DisCo: Disentangled Control for Referring Human Dance Generation in Real World

Tan Wang[*†§], **Linjie Li**[*‡], **Kevin Lin**[*‡], **Chung-Ching Lin**[‡], **Zhengyuan Yang**[‡], **Hanwang Zhang**[†], **Zicheng Liu**[‡], **Lijuan Wang**[‡]
[†]Nanyang Technological University [‡]Microsoft Azure AI
{TAN317,hanwangzhang}@ntu.edu.sg {lindsey.li,keli,chungching.lin,zhengyang,zliu,lijuanw}@microsoft.com
https://disco-dance.github.io

Abstract

Generative AI has made significant strides in computer vision, particularly in image/video synthesis conditioned on text descriptions. Despite the advancements, it remains challenging especially in the generation of human-centric content such as dance synthesis. Existing dance synthesis methods struggle with the gap between synthesized content and real-world dance scenarios. In this paper, we define a new problem setting: *Referring Human Dance Generation*, which focuses on real-world dance scenarios with three important properties: *(i) Faithfulness*: the synthesis should retain the appearance of both human subject foreground and background from the reference image, and precisely follow the target pose; *(ii) Generalizability:* the model should generalize to unseen human subjects, backgrounds, and poses; *(iii) Compositionality:* it should allow for composition of seen/unseen subjects, backgrounds, and poses from different sources. To address these challenges, we introduce a novel approach, DisCo, which includes a novel model architecture with disentangled control to improve the faithfulness and 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.

1 Introduction

Generative AI has garnered significant interests in computer vision community. Recent advancements in text-driven image and video synthesis (T2I/T2V) [47, 44, 21, 51, 72, 19, 57, 10, 71] bolstered by the advent of diffusion models [20, 9, 59, 60] exhibited remarkable ingenuity and generative quality, demonstrating considerable potential in image and video synthesis, editing, and animation. However, the synthesized images/videos are still not satisfying, especially for human-centric applications, *e.g.*, human dance synthesis.

Despite the long history of human dance synthesis, existing methods greatly suffer from the gap between the synthesized content and real-world dance scenarios. Starting from the era of GAN [15, 4, 28], researchers [67, 66, 12, 7] try to extend the video2video style transfer for transferring dance movements from a source video to a target individual, which often requires human-specific fine-tuning on the target person. Recently, a line of work [30, 39] leverages pre-trained diffusion-based T2I/T2V models to generate dance images/videos conditioned on text prompts. Such coarse-grained condition greatly limits the degree of the controllability, making it almost impossible for users to precisely specify the anticipated subjects (*i.e.*, human appearance) as well as the dance-moves (*i.e.*, human pose). Though the introduction of ControlNet [73] partially alleviates this problem by

[*]Equal Contribution [§]Work done during internship at Microsoft

Preprint. Under review.

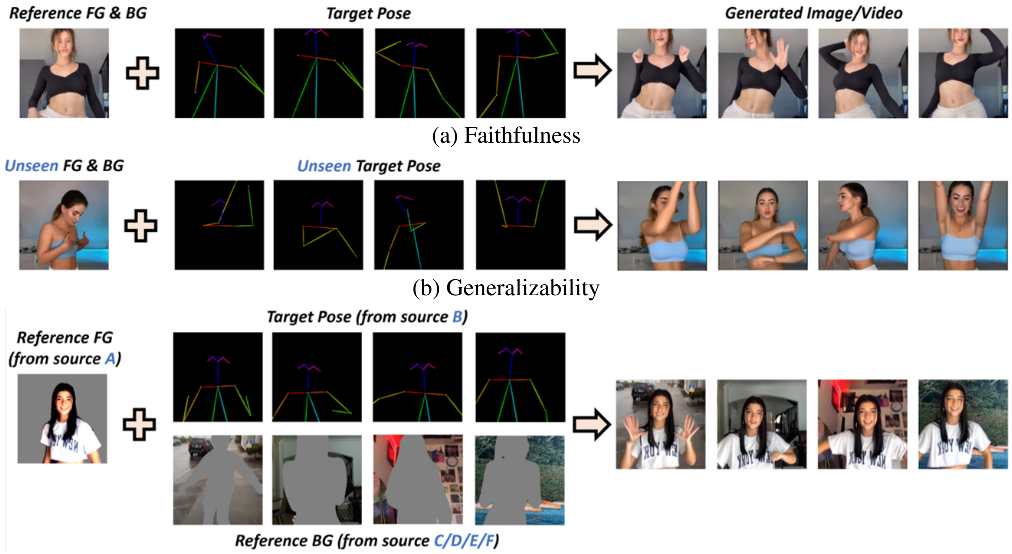


Figure 1: We propose DisCo for referring human dance generation, which can generate human dance images/videos with the following three properties: (a) **Faithfulness**: retaining the appearance of foreground (FG) and background (BG) in consistent to the reference image while precisely following the pose; (b) **Generalizability**: generalizable to unseen human subject FG, BG and pose; (c) **Compositionality**: adapting to arbitrary composition of human subject FG, background and pose, each from a different source.

incorporating pose control with geometric human keypoints, it remains unclear how ControlNet can ensure the consistency of rich semantics (such as human appearance) in the reference image, due to its dependency on text prompt. Moreover, almost all existing methods trained on limited dance video datasets suffer from either limited subject attributes [67, 7, 77] or excessively simplistic scenes and backgrounds [34, 38, 27, 26], leading to the poor zero-shot generalizability to unseen composition of human subjects, poses and backgrounds.

In order to support real-life applications, such as user-specific short video content generation, we define a new problem setting: *referring human dance generation*, with the focus on *real-world* dance scenarios. Given a reference image designating a human subject foreground and the background together (or two images depicting the human foreground and background respectively), and the dance movement specified by a pose or a sequence of poses, a model should be able to synthesize human dance images/videos with the following three properties:

- *Faithfulness*: The generated images/videos should retain the appearance of human subjects and backgrounds, in consistent to the reference images (Figure 1a). In addition, the generated human subject should precisely follow the provided pose.
- *Generalizability*: The model should be able to generalize to unseen human subject, background and pose (Figure 1b), without the need of human-specific fine-tuning.
- *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 1c).

