Alignment is All You Need: A Training-free Augmentation Strategy for Pose-guided Video Generation

Xiaoyu Jin [*1] **Zunnan Xu** [*1] **Mingwen Ou** [1] **Wenming Yang**[†1]

Abstract

Character animation is a transformative field in computer graphics and vision, enabling dynamic and realistic video animations from static images. Despite advancements, maintaining appearance consistency in animations remains a challenge. Our approach addresses this by introducing a training-free framework that ensures the generated video sequence preserves the reference image's subtleties, such as physique and proportions, through a dual alignment strategy. We decouple skeletal and motion priors from pose information, enabling precise control over animation generation. Our method also improves pixel-level alignment for conditional control from the reference character, enhancing the temporal consistency and visual cohesion of animations. Our method significantly enhances the quality of video generation without the need for large datasets or expensive computational resources.

1. Introduction

Character animation is a task in the fields of computer graphics and computer vision to enable the shift from static images to dynamic, realistic video animations. This technology has significant implications for various industries such as entertainment, social media, virtual reality, and other immersive digital experiences, providing more engaging and customized visual experiences. A key challenge in this area is maintaining appearance consistency and fidelity in animated sequences, as these aspects are essential for the realism and overall quality of the produced content.

Previous endeavors in character animation have advanced the transformation of static images into dynamic content. Traditional graphic techniques have been enhanced by data-driven models leveraging extensive visual datasets for more cost-effective solutions. While GAN-based methods (Goodfellow et al., 2014; Arjovsky et al., 2017; Karras et al., 2019) show potential in creating realistic details, they face challenges with motion transfer and maintaining subject identity across poses. Conversely, diffusion-based models (Ho et al., 2020; Karras et al., 2023; Rombach et al., 2022), while capable of producing visually plausible animations, are susceptible to appearance inconsistencies, resulting in unnatural limb proportions and sub-optimal effects when there are significant differences between the reference image physique and the pose used for generation (Guo et al., 2023; Xu et al., 2023; Li et al., 2023a; Hu et al., 2023). Our approach introduces a training-free framework that prioritizes appearance consistency. Unlike existing methods that ignore appearance details during animation, our method ensures the generated video sequence stays true to the motion while preserving subtleties of the reference image, like the subject's physique—accurately reflecting their height, build, and proportions in the image. Through a dual alignment strategy, our method can create animations that show both appearance consistency and fidelity to the reference image.

Building on the insights from previous research, our method introduces a dual alignment strategy that re-envisions the relationship between reference images and pose data. A core element of our innovation lies in the separation of skeletal and motion priors from the pose information itself. We identify the essential cues present in the key points representations, such as skeletal position, length and angular variances, which reflect an individual's body information and motion tendencies. By utilizing efficient linear matrix operations, our approach distinguishes the identity information and the movement information of the skeletal sequences. This enables the transfer of skeletal data from a reference image to the driving pose sequences while preserving the intrinsic motion characteristics of poses. This allows for precise control over the generation, ensuring that the animation faithfully reproduces the physique of reference character while maintaining a resemblance to the motions of the pose sequences. Furthermore, acknowledging the importance of accurate pixel-level alignment for conditional control, we improve the reference image to kickstart an animation that closely aligns with the initial frame of the driving pose video. This enhancement utilizes the information stored in

[*]Equal contribution [1]Shenzhen International Graduate School, Tsinghua University. Correspondence to: Wenming Yang <yang.wenming@sz.tsinghua.edu>.

First Workshop on Controllable Video Generation at ICML 2024.

current diffusion models to direct the reference image to mimic the motion of the starting pose. The outcome is an improved alignment between the reference image and the driving pose video, establishing the foundation for a temporally consistent and visually cohesive animation sequence. Our main contributions are as follows:

- We introduce a training-free augmentation strategy for pose-guided animation generation that avoid the need for using large video datasets and expensive GPU resources.
- We propose a novel dual alignment method that can be seamlessly integrated into pose-guided generative models to enhance the quality of generated videos.
- Experiments demonstrate that our method can effectively enhance the quality of character animation generation.

