

Dress&Dance: Dress up and Dance as You Like It

Jun-Kun Chen¹ Aayush Bansal² Minh Phuoc Vo² Yu-Xiong Wang¹

¹University of Illinois Urbana-Champaign ²SpreeAI

{junkun3, yxw}@illinois.edu {aayush.bansal, minh.vo}@spreeai.com

immortalco.github.io/DressAndDance View demo

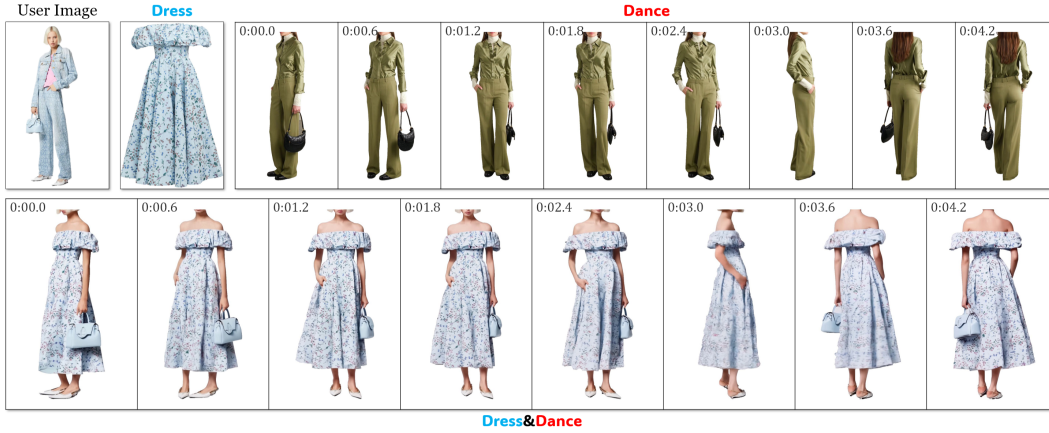


Figure 1. Given a single image of a user, a garment (*dress*) that they would like to wear, and an example video showing how they would like to animate themselves (*dance*) as shown in the first row, our method **Dress&Dance** generates a high-quality 5s video (1152×720 , 24 FPS) of the user wearing the target garment and following the desired motion while maintaining their accessories such as bag and shoes. Video results are provided in our **Suppl. Video (SV, demo.mp4)**.

Abstract

We present **Dress&Dance**, a video diffusion framework for high-resolution virtual try-on with simple motion control and flexible conditioning. Given a single user photo, garment image(s), and a reference motion video, **Dress&Dance** generates a video of the user wearing the target garment(s) while following the reference motion, allowing users to control motion by selecting or swapping motion templates instead of recording personal full-body videos. To handle heterogeneous and changing conditions such as multi-garment try-on, garment transfer, and text refinement, we design *CondNets*, which keep visual conditions in condition-specific layer-wise streams and exchange information with text at every denoising block, together with garment-aware target steering for accurate placement. We also introduce a high-resolution dataset and benchmark, combine video-image hybrid training with synthetic unpaired supervision, and use a progressive curriculum over resolution and frame rate. **Dress&Dance** generates 5-second videos at 24 FPS and 1152×720 resolution; evaluations show strong garment and temporal fidelity.

1. Introduction

In a fitting room, people turn, lift their arms, and take steps to inspect drape, seams, prints, textures, and self-occlusions that a static image cannot show. A practical virtual try-on system should therefore support (i) *user-friendly, low-effort motion control*: ideally, a user can try on a garment from a single photo and simply choose a motion template, instead of recording and uploading a full-body video. At the same time, users expect (ii) *flexibility in how they try on clothes* – such as simultaneously trying on multiple garments, transferring clothes from another person’s photo, or adjusting accessories with text – which calls for a model that can gracefully handle changing combinations of conditions. Finally, for virtual try-on to be truly useful, it must generate videos *natively* at (iii) *high resolution and high frame rate*, so that garment details remain visible and dynamics remain faithful. However, existing video try-on systems [16, 25, 35, 53, 74, 80] often fall short on these fronts: they commonly require a full-body video for both appearance and motion, rely on input-specific conditioning designs that are difficult to extend, and operate at low resolution and frames per second (FPS). As a result, there remains a substantial gap between practical user needs and



Figure 2. Dress&Dance allows for flexible virtual try-on: (a) complicated dancing motions, (b) simultaneous try-on of tops and bottoms, and (c) transferring a garment from another user.

what current models can deliver. Dress&Dance generates each preview in 5–10 minutes.

To bridge this gap, we formulate a user-centric video try-on setting that *decouples user appearance from motion control*. Specifically, the user provides a single photo of themselves, one or more garment images (typically product photos), and a motion reference video, with an optional text prompt for lightweight refinements (e.g., adjusting accessories or minor garment edits). The goal is to synthesize a try-on video where the user appears to wear the target garment(s) while faithfully following the motion from the reference. Crucially, the reference video can be chosen from a template library or sourced from any example clip, so users can control motion by simply selecting a template, rather than recording a full-body personal video. We instantiate this setting with **Dress&Dance**: *dress* – what to wear via garment images, and *dance* – how to move via the motion reference, as illustrated in Fig. 1.

