TCAN: Animating Human Images with Temporally Consistent Pose Guidance using Diffusion Models

Jeongho Kim[1*], Min-Jung Kim[1*], Junsoo Lee[2], and Jaegul Choo[1]

[1]KAIST [2]NAVER WEBTOON AI

{rlawjdghek, emjay73, jchoo}@kaist.ac.kr, ljs93kr@gmail.com

Abstract. Pose-driven human-image animation diffusion models have shown remarkable capabilities in realistic human video synthesis. Despite the promising results achieved by previous approaches, challenges persist in achieving temporally consistent animation and ensuring robustness with off-the-shelf pose detectors. In this paper, we present TCAN, a pose-driven human image animation method that is robust to erroneous poses and consistent over time. In contrast to previous methods, we utilize the pre-trained ControlNet without fine-tuning to leverage its extensive pre-acquired knowledge from numerous pose-image-caption pairs. To keep the ControlNet frozen, we adapt LoRA to the UNet layers, enabling the network to align the latent space between the pose and appearance features. Additionally, by introducing an additional temporal layer to the ControlNet, we enhance robustness against outliers of the pose detector. Through the analysis of attention maps over the temporal axis, we also designed a novel temperature map leveraging pose information, allowing for a more static background. Extensive experiments demonstrate that the proposed method can achieve promising results in video synthesis tasks encompassing various poses, like chibi. Project Page: https://eccv2024tcan.github.io/

Keywords: Image-to-Video Generation· Human Animation · Diffusion Model

1 Introduction

Pose-driven human-image animation is a task that breathes life into a static human in an image. Given a source image and a driving video, the task aims to transfer the motion in the driving video to a human in the source image. Although it is an intriguing task because of its wide impact on many applications such as social media, entertainment, and movie industry, several factors make this task challenging. First, the identity in the source image should remain the same over time, changing only the pose. In other words, the appearance of a human in the generated video should not be affected by the appearance of a human

---

[*]Authors contributed equally to this work.

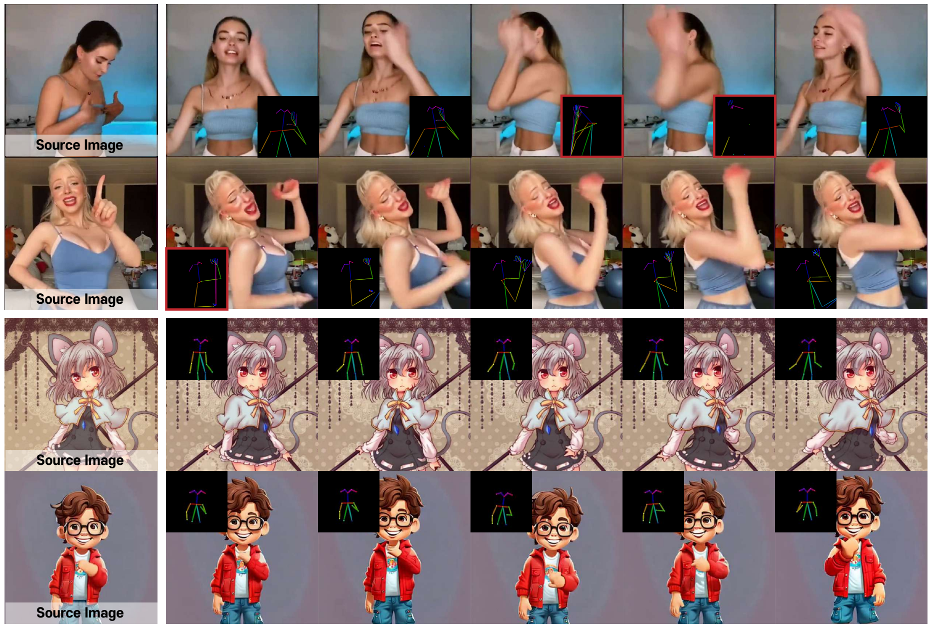

Fig. 1: Generated results of TCAN: The first two rows show the results on the TikTok dataset, and the last two rows show the results on chibi animation characters. All results are generated using TCAN trained on the TikTok dataset. Severely erroneous input poses are highlighted in red. Note that the proposed TCAN can generalize to poses with outliers and unusual ratios, such as those of chibi characters.

in the driving video. Secondly, the occluded regions in the source image should be naturally inpainted as the pose changes, while ensuring that the inpainted regions remain unchanged over time.

One possible approach is to impose a pose-guided image generation model frame-by-frame [33,34,36]. Upon the success of the Text-to-Image (T2I) diffusion model, researchers made an effort to control the diffusion model using additional conditions in addition to the text prompts [9,10,19,33,34,36,37]. ControlNet [34] is the one representative work that enriches the controllability of the stable diffusion model, embracing various conditions, including pose, depth, and edge map. Uni-ControlNet [36] extends ControlNet to handle multiple conditions at a time, enabling the generation of an image conditioned on a source image and a pose image. Meanwhile, in an effort to adapt the appearance of the source image along with text prompt, IP-Adapter [33] introduces an image prompt adapter that embeds the fine-grained image feature into a diffusion model. As the IP-Adapter is compatible with pose ControlNet [34], a human image with a given pose can be generated using both. However, the aforementioned methods are oriented towards a single image generation, not considering temporal consistency.

