TPC: Test-time Procrustes Calibration for Diffusion-based Human Image Animation

Sunjae Yoon **Gwanhyeong Koo** **Younghwan Lee** **Chang D. Yoo***
Korea Advanced Institute of Science and Technology (KAIST)
{sunjae.yoon,cd_yoo}@kaist.ac.kr

Abstract

Human image animation aims to generate a human motion video from the inputs of a reference human image and a target motion video. Current diffusion-based image animation systems exhibit high precision in transferring human identity into targeted motion, yet they still exhibit irregular quality in their outputs. Their optimal precision is achieved only when the physical compositions (i.e., scale and rotation) of the human shapes in the reference image and target pose frame are aligned. In the absence of such alignment, there is a noticeable decline in fidelity and consistency. Especially, in real-world environments, this compositional misalignment commonly occurs, posing significant challenges to the practical usage of current systems. To this end, we propose Test-time Procrustes Calibration (TPC), which enhances the robustness of diffusion-based image animation systems by maintaining optimal performance even when faced with compositional misalignment, effectively addressing real-world scenarios. The TPC provides a calibrated reference image for the diffusion model, enhancing its capability to understand the correspondence between human shapes in the reference and target images. Our method is simple and can be applied to any diffusion-based image animation system in a model-agnostic manner, improving the effectiveness at test time without additional training.

1 Introduction

Denoising diffusion models [5, 27, 26, 10] have transformed the generative landscape of artificial intelligence, leading to groundbreaking achievements [17, 20, 41, 39] in image, speech, and video generation. We explore the application of diffusion model in the specific context of image-to-video generation, focusing on the task of human image animation. The technology of image animation holds great promise, enabling immersive and interactive experiences in entertainment, virtual reality, and digital communication. The human image animation systems [33, 31, 15, 13] are designed to work with referential human image and target motion video, where they transfer the human identity into the target motion, ensuring seamless and unobtrusive integration. This process requires understanding the correspondence between human shapes in the reference image and the target pose frame.

Recent advancements [33, 31] of human image animation systems have demonstrated notable precision in adapting human identity into target motion. Despite advancements, these systems continue to suffer from irregular quality of image animation when the compositions (*i.e.*, scale and rotation) of human shapes are not aligned between reference image and target motion. To be specific, Figure 1 (a) presents exploratory experiments on the compositional misalignment of human shapes between the reference and target. As shown in the left experiments, for a given target pose, adjusting the composition of the same human in the reference image (*i.e.*, via scaling or rotating) results in inconsistent and low-fidelity output images, especially in terms of clothes and faces of the human.

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3768

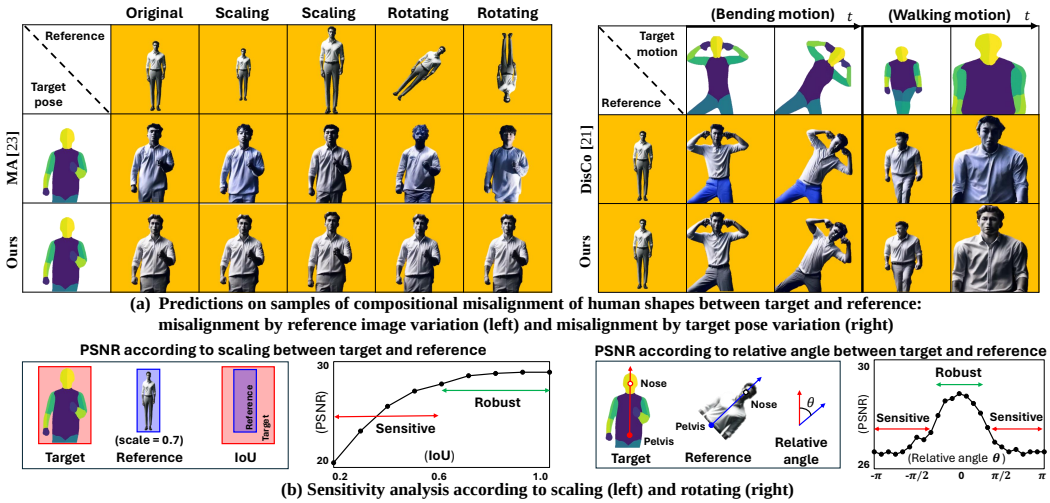


Figure 1: Illustration of compositional misalignment: (a) Results of current human image animation models [33, 31] on samples in compositional misalignment of human shapes between reference and target. (b) Sensitivity analysis according to variation of compositional misalignment by scaling and rotating (MA: MagicAnimate). Best viewed with zoom.

Furthermore, in the right experiments, when providing motion sequences that display various dynamic movements for a given reference image, the human image animation outputs consistently show low fidelity, especially evident in target poses (*e.g.*, bending or approaching close to viewpoint) that cause significant differences in the composition of the human shape. To quantitatively investigate these, Figure 1 (b) presents a sensitivity analysis evaluating how current image animation systems [33, 31] respond to varying degrees of the compositional misalignment of human shape. The left shows the fidelity (*i.e.*, PSNR) of the resulting human according to the relative scale difference of input human shapes between the target and reference. We employ the Intersection of Union (IoU) of bounding boxes of the shapes for relative scale. The right shows fidelity according to the variations of relative angle θ [2] of human shapes between the target and reference. Current systems demonstrate significant vulnerability to the variations of the compositions, indicating that output fidelity forms a robust region only in the areas with compositionally aligned conditions (*i.e.*, $-\frac{\pi}{6} < \theta < \frac{\pi}{6}$, IoU > 0.7).

In fact, diffusion-based systems are inevitably susceptible to this compositional misalignment. The diffusion model uses a reference human image as a condition to generate controlled output from noise based on the target pose. Here, the conditioning is performed through cross-attention based on a visual similarity between patch-wise features of the target pose frame and reference image. To be specific, Figure 2 shows the attention map about a single patch of the target frame (*i.e.*, shoulder denoted by blue point) from the reference image during the denoising process. The upper section displays a sample where the human shapes are relatively aligned between the target and the reference, while the lower section shows a case where this alignment is not present. Initially, the attention maps were blurry in both cases. However, as denoising continued, it became clear that samples with aligned human shapes correctly established a correspondence between the target frame and the reference image. (*i.e.*, Attention to the shoulder on the target frame is focused on the shoulder on the reference image.) However, when the shape is misaligned, attention to the target frame's shoulder is incorrectly focused on unrelated areas in reference, even up to the

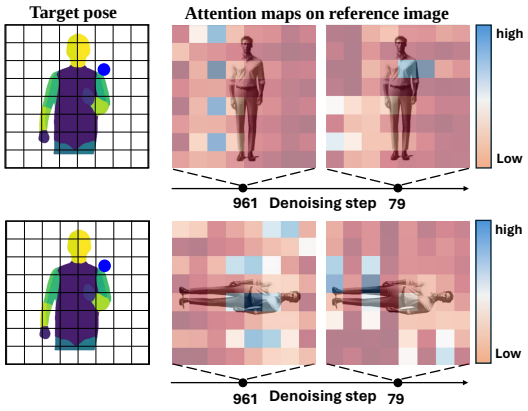
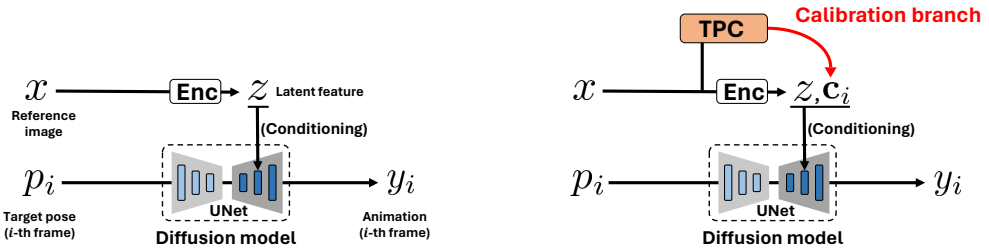


Figure 2: Attention maps on the reference image corresponding to the target human shape (*e.g.*, shoulder at blue point) according to denoising.