In this regard, we propose a novel approach, DisCo, for referring human dance generation in real world. DisCo consists of two key designs: (i) a novel *model architecture with disentangled control* for improved faithfulness and 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 propose an organic integration of conditions with cross-attention and ControlNet. Specifically, we substitute the text condition in T2I diffusion model with the CLIP image embeddings of the human subject, which is incorporated via the cross-attention modules of U-Net; while the background and human pose conditions are fed into two separate ControlNet branches. By disentangling the control from all three conditions, DisCo can not only achieve fidelity in human foregrounds and backgrounds but also enable arbitrary compositionality of human subjects, backgrounds, and dance-moves (as shown in Figure 1).

Human Attribute Pre-training (Section 3.3): We design a proxy task in which the model conditions on the separate foreground and background areas and must reconstruct the complete image. In this way, the model learns to 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 leverage 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 define a new problem setting of generating real-world dance content, referring human dance generation, to facilitate its potential application in the production of user-specific short videos.
- To address this problem, we propose DisCo framework with (i) a novel model architecture for disentangled control to ensure faithfulness and compositionality in generation; and (ii) human attribute pre-training to improve generalizability to unseen humans.
- We conduct a broad variety of evaluations and applications to demonstrate the effectiveness of DisCo, including human image editing and human dance video synthesis.

2 Related Work

Diffusion Models for Controllable Image/Video Generation. Different from GANs [15, 4, 28], recent diffusion probabilistic models [58, 20, 9, 59, 60] trained with totally new objectives and pipeline 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 [47, 44, 21, 51, 72]. Among these methods, Stable Diffusion [48] (SD) stands as the representative work to date, with high efficiency and competitive quality via diffusion over the latent space. For better controllability, ControlNet [73] introduces additional control mechanisms into SD beyond texts, such as sketch, human skeleton, and segmentation map. Compared to text-to-image synthesis, text-to-video synthesis [19, 57, 10, 71, 30], is relative new and still remains challenging, due to (i) lack of well-annotated video-text data, hence expensive to scale; (ii) difficulties in modeling temporal consistency, to ensure realistic actions, appearance and scenes; and (iii) lack of compositional understanding for human-object or object-object interaction generation. Thus, controllable video generation methods stem from pre-trained text-to-image models, especially SD. Follow-your-pose [39] tries to introduce pose condition into video generation with image-pose pairs and pose-free videos. Text2Video-Zero [30] propose to modify the pre-trained T2I model to enable the zero-shot human video generation. In this work, we look into a more challenging setting of conditional human image/video synthesis, specifically referring human dance generation, which requires precise control of both human attributes (such as identity, clothing, makeup, hairstyle, *etc.*) and the dance-moves (poses).

Human Dance Synthesis. Early work on this task includes video-to-video synthesis [67, 66, 12, 7], still image animation [56, 70, 22, 55, 68, 75, 3, 40] and motion transfer [77, 54, 33, 61, 13]. Nevertheless, these lines of 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 prediction, motion representation and occlusion map generation. The advances of diffusion models [48] greatly simplify the training of such generative models, inspiring follow-up diffusion models tailored for human dance generation. For example, [32] augments the video generation process with a set of user-specified strokes, and [43] synthesizes an optical flow sequence with the given text prompt for action generation. However, these methods require a separate motion prediction module and struggle to precisely control the human pose. DreamPose [27] 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 define the problem setting for **referring human dance generation**. 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 generated 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 & ControlNet

Latent Diffusion Models (LDM) is a type of diffusion model that operates in the encoded latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, therefore saving the computational and time complexity. An exemplary LDM is the popular Stable Diffusion (SD) [48] which consists an autoencoder VQ-VAE [65] and a time-conditioned U-Net [49] for noise estimation. Since SD is designed for T2I, a CLIP ViT-L/14 text encoder [46] 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 [48]:

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

where ϵ_θ represents the trainable modules, containing 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 [59]) to generate z_0 and pass it to the decoder $\mathcal{D}$ towards the final image.

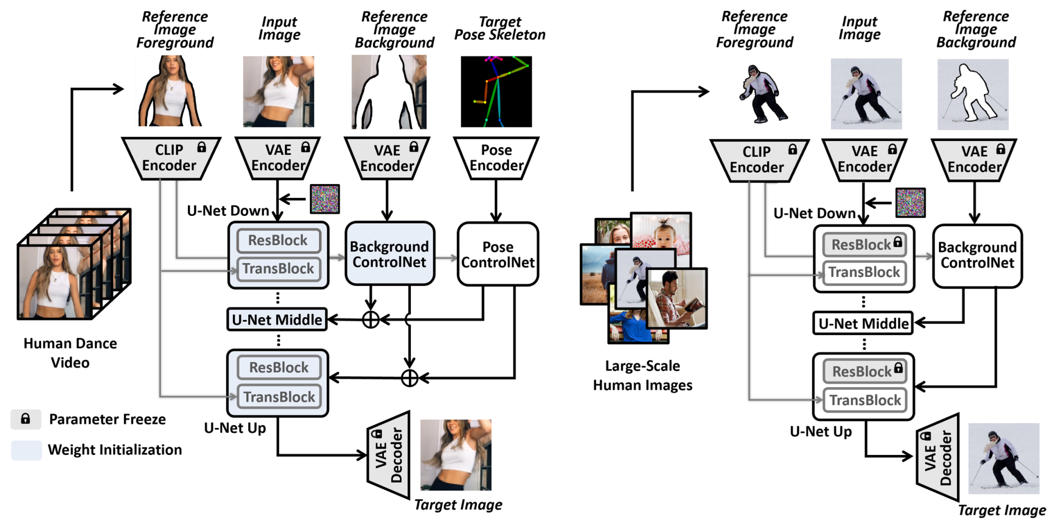
ControlNet [73], built upon SD, manipulates the input to the intermediate layers of the U-Net in SD so as to further control the overall behavior of SD. Specifically, it creates a trainable copy of the U-Net down/middle blocks and adds an additional "zero convolution" layer, *i.e.*, 1×1 convolution layer with both weight and bias initialized with zeros. 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} captured by cross-attention modules in SD, ControlNet is trained with an additional external condition vector c which can be many types of condition, such as edge map, pose, depth map and segmentation map. Training on a specific domain of c results in an effective conditional T2I generation.