To incorporate temporal consistency into the diffusion model, T2V models focus on the temporal alignment of latents by inserting temporal layers into the diffusion model. In general, T2V models can be extended to Image-to-Video (I2V) models by replacing text conditions with CLIP image embedding [3] or by concatenating an encoded initial frame to the temporal layer input along the channel dimension [4]. However, controlling the object in the generated video using a pose sequence is out of their research scope. To control humans in the video, there have been several attempts to generate video conditioned on a pose sequence [8,15,24,28,30,32]. MagicAnimate [30], which is the most similar work to ours, extracts pose features from fine-tuned ControlNet and feeds them into the skip connection of Stable Diffusion UNet. Since MagicAnimate fine-tunes the ControlNet with the dataset of the target domain, it tends to overfit to the target domain poses. Meanwhile, using the pre-trained ControlNet without fine-tuning does not guarantee generalizability or target domain performance, because there is no way to align the appearance and pose features when the ControlNet is frozen.

To leverage the pre-acquired knowledge of ControlNet, we keep ControlNet frozen and train additional LoRA layer to align the appearance and the pose feature without compromising the ControlNet's performance. Moreover, to ensure the background remains static while maintaining the dynamic motion of the foreground region, we introduce a pose-driven temperature map to obtain the static background. Finally, to enhance temporal robustness against the noise of off-the-shelf pose estimators, we integrate temporal layers into ControlNet for pose conditioning.

To sum up, our contributions are as follows:

- We propose TCAN, a novel human image animation framework based on the diffusion model that maintains temporal consistency and generalizes well to unseen domains.
- We are the first to freeze the pre-trained ControlNet and find it helpful for not only preventing the network from overfitting to the training domain but also disentangling appearance from the pose. By introducing the Appearance-Pose Adaptation (APPA) layer, we effectively address artifacts caused by frozen ControlNet adaptation.
- We propose Pose-driven Temperature Map (PTM) and Temporal ControlNet (T-CtrlN) to enhance stability at the video level, demonstrating superior performance improvement in the video quality metrics (FID-VID and FVD).

2 Related work

2.1 Controllable Diffusion Model for Image Generation

Thanks to the scalability and stable training of diffusion models [14, 26], generative foundation models [20, 22, 23] are now leading recent image generation research. Building on the foundation model, active research has been conducted on the image-based controllable diffusion model. In particular, several studies have expanded the T2I foundation models into an image-controllable generative model through modules using image conditioning [9, 10, 18, 19, 31, 33, 34]. Such modules, trained on large-scale datasets, work as plug-and-play components. They can be combined with other trained models and used across various applications. Despite their outstanding adaptation abilities at the image level, the aforementioned methods focus only on the controllability of a single frame, rather than a video. In this paper, we leverage modules trained on a vast amount of image datasets and extend them to the video level, presenting a video generation model that demonstrates enhanced generalization performance.

2.2 Pose-driven Video Diffusion Models

Several efforts [3, 4, 12, 25] have expanded image generation models to video generation models, with most employing a two-stage training scheme [4, 12]. In the first stage, the image generation model is fine-tuned using video frames. In the second stage, temporal layers are added and trained while keeping the base image generation model frozen.

Recent pose-driven video diffusion models follow the same two-stage training paradigm. In the first stage, the pose-driven image generation model is trained on individual video frames and corresponding pose images extracted using a pose estimator. Then, the model is inflated by inserting temporal layers to capture the temporal information of the pose sequence extracted from a driving video. Disco [28], for example, takes three properties as input — human subjects, backgrounds, and poses — to achieve not only arbitrary compositionality but also high fidelity. MagicAnimate [30] achieves state-of-the-art performance with a parallel UNet architecture conditioned on the source image through an attention mechanism. While both works show impressive video quality in human image animation, they address neither overfitting nor the incompleteness of pose sequences.

Unlike the aforementioned methods, we propose a more generalized model that performs well on datasets beyond the real human image domain by adapting prior knowledge from the pre-trained model. Additionally, we use the temporal ControlNet in the second stage to address incomplete pose information. To the best of our knowledge, both overfitting and pose incompleteness have not been explored in previous research.

3 Method

An overview of TCAN is presented in Fig. 2. Given a source image and a driving video with F frames, we aim to generate a video that: 1) incorporates the foreground appearance from the source image, 2) follows the pose of the driving video, and 3) maintains consistency in the background of the source image.