[2]We define the axis of the torso of a human from the pelvis to nose using estimated human body key points [4] and measure the relative angle of axes between target and reference.



(a) Diffusion-based human image animation

(b) Diffusion-based human image animation with Test-time Procrustes Calibration (TPC)

Figure 3: Illustration of (a) current diffusion-based human image animation systems and (b) Test-time Procrustes Calibration (TPC) on top of these systems. The TPC can be applied to diffusion-based models in a model-agnostic manner, enhancing the fidelity and consistency of the output video.

last stages of denoising. Consequently, the compositional misalignment between the target and the reference image hinders accurate attention correspondence throughout the denoising process.

To this end, we propose a diffusion guidance referred to as Test-time Procrustes Calibration (TPC). As shown in Figure 3 (a), the existing diffusion-based human image animation system takes inputs of reference image x and i-th target pose frame p_i, where it generates i-th output animation frame y_i. Here, the x is encoded into the latent feature z, which serves as a conditioning input for the denoising diffusion model (*i.e.*, UNet). As depicted in Figure 3 (b), our proposed TPC incorporates an auxiliary branch into this diffusion conditioning process. This branch is defined as a calibration branch that guides denoising UNet to properly capture visual correspondence between target and reference. To be specific, the TPC provides a calibrated reference image latent $\mathbf{c}_i$ that aligns with the human shape in the target pose p_i based on statistical shape analysis referred to as Procrustes analysis [7]. Figure 4 offers a qualitative understanding of the influence of this $\mathbf{c}_i$. Conceptually, in Figure 4 (a), when the humans in reference image x and the target pose sequence $p = \{p_{i-1}, p_i, p_{i+1}, p_{i+2}\}$ are projected onto the ideal 2D shape-style space, they show distinct locations in terms of style axis, where the system aims to generate image animation frame $y = \{y_{i-1}, y_i, y_{i+1}, y_{i+2}\}$ with the style of x and the shape of p. However, in cases (*e.g.*, x and p_{i+2}) where significant gaps exist along the shape axis (*i.e.*, compositional misalignment), the current diffusion model struggles to preserve the original style. This leads to low-fidelity outputs (*e.g.*, y_{i+2}) and results in temporal inconsistencies due to unstable fidelity across frames. Thus, as illustrated in Figure 4 (b), we bridge this gap in shapes between reference and target by providing correspondence guidance latent $\mathbf{c}$ by our designed TPC. With this guidance condition, diffusion-based animation systems achieve robustness to fidelity variations and maintain temporal consistency among frames. The TPC is simple and works in a model-agnostic manner without additional training and validates its effectiveness on human image animation benchmarks (*i.e.*, TikTok[14], TED-talks [25]) and even in unseen domain data of real environment scenarios.

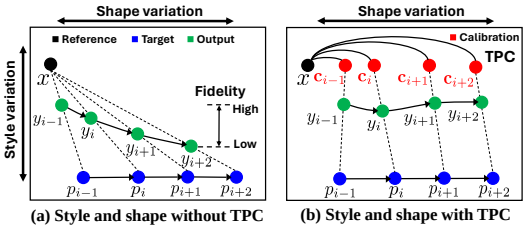


Figure 4: Conceptual illustration of the effectiveness of TPC in terms of style and shape variation.

2 Related Work

2.1 Diffusion-based Human Image Animation

Human image animation aims to provide a video about an animated version of an input human image. As a human-centric application of image-to-video technology, the human image animation systems [25, 24, 45] have previously been developed based on generative adversarial networks. The recent emergence of denoising diffusion models [27, 26, 10] has presented a new paradigm for generative models, where image animation has also faced fundamental innovations. The diffusion-based framework generates an image animation from noise using pre-trained denoising capabilities based on input human image and target motion video, where the motion video can be extracted

https://doi.org/10.52202/079017-3768

from various pose estimation models [4, 8]. Early work of diffusion-based image animation was made in DreamPose [15], which leveraged the pre-trained text-to-image diffusion model (*e.g.*, Stable Diffusion [23]) by conditioning on human image embeddings instead of text, rendering videos of humans in various outfits performing simple walking motions. AnimateAnyone [13] introduces a UNet-style reference image encoder that enhances the layered conditioning of the reference image following the encoding-decoding process of UNet. Furthermore, ControlNet [43] has been a popular choice with the diffusion model by offering more controlled guidance about target motion. To enhance background fidelity, DisCo [31] segments the background of the reference image and integrates it into ControlNet, along with the target motion. For the temporal consistency of output video, MagicAnimate [33] introduces temporal attention by inflating the original 2D UNet to 3D temporal UNet. However, current systems still suffer from quality irregularity issues when the human shapes are not aligned between the reference image and the target motion. Such misalignments frequently occur in real-world scenarios, prompting our proposed Procrustes calibration to address this challenge.

3 Preliminary

3.1 Procrustes Analysis

Procrustes[3] analysis (PA) [6, 7] is a statistical shape analysis technique used to compare the shapes of objects [11, 19]. The PA involves finding the best alignment between two shapes by scaling, translating, and rotating one shape to match the other as closely as possible. To formulate the process of PA, we are given two sets of n points as $X = \{x_1, \cdots, x_n\} \in \mathbb{R}^{n \times d}$ and $Y = \{y_1, \cdots, y_n\} \in \mathbb{R}^{n \times d}$, where d is the dimension of the point. We perform Procrustes transformation composed of scaling factor $s \in \mathbb{R}^1$, rotation matrix $r \in \mathbb{R}^{d \times d}$, and translation vector $t \in \mathbb{R}^{1 \times d}$ as given below:

$$\hat{Y} = s \cdot Xr + t, \tag{1}$$

where the vector t is added with broadcasting to all n points. The objective is to find the optimal transformation parameters s, r, t to minimize the sum of squared differences between $\hat{Y}$ and the Y.

$$\underset{s,r,t}{\operatorname{argmin}} ||\hat{Y} - Y||_F, \tag{2}$$

where $|| \cdot ||_F$ is Frobenius norm. The optimal value r^* is typically obtained via singular value decomposition and the s^* and t^* are calculated after the optimal rotation r^* is found. Here, we apply this PA to align the human shapes of a reference image and target motion frame.

4 Method

Given a human reference image R and a target pose sequence $P = [P_1, \cdots, P_L]$, a human image animation system generates image animation video $V = [V_1, \cdots, V_L]$ which follows the pose sequence by the human in the reference image, where L is the number of frames. Figure 5 shows the application of Test-time Procrustes Calibration (TPC) into the general diffusion-based human image animation system. The TPC aims to improve the quality of resulting animation by consistently ensuring the compositional alignment of human shapes between the reference image and the target poses. At each denoising step t, TPC takes an input reference image R and target poses P, and produces calibrated image C and its embedded latent $\mathbf{c}$ for conditioning in diffusion denoising. The calibrated latent $\mathbf{c}$ guides the conditioning module (*i.e.*, cross-attention) in denoising UNet with precise correspondence about human shapes between the reference and the target poses. To perform this, the TPC follows the sequential process of $R \rightarrow C \rightarrow \mathbf{c}$, and it comprises two main modules: (1) Procrustes Warping (Sec 4.1) and (2) Iterative Propagation (Sec 4.2). Using the input reference image R and target poses P, Procrustes warping produces a calibrated reference image C optimized to align the human shape with each target pose. After embedding this calibrated image C into calibrated latent feature $\mathbf{c}$, Iterative Propagation iteratively refines the $\mathbf{c}$ during denoising step to enhance temporal consistency among the features by applying our designed feature propagation method. This calibrated latent $\mathbf{c}$ is finally given to diffusion model's conditioning module (*i.e.*, cross-attention) as a condition by concatenating with the original latent feature $\mathbf{r}$ of the reference image.

[3]The name "Procrustes" comes from Greek mythology, where Procrustes was a bandit who would stretch or cut people to fit his bed, reflecting the idea of adjusting data to fit a common form.