Our Dress&Dance formulates virtual try-on as an inherently *multi-condition generation* problem: it must jointly condition on user images, target garment images, motion reference videos, and optional text prompts – and, in practice, users further expect flexible control that spans various try-on modes (e.g., multi-garment try-on, garment transfer, and text description). A core technical obstacle is therefore the *simultaneous* handling of these heterogeneous, multi-modal inputs in a single end-to-end model. Prior try-on approaches often adopt specific conditioning strategies for each modality, resulting in input-specific architectures that are difficult to extend. This challenge is compounded by the fact that these modalities are *rarely pixelwise-aligned* with the generated video – garment inputs are commonly flat product photos rather than images aligned to the user’s pose – making aligned control designs such as ControlNet [89] ill-suited in our setting. To address this, we design *Cond-*

Nets, which keep heterogeneous visual conditions in separate layer-wise streams, exchange information with text at each denoising block, and expose their tokens to the main denoising attention; related conditioning mechanisms are discussed in Sec. 2 and Suppl. Sec. B.2. During adaptation of a pretrained text-to-video model, we observed that it could underuse the geometrically harder garment input, producing semantically similar garments with insufficient instance fidelity. We counter this with *garment-aware target steering*, a training strategy that progressively increases the geometric difficulty of garment supervision. Together, these designs provide a unified interface that supports diverse try-on modes within one framework, as summarized in Fig. 2.

In addition to architectural improvements, native high-resolution, high-frame-rate try-on requires training signals that preserve fine-grained details; Fig. 6 shows how low-resolution inputs and outputs attenuate garment and user details. We therefore construct a new *high-resolution* training dataset and, in tandem, a high-resolution benchmark to evaluate garment fidelity under realistic motion. Beyond videos, we further improve data scaling by introducing *video-image hybrid batches*, which mix a video try-on batch with several image (as single-frame video) try-on batches during training to better leverage the large-scale image dataset. To mitigate the scarcity of unpaired supervision data (i.e., the same person in comparable poses wearing different garments), we also propose *synthetic triplet generation* that uses pretrained virtual try-on models to produce unpaired triplets for training, thereby reducing the mismatch between training and inference. For optimization, we design a *multi-stage training curriculum* that progressively increases condition variety and spatial resolution, enabling efficient scale-up to high-resolution.

Our contributions are threefold: (1) We formulate a

user-friendly video try-on setting where users provide a single photo, garment image(s), and a motion reference video (with optional text), enabling low-effort motion control. (2) We design CondNets to organize heterogeneous visual conditions as separate, text-aware layer-wise streams, together with garment-aware target steering for accurate garment placement under text prompts and various try-on modes. (3) We build a high-resolution training dataset and benchmark, and introduce video-image hybrid batches, synthetic triplet generation, and a multi-stage training curriculum to enable native high-resolution, high-FPS video synthesis. Our **Dress&Dance** achieves high-fidelity video try-on at 1152×720 resolution and 24 FPS across diverse try-on modes, with strong garment fidelity and temporal quality in the evaluated comparisons.

2. Related Work

Image-to-Image Virtual Try-On. The dominant line of virtual try-on research focuses on generating a single image output from a user image and one or more garment references [2, 5, 7–9, 14, 17, 22, 24, 31, 36, 39, 40, 43, 44, 46, 51, 52, 72, 73, 76, 77, 79, 82–86, 88, 99, 100]. Many adopt a two-step design: garment warping followed by synthesis using GANs [18] or diffusion models [56, 57, 61]. While effective, these methods are trained on *paired data*: data pairs consisting of a garment and an image of someone wearing the *same* garment. When training, they wipe the garment information from the image by extracting intermediates (*e.g.* agnostic masks, dense poses, *etc.*), which naturally introduces information leakage and propagates artifacts. Some methods aim to jointly modify poses and garments [12, 23, 59, 63, 64], but they typically lack temporal smoothness when extended to sequences. Our approach addresses these limitations by synthesizing unpaired training triplets, aligning training and inference settings, and avoiding intermediate representations required by paired training. **Video Virtual Try-On (VVT).** Moving beyond single images, video-based try-on aims to generate temporally coherent sequences conditioned on garments. Early approaches [13, 32, 38, 96] rely on GANs and frame warping to approximate motion. With the rise of diffusion models, subsequent works explore both image-to-video and video-to-video formulations. A straightforward strategy is to first synthesize a try-on image and then animate it using an image-to-video module [1, 4, 20, 21, 26, 28, 34, 66, 67, 70, 81, 92, 101]. However, such modular designs suffer from error accumulation across stages. More integrated methods [10, 16, 25, 35, 53, 74, 80, 95] employ end-to-end video generation to preserve garment details while enforcing temporal consistency. Recent DiT-based systems include CatV2TON [10], which temporally concatenates person and garment inputs, and MagicTryOn [42], which uses full spatiotemporal attention and coarse-to-fine gar-

ment cues. DreamVVT [102] uses keyframe try-on guidance in a two-stage pipeline; DPIDM [41] models human–garment pose interaction; and RealVVT [45] uses pose guidance with a video foundation model. VFR [6] generates arbitrarily long try-on videos from a single image, while SwiftTry [53] supports long-video inference through ShiftCaching over overlapping clips. These systems edit a person video and thereby inherit its source motion, whereas our setting composes a single user image with an independently selected motion reference.

Despite this progress, existing VVT systems still face several practical challenges. First, they struggle to effectively handle the heterogeneous nature of multi-modal inputs. Existing approaches to image or video try-on frequently design disparate conditioning strategies for each modality, leading to complex and non-scalable architectures. For instance, some use channel concatenation for pose maps while relying on attention for garments. More importantly, general conditioning models like ControlNet [89] were developed for UNet [62] feature maps and aligned spatial controls, and therefore do not directly address diffusion transformers (DiTs) [56] or pixelwise-unaligned flat garment inputs, whose placement requires complex, non-local transformations (more details in Suppl. Sec. B.1). Reference- and token-based mechanisms [28, 29, 69, 93] establish appearance-feature reuse, token concatenation, parameter reuse, and LoRA-based control; VACE [33] and Phantom [48] provide related video-conditioning designs. CondNet combines these mechanisms in a VVT-specific organization of separate layer-wise visual streams with per-block text exchange; Suppl. Tab. B.1 provides a detailed comparison.