2. Related Work

2.1. Diffusion Model for Image Generation

In the domain of text-to-image synthesis, diffusion-based models have indeed set new benchmarks for generation quality and have become a central focus of research. These models, such as DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), Latent Diffusion Model (LDM) (Rombach et al., 2022), Glide (Nichol et al., 2021), eDiffi (Balaji et al., 2022), and Composer (Huang et al., 2023), have demonstrated the ability to produce high-quality and diverse image outputs from textual descriptions (Lu et al., 2024; Li et al., 2024). The Latent Diffusion Model (LDM) has introduced a method of denoising in the latent space to reduce computational complexity while maintaining the quality of generated images. This offers an effective and efficient approach to image synthesis. Further advancements have been made in controlling the visual generation process. Models like ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2023) have integrated additional encoding layers to enhance the controllability of the generation process. This enhancement allows for conditional generation based on various factors such as pose, mask, edge, and depth information. Building upon these capabilities, some studies have explored image generation that is conditioned on specific image-related inputs. For instance, IP-Adapter (Ye et al., 2023) enables diffusion models to generate images that incorporate content specified by an image prompt. ObjectStitch (Song et al., 2023) and Paint-by-Example (Yang et al., 2023a) utilize the capabilities of CLIP to propose diffusion-based methods for image editing under given image conditions. In the context of fashion and virtual try-on applications, TryonDiffusion (Zhu et al., 2023) applies diffusion models to the task of virtual apparel try-on and introduces an innovative Parallel-UNet structure to enhance the process. These developments highlight the rapid progress and innovation in the field of text-to-image generation. Diffusion-based methods are not only pushing the boundaries of what is possible but also expanding the horizons of controllability and applicability in diverse scenarios.

2.2. Pose Guidance in Character Animation Generation

The success of diffusion models in text-to-image synthesis has significantly influenced text-to-video research (Khachatryan et al., 2023; QI et al., 2023; Hong et al., 2022; Wu et al., 2023; Yang et al., 2023b; Xu et al., 2024b; Esser et al., 2023; Singer et al., 2022; Ho et al., 2022), particularly in model structure and the incorporation of pose guidance. In the context of pose guidance, DWpose (Yang et al., 2023c) offers an enhanced alternative to OpenPose (Cao et al., 2017), providing more accurate and expressive skeletons that are beneficial for high-quality image generation. DensePose (Güler et al., 2018) establishes dense correspondences between images and surface representations, which is crucial for detailed pose guidance. The SMPL model (Loper et al., 2015), known for its realistic human representation, is widely used for pose and shape analysis in character animation (Li et al., 2023b; Yang et al., 2024; Xu et al., 2024a). It serves as essential ground truth for neural networks and is considered in our approach for reconstructing poses and shapes, providing a comprehensive foundation for appearance alignment and pose guidance in video generation. Our approach draws inspiration from these methods by focusing on decoupling the identity and the movement from DWpose/Openpose guidance in the video generation pipeline. This ensures that the generated animations maintain coherence in appearance with the reference image.

3. Methodology

To create pose-guided personalized videos in a training-free setting, we introduce a simple yet effective framework in Section 3. In Section 3.1, we describe the settings and architecture of our pipeline. Section 3.2 details our training-free Skeleton based Pose Adapter method. Finally, in Section 3.3, we present our Kickstart Alignment Strategy.

3.1. Settings and Framework

Given a reference character image I_r and a template pose sequence $\mathbf{P} = \{\mathbf{p}_a\}_{a=1}^{M}$ consisting of M frames, our goal is to generate a high-quality video $\mathbf{V} = \{\mathbf{v}_a\}_{a=1}^{M}$. This video should be of superior fidelity, exhibit exceptional faithfulness to I_r, and accurately align with $\mathbf{P}$.

As illustrated in Figure 1, our framework builds upon the existing pose-guided video generation model. Our contribution is entirely training-free, manipulating the input reference image and pose sequence without participating in the training of the main network. Specifically, the Skeleton based Pose Adapter decouples and embeds the identity in-