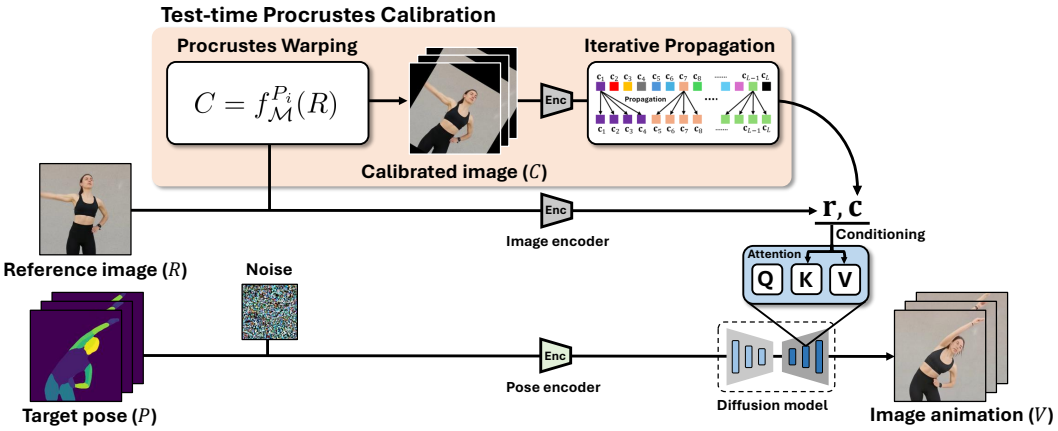
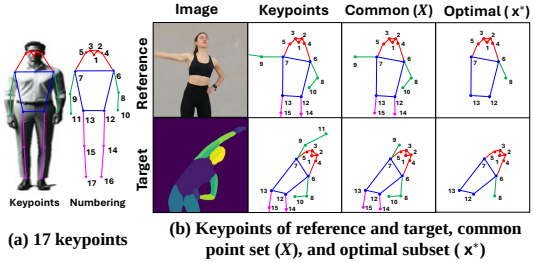


Figure 5: Illustration of Test-time Procrustes Calibration (TPC) on diffusion-based human image animation. TPC provides calibrated latent feature $\mathbf{c}$ to enhance shape correspondence between the reference image and target poses. Procrustes Warping aligns the reference image with the target pose shape, while Iterative Propagation improves temporal consistency among calibrated features.

4.1 Procrustes Warping

Procrustes Warping (PW) aims to align human shapes between reference and target pose. Thus, the PW takes a reference image R as input and produces a calibrated reference image $C = [C_1, \cdots, C_L]$ to the target human pose $P = [P_1, \cdots, P_L]$. To construct C, we first extract keypoint sets from both the reference and target humans. Then, we apply Procrustes analysis [4] between the two sets, determining transformation parameters: scaling, rotation, and translation. Using these parameters, we transform the reference image to align it with the target. To be specific, as shown in Figure 6 (a), we define 17 keypoints [5] in 2-dimensional space for human body and face using keypoint extractor (*e.g.*, OpenPose [4]). Figure 6 (b) illustrates that we obtain these keypoints from the reference image R and i-th target pose frame P_i. From these, we filter out commonly visible N points, defining $X = \{x_1, \cdots, x_N\}$ as reference set and $Y = \{y_1, \cdots, y_N\}$ as target set. Following Procrustes analysis (*i.e.*, Eq. (1,2)), we obtain optimal transformation parameters $s^* \in \mathbb{R}^1, r^* \in \mathbb{R}^{2\times2}, t^* \in \mathbb{R}^{1\times2}$ and warp all pixels $[u, v] \in \mathbb{R}^{n\times2}$ (n is the number of pixels) in the reference image by mapping as below:



(a) 17 keypoints

(b) Keypoints of reference and target, common point set (X), and optimal subset ($\mathbf{x}^*$)

Figure 6: Illustration of (a) pre-defined 17 keypoints and (b) keypoints in reference and target human, their common point set, and optimal set.

$$\mathcal{M} : [u, v] \rightarrow s^*[u, v] \cdot r^* + t^*. \tag{3}$$

Therefore, we obtain the i-th calibrated image $C_i = f_{\mathcal{M}}^{P_i}(R)$ using Procrustes warping $f_{\mathcal{M}}^{P_i}$, which aligns R with human shape in P_i using mapping $\mathcal{M}$. However, considering all N common visible points for warping is unreasonable since the reference and target shapes cannot perfectly overlap due to differing poses (*e.g.*, the arms in Figure 6 (b) cannot overlap). To address this, we apply Procrustes warping to subset $\mathbf{x} \subset X$ of the common points and select the most effective subset $\mathbf{x}^*$ based on our defined alignment score function h as below (subset $\mathbf{y}$ is also defined corresponding to the $\mathbf{x}$):

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmax}}\ h(P_i, C_i^{\mathbf{x}}), \tag{4}$$

where $C_i^{\mathbf{x}}$ denotes calibrated image using keypoints of subset $\mathbf{x}$. The h is alignment score function that computes pixel-wise IoU[6] of human shapes between pose P_i and calibrated image $C_i^{\mathbf{x}}$. Thus, the final i-th calibrated image is defined as $C_i = C_i^{\mathbf{x}^*}$ with optimal subset $\mathbf{x}^*$.

[4]Reason for choosing the Procrustes transform: Transformations are largely categorized into (1) shape-preserving (*e.g.*, linear, Procrustes) and (2) shape-distorting (*e.g.*, affine). To put the conclusion first, the Procrustes transform was the most effective. Shape-distorting methods were less effective than shape-preserving due to the loss of visual information by distortion. Detailed analysis is available in Table 2 and Figure 11.

[5]We place more points on informative regions of human identity (face: 5, torso: 4, arms: 4, and legs: 4).

[6]SAM [16] is used for segmenting each human shape to measure pixel-wise IoU.

https://doi.org/10.52202/079017-3768

4.2 Iterative Propagation

Conceptually, as shown in Figure 5, our proposed TPC aims to mitigate compositional misalignment by feeding calibrated image latent features $\mathbf{c} = [\mathbf{c}_1, \cdots, \mathbf{c}_L]$ into the conditioning module (*i.e.*, cross-attention) of denoising diffusion along with the original reference latent feature $\mathbf{r}$.[7] Thus, the $\mathbf{c}$ is designed to bridge the correspondence of the human shapes between the reference image and target pose. However, the pose variation within the target pose frames affects the degree of calibration with the reference image, which reduces temporal consistency in the output. To address this, we introduce Iterative Propagation (IP) to enhance consistency among calibrated latent features. Figure 7 illustrates the IP process during the diffusion denoising. The IP forms M groups of sequential features in calibrated features $\mathbf{c}$, randomly selects a feature within each group, and updates all features in the group with the selected one. This method enhances temporal consistency among calibrated latent features while maintaining compositional alignment with the target pose due to the continuity of target pose variation. The random selection of features ensures all features have an equal chance of being chosen during the denoising process.

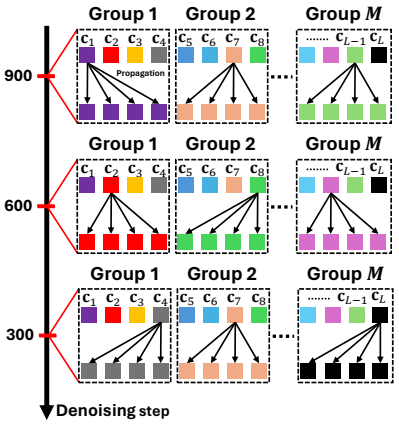


Figure 7: Illustration of iterative propagation on calibrated latent features. It shows M groups on L frame calibrated features and updates features in each group with randomly selected ones during the denoising process.

