DISCO: Disentangled Control for Realistic Human Dance Generation

Tan Wang[1*], Linjie Li[2*], Kevin Lin[2*], Yuanhao Zhai[3], Chung-Ching Lin[2], Zhengyuan Yang[2],
Hanwang Zhang[1,4], Zicheng Liu[5], Lijuan Wang[2]

[1]Nanyang Technological University [2]Microsoft [3]University at Buffalo [4]Skywork AI [5]Advanced Micro Devices
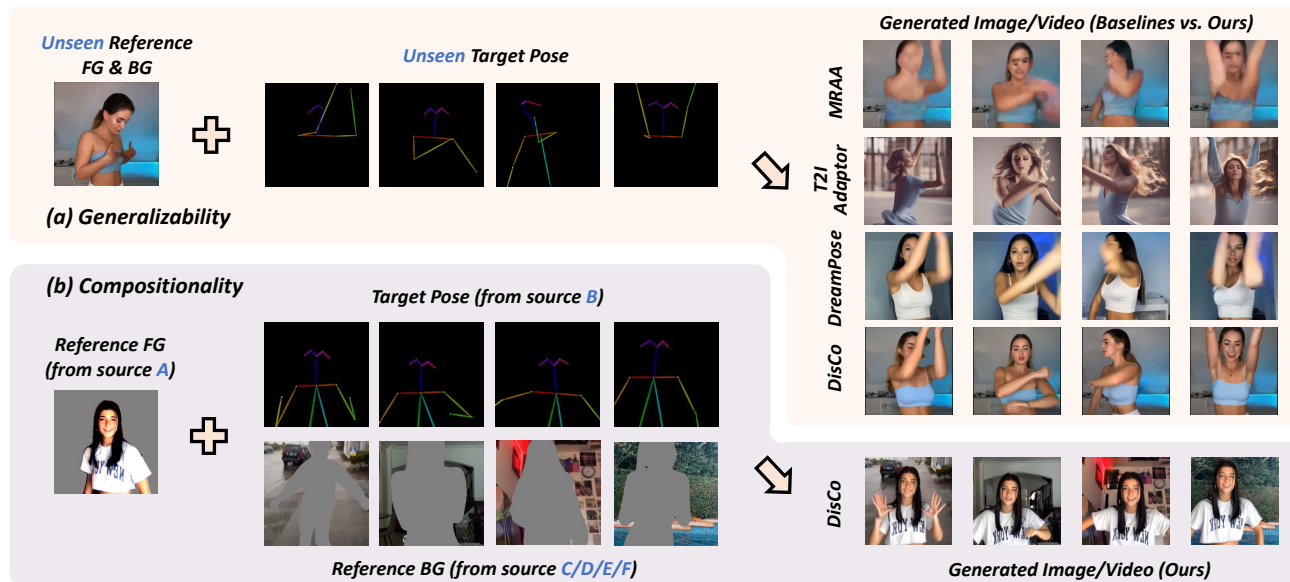
Figure 1. We propose DISCO for human dance generation on social media platforms, focusing on two key properties compared to conventional human motion transfer: (a) **Generalizability**: generalizable to unseen human subject (FG), background (BG) and pose; (b) **Compositionality**: adapting to the arbitrary composition of FG, BG and pose, each from a different source.

Abstract

Generative AI has made significant strides in computer vision, particularly in text-driven image/video synthesis (T2I/T2V). Despite the notable advancements, it remains challenging in human-centric content synthesis such as realistic dance generation. Current methodologies, primarily tailored for human motion transfer, encounter difficulties when confronted with real-world dance scenarios (e.g., social media dance), which require to generalize across a wide spectrum of poses and intricate human details. In this paper, we depart from the traditional paradigm of human motion transfer and emphasize two additional critical attributes for the synthesis of human dance content in social media contexts: (i) Generalizability: the model should be able to generalize beyond generic human viewpoints as well as unseen human subjects, backgrounds, and poses; (ii) Compositionality: it should allow for the seamless composition of seen/unseen subjects, backgrounds, and poses from different sources. To address these challenges, we introduce DISCO, which includes a novel model architecture with disentangled control to improve the compositionality of dance synthesis, and an effective human attribute pre-training for better generalizability to unseen humans. Extensive qualitative and quantitative results demonstrate that DISCO can generate high-quality human dance images and videos with diverse appearances and flexible motions. Code is available at https://disco-dance.github.io/.

1. Introduction

Starting from the era of GAN [3, 12], researchers [52, 53] have tried to explore the human motion transfer by transferring talking and Tai-Chi poses from a source image to a target individual. This requires the generated images/videos to precisely follow the source pose and retain the appearance of human subjects and backgrounds from the target image. However, when it comes to much more diverse and nuanced visual contents such as TikTok dancing videos, GANs tend to struggle due to mode collapse, capturing only a limited portion of the real data distribution (Figure 1a, MRAA [53]).

Recently, diffusion-based generative models [17, 56, 57] have significantly improved the synthesis in both diversity and stability. The introduction of ControlNet [71] further enhances the controllability by injecting geometric conditions (*e.g.*, human skeleton) into Stable Diffusion (SD) model [45], thus becomes possible to be utilized in human dance generation. However, prevailing ControlNet-based methods either rely on guidance from coarse-grained text prompts (Figure 1a, T2I Adaptor [40]) or simply substitute the text condition in T2I/T2V models with the referring image (Figure 1a, DreamPose [25]). It remains unclear how to ensure the consistency of rich human semantics and background in real-world dance scenarios. Moreover, almost all existing methods are trained on **insufficient dance video data**, hence suffer from either limited human attributes [5, 63, 75] or excessively simple poses and scenes [31, 36], leading to the poor zero-shot generalizability to unseen dance scenarios.