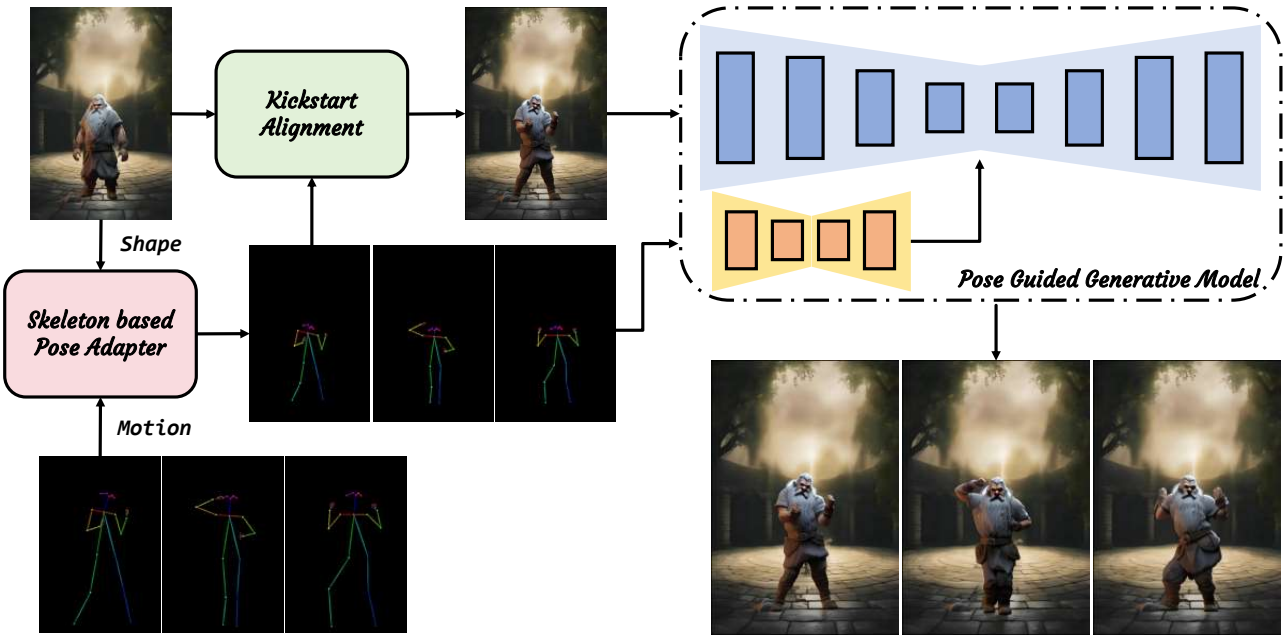

Figure 1. The overall framework of our method. The architecture of our method is composed of two key components: Skeleton based Pose Adapter and Kickstart Alignment. These components work together to refine the pose used for driving video generation and the reference image. These refined control conditions are then inputted into the existing pose-guided video generation model, enabling dynamic and realistic video animations with increased consistency and fidelity.

formation from the input pose sequence. The aligned pose is then used to transfer the input image into another one with the gesture of the initial pose, while preserving the identity information. Similar to other diffusion-based video generation methods, the U-net employs multiple frames of noise, along with a reference image and pose sequence with detailed identity information, to generate vivid personalized videos.

3.2. Skeleton based Pose Adapter

In pose-guided video generation tasks, a skeleton-based human pose estimation model (Yang et al., 2023c; Cao et al., 2017) is usually employed to extract the pose sequence $\mathbf{P}$ from a template video. This sequence combines action sequence information with the identity information, such as the physique, position, and distance from the camera.

However, the identity information embedded within the template video is irrelevant and even harmful to our purpose. Therefore, a Skeleton-based Pose Adapter method is proposed to get the aligned pose sequence $\mathbf{Q}$, which decouples the action information from the template pose sequence and embeds the identity information into the aligned $\mathbf{Q}$. The overall logic of our algorithm is presented in Algorithm 1.

Formally, the extracted template pose sequence $\mathbf{P} = \{\mathbf{p}_a\}_{a=1}^{M}$, where $\mathbf{P}$ includes the position information of the human keypoints. And the skeletal pose information $\mathbf{q_0}$ estimated from the reference image is denoted as $\mathbf{q_0}$. Let $\mathbf{C_1} = \{\mathbf{c}_{1i}\}_{i=1}^{n}$ and $\mathbf{C_2} = \{\mathbf{c}_{2i}\}_{i=1}^{n}$ be two sets of coordinates representing the key points of $\mathbf{p}_a$ and $\mathbf{q_0}$, where n is the number of points in each set. Let $\mathbf{L} = \{(i_k, j_k)\}_{k=1}^{m}$ be a sequence of limb connections between points in $\mathbf{C_1}$ and $\mathbf{C_2}$. For each pair of connected points $(i_k, j_k) \in \mathbf{L}$, we calculate the Euclidean distances d_{1_k} and d_{2_k} in $\mathbf{C_1}$ and $\mathbf{C_2}$ to get the ratio r_k:

$$r_k = \frac{d_{2_k}}{d_{1_k}} = \frac{\|\mathbf{c}_{1j_k} - \mathbf{c}_{1i_k}\|_2}{\|\mathbf{c}_{2j_k} - \mathbf{c}_{2i_k}\|_2}, \tag{1}$$

so the vector of all ratios is expressed as $\mathbf{r} = \frac{\mathbf{d_2}}{\mathbf{d_1}} = \{r_k\}_{k=1}^{m}$.