3.2 Model Architecture with Disentangled Control

The direct application of ControlNet to referring human dance generation presents challenges due to the missing of reference human image condition, which is critical for keeping the human identity and attribute consistent in the synthesized images. Recent explorations in image variations [1] replace the CLIP text embedding with the CLIP image embedding as the SD condition, which can retain some high-level semantics from the reference image. Nevertheless, the geometric/structural control onto the generated image is still missing.

Taking 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, accommodating any combination of human foreground, pose, and background (Figure 1c). Specifically, given a reference human image, we can first utilize an existing human matting method (*e.g.*, SAM [31, 37]) 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.

Referring Foreground via Cross Attention. To help model easily adapt to the CLIP image feature space, we first use the pre-trained image variation latent diffusion model [1] for the U-Net parameter



(a) Model Architecture with Disentangled Control (b) Human Attribute Pre-training

Figure 2: The proposed DISCO framework for referring human dance generation.

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 $c_{\text{text}} \in \mathbb{R}^{l \times d}$ 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 with 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 ContorlNet branch follows the original ControlNet. As we replace the text condition with 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 latent diffusion modeling objective [48]:

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

where ϵ_θ, τ_θ and μ_θ are the trainable network modules. Specifically, ϵ_θ contains the U-Net architecture composed of the convolution (ResBlock) and self-/cross-attention (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.

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, 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 with DISCO, human attribute pre-training, to improve the generalizability and the faithfulness in generation when encountering unseen human subjects. Rather than directly retrieving and constructing the high-quality human dance video dataset, we explore a much more efficient alternative approach, namely *Human Attribute Pre-training*, to learn diverse human attributes from large-scale human images. Figure 2b shows the details. 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:

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

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

Upon finishing, 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 [73], for human dance generation fine-tuning.

4 Experiments

4.1 Experimental Setup

We train the models on the public TikTok dataset [25] 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 [31] and OpenPose [6] on each frame to infer the human subject mask for separating the foreground from the background 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. 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] [25], COCO [35], SHHQ [11], DeepFashion2 [14], LAION [52]). We first run Grounded-SAM [31] 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 for pre-training. All pre-training experiments are conducted on 4x8 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 Stable Diffusion Image Variations [1]. The ControlNet branches are initialized with the same weight as the U-Net model, except for the zero-convolution layers, following [73]. After human attribute pre-training, we initialize the U-Net and ControlNet branch for background control 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 controlability 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 3 for **human image editing**: (i) *Human Subject / Pose Re-targeting*: the model have been exposed to training instances of the human subject or the ones of 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

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

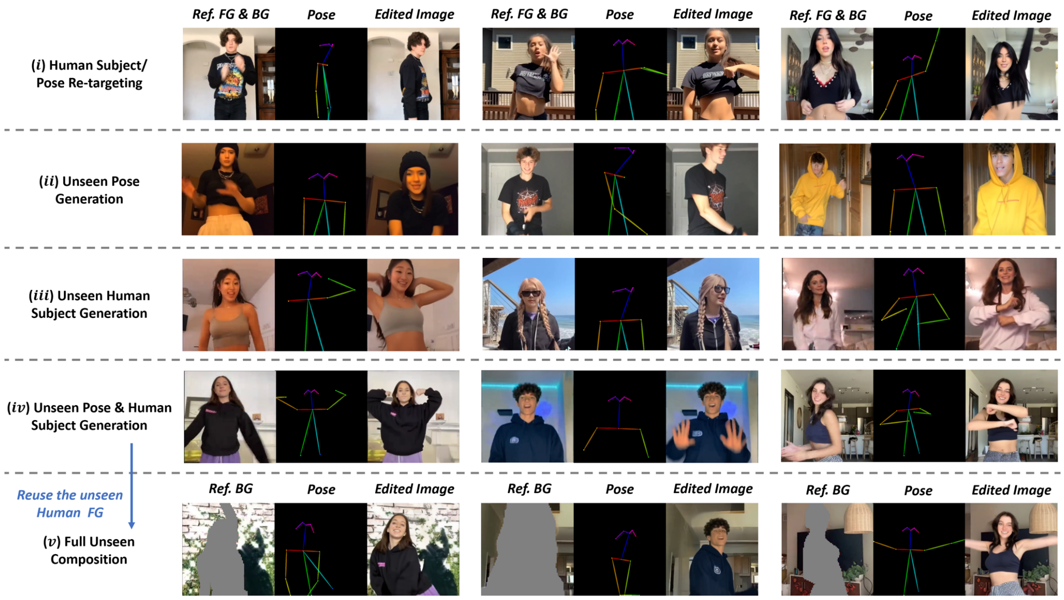


Figure 3: Visualizations of 5 representative scenarios for **human image editing** (best viewed when zoomed-in).

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 Composition*: sampling a novel background from another unseen image/video based on the "unseen pose & human subject generation". Examples in Figure 3 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 4 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 [24, 71, 27] for subject-specific fine-tuning on one or multiple images of the same human subject. We leave discussions and results of this setting to Appendix.