In order to support real-life applications, such as user-specific short video generation, we step from the conventional human motion transfer to realistic human dance synthesis and further highlight two properties:

- *Generalizability*: The model should be able to generalize to hard cases, *e.g.*, non-generic human view as well as unseen human subject, background and pose (Figure 1a).
- *Compositionality*: The generated images/videos can be from an arbitrary composition of seen or unseen human subject, background and pose, sourced from different images/videos (Figure 1b).

In this regard, we propose a novel approach, DISCO, for realistic and faithful human dance generation in social media scenarios. DISCO consists of two key designs: (*i*) a novel *model architecture with disentangled control* for improved compositionality; and (*ii*) an effective pre-training strategy with DISCO for better generalizability, named *human attribute pre-training*.

Model Architecture with Disentangled Control (Section 3.2). We attribute the failure of existing ControlNet-based methods to the inappropriate integration of various conditions. In this paper, Instead of naively combining multiple ControlNet, we utilize the VAE as the background encoder to fully leverage the prior of semantically rich images, while a tiny convolutional encoder for highly abstract skeleton. For the human subject, we incorporate its CLIP image embedding with the denoising U-Net as well as all other conditions via the cross-attention modules, to help dynamic foreground synthesis. By disentangling the control of three conditions, DISCO can not only enable arbitrary compositionality of human subjects, backgrounds, and dance-moves (Figure 1b), but also achieve high fidelity via the thorough utilization of the various input conditions (check Table 3 & 6 for the ablation of the condition mechanism).

Human Attribute Pre-training (Section 3.3). More importantly, we design a novel proxy task in which the model

conditions on the separate foreground and background region features and must reconstruct the complete image. This task is non-trivial as it enables the model to (*a*) effectively distinguish the dynamic human subject and static background for the ease of following pose transfer; (*b*) better encode-and-decode the complicated human faces and clothes during pre-training, and leaves the pose control learning to the fine-tuning stage of human dance synthesis. Crucially, without the constraint of pairwise human images for pose control, we can overcome the insufficiency of high-quality dance video data by leveraging large-scale collections of human images to learn diverse human attributes, in turn, greatly improve the generalizability of DISCO to unseen humans.

Our contributions are summarized as three-folds:

- We highlight two properties, generalizability and compositionality, that are missing from the conventional human motion transfer by introducing a more challenging social media dance synthesis problem, to facilitate its potential in the production of user-specific short videos.
- To address these problems, we propose DISCO framework with (*i*) a novel model architecture for disentangled control to ensure compositionality in generation; and (*ii*) an effective human attribute pre-training to improve generalizability to unseen humans and non-generic views.
- We conduct a broad variety of evaluations together with a user study, to demonstrate the effectiveness of DISCO. Notably, even without temporal consistency modeling, DISCO can already achieve superior FID (**28.31** v.s 53.78) and FID-VID (**55.17** v.s 66.36) scores over the state-of-the-art approaches. Adding temporal modeling further boosts FID-VID scores of DISCO to **29.37**.

2. Related Work

Diffusion Models for Controllable Image/Video Generation. Diffusion probabilistic models [7, 55] have shown great success in high-quality image/video generation. Towards user-specific generation, text prompts are first utilized as the condition for image generation [16, 44, 48, 69]. Among them, Stable Diffusion [45] (SD) stands as the representative work to date, with high efficiency and competitive quality via diffusion over the latent space. For better controllability, ControlNet [71] introduces additional control mechanisms into SD beyond texts, such as sketch, human skeleton, and segmentation map. Compared to image, text-to-video synthesis [8, 18, 27, 54, 68], is more challenging due to the lack of well-annotated data and difficulties in temporal modeling. Thus, existing controllable video generation methods [27, 37] mainly stem from pre-trained text-to-image model and try to introduce motion/pose prior to the text-to-video synthesis. In this work, we look into a more challenging setting of conditional human image/video synthesis which requires precise control of both human at-

tributes (such as identity, clothing, hairstyle, *etc*.) as well as the dance-moves (poses) in social media dance scenarios.

Human Dance Synthesis. Early work on this task includes video-to-video synthesis [5, 10, 63, 64], still image animation [2, 19, 65, 67, 70] and motion transfer [30, 51, 58, 75]. Nevertheless, these methods require either a several-minute-long target person video for human-specific fine-tuning, or multiple separate networks and cascaded training stages for background, motion and occlusion map prediction. The advances of diffusion models [45] greatly simplify the training of such generative models, inspiring follow-up diffusion models [29, 41] tailored for human dance generation. Still, these methods require a separate motion prediction module and struggle to precisely control the human pose. Dream-Pose [25] is perhaps the most relevant study to ours, which proposes an image-and-pose conditioned diffusion method for still fashion image animation. However, as they consider only fashion subjects with easy catwalk poses in front of an empty background, their model may suffer from limited generalization ability, prohibiting its potential for more intricate human dance synthesis in real-world scenarios.

3. DISCO

We start by first formally introducing the social media human dance generation setting. Let f and g represent human foreground and background in the reference image. Given a specific (or a sequence of) pose keypoint $p = p_t$ (or $p = \{p_1, p_2, ..., p_T\}$), we aim to generate realistic images I_t (or videos $V = \{I_1, I_2, ..., I_T\}$) conditioned on f, g, p. The generated images (or videos) should be 1) *faithful*: the human attribute and background of the synthesis should be consistent with f and g from the reference image and the human subject should be aligned with the pose p; 2) *generalizable*: the model should be able to generalize to unseen humans, backgrounds and poses, without the need of human-specific fine-tuning; and 3) *composable*: the model should adapt to arbitrary composition of f, g, p from different image/video sources to generate novel images/videos. In what follows, Section 3.1 briefly reviews the latent diffusion models and ControlNet, which are the basis of DISCO. Section 3.2 details the model architecture of DISCO with disentangled control of human foreground, background, and pose to enable faithful and fully composable human dance image/video synthesis. Section 3.3 presents how to further enhance the generalizability of DISCO, as well as the faithfulness in generated contents by pre-training human attributes from large-scale human images. The overview of DISCO can be found in Figure 2.