Given two vectors $\mathbf{c_{1j_k}}$ and $\mathbf{c_{1i_k}}$ representing the start and end coordinates of a limb segment, we can calculate the length of the limb and the angle between the vectors using the following formula:

$$\theta_k = \arctan 2\left(\mathbf{c}_{1j_k} - \mathbf{c}_{1i_k}, \mathbf{c}_{2j_k} - \mathbf{c}_{2i_k}\right). \tag{2}$$

Considering the process of moving a point $\mathbf{P}$ in a two-dimensional space along a direction determined by an angle θ_k by an updated distance $l_k = r_k * d_{1k}$, we first define a vector $\mathbf{v_k}$ whose magnitude and direction are determined by l_k and θ_k:

$$\mathbf{v_k} = l_k \cdot \begin{bmatrix} \cos(\theta_k) \\ \sin(\theta_k) \end{bmatrix}. \tag{3}$$

Algorithm 1 Pose Adapter

Input: Template pose sequence: $\mathbf{pose1}$, reference image pose: $pose2$.
Output: Transformed pose image $kps\_results$, transformed pose array $\mathbf{pose2}$.

Scale $pose2$ and $\mathbf{pose1}$ by their image shape to match the real size.
Calculate $edge\_ratios$ using the coordinates and limbSeq of $\mathbf{pose1}$ and $pose2$.
for all pose in $\mathbf{poses1}$ **do**
    Update body positions of $pose2$ with $edge\_ratios$.
    Update hand positions of $pose2$ with updated body positions and $edge\_ratios$.
    Normalize $\mathbf{pose2}$ to draw the pose on a canvas to get transformed pose image.
    Add pose image to $kps\_results$ and add $pose2$ to $\mathbf{pose2}$.
end for
return $kps\_results$, $\mathbf{pose2}$

Finally, the coordinates of the new point $\mathbf{c_{2i}}'$ can be obtained by adding the original point $\mathbf{c_{1i}}$ to the vector $\mathbf{v_k}$, which can be represented as:

$$\mathbf{c_{2i}}' = \mathbf{c_{1i}} + \mathbf{v_k} + \epsilon, \tag{4}$$

where ϵ is the offset of the base coordinate. When set to 0, the position of aligned pose stays the same of the template pose. When set it to the difference between the base coordinates of $\mathbf{C_1}$ and $\mathbf{C_2}$, the position of the aligned pose is consistent with the reference character. Through the above process, we can obtain the aligned pose $\mathbf{C_2}'$ of every frame, to finally construct the aligned pose sequence Q.

3.3. Kickstart Alignment Strategy

Inspired by this concept, our approach further enhances the alignment of the input reference image through a similar kickstart alignment technique. We achieve this by employing pose-guided image synthesis models, specifically PCDMs (Shen et al., 2023). By doing this, we make a more accurate and natural depiction of the initial frame of generated video. This strategy ensures that the animated character's starting position is poised to transition seamlessly into the animated sequence, much like a dancer's initial stance before an expressive performance.

The kickstart alignment involves an initial alignment using the first frame of pose sequences to identify key points and skeletal structures from the reference image. This step lays the groundwork for the subsequent pose-guided generation, ensuring that the reference image's pose is conditioned on the first frame of an adjusted pose sequence. This frame serves as a control signal, guiding the reference image to mimic the specific action depicted in the pose. The selection of the initial pose frame is motivated by its role in setting the tone for the entire animation, much like a dancer's initial stance sets the stage for their performance.

Our method's utilization of pose-controlled generation models enables a high degree of control over pixel-level alignment, ensuring that the animated output is not only consistent with the motion sequence but also preserves the visual integrity of the reference image. This dual emphasis on pose and pixel alignment leads to a more natural and seamless animation.

4. Experiments

4.1. Experiment Settings

We implement the experiments based on the existing pose guided video generation Model. For each character animation, we set the reference image with a unified 768×512 resolution. The template videos can be a different resolution. All experiments are performed on a single NVIDIA A100 GPU. Since our method only aligns the input conditions and is training-free, the experiments we conduct are all ablation studies to verify the effectiveness of the proposed method.