4.3 Main Results

We provide quantitative and qualitative comparisons against DreamPose [27], an image-to-video model designed for fashion domain with densepose control, which is the most relevant work to ours. DreamPose replaces the CLIP text feature with the image embedding with a dual CLIP-VAE encoder and adapter module. Instead of adopting ControlNet, it utilizes a sequence of denseposes [16] as the pose condition. It is worth noting that the TikTok dancing videos we evaluate on are all real-world user-generated content, which are much more complicated than those in existing dancing video datasets [67, 38, 34, 7, 27, 26], with clean background, same/similar clothing or fixed camera angles.

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 [18], SSIM [69], LISPIS [74], PSNR [23] and L1; while for videos, we concatenate every consecutive 16 frames to form a sample, to report FID-VID [2] and FVD [63]. 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 ~ 440. Not surprisingly, classifier-free guidance gives additional advantages to the generation quality of DisCo. The substantial performance gain against the recent SOTA model DreamPose evidently demonstrates

Table 1: Quantitative comparisons of DisCo with the recent SOTA method DreamPose [27]. "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.

Method	Image					Video	
	FID ↓	**SSIM** ↑	**PSNR** ↑	**LISPIS** ↓	**L1** ↓	**FID-VID** ↓	**FVD** ↓
DreamPose [27]	79.46	0.509	28.04	0.450	6.91E-04	73.42	781.88
DreamPose [27] (CFG)	72.62	0.511	28.11	0.442	6.88E-04	53.36	671.50
DisCo (w/o HAP)	61.06	0.631	28.78	0.317	4.46E-04	32.56	555.44
DisCo (w/. HAP)	38.19	0.663	**29.33**	0.291	3.69E-04	20.24	439.94
DisCo (w/. HAP, CFG)	30.75	0.668	29.03	0.292	3.78E-04	18.86	393.34
DisCo † (w/. HAP, CFG)	**28.31**	**0.674**	29.15	**0.285**	**3.69E-04**	**15.77**	**348.04**

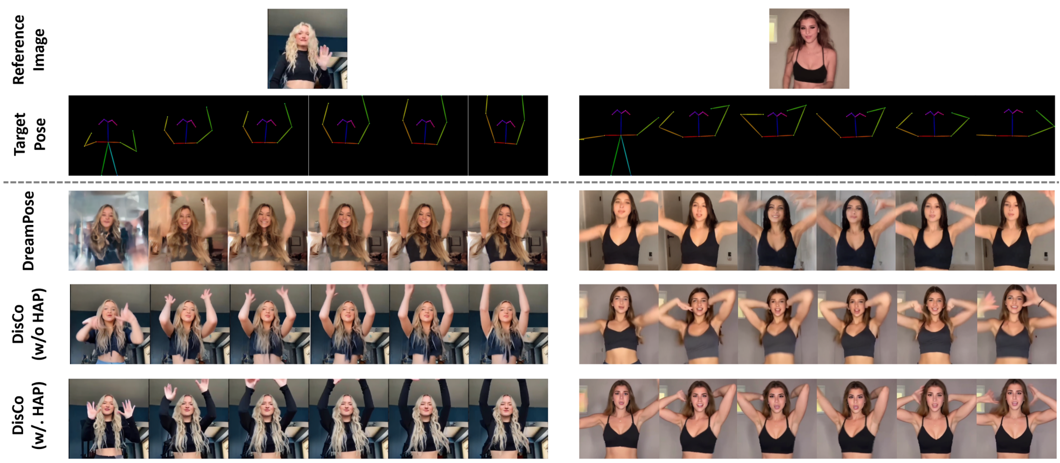


Figure 4: Qualitative comparison between our DisCo (w/ or w/o HAP) and DreamPose [27] on **referring human dance video generation** with the input of a reference image and a sequence of target poses. Note that the reference image and target poses are from the testing split, where the human subjects, backgrounds, poses are not available during the model training. Best viewed when zoomed-in.

the superiority of DisCo. 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.

Qualitative Comparison. We qualitatively compare DisCo to DreamPose in Figure 4. DreamPose obviously suffers from inconsistent human attribute and unstable background. Without HAP, DisCo can already reconstruct the coarse-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.

4.4 Ablation Study

Architecture Design. Table 2 quantitatively analyzes the impact of different architecture designs in DisCo. First, to ablate the control with reference image, we observe that either ControlNet or the cross-attention module struggle haodto handle the control of the whole reference image without 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 5 show that such architecture still struggles to maintain the consistency of human attributes and the stability of the background. For the encoding of reference foreground, DisCo with

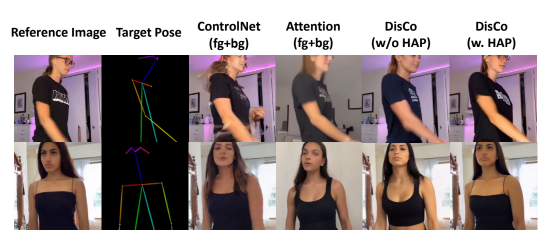


Figure 5: The qualitative comparison between different architecture designs.

Table 2: 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.

Method	FID ↓	SSIM ↑	PSNR ↑	LISPIS ↓	L1 ↓	FID-VID ↓	FVD ↓
DisCo	61.06	**0.631**	**28.78**	0.317	**4.46E-04**	**32.56**	555.44
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	33.16	**527.74**
Attention (fg+bg)	80.50	0.474	28.01	0.485	7.50E-04	80.76	831.10
Ablation on reference foreground encoding (DisCo setting: CLIP Local)							
CLIP Global	63.92	0.621	28.61	**0.311**	5.00E-04	37.98	562.75
CLIP Local + VAE	**59.74**	0.623	28.52	0.331	4.79E-04	35.06	566.57

Table 3: 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	32.56	555.44
TikTok [25]	90K	50.68	0.648	28.81	0.309	4.27E-04	30.99	569.45
+ COCO [35]	110K	48.89	0.654	28.97	0.303	4.07E-04	26.68	516.89
+ SSHQ [11]	184K	44.13	0.655	29.00	0.300	3.93E-04	25.44	502.04
+ DpFashion2 [14] + LAION [52]	700K	**38.19**	**0.663**	**29.33**	**0.291**	**3.69E-04**	**20.24**	**439.94**

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 [27], which leads to a slight better FID score but worse performance on all other metrics.