3.1. Preliminary

Latent Diffusion Models (LDM) is a type of diffusion model that operates in the encoded latent space of an au-

toencoder $\mathcal{D}(\mathcal{E}(\cdot))$. An exemplary LDM is the popular Stable Diffusion (SD) [45] which consists an autoencoder VQ-VAE [62] and a time-conditioned U-Net [46] for noise estimation. A CLIP ViT-L/14 text encoder [43] is used to project the input text query into the text embedding condition c_{text}. During training, given an image I and the text condition c_{text}, the image latent $z_0 = \mathcal{E}(I)$ is diffused in T time steps with a deterministic Gaussian process to produce the noisy latent $z_T \sim \mathcal{N}(0, 1)$. SD is trained to learn the reverse denoising process with the following objective [45]:

$$L = \mathbb{E}_{\mathcal{E}(I), c_{\text{text}}, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_{\text{text}})\|_2^2 \right], \quad (1)$$

where $t = \{1, ..., T\}$, ϵ_θ represents the trainable module. It contains a U-Net architecture composed of the convolution (ResBlock) and self-/cross-attention (TransBlock), which accepts the noisy latents z_t and the text embedding condition c_{text} as the input. After training, one can apply a deterministic sampling process (*e.g.*, DDIM [56]) to generate z_0 and pass it to the decoder $\mathcal{D}$ towards the final image.

ControlNet [71], built upon SD, manipulates the input to the intermediate layers of the U-Net in SD, for controllable image generation. Specifically, it creates a trainable copy of the U-Net down/middle blocks and adds an additional "zero convolution" layer. The outputs of each copy block is then added to the skip connections of the original U-Net. Apart from the text condition c_{text}, ControlNet is trained with an additional external condition vector c which can be many types, such as edge map, depth map and segmentation.

3.2. Model Architecture with Disentangled Control

There are two typical control designs for the social media dance generation. The direct usage of ControlNet presents challenges due to the missing reference human image condition, which is critical for keeping the human identity and attributes consistent. Recent explorations in image variations [24] replace the CLIP text embedding with the image embedding as the condition, which can retain some high-level semantics from the reference image. Nevertheless, the geometric/structural control of the generated image is still missing. Meanwhile, simply combining these two designs does *not* work well in practice (Table 3 and 6).

To take the distinctive benefits of these two different control designs, we introduce a novel model architecture with disentangled control, to enable accurate alterations to the human pose, while simultaneously maintaining attribute and background stability. Meanwhile, it also facilitates full compositionality in the human dance synthesis (Figure 1b).Specifically, given a reference human image, we can first utilize an existing human matting method (*e.g.*, SAM [28, 34]) to separate the human foreground from the background. Next, we explain how all three conditions, the human foreground f, the background g and the desired pose p, are incorporated into DISCO.
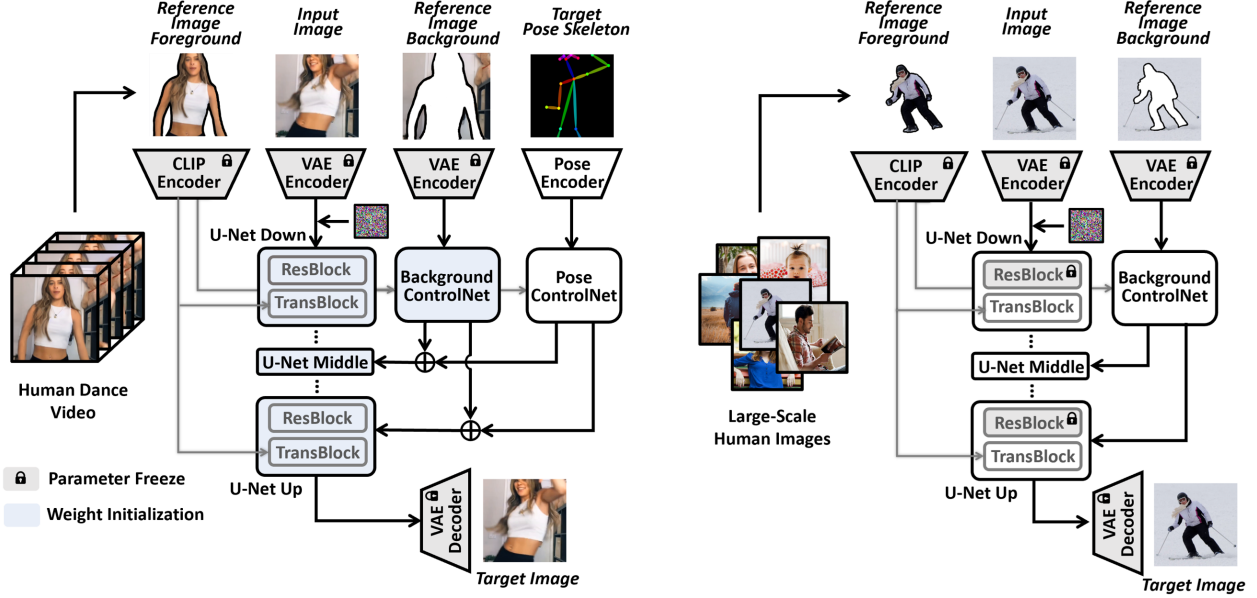
(a) Model Architecture with Disentangled Control   (b) Human Attribute Pre-training

Figure 2. The proposed DISCO framework for social media human dance generation.