4.3 Plug-and-Play Test-time Procrustes Calibration

We integrate calibrated latent feature $\mathbf{c}$ into human image animation systems by applying it into a conditioning module (*i.e.*, cross-attention) of video diffusion UNet. For i-th frame m patch-wise image feature $\mathbf{p}_i \in \mathbb{R}^{m \times d}$, reference feature $\mathbf{r} \in \mathbb{R}^{m \times d}$, and calibrated latent feature $\mathbf{c}_i \in \mathbb{R}^{m \times d}$ the cross attention is formulated as $Q \leftarrow \text{Softmax}(QK^T/d)V$, where it satisfies $Q = \mathbf{p}_i$ and $K = V = [\mathbf{r}, \mathbf{c}_i] \in \mathbb{R}^{2m \times d}$ is concatenated latent condition.

5 Experiments

5.1 Experimental Settings

Implementation Details. SAM [16] is used for screening out the background in calibrated images. VQ-VAE [30] is used for encoding images of the video. The number of groups in iterative propagation is chosen as $M = 30$ under ablation studies in Table 2. The average number of video frames is about 120. We use Stable Diffusion 1.5 [23] for all baselines on 4 NVIDIA A100 GPUs. We follow the same pose encoders and image encoders of baseline models.

Data and Baselines. We validate human image animation on two popular benchmarks (*i.e.*, TikTok [14], TED-talks [25]) about a test split. Due to no validation splits, we provide valid sets matching the test set sizes for the ablation study. We further collected 114 samples [8] from TikTok and TED-talks as another test split. These samples contain compositional misalignment about rotation and scaling between human shapes in reference images and motion videos. The criteria for misalignment include relative angle and scaling differences, with samples having a relative angle $\theta > \pi/6$ or relative scale IoU < 0.7. As shown in Figure 1 (a), the relative angle measures an angle between the straight lines from the pelvis to the nose in humans of reference and target poses using keypoints estimator [4]. The relative scaling measures the IoU (Intersection of Union) between bounding boxes of humans in the reference and the target. Procrustes Calibration is validated on recent diffusion-based human image animation models including MagicAnimate [33], DisCo [31], AnimateAnyone[9] [13], DreamPose [15] on their public codes and papers.

[7] The dimensions are $\mathbf{c} \in \mathbb{R}^{L \times m \times d}$ and $\mathbf{r} \in \mathbb{R}^{m \times d}$, by d-dimensional patch-wise image encoder $\text{Enc}(\cdot)$, where m is the number of image patches and the L is the number of frames.

[8] Supplementary provides all the video links of these samples.

[9] As the code is not available, we use the work at: https://github.com/MooreThreads/Moore-AnimateAnyone

Table 1: Quantitative evaluations of Test-time Procrustes Calibration (TPC) with recent diffusion-based human image animation models. A-Anyone: AnimateAnyone, M-Animate: MagicAnimate. It is reported in a format of (original test set / compositional misalignment test set).

Method	Image					Video		Human
	$L1\downarrow_{\times \text{E-04}}$	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓	
TikTok [14]								
DreamPose	7.22/9.64	27.31/25.17	0.532/0.481	0.449/0.529	55.4/87.2	61.1/93.1	568/738	0.04
DreamPose + TPC	5.15/5.31	28.47/28.01	0.620/0.613	0.406/0.412	48.4/49.3	54.7/56.3	426/441	0.96
DisCo	4.09/5.23	28.43/24.97	0.641/0.512	0.312/0.492	37.1/71.4	58.3/82.1	339/522	0.28
DisCo + TPC	3.49/3.82	28.92/28.87	0.689/0.673	0.283/0.287	34.3/36.2	51.2/52.4	281/297	0.72
A-Anyone	3.77/4.82	29.06/26.52	0.670/0.584	0.289/0.392	32.9/65.2	54.2/59.1	296/442	0.24
A-Anyone + TPC	3.32/3.53	29.27/29.01	0.705/0.688	0.264/0.273	31.3/32.6	48.7/49.3	254/269	0.76
M-Animate	3.17/4.36	29.11/27.82	0.717/0.641	0.241/0.321	31.8/49.2	22.4/52.3	182/362	0.34
M-Animate + TPC	**2.98**/3.19	**29.43**/29.21	**0.753**/0.731	**0.232**/0.249	**29.2**/30.4	**21.0**/21.9	**158**/164	0.66
TED-talks [25]								
DreamPose	6.65/7.41	27.63/26.11	0.559/0.482	0.421/0.521	63.4/86.2	43.6/64.2	411/532	0.24
DreamPose + TPC	6.17/6.31	28.11/28.03	0.593/0.589	0.393/0.404	53.2/55.2	38.2/38.9	369/372	0.76
DisCo	3.52/6.91	28.51/26.71	0.661/0.511	0.309/0.451	34.5/71.1	28.2/52.6	332/471	0.24
DisCo + TPC	3.18/3.23	28.93/28.87	0.704/0.692	0.283/0.294	31.8/32.6	25.1/25.8	298/304	0.76
A-Anyone	3.12/6.61	28.93/27.07	0.712/0.613	0.267/0.361	29.3/54.2	20.3/39.5	192/372	0.24
A-Anyone + TPC	2.81/2.91	29.31/29.25	0.753/0.742	0.254/0.269	27.8/28.6	**18.8**/19.5	173/178	0.76
M-Animate	2.92/4.31	29.17/27.47	0.734/0.661	0.239/0.312	25.7/47.1	20.2/39.1	136/331	0.24
M-Animate + TPC	**2.77**/2.87	**29.52**/29.41	**0.782**/0.764	**0.233**/0.249	**24.3**/25.6	19.2/20.4	**128**/137	0.76

5.2 Evaluation Metrics

We evaluate videos in terms of single-frame quality and video quality. For the single-frame, we measure Peak Signal-to-Noise Ratio (PSNR) [12], Structural Similarity Index Measure (SSIM) [32], Learned Perceptual Image Patch Similarity (LPIPS) [44], FID (Fréchet Image Distance) [9], and L1 error between output and ground-truth images. For the video quality, we measure Fréchet Video Distance (FVD) [29] and FID-VID [1]. All automatic metrics are averaged over 10 runs with different seeds. We also analyze human preferences for results from the baselines with and without our TPC.

5.3 Experimental Results

Quantitative Comparisons. Table 1 presents evaluations of image animation on two benchmark datasets (*i.e.*, TikTok, TED-talks) using recent diffusion image animation baselines (MagicAnimate, DisCo, AnimateAnyone, DreamPose) with TPC across three assessments (*i.e.*, image, video, human). Evaluations are conducted on two test splits: the original set and the compositional misalignment set. Initially, all baselines use a single reference image latent for conditioning all poses. After integrating TPC, they use calibrated latents corresponding to each pose. Consistent improvements in image and video quality are observed in both test sets. All baselines struggle with the compositional misalignment set, but when integrated with TPC, they achieve quality close to the original test set. This demonstrates that morphological similarity affects the current diffusion model's ability to map human shapes from reference to target.

Qualitative Comparisons. To validate our proposed TPC, we applied it to four recent baselines and compared the original results. We used the same types of human pose inputs (e.g., OpenPose, DensePose) for each baseline to prepare target motion videos. Figure 8 shows predictions on four different samples exhibiting compositional misalignment. The top left results show predictions on temporal misalignment between the reference and target due to the target's bending motion. MagicAnimate's fidelity diminishes with increased bending (red box in the last frame), whereas the model with TPC maintains high fidelity. The top right results display predictions on temporal misalignment due to a walking motion towards the front, with TPC similarly enhancing DisCo's performance. The bottom left results show consistent misalignment across all frames. DreamPose struggles with low fidelity, causing unwanted stripes on the pants (red box), which are clearly removed with TPC. The results in the bottom right exhibit consistent misalignment due to a scale difference. In AnimateAnyone, the reference human's yellow pants were incorrectly mapped onto the target human's arms, making them appear yellow. However, the TPC completely mitigates this incorrect mapping. This occurs because the calibrated image filters out the pants and enhances the correspondence of each body part. (Please, refer to the calibrated images also in the Appendix.)

