Follow Your Pose: Pose-Guided Text-to-Video Generation Using Pose-Free Videos

Yue Ma[1*†], Yingqing He[2*], Xiaodong Cun[3], Xintao Wang[3], Siran Chen[4], Xiu Li[1‡], Qifeng Chen[2‡]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] The Hong Kong University of Science and Technology, Hong Kong
[3] Tencent AI Lab, Shenzhen, China
[4] Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, China
{y-ma21, lixiu}@mails.tsinghua.edu.cn, yhebm@connect.ust.hk, {vinthony,xintao.alpha}@gmail.com
Chensiran17@mails.ucas.ac.cn, cqf@ust.hk

Abstract

Generating text-editable and pose-controllable character videos have an imperious demand in creating various digital human. Nevertheless, this task has been restricted by the absence of a comprehensive dataset featuring paired video-pose captions and the generative prior models for videos. In this work, we design a novel two-stage training scheme that can utilize easily obtained datasets (i.e., image pose pair and pose-free video) and the pre-trained text-to-image (T2I) model to obtain the pose-controllable character videos. Specifically, in the first stage, only the pose-image pairs are used only for controllable text-to-image generation. We learn a zero-initialized convolutional encoder to encode the pose information. In the second stage, we finetune the motion of the above network via a pose-free video dataset by adding the learnable temporal self-attention and reformed cross-frame self-attention blocks. Powered by our new designs, our method successfully generates continuously pose-controllable character videos while keeping the concept generation and composition ability from the pre-trained T2I model. The code and models are available on https://follow-your-pose.github.io/.

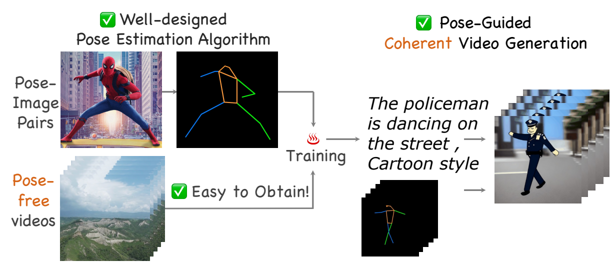

Figure 1: High-level overview of training scheme. Images with diverse characters are easy to get from the public large-scale captioned image dataset (Schuhmann et al. 2021). We can exploit the off-the-shelf pose estimation algorithm to obtain accurate pose-image pairs. However, captioned video datasets of diverse characters are absent. Therefore, we leverage the in-the-wild videos which are easy to obtain to learn temporal coherence. Through this separate learning strategy, our tunable blocks also learn to synthesize creative and diverse videos with pose-specific motion while maintaining realistic temporal coherence.

Introduction

Recent advances in text-to-image (T2I) synthesis (Rombach et al. 2022; Ho et al. 2022a; Ma, Jia, and Zhou 2023) have demonstrated impressive performance and creativity, enabling the conversion of textural descriptions into visually compelling creations. Such AI generative systems excel in various aspects, including rendering realistic and diverse appearances (Mou et al. 2023; Zhang and Agrawala 2023; Ma et al. 2022; Gao et al. 2023), powerful editing (Hertz et al. 2022; Qi et al. 2023; Ma et al. 2023c), composition (Avrahami, Fried, and Lischinski 2022), generalization capabilities, and creating new art that satisfies people's imagination (Mou et al. 2023; Zhang and Agrawala 2023). However, in text-to-video generation, there still have limited applications due to the restriction of the high-quality video dataset and the video generative prior model.

In this work, we aim to create high-quality character videos from text descriptions and also control their pose via the given control signal, *i.e.*, human skeletons. To achieve this, we design an efficient training scheme using easily obtained image datasets and the pre-trained T2I models for controllable text-to-video generation. Specifically, given image-pose pairs and pose-free videos, we design a novel two-stage training strategy with carefully tuned blocks through a pretrained text-to-image model, *i.e.*, stable diffusion (Rombach et al. 2022). In the first stage, we add the control ability into the pre-trained text-to-image model via an additional multi-layer pose encoder. In the second stage, we train our model on pose-free videos to learn temporal consistency, such as coherent backgrounds, and avoid flickering. To make the pre-trained T2I model suitable for video inputs, we make several key modifications. Firstly, we add extra temporal self-attention layers (Singer et al. 2022; Wu et al. 2022; Qi et al. 2023; Ma et al. 2023a,b; He et al. 2023; Li et al. 2023) to the stable diffusion network. Secondly, inspired by the recent one-shot video generation model (Wu et al. 2022), we reshape the attention to cross-frame attention

[*]Equal contributions.
[†]Work done during an internship at Tencent AI Lab.
[‡]Corresponding authors.