Referring Foreground via Cross Attention. To help model easily adapt to the CLIP image feature space, we first use the pre-trained image variation diffusion model [24] for the U-Net initialization. However, in contrast to using the global CLIP image embeddings employed by image variation methods, here we adopt the local CLIP image embeddings right before the global pooling layer, for more fine-grained human semantics encoding. Consequently, the original text embedding $c_{text} \in \mathbb{R}^{l \times d}$ is superseded by the local CLIP image embeddings of the human foreground $c_f \in \mathbb{R}^{hw \times d}$ to serve as the key and value feature in cross-attention layer, where l, h, w, d represent the caption length, the height, width of the visual feature map and the feature dimension.

Controlling Background and Pose via ControlNets. For pose p, we adopt the vanilla design of ControlNet. Specifically, we embed the pose image into the same latent space as the Unet input via four convolution layers, and dedicate a ControlNet branch τ_θ to learn the pose control. For background g, we insert another ControlNet branch μ_θ to the model. Notably, we propose to use the pre-trained VQ-VAE encoder $\mathcal{E}$ in SD, instead of four randomly initialized convolution layers, to convert the background image into dense feature maps to preserve intricate details. The remainder of the architecture for the background ControlNet branch follows the original ControlNet. As we replace the text condition with the referring foreground in the cross-attention modules, we also update the condition input to ControlNet as the local CLIP image feature of the referring foreground. As shown in Figure 2a, the outputs of the two ControlNet branches are combined via addition and fed into the middle and up block of the U-Net.

With the design of the disentangled controls above, we fine-tune DISCO with the same diffusion objective [45]:

$$L = \mathbb{E}_{\mathcal{E}(I), c_f, \tau_\theta(p), \mu_\theta(g), \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon \\ - \epsilon_\theta(z_t, t, c_f, \tau_\theta(p), \mu_\theta(g))\|_2^2], \quad (2)$$

where ϵ_θ, τ_θ and μ_θ are the trainable network modules. Specifically, ϵ_θ represents the U-Net architecture composed of the convolution block (ResBlock) and self-/cross-attention block (TransBlock), which accepts the noisy latents z_t and the referring foreground condition c_f as the inputs. τ_θ and μ_θ represent the two ControlNet branches for pose condition p and background condition g, respectively.
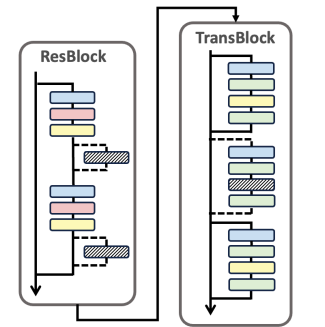

Figure 3. The detailed architecture of the ResBlock and TransBlock. The temporal module (dotted line) is optional.

Temporal Modeling (TM). To achieve better temporal continuity for video output, we follow [8, 54] to introduce a 1D temporal convolution/attention layer after the existing 2D spatial ones in ResBlock and TransBlock (Figure 3). Notably, only the last yellow spatial attention module in TransBlock accept the CLIP image features for cross-attention operation. Besides, the temporal layers are zero-initialized, and connected via residual connections. Simultaneously, we adjust the model input from image to video by concatenating the pose p from n consecutive frames, duplicating the background g and the initial noise for n times.

3.3. Human Attribute Pre-training

In utilizing the disentangled control architecture for DISCO, although it shows promises in pose control and background reconstruction, we find it remains challenging to have faithful generations with unseen human subject foregrounds and non-generic human views, demonstrating poor generalizability. The crux of this matter lies in the current training pipeline. It relies on high-quality human videos to provide training pairs of human images, with the same human foreground and background appearance, but different poses. Yet, we observe that current training datasets used for human dance generation confront a dilemma of "mutual exclusivity" — they cannot ensure both diversity in human attributes (such as identity, clothing, makeup, hairstyle, *etc.*) and the complicated poses due to the prohibitive costs of collecting and filtering human videos. As an alternative, human images, which are widely available over the internet, contain diverse human subject foregrounds and backgrounds, despite of the missing paired images with pose alterations.

This motivates us to propose a pre-training task, human attribute pre-training, to improve the generalizability and faithfulness. Rather than directly constructing the high-quality human dance video dataset, we explore an "*easy-to-hard*" training schema to step from image reconstruction to motion editing. Figure 2b shows the details. Specifically, we propose to reconstruct the whole image given the human foreground and background features. In this way, model effectively learns the distinguishment between human subject and foreground, as well as diverse human attributes from large-scale images. Compared to the human dance generation fine-tuning, the ControlNet branch for pose control is removed while the rest of the architecture remains the same. Consequently, we modify the objective as:

$$L = \mathbb{E}_{\mathcal{E}(I), c_f, \mu_\theta(g), \epsilon \sim \mathcal{N}(0,1), t}[\|\epsilon - \epsilon_\theta(z_t, t, c_f, \mu_\theta(g))\|_2^2]. \quad (3)$$

Empirically, we find that freezing the ResNet blocks in U-Net during pre-training can achieve better reconstruction quality of human faces and subtleties.

For human dance generation fine-tuning, we initialize the U-Net and ControlNet branch for background control (highlighted with blue in Figure 2a) by the pre-trained model, and initialize the pose ControlNet branch with the pre-trained U-Net weight following [71].