https://doi.org/10.52202/079017-3768

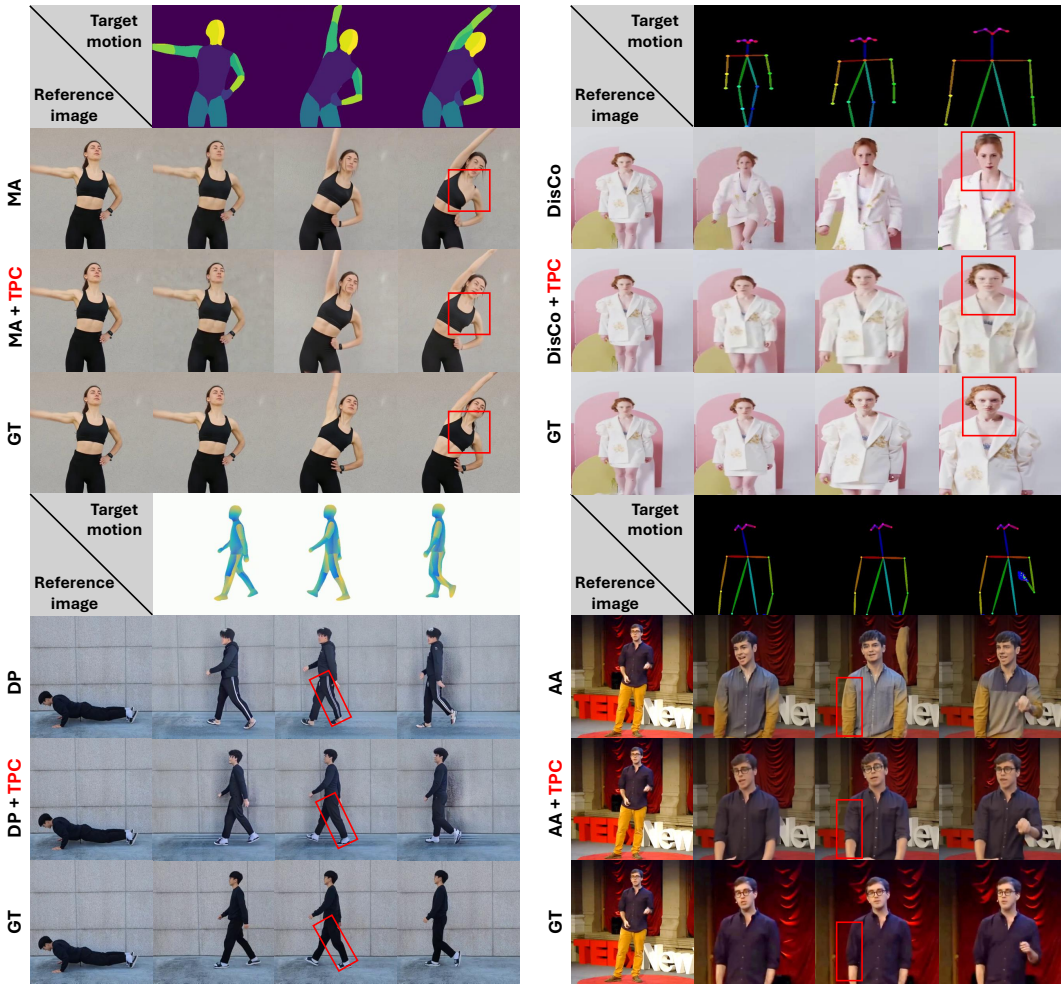


Figure 8: Qualitative results about applying TPC on diffusion-based human image animation systems (*i.e.*, MA: MagicAnimate, DisCo, AA: AnimateAnyone, DP: DreamPose) on cases about compositional misalignment of human shape between a reference image and target motion: Temporal misalignment by motion affecting factor rotation (top left) and scale (top right). Consistent misalignment affecting rotation (bottom left) and scale (bottom right). Calibrated images and predictions of compositional aligned samples are available in Appendix. Please see also video in the supplementary.

Figure 10 shows the results on reference images from an unseen domain, generated by the T2I model [21], applied in a compositional misalignment scenario. The left displays a temporal misalignment sample due to bending motions, with improved fidelity in both baselines, especially in DisCo. The right side displays a consistent misalignment sample, where the baselines show low precision in transferring identity when the reference is flipped upside down. However, with the integration of TPC, their performance is significantly improved.

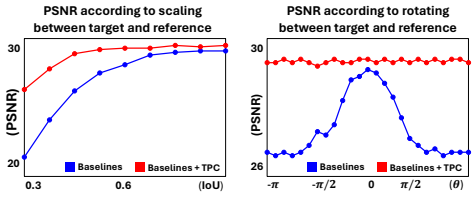


Figure 9: Robustness analysis of variations in compositional misalignment. Measurements of scale and rotation are in Figure 1 (b).

Robustness analysis To assess the robustness of baselines with TPC, we measured fidelity (PSNR) by varying the scale and rotation of the human reference, as shown in Figure 9. In the case of scaling, the baselines show a drop in performance when the size difference between the reference and target human shapes falls below an IoU of 0.6. However, TPC maintains optimal performance even down to an IoU of 0.4. Notably, TPC enhances baseline robustness to all rotational variations by ensuring the calibrated image aligns well with shape differences caused by rotation.

https://doi.org/10.52202/079017-3768

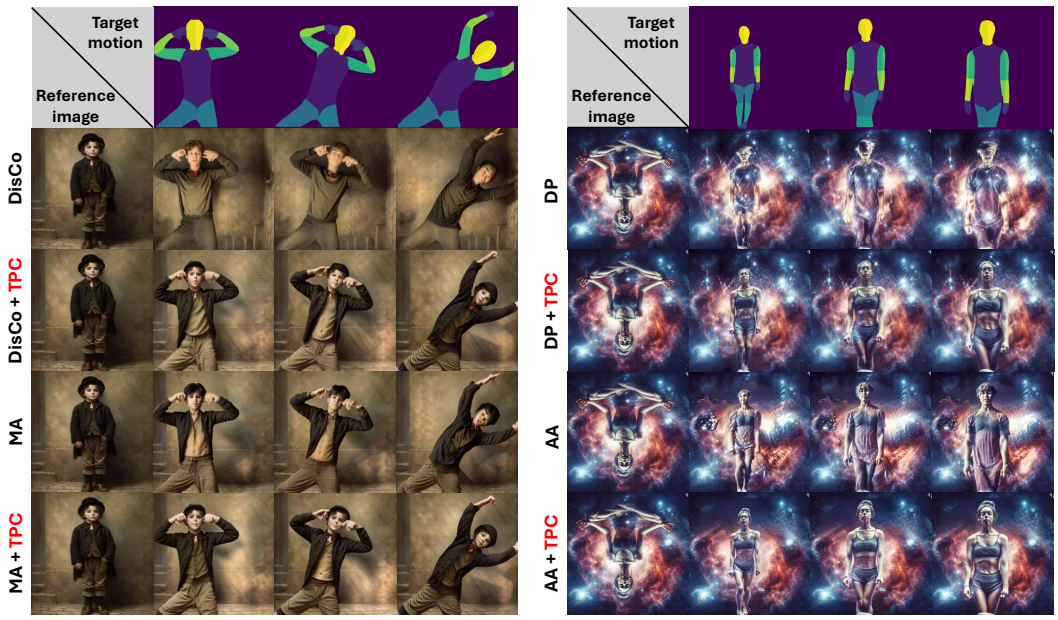


Figure 10: Qualitative results about unseen domain reference image. TPC enhances human image animation models' robustness on samples about compositional misalignment in the unseen domain.

Table 2: Ablation studies on transformation methods for reference image calibration and iterative propagation (IP) on TED-talks and TikTok. (validation splits, average score compositional alignment/misalignment).

Method	Foreground		Background	
	SSIM↑	FVD↓	SSIM↑	FVD↓
Linear	0.702	191	0.751	170
Affine	0.704	193	0.754	171
Procrustes	**0.734**	**162**	**0.782**	**142**
w/o IP	0.709	184	0.728	162
w/ IP (M=20)	0.731	164	0.776	145
w/ IP (M=30)	**0.734**	**162**	**0.782**	**142**
w/ IP (M=40)	0.728	165	0.777	145

Table 3: Inference time of baselines with TPC. DC: DisCo.