Figure 2: Pose-Guided Text-to-Video Generation. We propose an efficient training scheme to empower the ability of the pre-trained text-to-image model (*i.e.*, Stable Diffusion (Rombach et al. 2022)) to generate pose-controllable character videos with minimal data requirements. Thanks to the proposed method, we can generate various high-definition pose-controllable character videos that are well-aligned with the pose sequences and the semantics of text prompts.

for better content consistency. We then fine-tune the video model for text-to-video generation tasks without the pose control. During the training of the second stage, we only update the parameters related to the temporal consistency, *i.e.*, the newly introduced temporal self-attention and the cross-frame attention blocks, while fixing other parameters. Since the pose control is ensured by our pose encoder and the temporal information is learned via the video dataset and well-designed temporal modules, after the training of two stages, our method manages to generate pose-controllable character videos with text-editable appearance.

Overall, our proposed method is equipped with delicate designs to generate videos that offer flexible control through pose sequence and textual descriptions. Moreover, our model inherits the robust concept generation and composition capabilities of the pre-trained T2I model. As a result, it manages to generate diverse characters, backgrounds, appealing styles, and multiple characters. Although our main focus is on pose control, our approach can be easily extended to other control modalities (*i.e.*, depth, edges, sketch, *e.g.*) as demonstrated in concurrent works including T2I adapter (Mou et al. 2023), and ControlNet (Zhang and Agrawala 2023).

In summary, our contributions are:

- We tackle a new task of pose-controllable text-to-video generation, and introduce the LAION-Pose dataset to facilitate the learning of pose-content alignment.
- To overcome the challenge of lacking diverse character videos, we decouple the task into two subproblems: pose alignment and temporal coherence. We then propose a two-stage training mechanism by carefully tuning different sets of parameters in different datasets.
- Extensive experiments compared with various baselines demonstrate the superiority of our approach, in terms of generation quality, text-video alignment, pose-video alignment, and temporal coherence.

Related Work

Text-to-Video Generation

Synthesizing natural videos is a challenging task due to the complex and high-dimensional structural characteristics involved. Early works focus on Generative adversarial networks (GAN (Goodfellow et al. 2020)), which generate samples from the Gaussian distribution via a two-player minimax game. However, GAN-based frameworks are not stale to train and might be hard to model the large-scale dataset. Thanks to the ability of large language models (Radford et al. 2021) and transformer (Vaswani et al. 2017), more current works generate video from the text description. *e.g.*, GODIVA (Wu et al. 2021a) extends VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2018) to text-to-video generation by mapping text tokens to video tokens. NÜWA (Wu et al. 2021b) proposes an auto-regressive framework that can be used for both text-to-image and video generation tasks. CogVideo (Hong et al. 2022) improves video generation quality by extending CogView-2 (Ding et al. 2022) to T2V generation through the incorporation of temporal attention modules and pre-trained text-to-image models. As for the popularity of the diffusion-based method, Video Diffusion Models (VDM) (Ho et al. 2022b) utilizes a factorized space-time U-Net to directly perform the diffusion process on pixels. Imagen Video (Ho et al. 2022a) enhances VDM by implementing cascaded diffusion models and v-prediction parameterization. Similar to CogVideo, Make-A-Video (Singer et al. 2022) extends the significant progress