Pre-training Data Size. Table 3 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. The final success of HAP comes from two-sides: 1) learning diverse human attributes from large-scale human image data; 2) the "easy-to-hard" training schema, with HAP focusing on reconstructing human images without pose editing, then learning pose control and implicit appearance distortions brought by motion in the fine-tuning stage.

5 Conclusion

Contributions. We revisit human dance synthesis in real-world for a more practical application and define a new problem setting: referring human dance generation with three key properties, *faithfulness*, *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 for user-specific short video content generation.

Limitations. The limitation of DisCo sheds light on potential future directions: (1) incorporating hand keypoints for more fine-grained control of hand pose; (2) explicit temporal modeling to improve temporal consistency, especially when there are large motions; and (3) extending to more complicated scenarios, such as multi-person dance generation and human-object interaction.

Broader Impact. The proposed DisCo can be served as a strong starting point for *real-world* referring human dance generation, showing great potential for various applications, *e.g.*, human image editing and video content creation. We believe that our proposed method will open up new possibilities for human-centric generation and editing, but also requires proper regulations to mitigate potential harm, such as the creation of deceptive content and infringement of human rights.

Appendix - DisCo: Disentangled Control for Referring Human Dance Generation in Real World

This appendix is organized as follows:

- Section A includes comprehensive analysis and comparison between our proposed DisCo and related works.
- Section B demonstrates how DisCo can be readily combined with subject-specific fine-tuning.
- Section C provides more qualitative and quantitative results to supplement the main paper.

A Detailed Discussion on Related Work

We include additional discussions with the related visual-controllable image/video generation methods, especially the more recent diffusion-based models, due to the space limitation of the main paper. To fully (or partly) maintain the visual contents given a reference image/video, existing diffusion-based synthesis methods can be broadly divided into the following two categories based on their immediate applications:

Image/Video Editing Conditioned on Text. The most common approach for preserving specific image information is to edit existing images [41, 17, 5, 29, 62] and videos [71, 36, 45, 53] with text, instead of unconditioned generation solely reliant on text descriptions. For example, Prompt-to-Prompt [17] control the spatial layout and geometry of the generated image by modifying the cross-attention maps of the source image. SDEdit [41] corrupts the images by adding noise and then denoises it for editing. DiffEdit [8] first automatically generates the mask highlighting regions to be edited by probing a diffusion model conditioned on different text prompts, then generates edited image guided by the mask. Another line of work requires parameter fine-tuning with user-provided image(s). For example, UniTune [64] tries to fine-tune the large T2I diffusion model on a single image-text pair to encode the semantics into a rare token. The editing image is generated by conditioning on a text prompt containing such rare token. Similarly, Imagic [29] optimizes the text embedding to reconstruct reference image and then interpolate such embedding for image editing.

For video editing, in addition to the Follow-your-pose [39] and Text2Video-Zero [30] discussed in the main text, Tune-A-Video [71] fine-tunes the SD on a single video to transfer the motion to generate the new video with text-guided subject attributes. Video-P2P [36] and FateZero [45] extend the image-based Prompt-to-Prompt to video data by decoupling the video editing into image inversion and attention map revision. However, all these methods are constrained, especially when the editing of the content cannot be accurately described by the text condition. We notice that a very-recent work Make-A-Protagonist [76] tries to introduce visual clue into video editing to mitigate this issue. However, this approach, while innovative, is still met with limitations. On the one hand, it still struggles to fully retain the fine-grained human appearance and background details; on the other hand, it still requires a specific source video for sample-wise fine-tuning which is labor-intensive and time-consuming. In contrast, our DisCo not only readily facilitates human dance synthesis given any human image, but also significantly improves the faithfulness and compositionality of the synthesis. Furthermore, DisCo can also be regarded as a powerful pre-trained human video synthesis baseline which can be further integrated with various subject-specific fine-tuning techniques (see section B for more details).

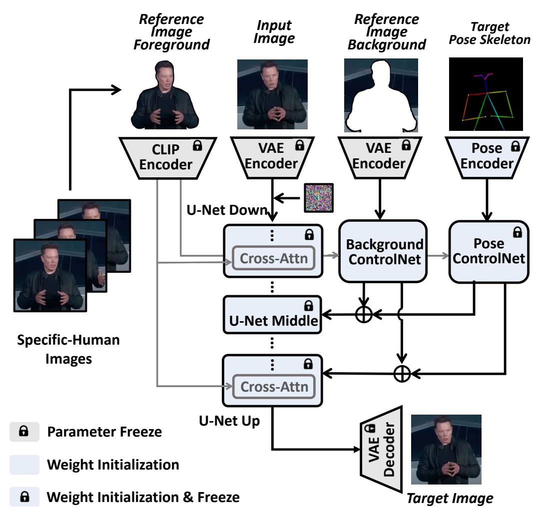


Figure 6: The model architecture for further subject-specific fine-tuning.

Visual Content Variation. For preserving the visual prior, another line of work [1, 47, 10] directly feeds the CLIP image embedding into the diffusion model to achieve image/video variation. However, these approaches struggle to accurately control the degree as well as the area of the variation. To partially mitigate this problem, DreamBooth [50] and DreamMix [42] necessitate multiple images to fine-tune the T2I and T2V models for learning and maintaining a specific visual concept. However, the precise visual manipulation is still missing. In this paper, we propose a disentangled architecture to accurately and fully control the human attribute, background and pose for referring human dance generation.