Method	sec/frame
DP	18.3
DP + TPC	18.9
DC	5.8
DC + TPC	6.3
AA	2.3
AA + TPC	2.6
MA	1.4
MA + TPC	1.6

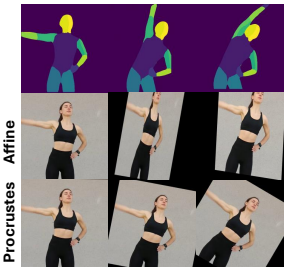


Figure 11: Comparisons of calibrated images using an affine transform and Procrustes transform.

Ablation Study. Table 2 presents ablation studies on transformation methods and iterative propagation. Transformations are categorized as shape-preserving (*e.g.*, Linear, Procrustes) and shape-distorting (*e.g.*, Affine). Linear transformation performs scaling and rotation without translation, based on the object's bounding box coordinates. The Procrustes method is the most effective, transforming by selecting the optimal keypoint. The affine transform is less effective due to shearing, which can cause distortion and information loss by moving the subject out of the frame, as shown in Figure 11. The second section presents an ablation study on iterative propagation, showing significant enhancements in foreground and background, with M=30 being the most effective. Table 3 provides inference time on baselines with TPC. Our method requires only minimal additional time because it processes all frames batch-wise, generating the calibrated image in a single iteration.

5.4 Broader Impacts and Ethic Statements

Visual generative models present a range of ethical dilemmas, including the creation of unauthorized counterfeit content, potential privacy breaches, and challenges related to fairness. Due to our reliance on the architecture of these models, our work inherently adopts these ethical vulnerabilities. Addressing these concerns is imperative and requires the establishment of comprehensive regulations and technical countermeasures. We are exploring advanced measures such as learning-based digital forensics and digital watermarking to ethically navigate the complexities of visual generative models.

https://doi.org/10.52202/079017-3768

5.5 Limitation and Future work

We provide an overview of the various limitations and potential development directions identified during this study. Human image animation systems transfer reference images to target poses. However, some frames in a target pose video may lack proper specification, leading to flicker or low fidelity in model predictions. Thus, integrating high-quality pose estimation is crucial. Another limitation is that significant differences in body shape between the reference and target poses result in awkward transfers, such as transferring a skinny person to a fat target pose. To achieve natural results, a system or module that aligns body shapes is required. For the future extension of human image animation system, it is essential to ensure robust operations under multiple individuals. The system should ensure consistent identity transfer across multiple individuals and accurately adapt to those appearing at specific times. To achieve this, video technologies about recognition [22, 46] and perception [37, 40] can be further incorporated. Currently, conditioning relies solely on images, but it is anticipated that various other modalities, such as text and audio, could also be incorporated. Especially, the integration of the text modality with emerging large language models [3, 28] is expected to drive innovative industrial advancements, with the introduction of video/image-based conversational systems [36, 42, 38] serving as a compelling utility. Lastly, integrating technologies focused on speed [18, 2], resource efficiency [34], and test-time calibration [35] is expected to improve human image animation systems' applicability into real-world environments.

6 Conclusion

Current diffusion-based human image animation systems are facing challenges on samples of compositional misalignment of human shapes between a reference image and target pose frames. To this end, this paper presents Test-time Procrustes Calibration (TPC) which improves the robustness of image animation models on the compositional misalignment samples in a model-agnostic manner. Extensive experiments demonstrate the effectiveness of TPC.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2021-II211381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220184, 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics)

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019.

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[6] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[7] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[11] Ji Woo Hong, Sunjae Yoon, Junyeong Kim, and Chang D Yoo. Joint path alignment framework for 3d human pose and shape estimation from video. *IEEE Access*, 11:43267–43275, 2023.

[12] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[13] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[14] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021.

[15] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[17] Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D Yoo. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. *arXiv preprint arXiv:2407.17850*, 2024.

[18] Gwanhyeong Koo, Sunjae Yoon, and Chang D Yoo. Wavelet-guided acceleration of text inversion in diffusion-based image editing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4380–4384. IEEE, 2024.

[19] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[20] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[24] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

[25] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

 https://doi.org/10.52202/079017-3768

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[29] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[30] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[31] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*, 2023.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[33] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

[34] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023.

[35] Hee Suk Yoon, Joshua Tian Jin Tee, Eunseop Yoon, Sunjae Yoon, Gwangsu Kim, Yingzhen Li, and Chang D Yoo. Esd: Expected squared difference as a tuning-free trainable calibration measure. *arXiv preprint arXiv:2303.02472*, 2023.

[36] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Kang Zhang, Yu-Jung Heo, Du-Seong Chang, and Chang D Yoo. Bi-mdrg: Bridging image history in multimodal dialogue response generation. *arXiv preprint arXiv:2408.05926*, 2024.

[37] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pages 185–200. Springer, 2022.

[38] Sunjae Yoon, Dahyun Kim, Eunseop Yoon, Hee Yoon, Junyeong Kim, and Chang Yoo. Hear: Hearing enhanced audio response for video-grounded dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11911–11924, 2023.

[39] Sunjae Yoon, Gwanhyeong Koo, Ji Woo Hong, and Chang D Yoo. Dni: Dilutional noise initialization for diffusion video editing. *arXiv preprint arXiv:2409.13037*, 2024.

[40] Sunjae Yoon, Gwanhyeong Koo, Dahyun Kim, and Chang D Yoo. Scanet: Scene complexity aware network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13576–13586, 2023.

[41] Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D Yoo. Frag: Frequency adapting group for diffusion video editing. *arXiv preprint arXiv:2406.06044*, 2024.

[42] Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang Yoo. Information-theoretic text hallucination reduction for video-grounded dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4182–4193, 2022.

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[45] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.

[46] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017.

A Appendix

A.1 Algorithm for Iterative Propagtion

Algorithm 1 Iterative Propagation

Input: Calibrated L frame features $\mathbf{c} = [\mathbf{c}_1, \cdots, \mathbf{c}_L]$, Number of frame groups M
Output: Denoising t-step propagated features $\mathbf{c}^t = [\mathbf{c}_1^t, \cdots, \mathbf{c}_L^t]$
Initialize the number of frames in each group: $m = \lfloor \frac{L}{M} \rfloor$
for $t = T$ **to** 1 **do** ▷ T is total denoising step.
  for $i = 1$ **to** M **do**
    $n \leftarrow (i-1) \times m$
    Sample $r \sim \mathcal{U}\{1, m\}$ ▷ $\mathcal{U}\{1, m\}$ is a discrete uniform distribution between 1 and m.
    $[\mathbf{c}_{n+1}^t, \mathbf{c}_{n+2}^t, \cdots, \mathbf{c}_{n+m}^t] \leftarrow [\mathbf{c}_{n+r}, \mathbf{c}_{n+r}, \cdots, \mathbf{c}_{n+r}]$ ▷ Update list.
  end
  $\mathbf{c}^t \leftarrow [\mathbf{c}_1^t, \cdots, \mathbf{c}_m^t] \cup [\mathbf{c}_{m+1}^t, \cdots, \mathbf{c}_{2m}^t] \cup \cdots \cup [\mathbf{c}_{m(M-1)+1}^t, \cdots, \mathbf{c}_L^t]$ ▷ Concatenation
end

A.2 More qualitative results

Explanation

Figure 12 shows calibrated reference images corresponding to the samples in Figure 8.

Figure 13 shows results on identical motion with different reference images.

Figure 14 and 15 show results on identical references with different motion videos in the TikTok dataset.

Figure 16 shows results on multiple humans in reference and target motion.