4. Experiments

4.1. Experimental Setup

We train the models on the public TikTok dataset [22] for referring human dance generation. TikTok dataset consists of about 350 dance videos (with video length of 10-15 seconds) capturing a single-person dance. For each video, we first extract frames with 30fps, and run Grounded-SAM [28, 34] and OpenPose [4] on each frame to infer the human subject mask and the pose skeleton. 335 videos are sampled as the

training split. To ensure videos from the same person (same identity with same/different appearance) are not present in both training and testing splits, we collect 10 TikTok-style videos depicting different people from the web, as the testing split. We train our model on 8 NVIDIA V100 GPUs for 70K steps with image size 256×256 and learning rate $2e^{-4}$. During training, we sample the first frame of the video as the reference and all others at 30 fps as targets. For equipping TM, we set $n = 8$, and apply a learning rate of $5e^{-4}$ on the temporal convolution/attention layers. Both reference and target images are randomly cropped at the same position along the height dimension with the aspect ratio of 1, before resized to 256×256. For evaluation, we apply center cropping instead of random cropping.

For human attribute pre-training, we use a combination of multiple public datasets (TikTok [1] [22], COCO [32], SHHQ [9], DeepFashion2 [11], LAION [49]). We first run Grounded-SAM [28, 34] with the prompt of "person" to automatically generate the human foreground mask, and then filter out images without human. This results in over 700K images All pre-training experiments are conducted on 32 NVIDIA V100 GPUs for 25K steps with image size 256×256 and learning rate $1e^{-3}$. We initialize the U-Net model with the pre-trained weights of SD Image Variations [24]. The ControlNet branches are initialized with the same weight as the U-Net model, except for the zero-convolution layers, following Zhang and Agrawala [71]. After human attribute pre-training, we initialize the U-Net and background ControlNet branch by the pre-trained model, and initialize the pose ControlNet branch with the pre-trained U-Net weight, for human dance generation fine-tuning.

4.2. DISCO Applications

Benefiting from the strong human synthesis capability powered by the disentangled control design as well as human attribute pre-training, our DISCO provides flexible and fine-grained controllability and generalizability to arbitrary combination of human subject, pose and background. Given three existing images, each with distinct human subject, background and pose, there can be a total of 27 combinations. Here, we showcase 5 representatives scenarios in Figure 4 for **human image editing**: (*i*) *Human Subject / Pose Re-targeting*: the model have been exposed to training instances of the human subject or the pose, but the specific combinations of both are new to the model; (*ii*) *Unseen Pose Generation*: the human subject is from the training set, but the pose is novel, from the testing set; (*iii*) *Unseen Human Subject Generation*: the poses are sampled from the training set, but the human subject is novel, which is either from the testing set or crawled from the web; (*iv*) *Unseen Pose & Human Subject Generation*: both the human subject and pose are not present in the training set; (*v*) *Full Unseen*

---

[1] We only use the training split for pre-training to avoid data leak.

Table 1. Quantitative comparisons of DISCO with the recent SOTA method DreamPose. "CFG" and "HAP" denote classifier-free guidance and human attribute pre-training, respectively. ↓ indicates the lower the better, and vice versa. For DISCO †, we further scale up the fine-tuning stage to ∼600 TikTok-style videos. Methods with ∗ directly use target image as the input, including more information compared to the OpenPose.

| Method	Image					Video	
	FID ↓	SSIM ↑	PSNR ↑	LISPIS ↓	L1 ↓	FID-VID ↓	FVD ↓
FOMM* [52]	85.03	0.648	29.01	0.335	3.61E-04	90.09	405.22
MRAA* [53]	54.47	0.672	**29.39**	0.296	**3.21E-04**	66.36	284.82
TPS* [73]	53.78	0.673	29.18	0.299	3.23E-04	72.55	306.17
DreamPose [25]	79.46	0.509	28.04	0.450	6.91E-04	80.51	551.56
DreamPose (CFG)	72.62	0.511	28.11	0.442	6.88E-04	78.77	551.02
DISCO (w/o HAP)	61.06	0.631	28.78	0.317	4.46E-04	73.29	366.39
DISCO (w/. HAP)	38.19	0.663	29.33	0.291	3.69E-04	61.88	286.91
DISCO (w/. HAP, CFG)	30.75	0.668	29.03	0.292	3.78E-04	59.90	292.80
DISCO † (w/. HAP, CFG)	**28.31**	**0.674**	29.15	**0.285**	3.69E-04	**55.17**	**267.75**
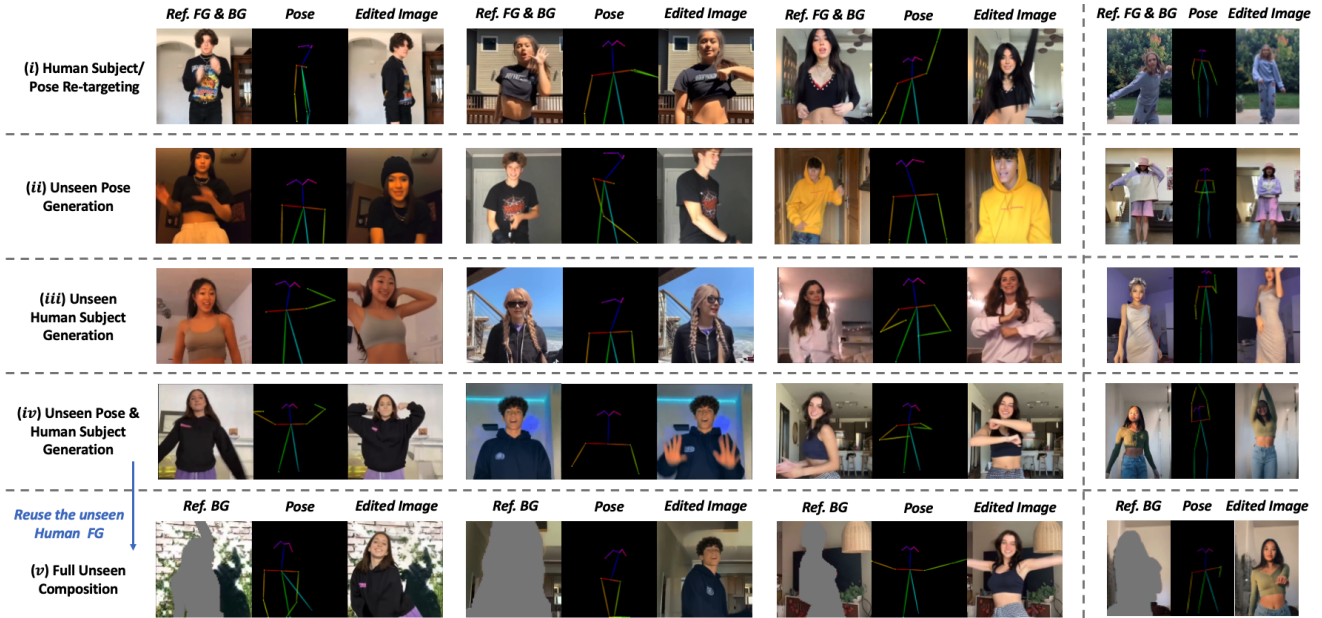
Figure 4. Visualizations of 5 representative scenarios for **human image editing**. In the last column, we show that DISCO also generalizes to different image ratios and full-body human views. Check Figure 12 in Appendix for more results.