B Subject-Specific Finetuning

As mentioned in the main paper, our DISCO can be flexibly integrated with existing efficient fine-tuning techniques for even more fine-grained human dance synthesis. This is particularly beneficial when facing out-of-domain reference images, which appear visually different to the TikTok style images. Figure 6 presents the framework for subject-specific fine-tuning, which is easily adapted from the framework presented in the main text (Figure 2a). Rather than utilizing a set of videos of different human subjects for training, subject-specific fine-tuning aims to leverage limited video frames of a specific human subject (*e.g.*, the video of Elon Mask talking about Tesla Model 3 in Figure 6 or even anime in Figure 7) for better dance synthesis. Compared to the standard fine-tuning, we additionally freeze the pose ControlNet branch and most parts of U-Net to avoid over-fitting to the limited poses in the subject-specific training video, only making the background ControlNet branch and the cross-attention layers in U-Net trainable. We also explored the widely-used LoRA [24] for parameter-efficient fine-tuning and observed similar generation results. As this is not the main focus of this paper, we leave other advanced techniques to future explorations along this direction.

Implementation Details. The model weights are initialized with the model fine-tuned on the general TikTok dancing videos. We train the model on 2 NVIDIA V100 GPUs for 500 iterations with learning rate $1e^{-3}$, image size 256×256 and batch size 64. The randomized crop is adopted to avoid over-fitting. The subject-specific training videos range from 3s to 10s, with relatively simple poses.

Qualitative Results. We test the subject-specific fine-tuning on various out-of-domain human subjects, including real-world celebrities and anime characters. After training, we perform the novel video synthesis with an out-of-domain reference image and a random dance pose sequence sampled in TikTok training set. As shown in Figure 7, upon additional fine-tuning, DISCO is able to generate dance videos preserving faithful human attribute and consistent background across an extensive range of poses. This indicates the considerable potential of DISCO to serve as a powerful pre-trained checkpoint.

C Addition Results

C.1 Quantitative Results

We show the full ablation results on architecture design in Table 4. In what follows, we focus on discussing results that are not present in the main text. In the first block of the table, we copy over the results from the full instances of DISCO, with or without HAP on TikTok Dance Dataset for reference.

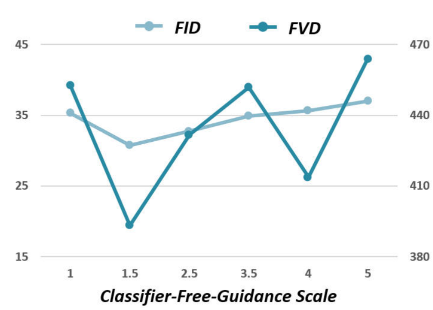


Figure 8: The effect of different classifier-free-guidance scale.

As mentioned in L176-178 of the main text, we propose to use the pre-trained VQ-VAE from SD, instead of four randomly initialized convolution layers in the original ControlNet for encoding the background reference image. In the second block of the table, we ablate this design by comparing two models, (1) "ControlNet (fg+bg)", inserting the control condition of reference image (containing both foreground and background) via a single ControlNet, with VQ-VAE encoding and (2) "ControlNet (fg+bg, no SD-VAE)", inserting the control condition of reference image via a single ControlNet with four randomly initialized convolution layers as the