As shown in Fig. 2, TCAN follows a two-stage training framework widely adopted in current diffusion-based video generation works [4, 12]: In the first stage, the source image is transformed to match the pose of each frame in the driving video at the image level. To achieve this, we tackle the human image

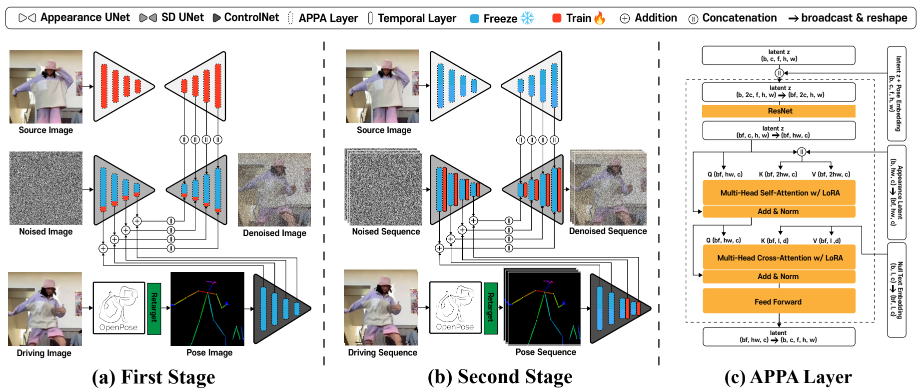

Fig. 2: Our method involves a two-stage training strategy. We randomly select two images from the training video and use them as the source image and the driving image, respectively. In the first stage, the appearance UNet and the APPA layer are trained conditioning on the source image while the ControlNet is frozen. In the second stage, we train the temporal layers in the denoising UNet and ControlNet.

animation task as a controllable image generation task under multiple image conditions (*i.e.*, a pose condition and a source image), aiming to generate images that follow the style of the source image while adhering to the pose of the driving frames. We utilize OpenPose [5] to provide the pose information from the driving frames as input to the ControlNet. For the source image conditioning, we use an attention mechanism [30, 37] to pass the intermediate feature maps of the appearance UNet to the denoising UNet, providing finer features of the source image to denoising UNet. In the second stage, we extend the image generation model to a video generation model by adding temporal layers [12].

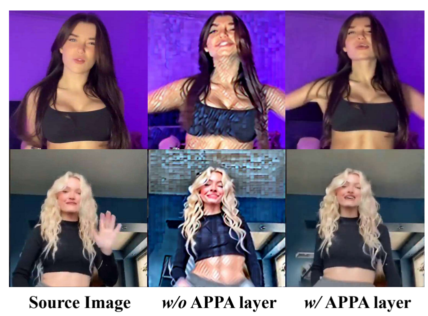

Fig. 3: Ablation on the APPA layer. Using only the frozen ControlNet significantly deteriorates texture quality, while the APPA layer accurately preserves the style of both the foreground and background from the source image in the output, showing notable differences.

3.1 Appearance-Pose Adaptation Layer

To leverage the well-generalized pose conditioning capability of the pre-trained ControlNet, we opted to freeze the ControlNet in the first stage, in contrast to previous studies. By adding the pose information from ControlNet to the denoising UNet's encoder feature z, we embed the pose information into the denoising UNet. This is similar to MagicAnimate [30], but we use the feature from the frozen ControlNet, unlike MagicAnimate. Along with pose information, we inject the appearance information z_a of the source image through the attention layer as in MagicAnimate [30], which is formulated as the following equation:

$$\begin{gathered}\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \\ \mathbf{Q} = \mathbf{W}_0^Q z, \mathbf{K} = \mathbf{W}_0^K (z \parallel z_a), \mathbf{V} = \mathbf{W}_0^V (z \parallel z_a),\end{gathered} \tag{1}$$

where $\parallel$ denotes the concatenation operation along the spatial dimension and $\mathbf{W}_0$ denotes the projection matrix of the denoising U-Net. We observe that solely training the appearance encoder with an adaptation of the pre-trained ControlNet allows for the successful reflection of the driving pose in the generated images. However, we also note a severe texture degradation in generation results, as shown in Fig. 3. We conjecture that the degradation of texture information is attributable to the misalignment between the pose information from the pre-trained ControlNet and newly trained appearance information. Therefore, we propose an APpearance-Pose Adaptation layer (APPA layer), which preserves the appearance of the source image while maintaining pose information from the frozen ControlNet by aligning the feature of two different properties. Our proposed APPA layer is illustrated in Fig. 2 (c). We implement the APPA layer by employing low-rank adaptation (LoRA) to the existing attention layers of the denoising UNet, specifically formulating the attention layer into the following equation, where appearance information z_a is injected:

$$\mathbf{Q} = (\mathbf{W}_0^{\mathbf{Q}} + \Delta\mathbf{W}^Q) z, \mathbf{K} = (\mathbf{W}_0^K + \Delta\mathbf{W}^K)(z \parallel z_a), \mathbf{V} = (\mathbf{W}_0^V + \Delta\mathbf{W}^V)(z \parallel z_a). \tag{2}$$

Here, $\Delta\mathbf{W}^{\{Q,K,V\}} = \mathbf{B}^{\{Q,K,V\}}\mathbf{A}^{\{Q,K,V\}}$. The matrices $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times c}$ denote low-rank-parameterized matrices with rank r. The appearance UNet and LoRA layers are trained in the first stage. By introducing the APPA layer, we effectively address artifacts caused by the frozen ControlNet adaptation. Moreover, we found that such an approach is helpful for preventing the network from overfitting to the training domain and disentangling appearance from the pose.

