fact that, as the density of input pose information increases, the conditional signal becomes more dominant. While moderate CFG scales help align generated outputs with the desired pose, excessively high CFG scales (e.g., 5.0) may over-constrain the model, reducing the diversity and stability of generated samples. This over-constraining effect manifests as lower PSNR and SSIM, along with elevated LPIPS and FVD scores, reflecting poorer visual quality and diminished sample diversity.

Table 3. Quantitative Result of Applying CFG

Scale	Setting	PSNR $\uparrow$	SSIM $\uparrow$	L1 $\downarrow$	LPIPS $\downarrow$	FVD $\downarrow$
5.0	dense	29.92	0.6768	9.77E-06	0.3471	100.22
3.0	dense	30.47	0.7126	7.82E-06	0.3241	99.78
1.5	dense	30.52	0.7172	7.67E-06	0.3226	99.94
1.2	dense	30.61	0.7245	7.42E-06	0.3192	99.29
5.0	sparse	29.84	0.6730	9.64E-06	0.3557	100.83
3.0	sparse	30.14	0.6900	8.73E-06	0.3441	98.91
1.5	sparse	30.56	0.7122	7.49E-06	0.3276	99.11
1.2	sparse	30.68	0.7183	7.28E-06	0.3245	99.01

**Qualitative Results.** Qualitative results are presented in Fig. 5. As observed, increasing the CFG scale enhances the model’s ability to generate the target pose, resulting in more accurate pose alignment. However, under strong skeletal conditioning—particularly in the dense injection setting—a higher CFG scale tends to introduce visual artifacts, such as unnatural limb shapes or distortions in body structure, which may negatively impact the overall visual quality.

In summary, the selection of CFG scale should be adapted to the density of pose conditioning. When the input pose density is high (i.e., dense setting), a lower CFG scale is preferable to prevent over-constraining the model and to maintain a balance between conditional fidelity and visual quality. Conversely, when the pose conditioning is sparse, increasing the CFG scale can effectively enhance the model’s ability to fit the provided pose information without significantly compromising generation stability or diversity. This adaptive strategy enables the model to achieve optimal performance across varying levels of conditioning strength, thereby providing practical guidance for tuning diffusion-based pose-guided generation systems.

## 8. More generation results of PoseAnything

In this section, we present more generated samples to further demonstrate the effectiveness and robustness of our model. As illustrated in Fig. 9, our approach PoseAnything consistently achieves precise pose adherence, accurately following the provided motion cues across diverse scenarios. Moreover, the results reveal a remarkable generalization capability, with the model maintaining high performance across a wide variety of scene backgrounds, subject appearances, and motion types. Importantly, our method preserves both visual and temporal coherence, ensuring smooth and realistic transitions throughout the generated sequences. These findings collectively underscore the reliability and adaptability of our model in handling complex and challenging motion generation tasks.

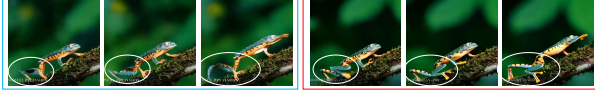
## 9. More Comparion Results

we benchmarked our method against UniAnimate-DiT, which uses the base model of equivalent capacity(Wan2.1-14B) with our method. As summarized below, the quantitative results conclusively validate the superiority and effectiveness of our proposed architecture.

Method	PSNR $\uparrow$	SSIM $\uparrow$	L1 $\downarrow$	LPIPS $\downarrow$	FID-VID $\downarrow$	FVD $\downarrow$	PSNR* $\uparrow$
UniAnimate-DiT	31.4	0.7517	8.14E-05	0.2261	18.13	285.68	20.66
<b>Ours</b>	<b>31.5</b>	<b>0.8362</b>	<b>2.79E-05</b>	<b>0.2240</b>	<b>11.97</b>	<b>133.95</b>	<b>21.08</b>

## 10. Qualitative Ablation Results on PTCM

The primary role of PTCM is to ensure high fidelity in details; consequently, the improvement in global quantitative metrics is not significant. Qualitative comparisons between PTCM and the Entire Object (EC) baseline demonstrates its effectiveness.