4.2. Comparison Result

To ensure a fair comparison, we employed the same base video generation network architecture, network weights, and test dataset. The following are comparative experiments for the proposed modules.

Comparison experiments of Skeleton based Pose Adapter. To evaluate the performance of our Skeleton based Pose Adapter, we conducted experiments driving by pose sequences from templates and pose sequences aligned by the Pose Adapter, respectively. The results are displayed in Figure 2 and Figure 3, which correspond to animations of anime characters and humankind in realword, respectively. On the left side, the poses sequence with the Pose Adapter and the generated animation video are presented, on the right side, there are template pose images and the output without the Pose Adapter. It is evident that when the template poses are not aligned with the input, the generated results are quite poor.

In Figure 2, as shown in the first and second sets. The base model is unable to address the discrepancy in body shape

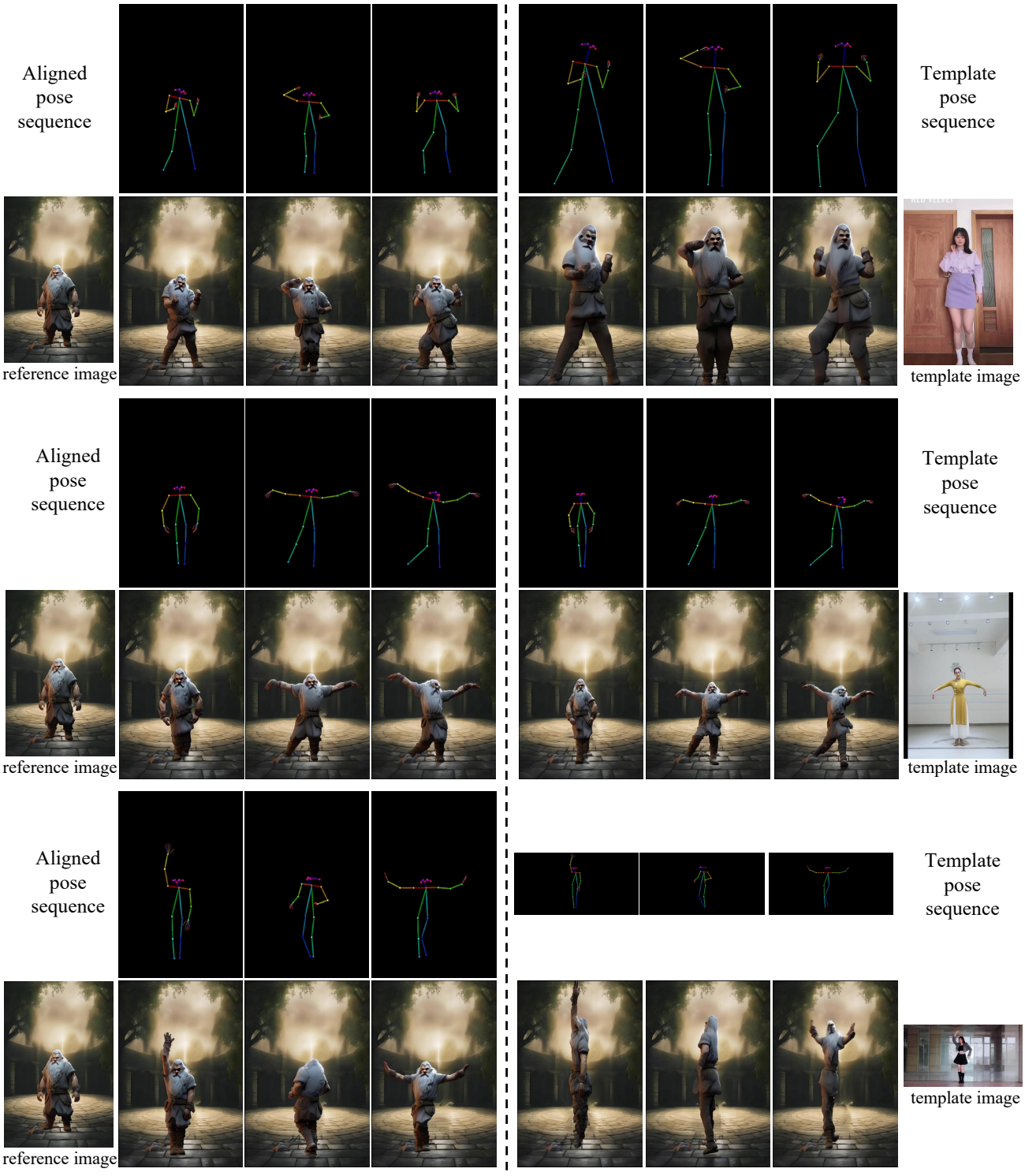

Figure 2. Ablation experiments of Pose Adapter on anime video generation.