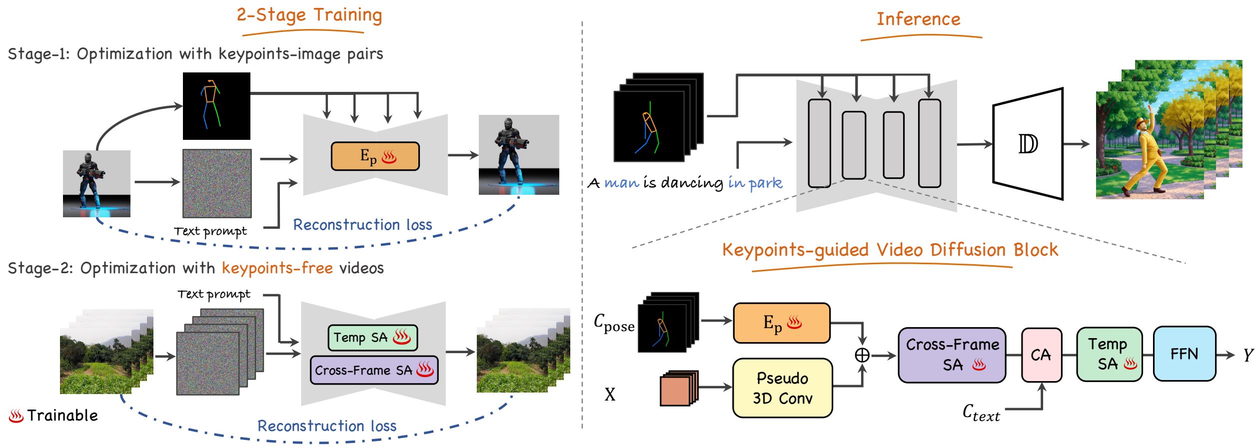

Figure 3: Overview. We propose a two-stage training strategy to effectively learn image-pose alignment from our proposed LAION-Pose dataset and learn temporal coherence from natural videos without pose annotations. During the first-stage training, only the pose encoder E_p is trainable to learn the pose control. During the second-stage training, only the temporal modules are trainable, including the temporal self-attention (SA) and the cross-frame self-attention. During inference, temporally coherent videos are generated by giving a text describing the target character and corresponding appearance and a pose sequence representing the motion. Most parameters from pre-trained stable diffusion are frozen, including the pseudo-3D convolution layers and the cross-attention (CA) and feed-forward network (FFN) modules.

made in the diffusion-based model in text-to-image generation (Rombach et al. 2022) to the T2V task. Similar to Make-A-Video (Singer et al. 2022), we extend image synthesis diffusion models to video generation by introducing temporal connections into a pre-existing image model. Differently, we decouple the motion and appearance in a way that allows for a controllable generation.

Pose-to-Video Generation

Generating the realistic human video from the driving signal, *e.g.*, key points, has been studied extensively during recent years, especially with the unprecedented success of GAN-based (Mirza and Osindero 2014) models for conditional video synthesis. Vid2vid (Wang et al. 2018) presents a video synthesis approach that utilizes conditional GANs, incorporating optical flow and temporal consistency constraints along with multiple discriminators to generate pose video sequences that are both realistic and diverse. Similar work has also been proposed by (Chan et al. 2019), where they over-fit two dance videos by transferring the movements from a source person to a target person through keypoints. However, these methods need to train for each pair. To solve this problem, fewshot-vid2vid (Wang et al. 2019) extends vid2vid to few-shot settings, where the new image can be animated through the learned model directly or a few-shot adaption. LiquidwarpingGAN (Liu et al. 2019) proposes a unified framework for human motion imitation using liquid warping and adversarial learning. FOMM (Siarohin et al. 2019) has designed a first-order motion representation method that is capable of driving any single image through well-decoupled motion fields. FRAA (Siarohin et al. 2021) extends FOMM by a novel motion representation and synthesis framework for articulated characters using pose-dependent motion embedding. Nevertheless, while these methods are capable of generating or driving images that resemble the original data distribution, they fall short in their ability to effectively edit videos through text.

Controllable Diffusion Models

Due to the ambiguity in generating videos or images from a text description, there has been an increasing focus on conditional image synthesis. Early work generates controllable results by image editing method (Meng et al. 2021) directly. More recently, several conditional strategies have been proposed (Mou et al. 2023; Zhang and Agrawala 2023). These methods train an additional encoder or adaptor to map the extracted information, *e.g.*, depth map, keypoints, and segmentation map to controlled features. However, since their methods are still image models, the results suffer from the flickering issue when applying these methods to video.

Method

Our text-guided pose controllable video generation task aims to generate realistic videos from pose sequences and the appearance description from the text. Our method is based on the pretrained text-to-image model, *i.e.*, latent-diffusion model (Rombach et al. 2022) where we make several modifications suitable to our task. In this section, we first give a brief introduction to the latent diffusion model in Sec. . Then, we show the details of our pose-guided controllable text-to-video generation network in Sec. .

Preliminary: Latent Diffusion Models

Latent diffusion models (LDM) (Rombach et al. 2022) is a type of diffusion model that models the distribution of the latent space of images and has shown remarkable image synthesis performance recently. It consists of two models: an