## 11. More Results on Generalization

We evaluate 100 newly collected OOD YouTube data to address sample size and generalization concerns. Our method maintains superior performance.

Method	PSNR $\uparrow$	SSIM $\uparrow$	L1 $\downarrow (\times 10^{-5})$	LPIPS $\downarrow$	FVD $\downarrow$	FID $\downarrow$
Tora	29.65	0.5893	7.19	0.3622	843.8	33.35
SG-I2V	29.59	0.5633	6.90	0.3490	827.5	26.42
ATI	30.32	0.6142	<b>6.25</b>	0.3398	826.0	25.61
<b>Ours</b>	<b>30.58</b>	<b>0.6262</b>	<u>6.65</u>	<b>0.3176</b>	<b>816.7</b>	<b>25.22</b>



Figure 6. Simple Subset of XPose

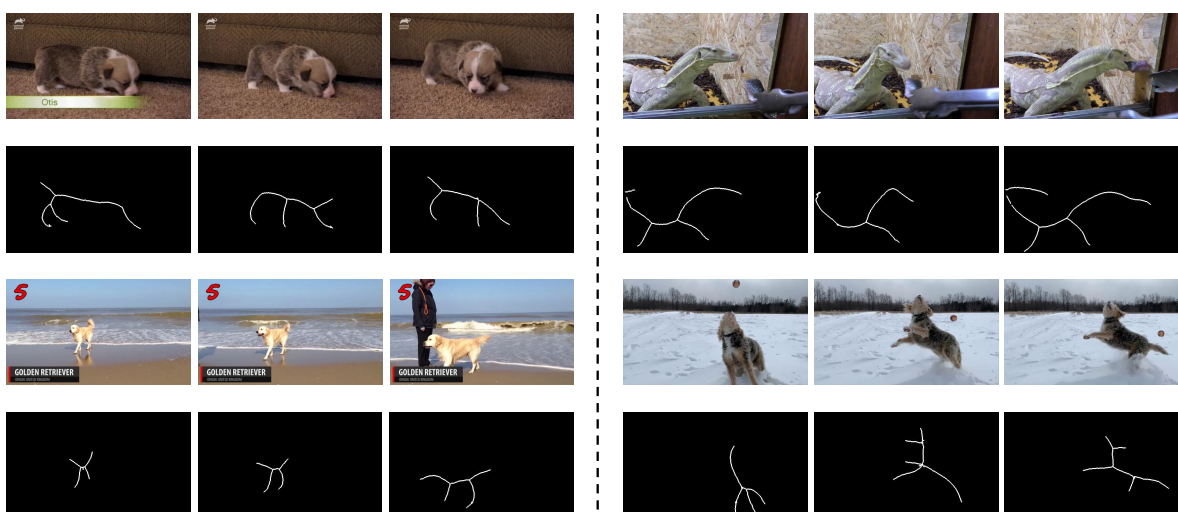


Figure 7. Medium Subset of XPose

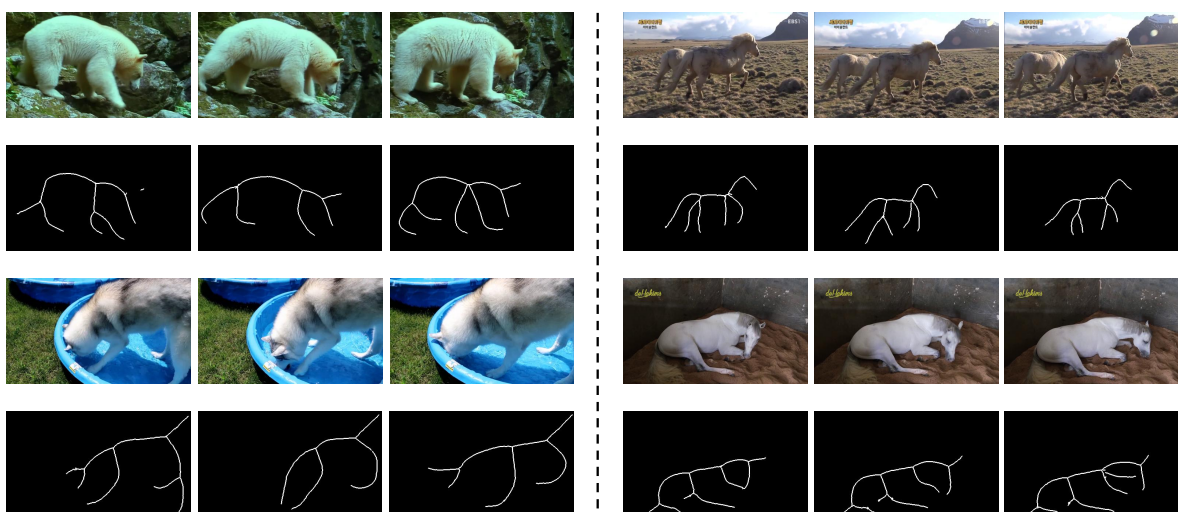
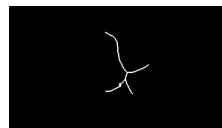
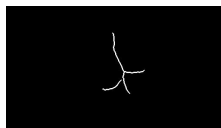
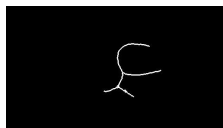
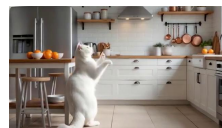
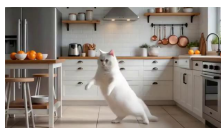
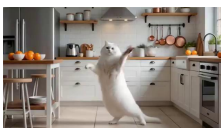
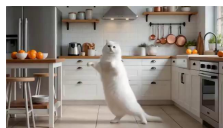
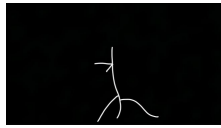
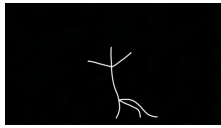
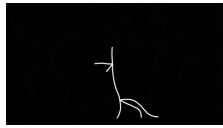
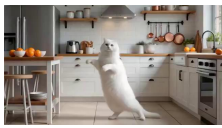
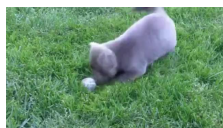
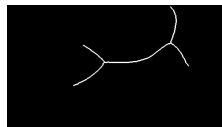
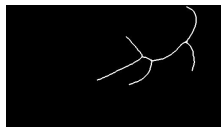
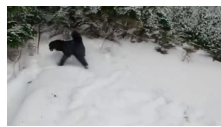
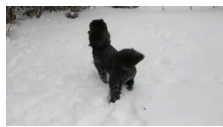
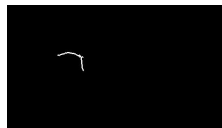
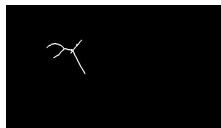
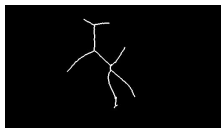
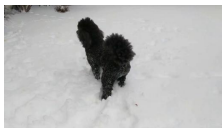
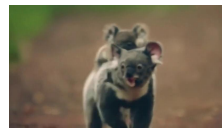
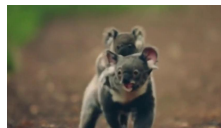
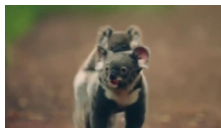
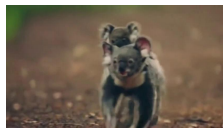
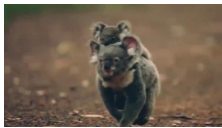


Figure 8. Complex Subset of XPose





## References

- [1] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints, 2024. 3
- [2] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fuyun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025. 3
- [3] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 3
- [4] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 3
- [5] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. SG-I2V: self-guided trajectory control in image-to-video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 4
- [6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. 1
- [7] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 4
- [8] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 4
- [9] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*, pages 313–331. Springer, 2024. 3
- [10] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: any trajectory instruction for controllable video generation. *CoRR*, abs/2505.22944, 2025. 4
- [11] Qiuhe Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024. 1
- [12] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoliang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unimate: Taming unified video diffusion models for consistent human image animation. *CoRR*, abs/2406.01188, 2024. 4
- [13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4
- [14] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao. Ultravideo: High-quality UHD video dataset with comprehensive captions. *CoRR*, abs/2506.13691, 2025. 1
- [15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4
- [16] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 2063–2073. Computer Vision Foundation / IEEE, 2025. 4