As in Figure 1, here we visualize the target pose in a format of DensePose for the visibility of target motions. The video samples in Figure 8 are publicly available from:

- https://www.pexels.com/video/a-woman-stretching-5510095/

- https://www.pexels.com/video/modeling-wedding-dresses-7305163/

- https://www.youtube.com/watch?v=8S0FDjFBj8o

- https://www.tiktok.com/@dbstjswo505/video/7371773469899476231

https://doi.org/10.52202/079017-3768

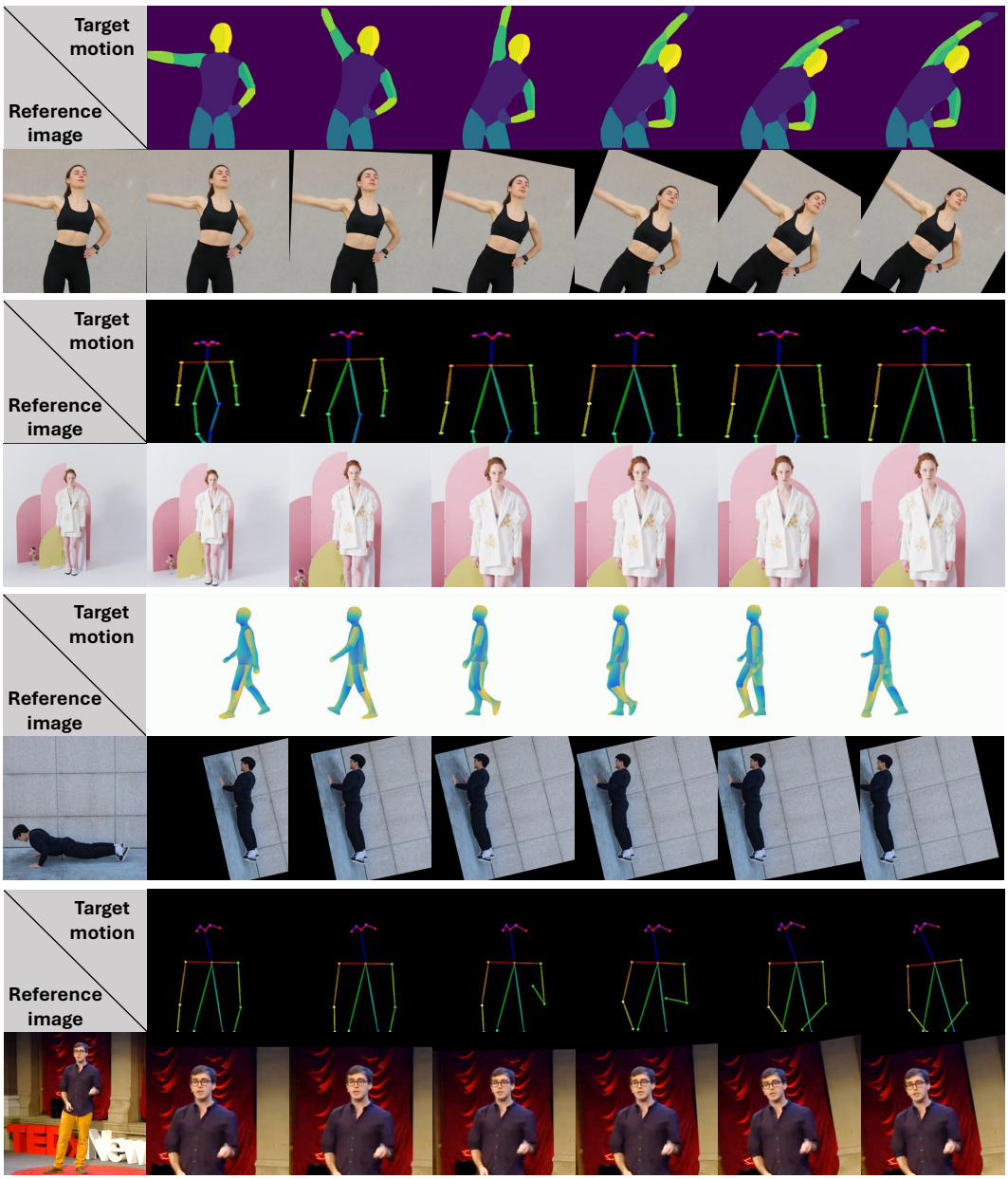


Figure 12: Qualitative results about calibrated images on the samples in Figure 8.

Figure 13: Qualitative results about applying TPC on recent diffusion-based human image animation systems (*i.e.*, DreamPose (DP), MagicAnimate (MA), DisCo, AnimateAnyone(AA)) on identical motions with different reference images. The images are obtained from T2I image generation model.

https://doi.org/10.52202/079017-3768

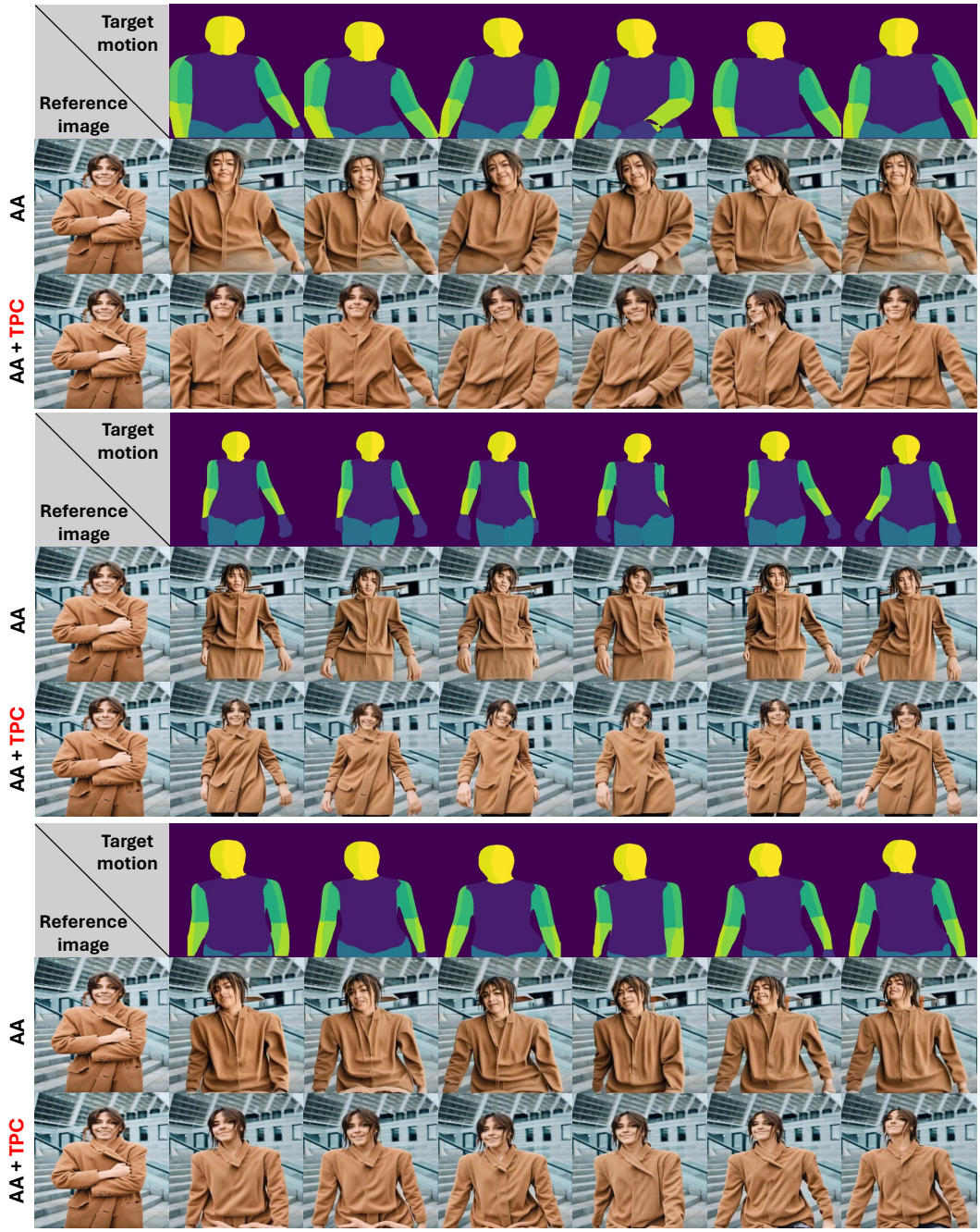


Figure 14: Qualitative results about applying TPC on recent diffusion-based AnimateAnyone (AA) on identical reference image with different motions. The motion videos are obtained from TikTok dance video dataset. (test-split)