Second, due to the scarcity of unpaired training data (the same person wearing different garments in the same pose), conventional methods rely on multiple intermediate representations to eliminate the original garment’s information. This process can lead to information loss, artifacts from imprecise masks, and a mismatch between the training and inference settings. Our synthetic triplets instead change the user-image garment condition while retaining a real video as the supervision target.

Image/Video-to-Video Generation. See Suppl. Sec. A.

3. Dress&Dance: Methodology

High-fidelity (1152×720 resolution, 24 frame-per-second) video virtual try-on requires synthesizing a dynamic video from a complex set of heterogeneous, pixelwise-unaligned inputs: one or more flat garment images g , a static user image $I_{u,g'}$ of user u wearing some other garments g' , a motion reference video m , and an optional text prompt p . At inference, the model maps $(g, I_{u,g'}, \text{skeleton}(m), p)$ to $\hat{V}_{u,g}$, a video of user u wearing g and following the motion in m , with text p providing optional refinements. The skele-



Figure 3. Our model inputs a user image, a desired garment, a reference video, and an optional text description. (a) Image-try-on-then-animation pipelines propagate errors from the intermediate image. (b) State-of-the-art commercial models, such as Kling [37], are able to perform try-on but still struggle in capturing the nuanced motion description, as *text* alone is not sufficient for motion description. When the hand occludes part of the garment in the intermediate try-on image, the separate video generator no longer receives the product image and cannot restore that pattern after the hand moves. (c) Our **Dress&Dance** generates high-quality virtual try-on results that faithfully preserve both garment and user appearance with reference-guided motion, even after the hand moves away from the garment. Video results are provided at [SV3:19](#).

ton contains body keypoints, orientation, and timing, but no RGB appearance, garment texture or color, or background. The core challenge lies in how to effectively and robustly integrate these distinct modalities and faithfully reflect them in the final high-resolution video.

We introduce **Dress&Dance** (Fig. 4), a video diffusion framework for high-fidelity video virtual try-on. Dress&Dance integrates heterogeneous controls – text, motion, garment, and user appearance – via *CondNets*, our conditioning networks, and improves garment controllability with *Garment-Aware Target Steering (GATS)*. To scale training to high-resolution, high-FPS generation under realistic motion, we curate high-resolution supervision, augment data with video-image hybrid batches and synthetic unpaired triplets, and adopt a multi-stage training curriculum that progressively increases *spatial* resolution while preserving full-length clips, followed by an FPS refinement stage to achieve high-resolution, high-FPS generation.

3.1. CondNet for Heterogeneous Try-On Conditions

Generating a high-fidelity virtual try-on video from multiple, distinct condition inputs necessitates a conditioning mechanism that can effectively integrate them. To address this, we introduce *CondNet*, a conditioning network that leverages attention to organize heterogeneous inputs as condition-specific, layer-wise streams before fusing their tokens with the main denoising sequence. CondNet pro-

vides a VVT-specific organization for heterogeneous high-resolution conditions.

Layer-wise Condition Streams. Our framework is built upon the Diffusion Transformer (DiT) architecture [56], which natively treats video and text as sequences of tokens and fuses them via attention. We extend this mechanism by converting each conditional input into a dedicated attention sequence that is incorporated into the DiT’s denoising process. As visualized in Fig. 4 and [Suppl. Fig. E.1](#), CondNets are implemented as a parameter-shared copy of the DiT’s layers, augmented with LoRA adaptors [27] for each condition. Each CondNet processes one conditional input, such as a user image, a garment image, or a motion skeleton video, and generates a corresponding attention sequence at each layer. As each CondNet mirrors the entire chain of DiT blocks, the token features of condition sequences obtained in each DiT block are aligned in their original feature space. Therefore, these conditional sequences can be directly concatenated with the video and text sequences within each DiT block for attention fusion. At each block, the CondNet updates its condition sequence through cross-attention with text; the resulting condition keys/values join the main video/text attention, which updates only the video and text tokens. This design allows the model to implicitly connect every pixel of the conditions to the output video frames through cross-attention, which supports the trans-