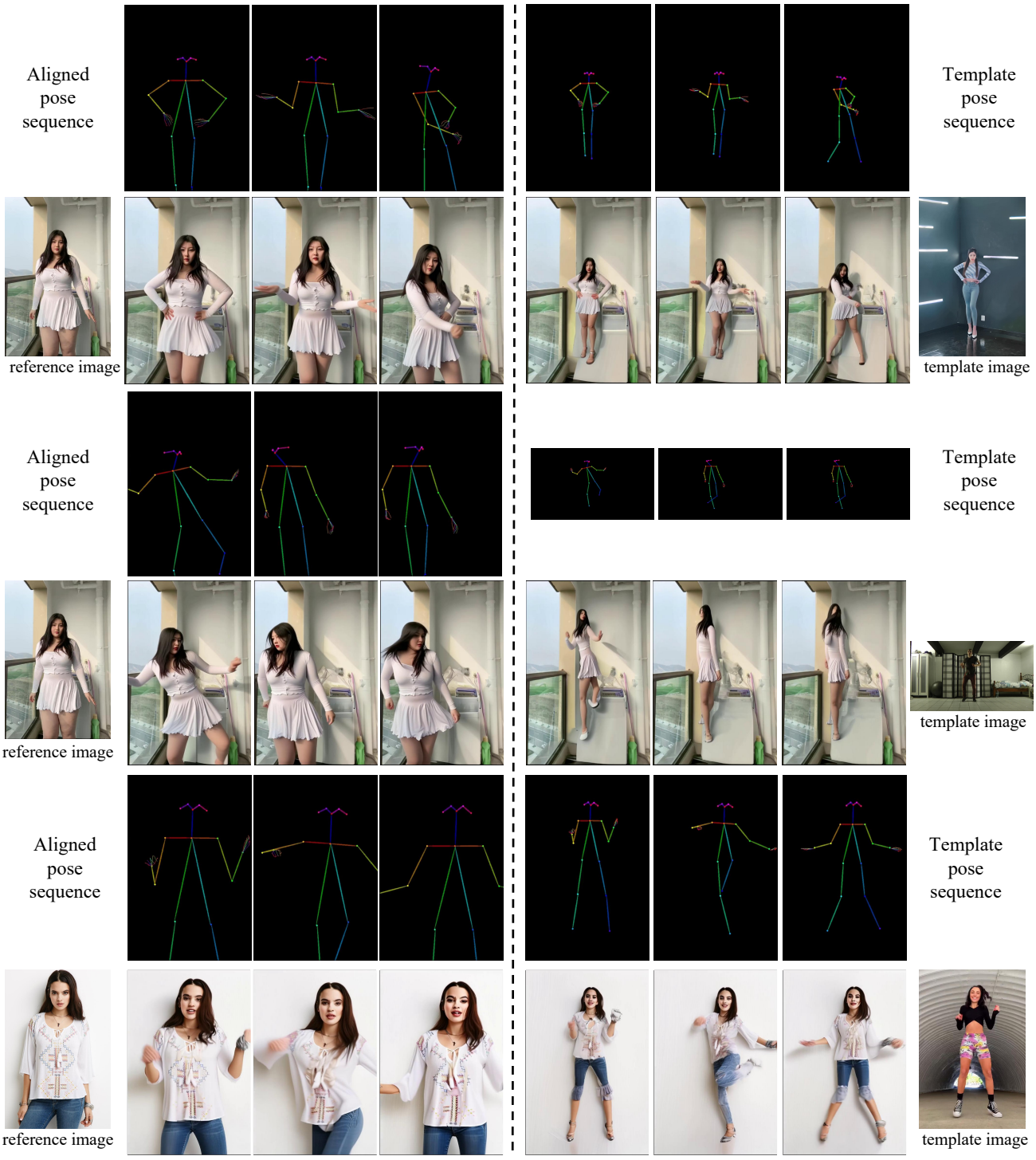

Figure 3. Ablation experiments of Pose Adapter on human video generation.