3.2 Temporal ControlNet

Pose-guided human image animation methods, including our TCAN, are conditioned on the pose sequence estimated from the driving video. Consequently, the quality of the generated outputs relies on the accuracy of the pose estimator, making them inherently susceptible to inaccuracies in the predicted poses. One potential method to mitigate the impact of errors from the pose estimator is to fix the false detections from the estimated poses before use. However, distinguishing noisy keypoints from true positives or adding missing keypoints is challenging and tedious. Moreover, this frame-by-frame noise removal approach fails to ensure temporal consistency across multiple frames of the pose sequence.

To prevent the generated video from collapsing due to abrupt and erroneous pose changes, we propose Temporal ControlNet. Temporal ControlNet incorporates temporal layers, referred to as the motion module in AnimateDiff [12],

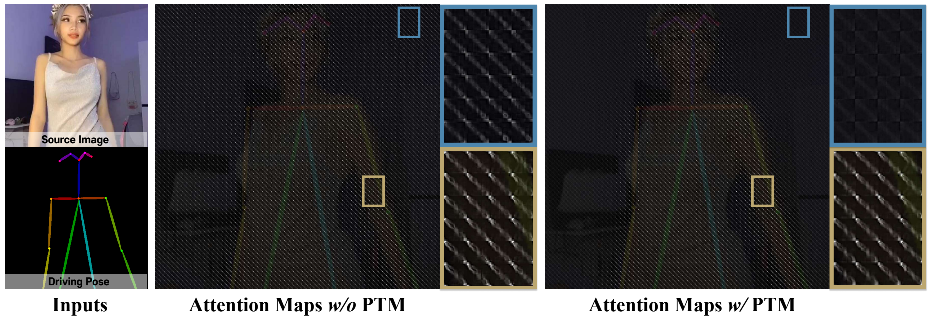

Fig. 4: Visualization of the attention maps from the temporal transformer block at a resolution of 64×64. Best viewed when zoomed in.

which converts a T2I diffusion model into a T2V diffusion model. When the input tensor $z \in \mathbb{R}^{b\times c\times f\times h\times w}$ of f frames is given, we reshape the tensor as $z \in \mathbb{R}^{bhw\times f\times c}$ and feed it to the temporal layer. The output tensor from the temporal layer is reshaped as $z \in \mathbb{R}^{b\times c\times f\times h\times w}$, so that the next layer can extract features with respect to the spatial dimension.

The rationale behind this strategy is that when the ControlNet, pre-trained at the image level, receives a sequence containing erroneous poses as input, we want ControlNet to refrain from directly conditioning the denoising UNet on the erroneous poses. Instead, by introducing temporal layers into ControlNet, we can mitigate the erroneous poses by referencing neighboring correct poses along the temporal axis. During training, only the temporal layers are trained in the second stage, while the pre-trained ControlNet remains frozen.

3.3 Pose-driven Temperature Map

Even with APPA layers and temporal ControlNet, the TCAN trained up to the second stage experiences flickering in the static region (i.e., background). To address this issue, we propose a novel Pose-driven Temperature Map (PTM). The PTM is motivated by the observation that, *compared to the foreground, the background tends to have a less specific focus over the temporal axis.* Fig. 4 supports the idea by visualizing the attention maps $m \in \mathbb{R}^{hf\times wf}$ of the temporal attention layer after the second stage training , where h and w are 64 and the resolution of each attention map is $f \times f$. Each attention map represents the attention of each location along the temporal axis across consecutive video frames. For ease of interpretation, we resized the generated frames and the driving pose frame to a size of attention maps m and overlaid them. Interestingly, we observe that the attention map corresponding to the moving parts, i.e., foreground objects, shows higher diagonal values compared to the static background. This indicates that dynamic objects stay at the same location *temporally*, thereby referencing adjacent frames more than the static background region. Consequently, we designed a temperature map to smooth the attention scores far from the dynamic object.

Given driving OpenPose sequence $\mathbb{P} = \{P^i \mid P^i \in \mathbb{R}^{3\times H\times W}, i = 1, \ldots, f\}$ of f frames, we first generate a binary mask $B \in \{0,1\}^{H\times W}$ indicating the presence of OpenPose in each pixel. Formally, this can be expressed as follows:

$$B(u,v) = \max_{i=1}^{f} B^i(u,v), \quad B^i(u,v) = \begin{cases} 1 & \text{if } P^i(u,v) \neq (0,0,0) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where u and v indicate the pixel coordinates. Let $B' = \{(u,v) | B(u,v) = 1\}$ be the set of pixel coordinates in B with a value of 1. Then, we can calculate the distance map $D \in \mathbb{R}^{H\times W}$ as follows:

$$D(x,y) = \frac{\min\{\sqrt{(x-u)^2 + (y-v)^2} | (u,v) \in B'\}}{\sqrt{(H/2)^2 + (W/2)^2}}, \tag{4}$$

where x and y indicate the pixel coordinates of distance map D. Based on the distance map, we smooth the attention map corresponding to regions far from the foreground objects , i.e., the background. We define the PTM $\mathcal{T} \in \mathbb{R}^{H\times W}$ as follows:

$$\mathcal{T} = \tau_{PTM} \cdot D + 1, \tag{5}$$

where τ_{PTM} serves as a hyper-parameter that adjusts the scale of the temperature values. The value of 1 is added in order to maintain the attention map of the region near the OpenPose unchanged. Experimentally, $\tau_{PTM} = 3$ shows reasonable performance, and we use this value throughout the paper. Finally, the PTM $\mathcal{T}$ is resized to match each temporal layer's spatial size in the denoising UNet and used as the temperature value for its attention operation. The formulation applied to the temporal layer of PTM $\mathcal{T}$ during inference stage is as follows (using `einops` [21] notation):

$$\begin{aligned} \mathbf{A}' &= \texttt{rearrange}(\mathbf{Q}_t\mathbf{K}_t^T, && \texttt{(b h w) f f} \rightarrow \texttt{b f f h w)} \\ \mathcal{T}' &= \texttt{rearrange}(\mathcal{T}, && \texttt{b h w} \rightarrow \texttt{b 1 1 h w)} \\ \mathbf{A} &= \texttt{rearrange}(\mathbf{A}'/(\mathcal{T}'\sqrt{d}), && \texttt{b f f h w} \rightarrow \texttt{(b h w) f f)} \\ z' &= \text{softmax}(\mathbf{A})\mathbf{V} \end{aligned}$$

where $\mathbf{Q}_t = \mathbf{W}_t^Q z, \mathbf{K}_t = \mathbf{W}_t^K z, \mathbf{V}_t = \mathbf{W}_t^V z$. Here, $\mathbf{W}_t^Q, \mathbf{W}_t^K$, and $\mathbf{W}_t^V$ are weight matrices of the temporal layer.

3.4 Training and Long-Term Video Prediction

Training We initialize the denoising UNet with the pretrained weights of Stable Diffusion 1.5[1] , the appearance UNet with those of Realistic Vision[2], and employ the weights of OpenPose ControlNet [3]. In the first stage, we freeze all modules except the appearance UNet and the LoRA layer within the APPA layer. In the

[1] Hugging Face SD 1.5: https://huggingface.co/runwayml/stable-diffusion-v1-5

[2] Civitai RealisticVision: https://civitai.com/models/4201?modelVersionId=130072

[3] Hugging Face OpenPose: https://huggingface.co/lllyasviel/sd-controlnet-openpose

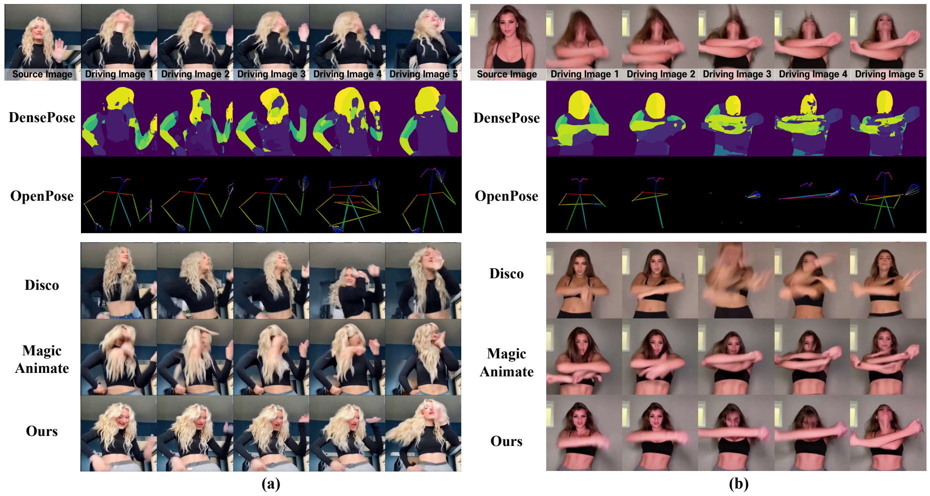

Fig. 5: Qualitative comparison with baselines on the TikTok [16] dataset. Frames from segments of the driving video are arranged sequentially.

second stage, we introduce temporal layers to the denoising UNet and ControlNet while keeping the rest of the weights frozen. The training objective function in both stages follows the simplified loss of the latent diffusion model [22].

Long-Term Video Prediction Due to the memory constraints, TCAN cannot generate long-term video all at once. However, generating the video of F frames by concatenating short videos of f frames does not take temporal consistency into account. To address this, during inference, we apply MultiDiffusion [2] along the temporal axis. This involves overlapping certain pose sequences in the model's input and averaging predicted noise for overlapping frames.

4 Experiments

Evaluation setting During training, we use the TikTok dataset [16] only. For evaluation, we utilize two different datasets. First, we use 10 TikTok-style videos collected from DisCo [28] as our test dataset, referring to it as the TikTok dataset evaluation. Second, to assess the generalization of TCAN, we evaluate the model trained on the TikTok dataset using animation characters. Note that the textures and body proportions of the animation characters are quite different from those of real human datasets. We select 10 images from the Bizarre Pose Dataset [7] as source images and use the 10 TikTok-style test videos as driving videos. We refer to this as the Bizarre dataset evaluation.