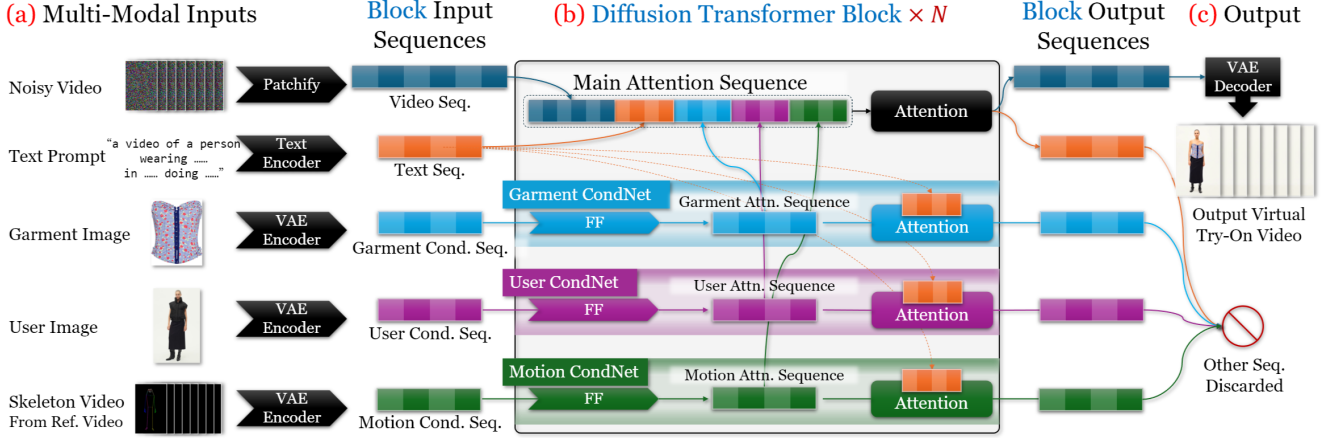


Figure 4. **Dress&Dance** accounts for multi-modal conditioning through our unified CondNets mechanism based on attention. The “Block Input Sequences” and “Block Output Sequences” represent the sequences passed recursively from one block to the next. At the final stage, only the video sequence is decoded as the generated try-on video, while other sequences are discarded. More details are in [Suppl. Sec. E](#).

fer of fine-grained garment and user details. For motion guidance, we preprocess the reference video to extract a sparse skeleton sequence, which significantly reduces computational overhead by focusing only on the human body parts. A mechanism-level comparison with related token-, adapter-, and reference-conditioning approaches is provided in [Suppl. Tab. B.1](#).

Observed Text Dominance in Garment Conditioning.

When new visual conditions were introduced into the pre-trained text-to-video model, we observed that optimization could preserve coarse text/video behavior while underusing the geometrically harder flat-garment input, producing semantically similar garments with insufficient instance fidelity. We refer to this observed training behavior as text dominance. Fitting a flat garment to a moving person is particularly demanding because it requires placement across poses and plausible generation of unseen views. Preliminary adaptations confirmed that normal video generation required the pretrained text pathway; fixed-prompt training left garment instance fidelity unchanged, motivating garment-focused supervision.

Garment-Aware Target Steering. To help the model bridge the large geometric gap from a flat garment image to its warped, posed, and occluded appearance in video, we design a garment-aware target steering strategy. We train with specialized synthetic target videos and a content-isolation operation that blends irrelevant regions as α original + $(1 - \alpha)$ gray with the fixed setting $\alpha = 0.1$ (Fig. 5). This strongly suppresses non-garment appearance while retaining coarse body and spatial context. Progressive targets divide the flat-to-posed correspondence into instance copying, spatial placement, and posed deformation before the complete try-on task.

The three phases are visualized in Fig. 5. In Phase 1, we construct the target video by placing the flat garment image as an overlay *in front of* each ground-truth frame,

while de-emphasizing the non-occluded surrounding background. This yields a direct copying objective that explicitly trains the model to preserve the garment’s instance-specific appearance. In Phase 2, we position the same flat garment overlay at the garment’s correct occurrence location and size (estimated using a segmentation model [75]) within each ground-truth frame, encouraging spatial association between the garment and the human body with controlled difficulty. In Phase 3, the target video uses the garment segmented from the ground-truth frame, with all other parts de-emphasized, teaching full garment placement under pose, deformation, and occlusion. This guidance converges quickly (approximately one to two hours per phase, measured as wall-clock time on 8 H100 GPUs).

3.2. Dataset and Data Processing

High-Resolution Video Dataset and Benchmark. Since existing video virtual try-on datasets [13, 16, 53] are limited to lower resolutions ($\leq 720 \times 540$), we curate two *high-resolution* ($\geq 1080P$) video datasets for training and evaluation. (1) The *Internet video dataset* is constructed by crawling publicly available fashion videos paired with flat garment images, resulting in approximately 80K garment-video pairs $(g, V_{u,g})$. (2) The *captured video dataset* is created by hiring 183 human models to record try-on videos for various sets of garments; each model records videos for around 100 garment sets, providing rich coverage of the same identity across different garments. As an empirical motivation for high-resolution supervision, Fig. 6 shows that fine-grained garment and user details are already severely attenuated in low-resolution inputs and further lost in low-resolution outputs. Alongside the training data, we construct a *high-resolution benchmark* to evaluate garment fidelity under realistic motion; we split each dataset into training and evaluation subsets, and report results on held-out users, garments, and motion references.



Figure 5. Our three-phase Garment-Aware Target Steering progressively changes the supervision target from flat-garment copying, through target placement, to posed and occluded garment synthesis. Content isolation suppresses unrelated appearance while retaining coarse spatial context.

Synthetic triplets are confined to training-data construction; all evaluation conditions are real held-out inputs.

Video-Image Hybrid Batches. Because image data are much easier to obtain than video data, we additionally collect a large-scale Internet image try-on dataset comprising approximately 4M image pairs $(g, I_{u,g})$. To leverage such rich data during video training, we define a *video-image hybrid batch* as a mini-batch that contains (i) one video try-on task (41-frame clips) and (ii) $k = 4$ image try-on tasks, where each image pair $(g, I_{u,g})$ is treated as a *single-frame video* to share the same input/output interface and losses. This strategy enables significantly augmented data diversity with minimal overhead compared to video-only training. We use the fixed setting $k = 4$ to add image diversity with modest compute overhead.