Composition: further sampling a novel background from another unseen image/video based on iv Examples in Figure 4 demonstrate that DISCO can flexibly update one of (or a composition of) human subject/background/pose in a given image to a user-specified one (or composition), either from existing training samples or novel images.

Observing satisfying human image editing results from DISCO, especially the faithfulness in edited images, we can further extend it to **human dance video generation** as it is. Given a reference image and a target pose sequence either extracted from an existing video or from user manipulation of a human skeleton, we generate the video frame-by-frame, with the reference image and a single pose as the inputs to DISCO. We delay the visualization of generated videos and relevant discussions to Figure 5 in the next section. More examples of the two applications above are included in Appendix.

Though it is not the focus of this paper, our final DISCO model can be readily and flexibly integrated with efficient fine-tuning techniques [21, 25, 68] for subject-specific fine-tuning on multiple images of the same human subject. We leave discussions and results of this setting to Appendix B.

4.3. Main Results

We provide quantitative and qualitative comparisons and test generalizability against both conventional motion transfer methods FOMM [52], MRAA [53], TPS [73], and most relevant work DreamPose [25], an image-to-video model designed for fashion domain with densepose control. It is worth noting that the TikTok dancing videos we evaluate on are all real-world content, which are much more complicated than those in existing dancing video datasets [5, 23, 63], with clean background and similar clothing.

Qualitative Comparison. We qualitatively compare DISCO to DreamPose in Figure 5. DreamPose obviously suffers from inconsistent human attribute and unstable background. Without HAP, DISCO can already reconstruct the coarse-
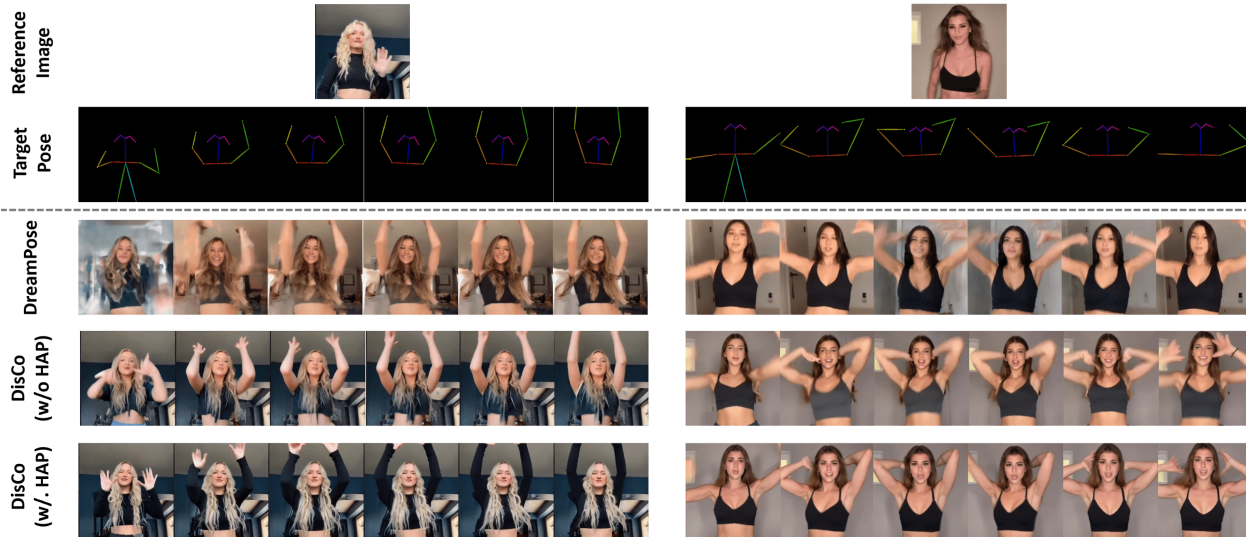
Figure 5. Qualitative comparison between our DISCO (w/ or w/o HAP) and DreamPose on **human dance video generation** with the input of a reference image and a sequence of target poses. Note that the reference image and target poses are from the testing split, where the human subjects, backgrounds, poses are not available during the model training. Best viewed when zoomed-in.

grained appearance of the human subject and maintain a steady background in the generated frames. With HAP, the more fine-grained human attributes (*e.g.*, black long sleeves in the left instance and the vest color in the right instance) can be further improved. It is worth highlighting that our DISCO can generate videos with surprisingly good temporal consistency, even without explicit temporal modeling.