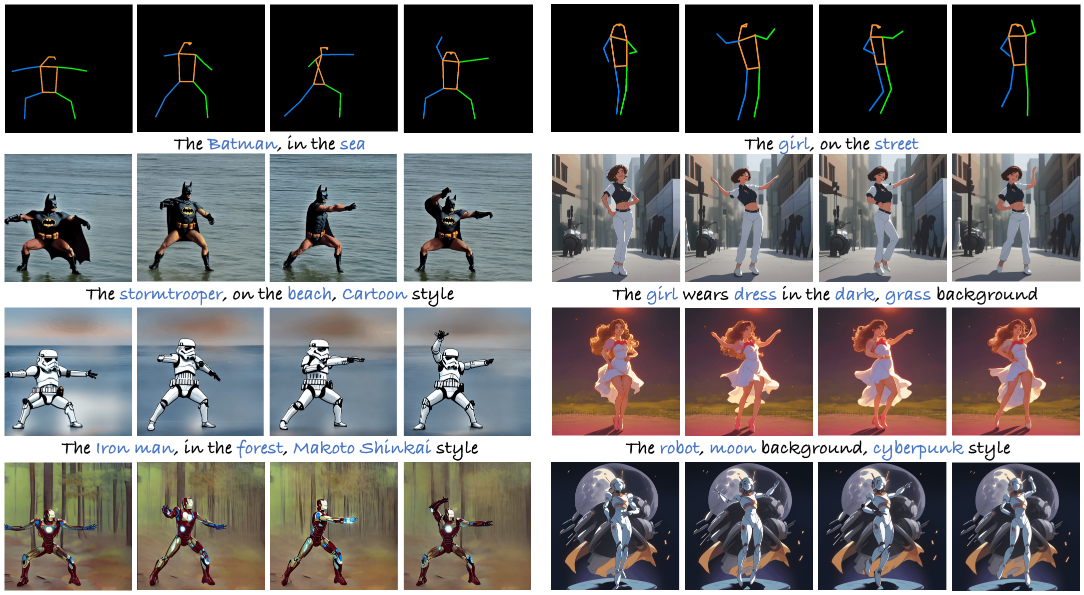

Figure 4: Results regarding various pose sequences and text prompts. Our method can generate videos with high content diversity and temporal coherence, following the pose guidance and semantics of text prompts.

autoencoder and a diffusion model. The autoencoder learns to reconstruct the images via an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder firstly projects the image $\mathbf{x}$ to a lower dimensional latent: $\mathbf{z} = \mathcal{E}(\mathbf{x})$, and the decoder reconstructs the original image from the latent: $\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{z})$.

The diffusion model learns the distribution of image latent $\mathbf{z}_0 \sim p_{data}(\mathbf{z}_0)$ via DDPM (Ho, Jain, and Abbeel 2020) and generates new samples in the latent space. The generation procedure is a gradual backward denoising process with T timesteps starting from pure gaussian noise $\mathbf{z}_t$ to a novel sample $\mathbf{z}_0$:

$$p_\theta(\mathbf{z}_{0:T}) \coloneqq p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t), \tag{1}$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \coloneqq \mathcal{N}(\mathbf{z}_{t_1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \tag{2}$$

The Markov chain is a gradually forward noising process via a predefined noise schedule $\beta_1, \ldots, \beta_T$:

$$q(\mathbf{z}_{1:T}|\mathbf{z}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}), \tag{3}$$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) \coloneqq \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}). \tag{4}$$

At each timestep, a random noise $\boldsymbol{\epsilon}$ is drawn from a diagonal Gaussian distribution, and a time-conditioned denoising model θ is trained to predict the added noise in each timestep with a simple MSE error:

$$\mathcal{L}_{\text{simple}}(\theta) \coloneqq \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) - \boldsymbol{\epsilon}\|_2^2, \tag{5}$$

Pose-guided Text-to-Video Generation

Due to the scarcity of qualified video-pose pairs in various datasets, we decide to decouple temporal and control conditions, whereas our model learns the pose control capability from images and the temporal consistency from videos. Therefore, we train our model in two different stages. As shown in Fig. 3, our method contains two different stages of training for different purposes. Below, we first briefly introduce the denoising network of the Latent Diffusion. Then, we give the details of each stage.

Base Model Architecture. The widely-used diffusion model (Rombach et al. 2022) for image synthesis employs U-Net (Ronneberger, Fischer, and Brox 2015) for denoising, which is a multi-stage neural network architecture that involves spatial downsampling followed by an upsampling. Each stage of the U-Net (Ronneberger, Fischer, and Brox 2015) consists of several attention blocks and layers of 2D convolutional residual blocks. The attention block is constructed from a spatial self-attention, a cross-attention, and a feed-forward network (FFN). The spatial self-attention is utilized for similar correlation by the locations of the latent in representation, while the cross-attention considers the correspondence between latent and conditional inputs (such as text).