Synthetic Unpaired Triplet Generation. Our supervised objective is defined on *unpaired triplets* $(g, I_{u,g'}, V_{u,g})$, where g is the *target* garment, $V_{u,g}$ is a target try-on video of user u wearing g , and $I_{u,g'}$ is a user image of the same user wearing a *different* garment $g' \neq g$ (used as the appearance condition at inference). Let $\text{skeleton}(\cdot)$ denote the operator that extracts a motion skeleton sequence from a video. Given a triplet, we train Dress&Dance with $(g, I_{u,g'}, \text{skeleton}(V_{u,g}))$ as input and $V_{u,g}$ as the supervision target (with the optional text condition omitted from the notation). For Internet-video and image-pair samples, we construct the required unpaired appearance condition synthetically using off-the-shelf pre-trained image try-on models [76, 79, 85]. Let $I_{u,g}^t$ denote a randomly sampled frame from $V_{u,g}$, and let $\mathcal{T}(\cdot)$ denote a pre-trained image try-on model. We sample a random garment g' and generate a synthetic unpaired user image $\tilde{I}_{u,g'} = \mathcal{T}(g', I_{u,g}^t)$, which yields the synthetic triplet $(g, \tilde{I}_{u,g'}, V_{u,g})$. In practice, we use *three* complementary try-on models [76, 79, 85] and randomly select one as $\mathcal{T}$ for each synthesis. This randomization increases condition diversity and reduces reliance on artifacts from any single synthesis model. The synthesized image $\tilde{I}_{u,g'}$ serves only as the appearance condition, while supervision remains the real video $V_{u,g}$. Since $\tilde{I}_{u,g'}$ has the same role as the inference-time appearance condition $I_{u,g'}$, this synthesis avoids extracting an agnostic intermediate from the target video.

3.3. Training

Scaling up Conditions and Resolution with Curriculum Learning. Generating high-resolution, high-frame-

rate videos from multi-modal conditions with diffusion models requires synthesizing a large number of tokens, making naive training at full resolution costly. We therefore adopt a multi-stage curriculum. We initialize Dress&Dance from CogVideoX-5B [87], a pre-trained text-to-video diffusion model that generates $T=41$ -frame clips at 8 frames per second (FPS). In this setting, aggressively injecting new try-on controls can degrade the foundation model’s pre-trained spatial-temporal priors (e.g., text understanding and long-range consistency), *i.e.*, *catastrophic forgetting*. Our curriculum progressively increases task complexity along two axes: *conditions* and *spatial resolution* (with FPS handled by a final refinement stage).

Condition scaling. We define the conditions $\mathcal{C} = \{\text{text}, \text{motion}, \text{garment}, \text{user}\}$ and introduce these signals stage by stage, starting from the text-to-video base and then progressively adding motion, garment, and user image CondNets and inputs. This staged injection stabilizes multi-condition adaptation. Each stage adds one difficulty at a time: condition correspondences, spatial detail, then 24-FPS refinement. Garment-aware target steering (Sec. 3.1) is used when adding garment condition support.

Resolution and temporal scaling. We progressively scale the *spatial* video resolution: Dress&Dance first learns the controlled generation task at lower resolutions (e.g., 768×480), and is then trained at the final resolution. At each stage, the associated conditions (e.g., garment and user images) are resized to match the current video resolution. To preserve the foundation model’s high-resolution spatial prior during early low-resolution stages, we further utilize video-image hybrid batches (Sec. 3.2): each mini-batch combines low-resolution, full-length 41-frame video clips with full-resolution (1152×720) image try-on pairs treated as single-frame videos. Since these additional image samples are single-frame, they add minimal overhead compared to multi-frame video training, while continuously reinforcing high-frequency spatial details. We retain 41-frame clips at every stage to preserve CogVideoX’s long-horizon temporal prior while scaling conditions and spatial resolution [19]. The full curriculum schedule is provided in Suppl. Sec. E.5.

Refiner. A final artifact-removing upsampler converts 8-FPS outputs into smooth 24-FPS videos; see Suppl. Sec. F.

Efficiency. Guidance distillation and few-step refinement accelerate inference to 5-10 minutes; see Suppl. Sec. G.