4.1 TikTok Dataset Evaluation

Qualitative Results As shown in Fig. 5, we qualitatively compare our method with two baselines, DisCo [28] and MagicAnimate [30], both of which demon-

Method	L1↓	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
IPA [33]+CtrlN [34]	7.38E-04	0.459	0.481	69.83	113.31	802.44
IPA [33]+CtrlN [34]-V	6.99E-04	0.479	0.461	66.81	86.33	666.27
DisCo [28]	3.78E-04	0.668	<u>0.292</u>	30.75	59.90	292.80
MagicAnimate	3.13E-04	0.714	**0.239**	32.09	<u>21.75</u>	<u>179.07</u>
TCAN (*w/ mm1*)	<u>8.85E-05</u>	<u>0.734</u>	0.299	**29.07**	29.74	189.77
TCAN (*w/ mm2*)	**8.81E-05**	**0.745**	0.294	<u>29.35</u>	**19.42**	**154.84**

Table 1: Quantitative results on TikTok dataset evaluation. Metrics for baseline methods are cited directly from MagicAnimate [30]. The best performance is highlighted in bold, and the second best is underlined. *mm1* and *mm2* refer to version 1 and version 2 of the pretrained motion module from AnimateDiff [12].

strate state-of-the-art performance in human image animation. Note that MagicAnimate utilizes DensePose, while the others use OpenPose. Due to errors in the off-the-shelf pose extractor, significant pose information loss is evident in the fourth frame of Fig. 5(a) and the third and fourth frames of Fig. 5(b). As a result, DisCo generates an anomalous frame with severe blurring and artifacts (third frame of Fig. 5(b)). Similarly, MagicAnimate, which is heavily dependent on DensePose, generates results that align with erroneous DensePose outputs (Fig. 5(a)). On the other hand, despite using the same OpenPose outputs as DisCo, TCAN generates temporally consistent outputs without significant distortion or artifacts, especially in the fourth frame of Fig. 5(a).

Quantitative Results In Tab. 1, we compare our TCAN with four diffusion-based generation models. The IPA + CtrlN is the model which combines IP-adapter [33] and pre-trained OpenPose ControlNet [34], and IPA + CtrlN-V is the model that extends IPA + CtrlN as a video generation model by inserting the temporal modules. DisCo [28] and MagicAnimate [30] are recently proposed human image animation baselines.

We conduct TikTok dataset evaluation of 10 test videos using both image-level metrics (*i.e.*, L1, SSIM [29], LPIPS [35], and FID [13]) and video-level metrics (*i.e.*, FID-VID [1] and FVD [27]). Given test videos, the first frame of each video is used as the source image, and the remaining frames are used as the driving sequence. We use all frames of test videos without sampling during evaluation. For image quality assessment, we measure L1, SSIM, LPIPS, and FID between the generated and real frames. For video quality assessment metrics, FID-VID and FVD, we use 3D ResNet-50 and Inflated 3D ConvNet [6] pre-trained on the Kinetics dataset [17] as feature extractors, respectively. For implementation, we follow the official DisCo repository [4].

As reported in Table 1, TCAN shows state-of-the-art performance in all metrics except LPIPS, without using the additional large-scale human dataset as in DisCo [28] and MagicAnimate [30]. We evaluate the performance using two versions of the motion module in AnimateDiff [12] and observe that utilizing the advanced motion module (*i.e.,* version 2) as the initial weight results in noticeably higher performance.

[4] DisCo Repository: https://github.com/Wangt-CN/DisCo

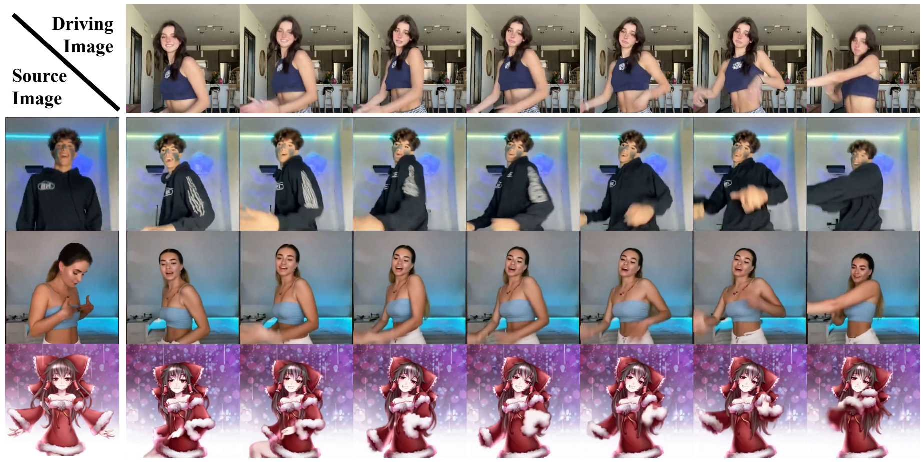

Fig. 6: Qualitative results of TCAN. This figure shows that our method effectively transfers motion information to various identities, including animated characters, despite differences in proportions between animation characters and humans.