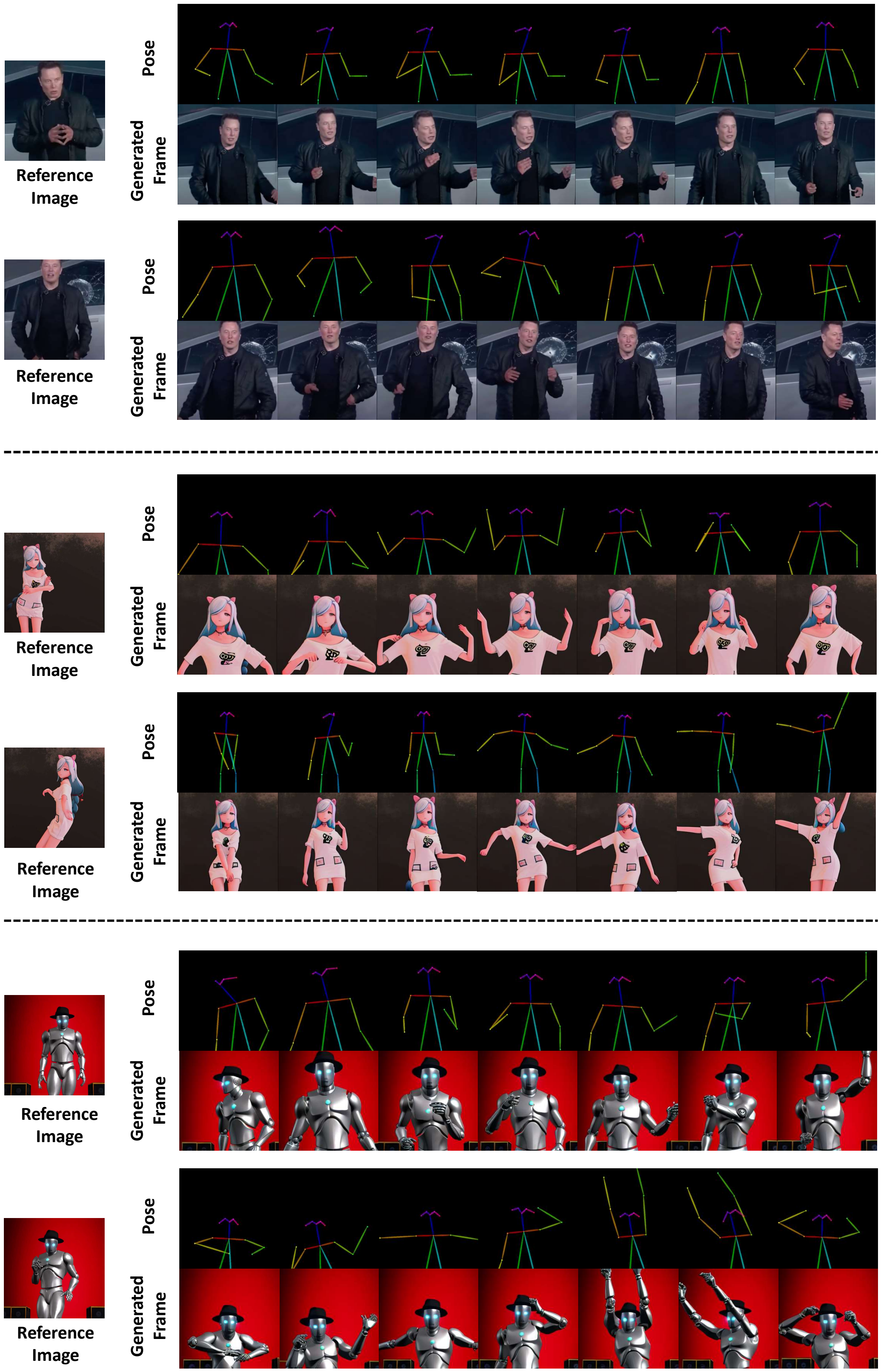


Figure 7: The synthesis frames for out-of-domain human subject after subject-specific fine-tuning guided by the pose sequence extracted from the TikTok dataset.

Table 4: Additional ablation results on architecture designs. “ControlNet (fg+bg)” and “Attention (fg+bg)” in the second block denote inserting the control condition of reference image (containing both foreground and background) via a single ControlNet or cross-attention modules. “HAP w/ pose” denotes adding pose ControlNet path with pose annotation into HAP. “ControlNet-Pose” Init. means initializing pose ControlNet of fine-tuning stage with the pre-trained ControlNet-Pose [73] checkpoint.

Method	**FID** ↓	**SSIM** ↑	**PSNR** ↑	**LISPIS** ↓	**L1** ↓	**FID-VID** ↓	**FVD** ↓
DisCo	61.06	0.631	28.78	0.317	4.46E-04	32.56	555.44
DisCo + TikTok HAP	50.68	0.648	28.81	0.309	4.27E-04	30.99	569.45
Ablation on control mechanism w/ reference image (DisCo setting: ControlNet (bg) + Attention (fg))							
ControlNet (fg+bg, no SD-VAE)	83.53	0.575	28.37	0.411	5.35E-04	47.69	639.11
ControlNet (fg+bg)	65.14	0.600	28.57	0.355	4.83E-04	33.16	527.74
Attention (fg+bg)	80.50	0.474	28.01	0.485	7.50E-04	80.76	831.10
Ablation on HAP w/ pose (DisCo setting: HAP w/o pose)							
TikTok HAP w/ pose	51.84	0.650	28.89	0.307	4.16E-04	29.59	547.14
Ablation on initializing w/ pre-trained ControlNet-Pose [73] (DisCo setting: initialize with U-Net weights)							
ControlNet-Pose Init.	62.18	0.633	28.37	0.320	4.46E-04	31.90	591.28
ControlNet-Pose Init.+TikTok HAP	55.81	0.641	28.69	0.316	4.43E-04	30.29	576.67

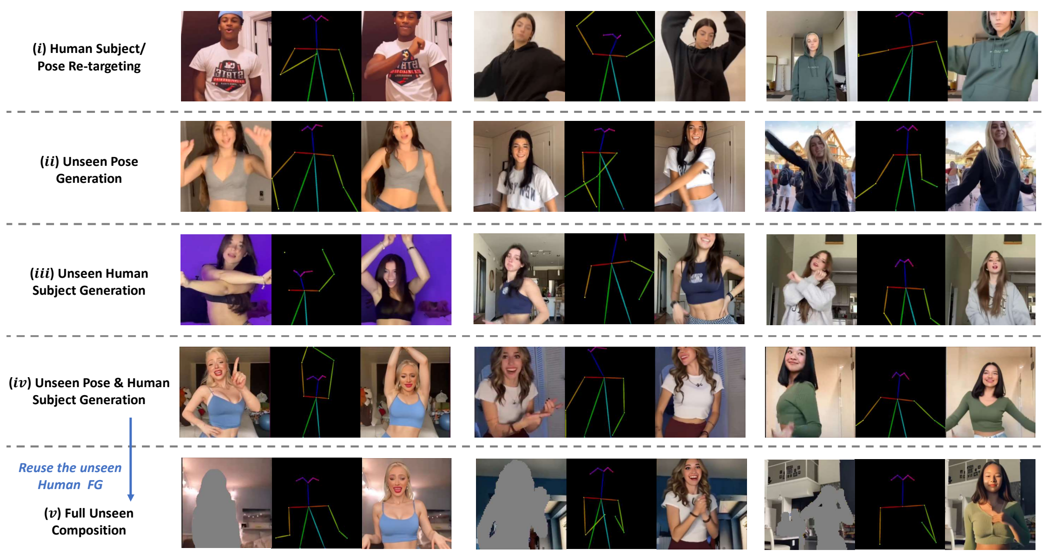


Figure 9: More visualizations for different scenarios of human image editing.

condition encoder. We note that the pre-trained VQ-VAE can produce a more descriptive dense representation of the reference image, contributing to better synthesis results (FID 65.14 v.s 83.59).

In the third block of the table, we investigate whether adding pose condition into human attribute pretraining is beneficial to the downstream performance. We observe that integrating pose into HAP leads in similar results, but requires additional annotation efforts on pose estimation.

Last but not least, we examine on the initialization of the pose ControlNet branch. Specifically, we try to initialize from the pre-trained ControlNet-Pose checkpoint [73] during fine-tuning. The results are shown in the last block of Table 4. Without HAP, the performance is comparable to DisCo, but it gets much worse than DisCo when both are pre-trained with HAP. This is because that the ControlNet-Pose is pre-trained with text condition and can not fully accommodate referring human dance generation with the reference image condition. After HAP, such gap is further enlarged, leading to even worse results. In Figure 8, we show the effect of varying the classifier-free-guidance scale. We can find that scale of 1.5 gives the best quantitative results for both image-wise and video-wise fidelity.