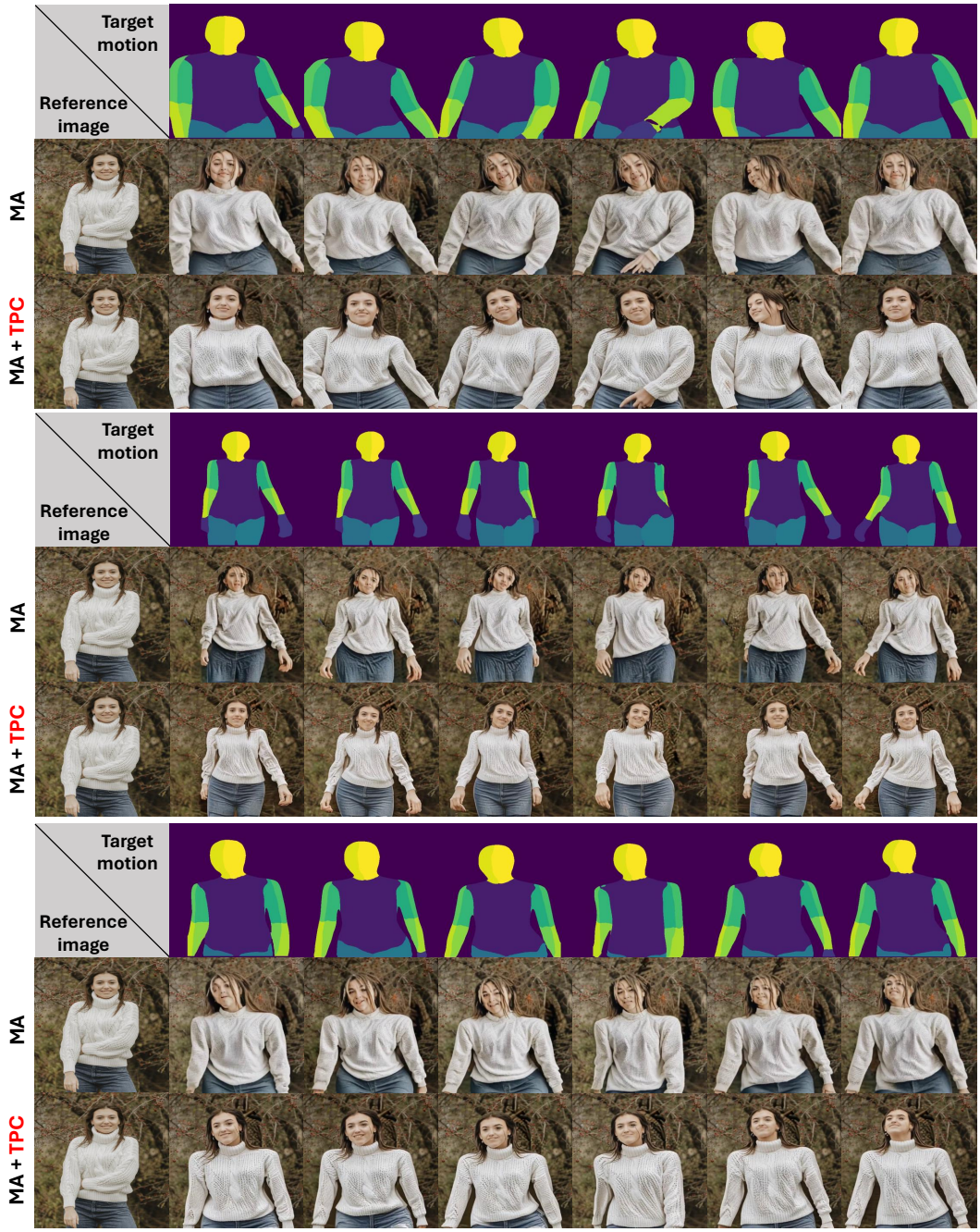


Figure 15: Qualitative results about applying TPC on recent diffusion-based MagicAnimate (MA) on identical reference image with different motions. The motion videos are obtained from TikTok dance video dataset. (test-split)

https://doi.org/10.52202/079017-3768

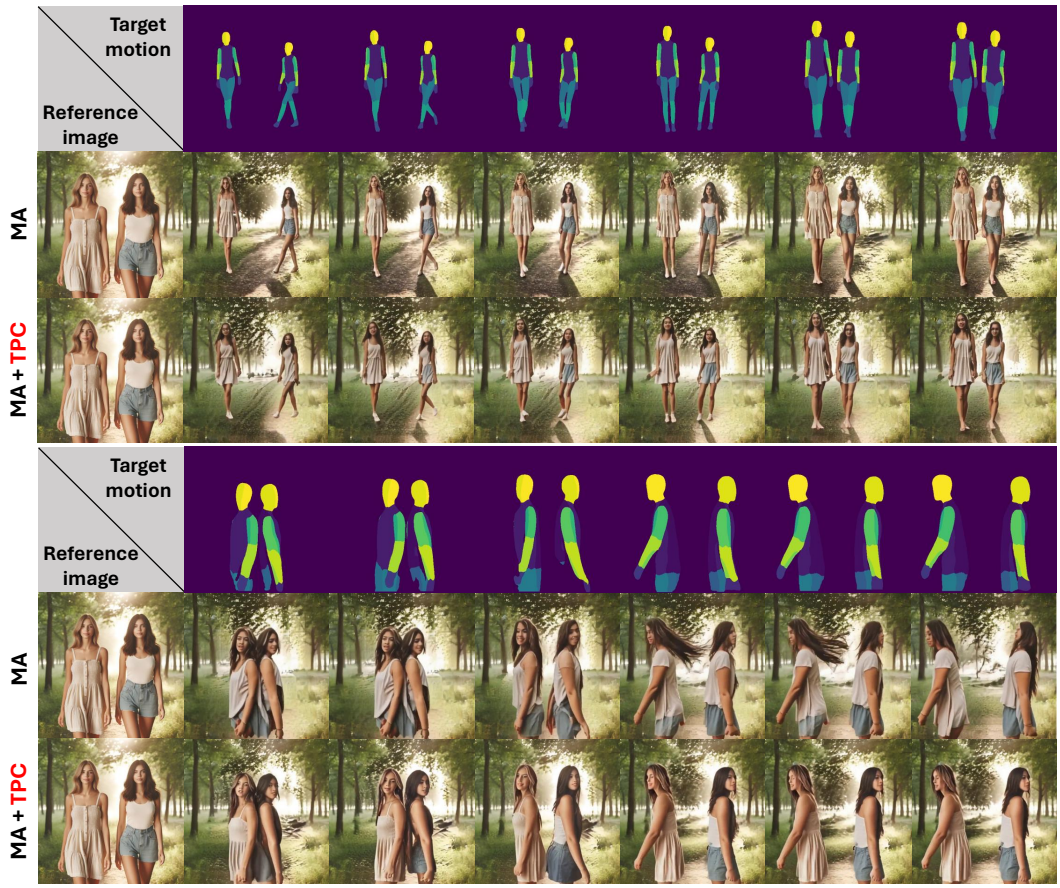


Figure 16: Qualitative results about applying TPC on recent diffusion-based MagicAnimate (MA) on multiple humans of reference and motion.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, in the line 15-17 in the Abstract and last paragraph in the Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is provided as a Limitation section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers DisCover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-3768

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Method section contains the details of our proposed framework.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit the source links for data used in our experiment and also a demo video.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: It is explained in the Implementation Details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: It is explained in the evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-3768

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about GPU resources in the Implementation Details section for implementing our framework.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read the NeurIPS Code of Ethics and follow all of it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a broader impact including societal negative impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The broader impact contains information about safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We explicitly provide all the references we used including url links.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-3768

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.