Quantitative Comparison. Since DISCO is applicable to both image and video generation for human dance synthesis, here we compare the models on both image- and video-wise generative metrics. To evaluate the image generation quality, we report frame-wise FID [15], SSIM [66], LISPIS [72], PSNR [20] and L1; while for videos, we concatenate every consecutive 16 frames to form a sample, to report FID-VID [1] and FVD [60]. As shown in Table 1, DISCO without human attribute pre-training (HAP) already significantly outperforms DreamPose by large margins across all metrics. Adding HAP further improves DISCO, reducing FID to ∼ 38 and FVD to ∼ 280.

The substantial performance gain against the recent SOTA model Dream-Pose evidently demonstrates the superiority of DISCO.

Table 2. Video generation comparisons by adding the Temporal Modeling (TM). We employ HAP and CFG by default.

Method	FID-VID ↓	FVD ↓
DISCO	59.90	292.80
DISCO (w/. TM)	34.19	260.34
DISCO †	55.17	267.75
DISCO † (w/. TM)	**29.37**	**229.66**

Furthermore, we additionally collect 250 TikTok-style short videos from the web to enlarge the training split to ∼600 videos in total. The performance gain has shown the potential of DISCO to be further scaled-up. As shown in Table 2, further incorporating temporal modeling to DISCO brings huge performance boosts to video synthesis metrics,
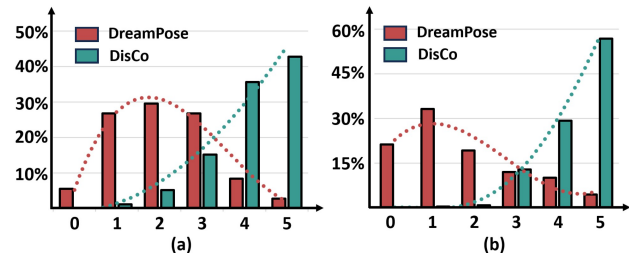

Figure 6. Results of User Study: the overall quality score distribution on both synthesis (a) images and (b) videos of DreamPose and DISCO.

e.g., improving FID-VID by ∼ 30 and FVD by ∼ 60.

Beyond standard metrics, a user study with 50 unique participants was conducted. Participants rated images and videos 0 (worst) to 5 (best). Figure 6 illustrates that DISCO (green histogram) clearly received higher user preference scores, with over 80% of users assigning a rating of 4 or above. Further details on the study methodology and additional results are provided in the Appendix C.4.

Generalizability. In addition to the TikTok data, we show strong **zero-shot** generation results of DISCO on several unseen out-of-distribution (OOD) data: Everybody Dance Now [5], AIST++ [31] and even wild youtube video in Figure 7. Furthermore, we quantitatively evaluate DISCO on two representative datasets in Table 5: conventional benchmark (Taichi) and OOD action recognition data (NTU-120). DISCO consistently outperforms the baselines.

4.4. Ablation Study

Architecture Design. Table 3 quantitatively analyzes the impact of different architecture designs in DISCO. First, to ablate the control with the reference image, we observe that either ControlNet or the cross-attention module struggles to handle the control of the whole reference image without

Table 3. Ablation on architecture designs without HAP. "ControlNet (fg+bg)" and "Attention (fg+bg)" in the first block denote inserting the control condition of reference image (containing both foreground and background) via a single ControlNet or cross-attention modules. "CLIP Global/Local" means using the global or local CLIP feature to represent the reference foreground. "CLIP Local + VAE" combines VAE features with CLIP Local features. Additional ablation results are included in Appendix C.2.

Method	FID ↓	SSIM ↑	PSNR ↑	LISPIS ↓	L1 ↓	FID-VID ↓	FVD ↓
DISCO	61.06	**0.631**	**28.78**	0.317	**4.46E-04**	**73.29**	**366.39**
Ablation on control mechanism w/ reference image (DISCO setting: ControlNet (bg) + Attention (fg))							
ControlNet (fg+bg)	65.14	0.600	28.57	0.355	4.83E-04	74.19	427.49
Attention (fg+bg)	80.50	0.474	28.01	0.485	7.50E-04	80.49	551.51
Ablation on reference foreground encoding (DISCO setting: CLIP Local)							
CLIP Global	63.92	0.621	28.61	**0.311**	5.00E-04	73.33	391.41
CLIP Local + VAE	**59.74**	0.623	28.52	0.331	4.79E-04	77.86	406.16

Table 4. Ablation analysis of image data size for human attribute pre-training.

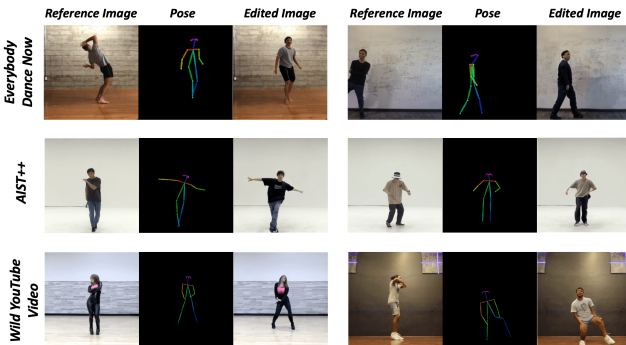| Pre-train Data	Data Size	FID ↓	SSIM ↑	PSNR ↑	LISPIS ↓	L1 ↓	FID-VID ↓	FVD ↓
N/A	0	61.06	0.631	28.78	0.317	4.46E-04	73.29	366.39
TikTok [22]	90K	50.68	0.648	28.81	0.309	4.27E-04	69.68	353.35
+ COCO [32]	110K	48.89	0.654	28.97	0.303	4.07E-04	62.15	326.88
+ SSHQ [9]	184K	44.13	0.655	29.00	0.300	3.93E-04	64.47	325.40
+ DpFashion2 [11] + LAION [49]	700K	**38.19**	**0.663**	**29.33**	**0.291**	**3.69E-04**	**61.88**	**286.91**

Figure 7. Zero-shot synthesis of DISCO on several OOD data.