Training Stage 1: Pose-Controllable Text-to-Image Generation. In this stage, we train the pose-controllable text-to-image models. However, the current method does not have pose-image-caption pair for pose generation. We, therefore, collect the human skeleton images in the LAION (Schuhmann et al. 2021) by MMpose (Contributors 2020), only retaining images that could be detected more than 50% of the key points. Finally, an image-text-pose dataset named LAION-Pose is formed. This dataset contains diverse human-like characters with various background contexts. To incorporate pose conditions into the pre-trained text-to-image denoising model, we fix all the parameters in the original denoising U-Net and propose a simple and lightweight approach that utilizes multiple 3D convolutional layers as the pose encoder and insert them into each block of U-Net.

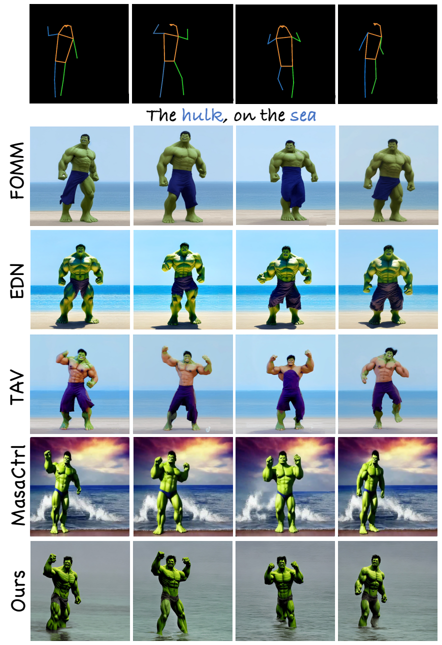

Figure 5: Qualitative comparison between evaluated methods. For FOMM (Siarohin et al. 2019) and EDN (Chan et al. 2019), we generate results by driving one image using source video. As for TAV (Wu et al. 2022), it learns the motion from source video. In masactrl (Cao et al. 2023), we leverage T2I-adapter to control the motion effectively. Our approach achieves consistent background and appearance in the generated video clip.

In detail, we use this convolutional layer in each block to extract pose features from the input pose sequence. After that, the pose features are downsampled to different resolutions (*i.e.* 64×64, 32×32, 16×16, 8×8). We inject this additional controlling information into pre-trained U-Net via a residual connection, by adding the feature into each layer of the U-Net model as shown in Fig 3. This residual injection has the advantage of preserving the generation ability of the pre-trained diffusion model, allowing us to update a few of the parameters to add extra pose-controlling information. Additionally, we can easily drop the condition information by setting the residual to all zero to restore the original diffusion model as demonstrated in Fig. 5. Meanwhile, we also try different ways to incorporate the pose condition into diffusion, such as concatenating it into the channel dimension of the input video, however, the performance is not good as ours as shown in the experiments. We use the simple noise reconstruction loss as the original stable diffusion to tune the injected parameters.

Training Stage 2: Video Generation via Pose-free Videos. However, the stage 1 model can generate similar pose videos yet the background is inconsistent. Thus, we further fine-tune the model from our first stage on the pose-free video dataset HDVLIA (Xue et al. 2022). This dataset contains continuous in-the-wild video text pairs. To generate temporal consistent video, we utilize the generative prior of the text-to-image model for our pose-to-video generation, following previous Video Diffusion Model baselines (Ho et al. 2022a; He et al. 2022; Ho et al. 2022b), we inflate first convolution layer of the pretrained U-Net to a $1 \times 3 \times 3$ convolution kernels as a pseudo-3D convolutional layer for video inputs and appending temporal self-attention for temporal modeling (Carreira and Zisserman 2017). To further maintain the temporal consistency, we leverage the cross-frame self-attention (SA) (Wu et al. 2022) between frames. Differently, we make it to generate longer video sequences by simply reusing the noise from each time step in the previous sampling process of DDIM. In detail, assuming that T frames are sampled each time, we add the noise of the last $\frac{T}{2}$ frames to the next loop as prior knowledge after the first sampling. Note that, throughout the denoising process, the noise at each time step is added to the prior knowledge to ensure the temporal consistency of the generated long videos. As shown in Fig. 3, the proposed method only tune the cross-frame self-attention and the temporal self-attention for video generation.

Discussion. Empowered by the above two-stage training with carefully-designed tune-able blocks for each stage, the proposed method can generate continuous pose-controllable video from the easily obtained datasets, *e.g.*, the image pairs of the human and pose and the random video. Our method can also be used in other related conditional video generation tasks as shown in concurrent conditional image generation works (Zhang and Agrawala 2023; Mou et al. 2023).