4.2 Bizarre Dataset Evaluation with Pose Re-targeting

Pose Re-targeting Beyond the TikTok dataset evaluation, where both the source image and driving sequence are derived from the same video, we extend our model to applications where the foreground object in the source image possesses body proportions different from those of an actual human. Specifically, when the source images are sampled from an animation domain, there are significant differences in body proportions compared to TikTok dataset, as depicted in Fig. 6 and Fig. 7. To address this issue, we propose a re-targeting methodology. The pose re-targeting is designed to transfer the body the proportions of an object in the source image to that of the target frame, regardless of the object proportion in the driving sequence. When used in practice, we re-target the pose sequence estimated from the driving video so that the retargeted poses fit the proportion of the animation character in the source image. As shown in Fig. 6, re-targeting enables the preservation of proportions of an object in the source image, regardless of whether the object features a human or an animated character, while faithfully following the motions in the driving video. Detailed explanations of the Pose Re-targeting are provided in the supplementary material.

Qualitative Results In Fig. 7, we qualitatively compare our TCAN with existing human image animation baselines, DisCo and MagicAnimate. For the sake of fairness, the re-targeted OpenPose sequence is used as a driving pose sequence for DisCo. We observe that the DisCo fails to maintain temporal consistency and character identity. Similarly, MagicAnimate, conditioned on DensePose [11], generates videos that compromise the source image's identity with the shape in driving video. In contrast, thanks to the APPA layer and pose re-targeting, our

Fig. 7: Qualitative comparison in a Bizarre dataset evaluation. As a source image, we use an animation character with different textures and body proportions compared to driving videos in TikTok dataset. Disco and our model, based on OpenPose are evaluated by applying the proposed pose re-targeting. To demonstrate the generalization performance of proposed method, we also visualize the results obtained when training ControlNet.

model animates following the driving video's motion while preserving the source image's identity. As demonstrated in the last two rows in Fig. 7, freezing the ControlNet plays a significant role in maintaining the identity. We conjecture that this occurs because freezing the ControlNet allows for the disentanglement of the network branches responsible for pose and appearance during training. Additionally, the pre-acquired knowledge of the ControlNet trained with OpenPose excels at understanding uncommon body postures and proportions.

User Study We further conduct a user study to evaluate the Bizarre dataset. A total of 27 participants were shown 10 videos generated by DisCo, MagicAnimate, and TCAN, respectively. Following this, participants were asked to rate the video on a scale from 1 to 5 based on four criteria: 1) Motion and identity, 2) Background, 3) Flickering, and 4) Overall preference. The detailed instructions for each criterion are as follows:

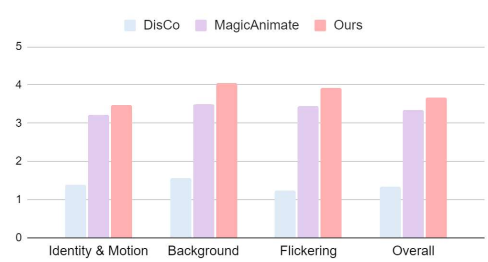

Fig. 8: User study results involving a total of 27 participants.

– Motion and Identity: *Please rate how well the motion of the driving video was reflected while maintaining the identity of the source image.*

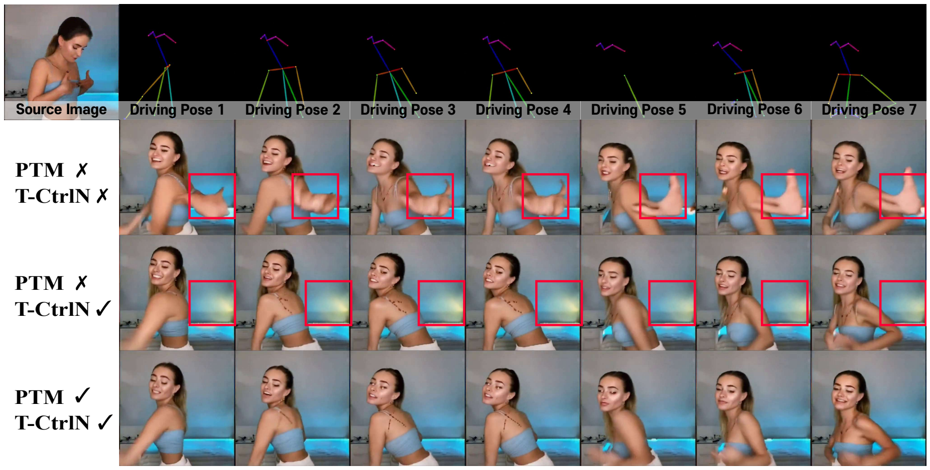

Fig. 9: Qualitative results from our ablation study. We visualized the results of incorporating the proposed temporal ControlNet and PTM into the model with APPA layer. Red boxes highlight areas with body artifacts and background flickering.