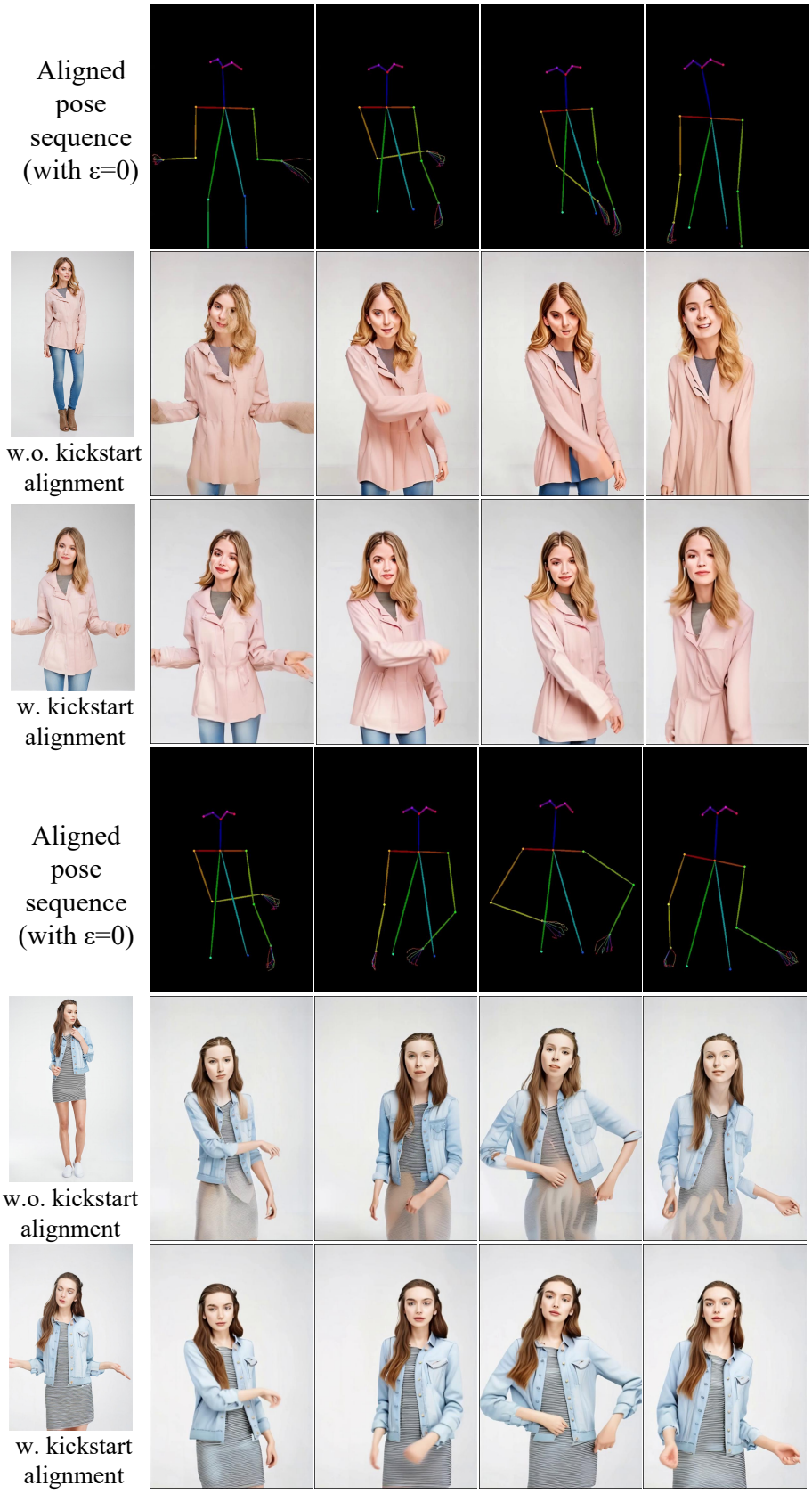

Figure 4. Ablation experiments of Kickstart Alignment on human video generation.

between the template and the reference image, resulting in generated frames altering the original identity characteristics of the input image. For instance, a dwarf loses its distinctive stocky physique and instead assumes a body shape similar to that of a human. And the third set, when the frame size is inconsistent, the pose image is squeezed and deformed, losing its control ability. Similar situation also appears in Figure 3. When the character's position in the frame is misaligned, the person and the background will become intertwined. When the frame size is inconsistent (the same as the third set of Figure 2), the pose image is deformed, and the results collapses entirely. When there is a mismatch between full-body poses and half-body reference, sometimes the animation result will be difficult to accept.