Experiments

Implementation Details

We implement our method based on the official codebase of Stable Diffusion (Rombach et al. 2022)[1] and the publicly available 1.4 billion parameter T2I model[2] and several LoRA models from CivitAI[3]. We freeze the image autoencoder to encode each video frame to latent representation individually. We first train our model for 100k steps on the Laion-Pose. Then, we train for 50k steps on the HDVILA(Xue et al. 2022). The eight consecutive frames at the resolution of 512×512 from the input video are sampled for temporal consistency learning. The training process is performed on 8 NVIDIA Tesla 40G-A100 GPUs and can be completed within two days. At inference, we apply DDIM sampler (Song, Meng, and Ermon 2020) with classifier-free guidance (Ho and Salimans 2022) in our experiments for Pose-guided T2V generation.

Applications

We provide several applications of our approach in generating videos with novel visual concepts (e.g., character, backgrounds, styles, etc.) and pose guided by the text prompt. Here, we showcase two various pose-guided training examples and their variations with edited prompts. More examples can be found in the supplementary material.

[1]https://github.com/CompVis/stable-diffusion.
[2]https://huggingface.co/CompVis/stable-diffusion-v1-4.
[3]https://civitai.com/

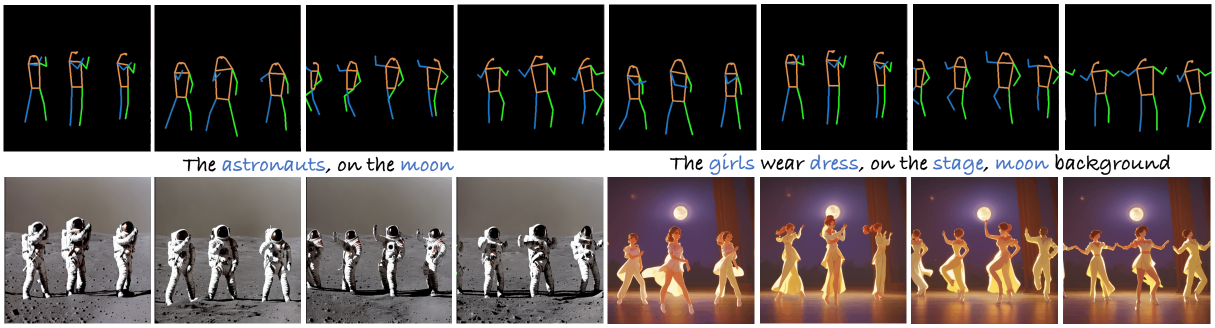

Figure 6: Results regarding multiple skeletons. Given a sequence of poses of multiple persons, our approach can also generalize well in this situation. Furthermore, as shown in the second row, the character identity can be specified via the input textual descriptions such as "the astronauts on the moon". Through this application, we can bring irrelevant characters into one video while specifying the background context.

Diverse character generation. One of the applications of our method is to edit the character with diverse appearances in the generated video. From the Fig. 4, we can see that our approach could generate videos with customized characters by changing the corresponding descriptions in the input text prompt. For example, given an input pose sequence, we manage to generate "batman", "stormtrooper" and "Iron man" etc., with the same motion by varying the input text prompt. The background and the character's appearance successfully keep consistent in different video frames.

Diverse background generation. Our approach also enables generate the natural video background (*i.e.*., the place where the character is), while keeping the pose of the character and the temporal information consistent. For example, we can synthesize the "sea" and "beach" background (the left 2nd and 3rd rows of Fig. 4). Meanwhile, we manage to generate special background such as "moon" and "grass" (the right 3rd and 4th rows of Fig. 4). The generated video maintains a high level of background consistency and effectively matches the pose and text prompts.

Stylized character video generation. Thanks to the broad knowledge of per-trained T2I models and LoRA (Hu et al. 2022), our method is able to generate videos with different impressive styles. By appending the style description (*e.g.*, cartoon style, makoto shinkai style, or cyberpunk style) to the sentence, our method produces videos in that style with consistent poses and semantics (the left 2nd and 3rd rows, the right 3rd row of Fig. 4).

Multiple-characters generation. We found that our approach even enables translating multiple skeletons to video, (*e.g.*, the astronauts and the girls in the 2nd row of Fig. 6). We note that the properties of generated video could nicely match input pose sentences and text prompts, indicating our approach powerful generalization ability.

Comparison with Baselines

We provide quantitative and qualitative comparisons with two concurrent works, including Tune-A-Video (Wu et al. 2022), and ControlNet (Mirza and Osindero 2014). ControlNet is based on a pre-trained Stable Diffusion (He et al. 2022) and mainly applied for various conditional image generation. It is extensively trained on a large-scale dataset of 200k pose-image-caption pairs collected from the internet.