- Background: *Please rate the extent to which the background of the source image is maintained in the generated video.*
- Flickering: *Please rate how minimal the flickering is in the video.*
- Overall: *Please rate the overall preference score for the video.*

As shown in Fig. 8, our TCAN consistently outperforms previous state-of-the-art baselines. DisCo receives the lowest scores on all of the criteria due to severe flickering, identity loss, and artifacts in the foreground character. MagicAnimate also suffers from identity loss by generating characters that fit the DensePose of the driving video. Thanks to our novel PTM, our model demonstrates superior performance with minimal background flickering while maintaining the best performance in other criteria. Consequently, our model achieves the highest overall quality preference.

5 Ablation Study

Qualitative Results We first investigate the generalization performance of APPA layer in the animation domain. We compare TCAN with a variant where ControlNet is unfrozen, and the APPA layer is omitted, For both models, we use re-targeted OpenPose as the input. As demonstrated in the Fig. 7, even when the ControlNet is initialized with pre-trained weights, the model exhibits a loss of prior pose knowledge if we unfreeze the ControlNet during training. Therefore, it fails to incorporate the variant poses in driving videos, leading to artifacts, particularly in the arm and torso regions. In contrast, our model, which keeps ControlNet frozen with the APPA layer, effectively maintains the appearance of the source image while accurately following the pose of the driving sequence.

APPA Layer	Temporal ControlNet	PTM	L1↓	SSIM↑	LPIPS↓	FID↓	FID-FVD↓	FVD↓
✗	✗	✗	1.13E-04	0.709	0.342	40.60	37.33	368.71
✓	✗	✗	9.52E-05	0.728	0.306	31.13	33.64	252.13
✓	✓	✗	9.38E-05	0.724	0.311	30.83	30.04	216.78
✓	✓	✓	**8.85E-05**	**0.734**	**0.299**	**29.07**	**29.74**	**189.77**

Table 2: Ablation results of our proposed training components on TikTok dataset. Our proposed TCAN utilizing all proposed modules shows the best performance across all image and video metrics. The highest performance is highlighted in bold.

Also, we examine the effect of our temporal ControlNet, and PTM. Fig. 9 shows eight consecutive frames from the generated video for each experimental scenario. First, we demonstrate the effect of the temporal ControlNet, which is shown in the second and third rows of Fig. 9. The ControlNet branch expanded with temporal layers smooth out erroneous poses compared to the original ControlNet, where the right(orange) and left(green) arms are switched after the second frame and get back to normal at the last frame. This improvement is attributed to the temporal ControlNet, which effectively addresses outlier poses compared to the time-independent pose conditions provided by the image-level ControlNet model. Furthermore, as observed in the last row of Fig. 9, the complete model with PTM effectively removes flickering artifacts caused by the yellow light, which is not present in the source image.

Quantitative Results Table 2 quantitatively demonstrates the effectiveness of each proposed method. The first two rows demonstrate that using the APPA layer to align appearance and pose significantly improves both image and video-level metrics, supporting the notable texture quality enhancement as evidenced in Fig. 3. Additionally, the temporal ControlNet and PTM have enhanced the temporal consistency of the generated videos, contributing to quantitative performance improvements, as shown in the last two rows of Table 2.

6 Conclusion

We present TCAN, a novel human image animation diffusion model with temporally consistent pose guidance. Our approach adapts the latent space between appearance and poses for effective utilization of the pre-trained pose ControlNet and alleviates errors from the off-the-shelf pose extractor by incorporating the temporal module into ControlNet. Moreover, during inference, our temperature map derived from the input pose sequence promises a consistent background of the output video. Experimental results on both challenging TikTok dataset and an extension to the unseen domain using Bizarre pose dataset promise the superiority and generalizability of our method in animating human images. Even so, the potential misuse of such videos as deep fakes highlights the ongoing need for research into verification models that can detect traces of generative models.

Acknowledgments This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)) and the National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0445). Finally, we thank all researchers at NAVER WEBTOON Corp.

References

1. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI. vol. 1, p. 2 (2019)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
5. Cao, Z., Simon, T., Wei, S.E., 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)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chen, S., Zwicker, M.: Transfer learning for pose estimation of illustrated characters. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 793–802 (2022)
8. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
9. Cheong, S.Y., Mustafa, A., Gilbert, A.: Upgpt: Universal diffusion model for person image generation, editing and pose transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4173–4182 (2023)
10. Cheong, S.Y., Mustafa, A., Gilbert, A.: Visconet: Bridging and harmonizing visual and textual conditioning for controlnet. arXiv preprint arXiv:2312.03154 (2023)
11. Güler, R.A., Neverova, N., 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)
12. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8153–8163 (2024)
16. Jafarian, Y., Park, H.S.: 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. pp. 12753–12762 (2021)
17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
18. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
19. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
20. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
21. Rogozhnikov, A.: Einops: Clear and reliable tensor manipulations with einstein-like notation. In: International Conference on Learning Representations (2021)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., 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)
23. 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)
24. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
25. 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)
26. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
27. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
28. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv e-prints pp. arXiv–2307 (2023)
29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
30. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)

31. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., 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 (2023)
32. Yang, Z., Zeng, A., Yuan, C., 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 (2023)
33. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
34. Zhang, L., Rao, A., 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)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
36. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
37. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., 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)