Figure 10: The qualitative comparison between different architecture designs for the video frame generation.

C.2 Qualitative Results

Figure 9 presents additional results from DisCo for different human image editing scenarios. Similar to Figure 3 in the main text, DisCo can handle novel human pose synthesis very well even with change in viewpoints and rotation in the human skeleton. More results for video generation is shown in Figure 11.

Figure 10 compares the video synthesis results with baseline architectures, to supplement Figure 5 of the main text. With a sequence of continuous poses, we can discern more clearly that both ControlNet-only and Attention-only baseline fail to maintain the consistency of human attributes and background, leading to less visually appealing generations than our DisCo.

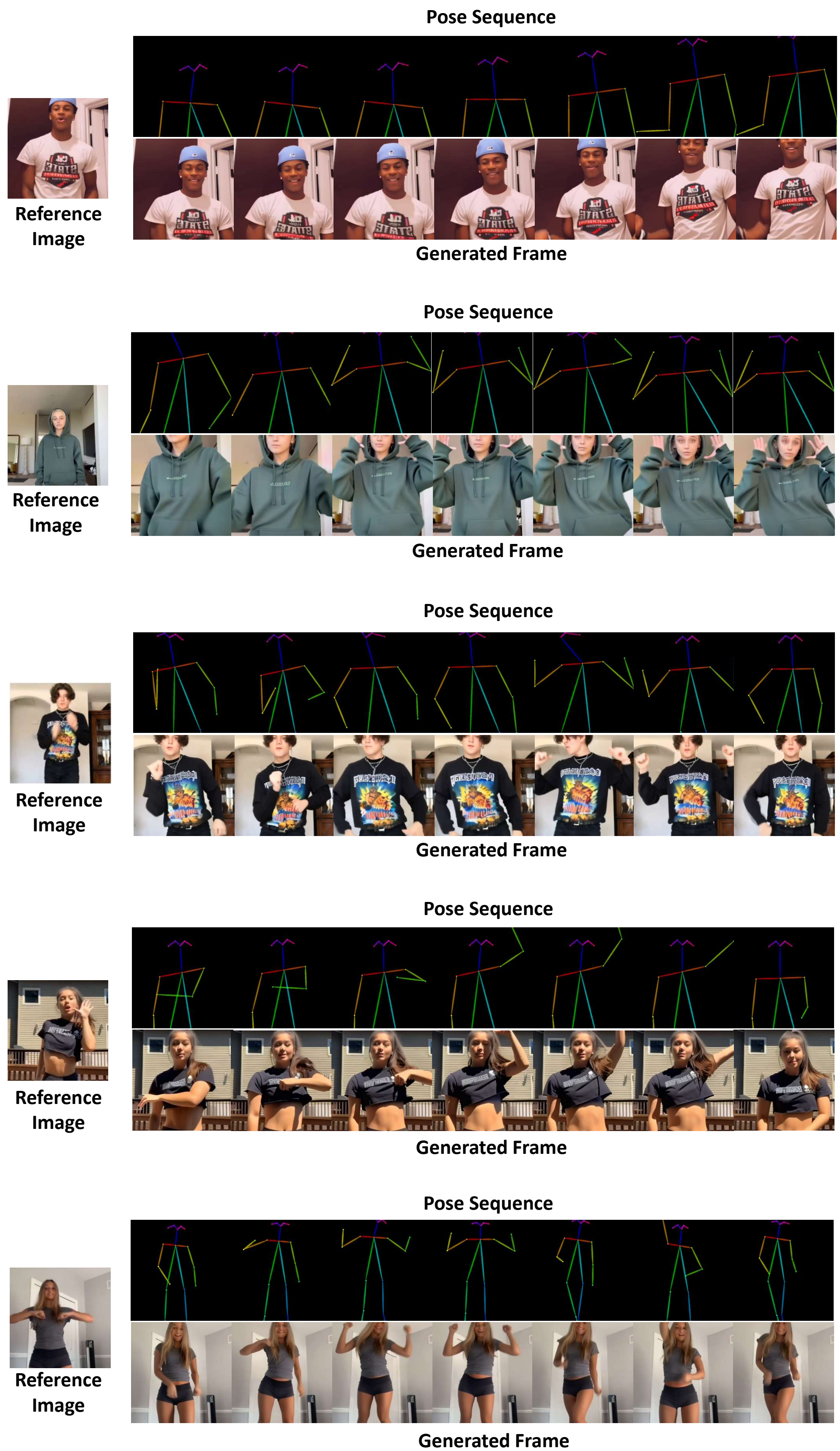


Figure 11: More qualitative examples for video generation.

References

[1] Stable Diffusion Image Variations. https://huggingface.co/lambdalabs/sd-image-variations-diffusers.

[2] 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.

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

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

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

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

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

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

[10] 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.

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

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

[13] Yuying Ge, Yibing Song, Ruimao Zhang, and Ping Luo. Metadance: Few-shot dancing video retargeting via temporal-aware meta-learning. *arXiv preprint arXiv:2201.04851*, 2022.

[14] 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.

[15] 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.

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018.

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[18] 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.

[19] 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.

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

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

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

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

[24] 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. *arXiv preprint arXiv:2106.09685*, 2021.

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

[26] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint arXiv:2304.08483*, 2023.

[27] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[29] 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.

[30] 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.

[31] 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. *arXiv:2304.02643*, 2023.

[32] 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.

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *CVPR*, 2022.

[41] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[42] 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.

[43] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. *arXiv preprint arXiv:2303.13744*, 2023.

[44] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[45] 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.

[46] 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.

[47] 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.

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

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

[50] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[51] 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.

[52] 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.

[53] 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.

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

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

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

[57] 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.

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

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

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

[66] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

[67] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[68] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[69] 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.

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

[71] 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.

[72] 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.

[73] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[74] 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.

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

[76] 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.

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