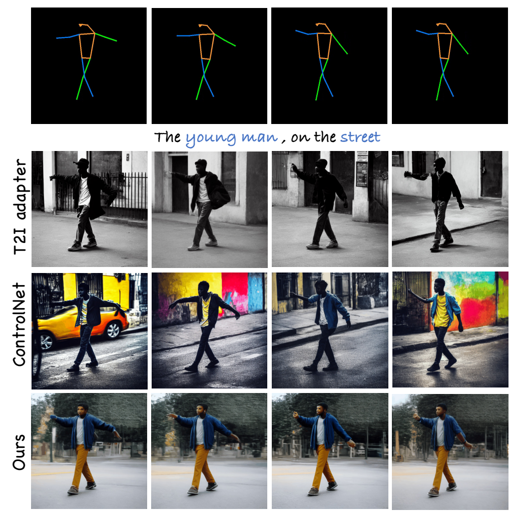

Figure 7: Comparison with ControlNet (Zhang and Agrawala 2023) and T2I adapter (Mou et al. 2023). These two approaches can translate poses into realistic images. However, they tend to generate different character appearances with different backgrounds. Our approach achieves consistent appearance in the generated video clip.

Quantitative results. 1) *CLIP score*: We follow (Ho et al. 2022a) to evaluate our approach on CLIP score (Nguyen et al. 2021; Park et al. 2021) or video-text alignment. We compute CLIP score for each frame and then average them across all frames. The final CLIP score is calculated on 1024 video samples. The results are reported in Tab. 1. Our approach produces a higher CLIP score, demonstrating better video-text alignment than the other two approaches. 2) *Quality*: Following Make-A-Video (Singer et al. 2022), We conduct the human evaluation of videos' quality across a test set containing 32 videos. In detail, we display three videos in random order and request the evaluators to identify the one with superior quality. Our observations indicate that the

Method	CS	QU (%)	PA (%)	FC (%)
FOMM	22.93	0.8	11.7	81.25
Everybodydance	23.04	1.3	13.7	79.83
Tune-A-Video	23.57	23.81	27.74	**93.78**
ControlNet	22.31	6.69	33.23	54.35
T2I-adapter	22.42	8.27	33.47	53.86
Masactrl	23.64	19.17	33.19	87.64
Ours	**24.09**	**39.96**	**34.92**	93.36

Table 1: Quantitative comparisons with related works. CS, QU, PA, and FC represent clip score, quality, pose accuracy, and frame consistency, respectively.

raters favor the videos generated by our approach over Tune-A-Video and ControlNet in video quality. 3) *Pose Accuracy*: we regard the input pose sentence as ground truth video and evaluate the average precision of the skeleton on 1024 video samples. For a fair comparison, we adopt the same pose detector (Sun et al. 2019) for both the processing of LAION-Pose collecting and evaluation. The results show that our model achieves comparable performance with ControlNet. 4) *Frame Consistency*: Following (Esser et al. 2023), we report frame consistency measured via CLIP cosine similarity of consecutive frames, and the results are shown in Tab. 1. Our model outperforms the ControlNet regarding temporal consistency, which demonstrates the necessity of our temporal designs. In addition, we obtain a comparable score with Tune-A-Video. However, Tune-A-Video relies on an input video and needs to overfit the input video, which is hard to sever as a general video generation model.

Qualitative results. We first compare our approach with other four methods with the same pose sequences and text prompt in Fig. 5. Our approach obtains the better performance in consistency and artistry. In similar setting, We also compare our approach and ControlNet and T2I-adapter in Fig. 7. It is apparent to discover that there is an inconsistent background in ControlNet (*e.g.*, the color of the street and the shirt of the boy). This phenomenon is common and can also be encountered in the 3rd row of Fig. 7 (the wall color). In contrast, our approach could effectively address the issue of temporal consistency and learn good inter-frame coherence.

Ablation Study

Effect of residual pose encoder. We ablate the feature residual design of the proposed pose encoder. To add controlling information to the diffusion model, one natural way is to directly concatenate the condition onto the input noisy latent of the model (Rombach et al. 2022). However, we discover that the concatenation approach yields worse performance than the residual approach. We compare the generation results of residual and concatenation approaches for injecting the pose information in the first row and the second row of Fig. 8. We discover that the feature residual manner for adding extra controls can preserve more of the generation ability of pre-trained stable diffusion. This is because the concatenation mechanism needs to retrain the first convolutional layer to meet the number of input channels which sacrifices the pretrained high-quality image synthesis prior.