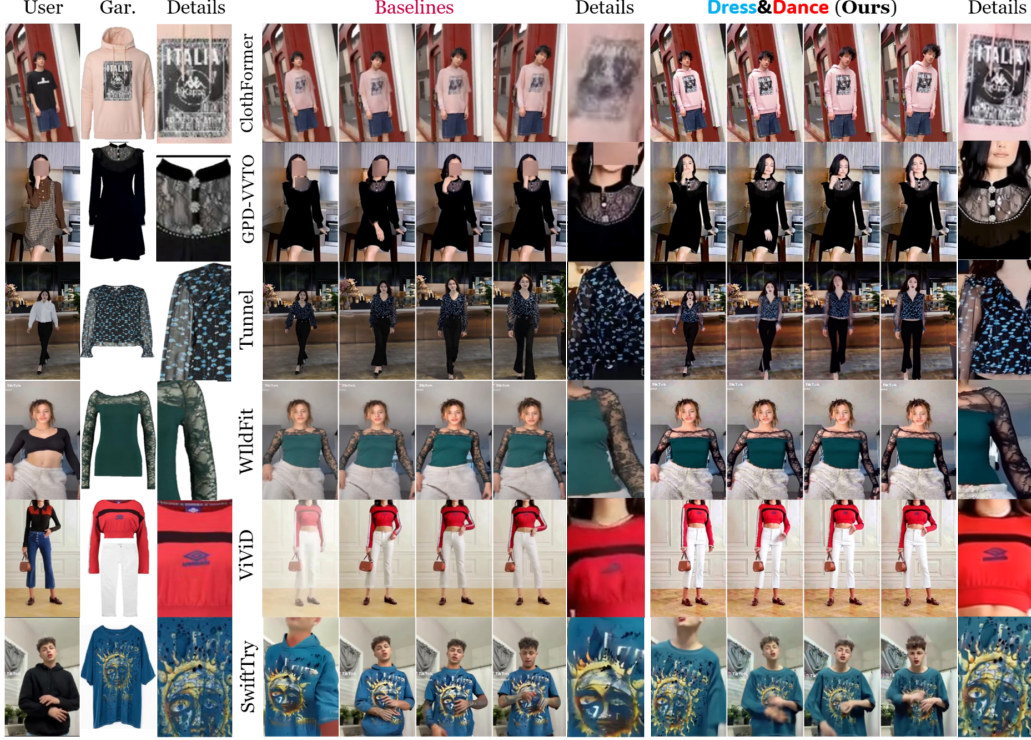


Figure 6. Qualitative VVT comparison under each method’s supported input interface and native output resolution. Dress&Dance renders fine textures and semi-transparent garments at 1152×720 .

4. Experiments

We evaluate the performance of Dress&Dance across three different try-on modes (Fig. 2): (1) the *single garment mode* allows a user to try on a garment from a single flat image; (2) the *multiple garment mode* supports the simultaneous try-on of all provided flat garment images; and (3) the *garment transfer mode* enables transferring a garment from an existing image via segmentation. We use the training data and benchmarks introduced in Sec. 3.2.

Baselines. We evaluate two complementary baseline categories. (1) *Video virtual try-on (VVT) baselines.* We first compare Dress&Dance against VVT methods, ViViD [16], WildVidFit [25], Tunnel Try-On [80], GPD-VVTO [74], and ClothFormer [32] in the single-garment setting. For the garment-transfer task, we compare with Fashion-VDM [35], whose input is a segmented garment from a person image rather than a flat product garment required by the single garment mode. For methods without released models, we compare qualitatively with cases shown by their authors using matched inputs. Released-model comparisons in Suppl. Sec. C.1 evaluate CatV2TON [10] and MagicTryOn-Turbo/14B [42] with their official preprocessing and reported or default resolutions. At the time of evaluation, we were unable to compare DreamVVT [102], DPIDM [41], or RealVVT [45] because complete official custom-input inference packages with the required weights were not available.

(2) *Image try-on + Image-to-video generation.* We compose baselines by applying an image try-on method to the user image and then animating the result with a separate video generation model. For image try-on, we use representative open-source and commercial methods, including Kling Try-On [37], CatVTON [9], Leffa [98], TPD [85], OOTD [79], and ML-VTON (a combination of GP-VTON [76] and HR-VTON [39]), and Dress&Dance in image-try-on mode. The animated videos are then generated by CogVideoX 1.5 I2V [87] for text-guided animation, or our Dress&Dance for motion-guided animation. We also apply this strategy with commercial animation models Kling 1.6 [37] and Ray2 [60].

Qualitative Comparison. We present qualitative results in Fig. 3 and Suppl. Figs. I.1, I.2, and I.3, with additional examples in SV0:15, SV2:02, and SV2:48. In single-garment try-on, Fig. 3 highlights a key limitation of “image try-on followed by animation” pipelines: once an occluded pattern is absent from the intermediate image, a separate animation stage receives no product-image evidence for restoring it after the hand moves. In contrast, Dress&Dance conditions on the garment throughout video generation, preserving accurate patterns, temporal consistency, and garment fidelity. Dress&Dance also handles complex reference-guided motion. As shown in Suppl. Fig. I.1 and SV3:14, text-conditioned baselines fail on challenging dance motions that are difficult to describe, whereas our method follows skeleton guidance extracted from the reference video

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS _{VGG} $\downarrow$	LPIPS _{AlexNet} $\downarrow$
Image-animation pipelines (without skeleton control)				
TPD [85] + CogVideoX I2V [87]	14.47	0.8305	0.2840	0.2461
OOTD [79] + CogVideoX I2V	14.68	0.8282	0.2779	0.2457
ML-VTON [39, 76] + CogVideoX I2V	14.49	0.8270	0.2961	0.2520
Dress&Dance Image Try-On + CogVideoX I2V	17.26	0.8515	0.2812	0.1635
Kling Image Try-On + Kling Video 1.6 [37]	17.33	0.8651	0.2296	0.1683
Dress&Dance Image Try-On + Kling Video 1.6	15.21	0.8297	0.2835	0.2066
Dress&Dance Image Try-On + Ray2 [60]	15.16	0.8290	0.2938	0.1990
Dress&Dance models (with target-derived skeleton input)				
Dress&Dance , Direct Training	17.14	0.8678	0.2854	0.1338
Dress&Dance	22.41	0.9038	0.2382	0.0624

Table 1. Reconstruction results on the captured benchmark. The lower panel is a pose-aligned appearance reconstruction diagnostic: Dress&Dance and Direct Training use the skeleton extracted from each scored target video. The upper panel provides contextual image-animation results for pipelines without this motion interface. “Dress&Dance Image Try-On +” denotes pipelines using our image try-on result.

and produces smooth motion with faithful try-on synthesis.

We further compare with video try-on baselines in single-garment mode in Fig. 6 and garment-transfer mode in Suppl. Fig. I.2. Dress&Dance generates native high-resolution videos at 1152×720 , preserving fine garment textures and challenging materials such as semi-transparent fabrics, while the baselines shown in Fig. 6 use lower native resolutions (*e.g.*, 720×540) with blurrier details. For flat inputs showing only the garment front, the model generates plausible side and back appearances consistent with the visible evidence. In the multi-garment setting (Suppl. Fig. I.3), Dress&Dance correctly composes top and bottom garments without requiring explicit garment-type labels or fixed input order. By contrast, Kling AI, despite supporting multi-garment try-on, incorrectly renders the trousers as a skirt in this challenging case.