Comparison experiments of Kickstart Alignment. In this section, we further incorporated Kickstart Alignment, which aligns the reference image with the gesture of the first pose of the template sequence. Figure 4 presents the results of two sets of Kickstart Alignment on human video generation, where the first row of each set is the aligned pose sequence. In this part, we set ϵ to 0 to prevent misalignment effects between half-body and full-body poses. The first set of results clearly illustrates that the absence of Kickstart Alignment results in the collapse of facial features and hair in the generated images. Moreover, the subsequent set of results indicates that the lack of Kickstart Alignment may also give rise to undesirable texture alterations. Generative models are tasked with the formidable challenge of extracting essential human body information from unaligned reference images and subsequently incorporating this information into the generation process. However, current models find this task to be overly demanding. Our approach, which incorporates alignment at the outset, effectively alleviates this challenge, resulting in substantial improvements in the quality of generated results.

5. Conclusion

In this paper, we present a novel training-free augmentation strategy for generating pose-guided personalized videos. To tackle the misalignment between original videos and reference characters, we introduce two critical algorithms: Skeleton-based Pose Adapter and Kickstart Alignment strategy. The visualization results indicate that our method exhibits a significant improvement on image fidelity to the source image while preserving intricate fine-grained appearance details. Our approach relies solely on input control conditions and does not require extra training, enabling straightforward integration into a wide variety of pose-guided video generation models. Moreover, our method involves only basic linear matrix operations and the creation of single-frame images, making it highly efficient.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In International conference on machine learning, pp. 214–223. PMLR, 2017.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291–7299, 2017.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7346–7356, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

Güler, R. A., Neverova, N., and Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7297–7306, 2018.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. Advances in Neural Information Processing Systems, 35:8633–8646, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In The Eleventh International Conference on Learning Representations, 2022.

Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., and Bo, L. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117, 2023.

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., and Zhou, J. Composer: creative and controllable image synthesis with composable conditions. In Proceedings of the 40th International Conference on Machine Learning, pp. 13753–13773, 2023.

Karras, J., Holynski, A., Wang, T.-C., and Kemelmacher-Shlizerman, I. Dreampose: Fashion video synthesis with stable diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22680–22690, 2023.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401–4410, 2019.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15954–15964, October 2023.

Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., and Zou, Y. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In International Conference on Computer Vision (ICCV), Oral, 2023a.

Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., and Zou, Y. Exploiting auxiliary caption for video grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 18508–18516, 2024.

Li, J., Yang, Z., Wang, X., Ma, J., Zhou, C., and Yang, Y. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9110–9121, 2023b.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia), 34(6): 248:1–248:16, October 2015.

Lu, Y., Zhang, M., Ma, A. J., Xie, X., and Lai, J. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6420–6429, 2024.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In International Conference on Machine Learning, 2021.

QI, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15932–15942, October 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35:36479–36494, 2022.

Shen, F., Ye, H., Zhang, J., Wang, C., Han, X., and Wei, Y. Advancing pose-guided image synthesis with progressive conditional diffusion models. In The Twelfth International Conference on Learning Representations, 2023.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.

Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S. Y., and Aliaga, D. Objectstitch: Object compositing with diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18310–18319, 2023.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7623–7633, 2023.

Xu, Z., Zhang, J., Liew, J. H., Yan, H., Liu, J.-W., Zhang, C., Feng, J., and Shou, M. Z. Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498, 2023.

Xu, Z., Lin, Y., Han, H., Yang, S., Li, R., Zhang, Y., and Li, X. Mambatalk: Efficient holistic gesture synthesis with selective state space models. arXiv preprint arXiv:2403.09471, 2024a.

Xu, Z., Zhang, Y., Yang, S., Li, R., and Li, X. Chain of generation: Multi-modal gesture synthesis via cascaded conditional control. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 6387–6395, 2024b.

Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., and Wen, F. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18381–18391, 2023a.

Yang, S., Zhou, Y., Liu, Z., and Loy, C. C. Rerender a video: Zero-shot text-guided video-to-video translation. arXiv preprint arXiv:2306.07954, 2023b.

Yang, S., Xu, Z., Xue, H., Cheng, Y., Huang, S., Gong, M., and Wu, Z. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. arXiv preprint arXiv:2401.03476, 2024.

Yang, Z., Zeng, A., Yuan, C., and Li, Y. Effective whole-body pose estimation with two-stages distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4210–4220, 2023c.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847, 2023.

Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., and Kemelmacher-Shlizerman, I. Tryondiffusion: A tale of two unets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4606–4615, 2023.