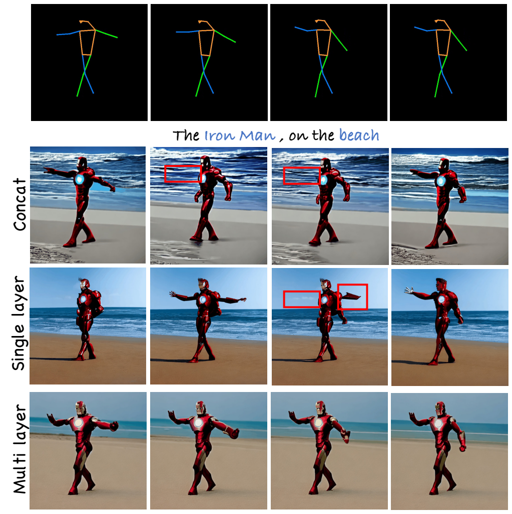

Figure 8: Ablation study. Including the condition injection mechanism (feature residual and channel-wise concatenation) and the number of layers with conditional signals.

Number of layers of condition control. We also ablate the number of layers injecting the controlling signals in order to achieve better controllability (see comparisons between the second row and the third row in Fig. 8). Our results indicate that adding controls to more layers can lead to improved pose-frame alignment. Specifically, adding the pose into one single layer results in the mismatch between the target arm and the generated arm of Iron man.

Conclusion

In this paper, we tackle the problem of generating text-editable and pose-controllable character videos. To achieve this, we reform and tune from the pre-trained text-to-image model due to its powerful semantic editing and composition capacities. We thus design a new two-stage training scheme that can utilize the large-scale image pose pairs and the diverse pose-free dataset. In detail, in the first training stage, we use a pose encoder to inject the pose information into the network structure and learn from the image-pose pair for the pose-controllable text-to-image generation. In the second training stage, we inflate the image model to a 3D network to learn the temporal coherence from the pose-free video. Equipped with our several new designs, we can generate novel and creative temporally coherent videos with the preservation of the conceptual combination ability of the original T2I model.

Acknowledgments

This project was supported by the National Key R&D Program of China under grant number 2022ZD0161501, Shenzhen Key Laboratory of next generation interactive media innovative technology(No:ZDSYS20210623092001004).

References

Avrahami, O.; Fried, O.; and Lischinski, D. 2022. Blended Latent Diffusion. *arXiv preprint arXiv:2206.02779*.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. arXiv:2304.08465.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In *IEEE International Conference on Computer Vision (ICCV)*.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217*, 2(3): 8.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. *arXiv preprint arXiv:2302.03011*.

Gao, K.; Bai, Y.; Gu, J.; Yang, Y.; and Xia, S.-T. 2023. Backdoor Defense via Adaptively Splitting Poisoned Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4005–4014.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, C.; Li, K.; Zhang, Y.; Zhang, Y.; Guo, Z.; Li, X.; Danelljan, M.; and Yu, F. 2023. Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects. *arXiv preprint arXiv:2308.03166*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *arXiv preprint arXiv:2211.13221*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *arXiv preprint arXiv:2204.03458*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-Scale Pretraining for Text-to-Video Generation via Transformers. (3).

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5904–5913.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023a. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023b. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Ma, Z.; Jia, G.; and Zhou, B. 2023. AdapEdit: Spatio-Temporal Guided Adaptive Editing Algorithm for Text-Based Continuity-Sensitive Image Editing. arXiv:2312.08019.

Ma, Z.; zhihuan yu; Li, J.; and Zhou, B. 2023c. LMD: Faster Image Reconstruction with Latent Masking Diffusion. arXiv:2312.07971.

Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*.

Nguyen, H.-L.; Kim, J.; Yeo, H.; Lee, D.; and Yoon, S.-E. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3386–3395.

Park, D. H.; Azadi, S.; Liu, X.; Darrell, T.; and Rohrbach, A. 2021. Benchmark for Compositional Text-to-Image Synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. arXiv:2303.09535.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13653–13662.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792*, 2(3): 4, 7, 8.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. HRNet: Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2018. Neural discrete representation learning. In *NeurIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.-C.; Liu, M.-Y.; Tao, A.; Liu, G.; Kautz, J.; and Catanzaro, B. 2019. Few-shot Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; and Duan, N. 2021a. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; and Duan, N. 2021b. Nüwa: Visual Synthesis Pretraining for Neural Visual World Creation.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565*.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.