Quantitative Comparison. Tab. 1 reports quantitative results on tasks constructed from the captured dataset, evaluated against ground truth using PSNR, SSIM, and LPIPS [90]. For each reconstruction, the inputs are a target garment and an image of the same user wearing a different garment. Dress&Dance and Direct Training (DT) additionally receive the skeleton extracted from the corresponding ground-truth video as the motion condition. This skeleton supplies only pose and timing; it contains no RGB or garment appearance. The lower panel compares the two models with the same target motion. Because the image-animation pipelines in the upper panel do not accept skeleton control, we report them under their native interfaces as contextual results. The captured benchmark contains 32 held-out reconstruction tasks; the Internet evaluation contains 32 single-garment, 32 multiple-garment, and 16 garment-transfer tasks. For the Internet dataset, its open-ended user–garment–motion combinations motivate reference-free perceptual evaluation. Following VQAScore [47], we use GPT [54] to evaluate garment fidelity ($\text{GPT}_{\text{Try-On}}$), user fidelity (GPT_{User}), motion quality ($\text{GPT}_{\text{Motion}}$), visual quality ($\text{GPT}_{\text{Visual}}$), and overall quality ($\text{GPT}_{\text{Overall}}$). The grading prompts and rubrics are pro-

vided in Suppl. Figs. I.4 and I.5; each video is graded 40 times and averaged to reduce sampling variance. In Suppl. Tab. C.1, GPT assigns Dress&Dance the highest garment-fidelity score in the reported comparisons. We additionally report Fréchet Video Distance (FVD) [71] on the captured and Internet sets in Suppl. Tab. C.4, and automatic metrics for temporal consistency, motion smoothness, and garment-region similarity in the released-model comparison in Suppl. Tab. C.5. We treat the GPT scores and the 40-participant user study in Suppl. Sec. C.4 as supporting perceptual evidence. These results support Dress&Dance’s ability to preserve both garment appearance and user identity under complex motion. Detailed analysis is provided in Suppl. Sec. C.

Direct Training Baseline. DT uses the same architecture and training data as Dress&Dance but trains with all conditions and full-resolution targets from the outset. As shown in Tab. 1 and SV3:14, DT remains substantially worse after roughly 200 wall-clock hours on eight H100 GPUs, often failing to preserve user identity and garment appearance. Because the two rows share architecture, data, and pose-aligned inputs, their gap supports the complete staged training recipe used by Dress&Dance.

5. Conclusion

We presented Dress&Dance, a high-resolution VVT framework driven by a single user photo, garment image(s), and reference motion. For diverse try-on settings, ConDNets organize heterogeneous visual conditions in separate layer-wise streams with per-block text exchange, alongside *garment-aware target steering* for garment placement. Together with our high-resolution dataset and benchmark, video-image hybrid training, synthesized unpaired supervision, and progressive curriculum, Dress&Dance produces 1152×720 , 24-FPS videos with strong garment fidelity and motion consistency across evaluated settings.

References

- [1] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-An-Animation: Large-scale text-conditional 3D human motion generation. In *ICCV*, pages 15039–15048, 2023. 3
- [2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *ECCV*, 2022. 3
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 13
- [4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, pages 5933–5942, 2019. 3
- [5] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *ICCV*, pages 7513–7522, 2023. 3
- [6] Jun-Kun Chen, Aayush Bansal, Minh Phuoc Vo, and Yu-Xiong Wang. Virtual Fitting Room: Generating arbitrarily long videos of virtual try-on from a single image. In *NeurIPS*, 2025. 3
- [7] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 3
- [8] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024.
- [9] Zheng Chong, Xiao Dong, Haoxiang Li, Wenqing Zhang, Hanqing Zhao, Dongmei Jiang, Xiaodan Liang, et al. Catv-ton: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, 2025. 3, 7
- [10] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025. 3, 7, 16
- [11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, abs/2210.11427, 2022. 18, 19
- [12] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *ICCV*, 2021. 3
- [13] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. FW-GAN: Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019. 3, 5
- [14] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *CVPR*, 2022. 3
- [15] Xiang Fan, Anand Bhattad, and Ranjay Krishna. Videoshop: Localized semantic video editing with noise-extrapolated diffusion inversion, 2024. 14
- [16] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models, 2024. 1, 3, 5, 7, 15, 16
- [17] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493, 2021. 3
- [18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3
- [19] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka diffusion models, 2024. 6
- [20] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, pages 5152–5161, 2022. 3
- [21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 3
- [22] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 3
- [23] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019. 3
- [24] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, 2022. 3
- [25] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. WildVidFit: Video virtual try-on in the wild via image-based controlled diffusion models. In *ECCV*, pages 123–139, 2024. 1, 3, 7, 15, 16
- [26] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *CVPR*, pages 5810–5819, 2021. 3
- [27] 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. In *ICLR*, 2022. 4, 17, 23
- [28] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 3, 14, 15
- [29] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 3, 14, 15
- [30] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024. 14
- [31] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 2020. 3