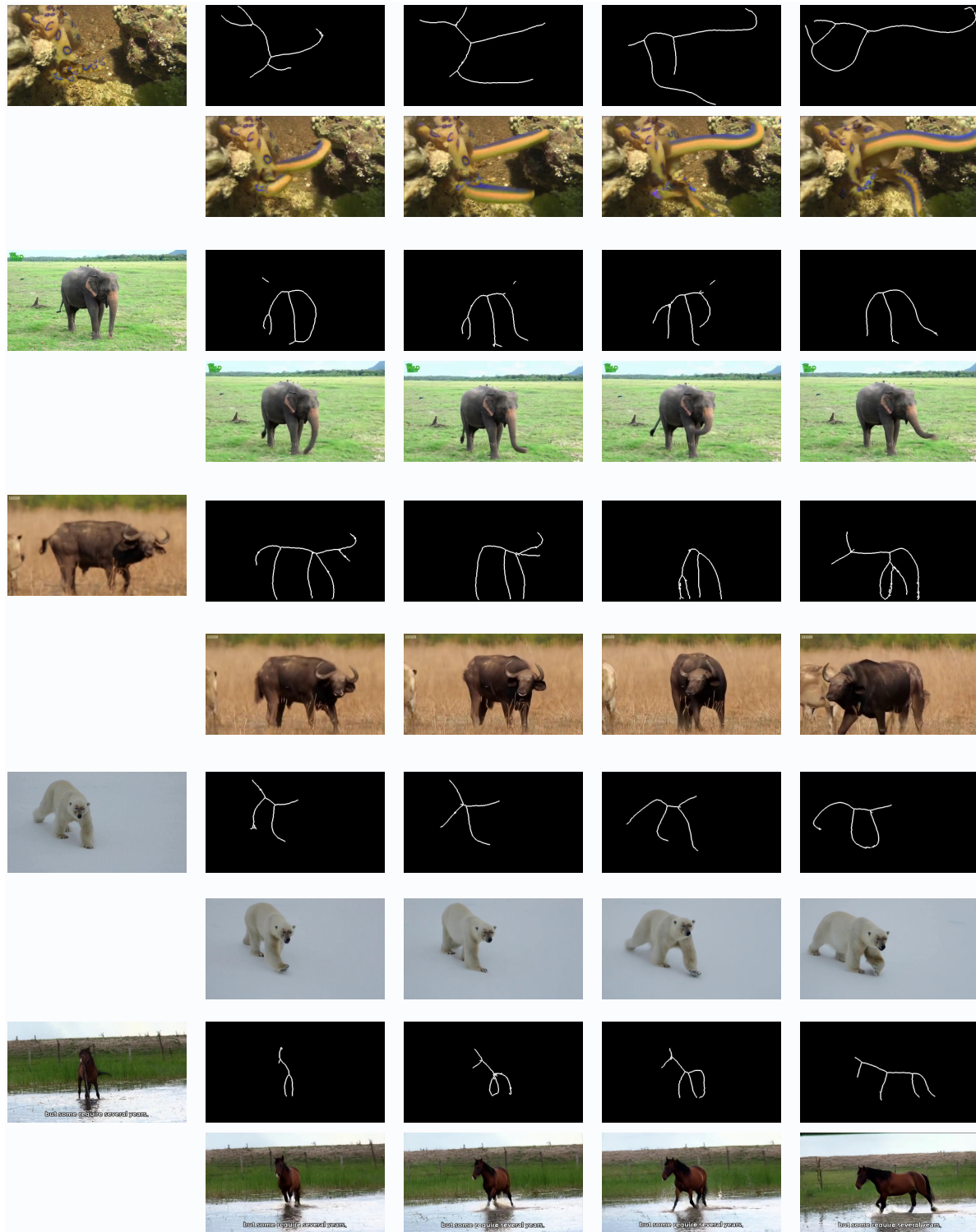


Figure 9. More Results of PoseAnything