Table 5. Quantitative results on OOD human motion dataset.

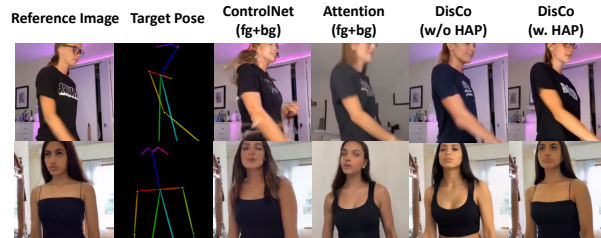| Method	TaiChi [52]			NTU-120 [33]		
	FID ↓	FID-VID ↓	FVD ↓	FID ↓	FID-VID ↓	FVD ↓
FOMM	24.43	24.50	374.26	80.29	40.34	1439.50
MRAA	17.32	21.55	312.57	97.07	58.19	1441.79
DISCO	**15.89**	**10.45**	**299.51**	**68.53**	**26.21**	**458.92**

Figure 8. The qualitative comparison between different architecture designs.

disentangling the foreground from the background, leads to inferior quantitative results on most metrics. Though ControlNet-only baseline achieves better FVD scores, our visualizations in Figure 8 show that such architecture still struggles to maintain the consistency of human attributes and the stability of the background. For the encoding of the reference foreground, DISCO with CLIP local feature produces better results than the one with CLIP global feature on 6 out of 7 metrics. We also explore to complement the CLIP local feature with VAE feature with a learnable adaptor following DreamPose [25], which leads to a slight better FID score but worse performance on all other metrics.

Pre-training Data Size. Table 4 investigates the effect of the data size in HAP stage by incrementally augmenting the pre-training data from open-source human image datasets. It is evident that a larger and more diverse pre-training data can yield better downstream results for referring human dance generation. Moreover, compared with "without pre-training" (1st row), adopting HAP on the same TikTok dataset as a self-supervised learning schema (2nd row) can already bring out significant performance gains.

5. Conclusion

We revisit human dance synthesis for the more practical social media scenario and emphasize two key properties, *generalizability* and *compositionality*. To tackle this problem, we propose DISCO, equipped with a novel architecture for disentangled control and an effective human attribute pre-training task. Extensive qualitative and quantitative results demonstrate the effectiveness of DISCO, which we believe is a step closer towards real-world applications. **Limitations.** Though DISCO has shown promising results and decent human dance generation quality, it cannot handle hand posture well without the fine-grained hand pose control. Moreover, it is hard to be applied to multi-person scenarios and human-object interactions.

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*, 2019. 7

[2] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *ICCV*, 2021. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 1

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 2, 3, 6, 7

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ICLR*, 2023. 12

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2

[8] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 2, 4, 12

[9] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Styleganhuman: A data-centric odyssey of human generation. In *ECCV*, 2022. 5, 8

[10] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. *arXiv preprint arXiv:1904.08379*, 2019. 3

[11] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, 2019. 5, 8

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 17

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ICLR*, 2023. 12

[15] 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. *NeurIPS*, 2017. 7

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[19] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *CVPR*, 2021. 3

[20] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010. 7

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 6, 14

[22] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021. 5, 8

[23] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *ICCV*, 2023. 6

[24] Pinkney Justin and Lambda. Stable Diffusion Image Variations. https://huggingface.co/lambdalabs/sd-image-variations-diffusers, 2022. 3, 4, 5, 12

[25] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *ICCV*, 2023. 2, 3, 6, 8, 17

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 12

[27] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 12

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *ICCV*, 2023. 3, 5

[29] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 3

[30] Jessica Lee, Deva Ramanan, and Rohit Girdhar. Metapix: Few-shot video retargeting. *arXiv preprint arXiv:1910.04742*, 2019. 3

[31] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 2, 7

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 8

[33] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701, 2019. 8

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 5

[35] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 12

[36] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 2

[37] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2, 12

[38] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *ICLR*, 2022. 12

[39] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 12

[40] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[41] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. *CVPR*, 2023. 3

[42] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 12

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 12

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*, 2023. 12

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5, 8

[50] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 12

[51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 3

[52] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 1, 6, 8

[53] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 6

[54] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 4

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[57] 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. 2

[58] Yucheng Suo, Zhedong Zheng, Xiaohan Wang, Bang Zhang, and Yi Yang. Jointly harnessing prior structures and temporal consistency for sign language video generation. *arXiv preprint arXiv:2207.03714*, 2022. 3

[59] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 12

[60] 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. 7

[61] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 12

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 3

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *NeurIPS*, 2018. 2, 3, 6

[64] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *NeurIPS*, 2019. 3

[65] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *ICLR*, 2022. 3

[66] 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. 7

[67] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *CVPR*, 2019. 3

[68] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 2, 6, 12

[69] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 2

[70] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *CVPR*, pages 15039–15048, 2021. 3

[71] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023. 2, 3, 5, 14, 15

[72] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[73] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 6

[74] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023. 12

[75] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance dance generation: Motion transfer for internet videos. In *ICCV Workshops*, 2019. 2, 3