- [32] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. ClothFormer: Taming video virtual try-on in all module. In *CVPR*, 2022. 3, 7, 15, 16
- [33] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3, 14, 15
- [34] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion image-to-video synthesis via Stable Diffusion. In *ICCV*, pages 22680–22690, 2023. 3
- [35] Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. Fashion-vdm: Video diffusion model for virtual try-on. In *Proceedings of ACM SIGGRAPH Asia 2024*, 2024. 1, 3, 7, 16, 25
- [36] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024. 3
- [37] Kling. Kling ai: Next-generation ai creative studio, 2024. 4, 7, 8, 13, 16, 24, 25
- [38] Gaurav Kuppaa, Andrew Jong, Xin Liu, Ziwei Liu, and Teng-Sheng Moh. ShineOn: Illuminating design choices for practical video-based virtual clothing try-on. In *WACV Workshops*, pages 191–200, 2021. 3
- [39] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pages 204–219. Springer, 2022. 3, 7, 8
- [40] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACMTOG*, 40(4):1–10, 2021. 3
- [41] Dong Li, Wenqi Zhong, Wei Yu, Yingwei Pan, Dingwen Zhang, Ting Yao, Junwei Han, and Tao Mei. Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In *CVPR*, pages 22648–22657, 2025. 3, 7
- [42] Guangyuan Li, Siming Zheng, Hao Zhang, Jinwei Chen, Junsheng Luan, Binkai Ou, Lei Zhao, Bo Li, and Peng-Tao Jiang. Magictryon: Harnessing diffusion transformer for garment-preserving video virtual try-on. *arXiv preprint arXiv:2505.21325*, 2025. 3, 7, 16
- [43] Kedan Li, Jeffrey Zhang, and David Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *TPAMI*, 45(10):12222–12235, 2023. 3
- [44] Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In *WACV*, 2024. 3
- [45] Siqi Li, Zhengkai Jiang, Jiawei Zhou, Zhihong Liu, Xiaowei Chi, and Haoqian Wang. Realvvt: Towards photorealistic video virtual try-on via spatio-temporal consistency. *arXiv preprint arXiv:2501.08682*, 2025. 3, 7
- [46] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *ICCV*, 2023. 3
- [47] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024. 8, 14
- [48] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 3, 14, 15
- [49] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 14
- [50] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models, 2023. 19
- [51] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, 2022. 3
- [52] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACMMM*, 2023. 3
- [53] Hung Nguyen, Quang Qui-Vinh Nguyen, Khoi Nguyen, and Rang Nguyen. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *AAAI*, 2025. 1, 3, 5, 15, 16
- [54] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8, 17
- [55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 14
- [56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3, 4
- [57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 14
- [59] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *ECCV*, 2018. 3
- [60] Ray2. Ray 2 - advanced ai video generation tool, 2025. 7, 8, 13, 16, 24
- [61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

- synthesis with latent diffusion models. In *CVPR*, 2022. 3, 19
- [62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, Cham, 2015. Springer International Publishing. 3
- [63] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *CVPR*, 2020. 3
- [64] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 3
- [65] Fengyuan Shi, Jiayi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models, 2024. 14
- [66] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 3
- [67] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662, 2021. 3
- [68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 19
- [69] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 3, 14, 15
- [70] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *ICLR*, 2023. 3
- [71] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *DGS@ICLR*, 2019. 8, 16
- [72] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 3
- [73] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stable-garment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024. 3
- [74] Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. GPD-VVTO: Preserving garment details in video virtual try-on. In *ACMMM*, 2024. 1, 3, 7, 15, 16
- [75] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 5
- [76] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, pages 23550–23559, 2023. 3, 6, 7, 8
- [77] Jiazhen Xing, Chao Xu, Yijie Qian, Yang Liu, Guang Dai, Baigui Sun, Yong Liu, and Jingdong Wang. TryOn-Adapter: Efficient fine-grained clothing identity adaptation for high-fidelity virtual try-on. *International Journal of Computer Vision*, pages 1–22, 2025. 3
- [78] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in NeurIPS*, 2022. 17
- [79] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. OOT-Diffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI*, pages 8996–9004, 2025. 3, 6, 7, 8, 15
- [80] Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. In *ACMMM*, pages 3199–3208, 2024. 1, 3, 7, 15, 16
- [81] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Han-shu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *CVPR*, pages 1481–1490, 2024. 3
- [82] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *CVPR*, 2023. 3
- [83] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020.
- [84] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *CVPR*, 2022.
- [85] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *CVPR*, pages 7017–7026, 2024. 6, 7, 8, 15
- [86] Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D4-VTON: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *ECCV*, pages 36–52. Springer, 2024. 3
- [87] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 6, 7, 8, 13, 14, 17, 18, 23
- [88] Jeffrey Zhang, Kedan Li, Shao-Yu Chang, and David Forsyth. Acvg-vton: Accurate and contained diffusion generation for virtual try-on, 2024. 3
- [89] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 14, 15
- [90] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 8

- [91] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. [13](#)
- [92] Yuang Zhang, Jiayi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. [3](#)
- [93] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. [3](#), [14](#), [15](#)
- [94] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023. [19](#)
- [95] Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint arXiv:2405.18326*, 2024. [3](#)
- [96] Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. MV-TON: Memory-based video virtual try-on network. In *ACMMM*, 2021. [3](#)
- [97] Yuanyi Zhong, Haoran Tang, Jun-Kun Chen, and Yu-Xiong Wang. Contrastive learning relies more on spatial inductive bias than supervised learning: An empirical study. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16281–16290, 2023. [18](#)
- [98] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, Aditya Patel, Tao Xiang, Miaoqing Shi, and Sen He. Learning flow fields in attention for controllable person image generation. In *CVPR*, pages 2491–2501, 2025. [7](#)
- [99] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *CVPR*, pages 4606–4615, 2023. [3](#)
- [100] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&M VTO: Multi-garment virtual try-on and editing. In *CVPR*, pages 1346–1356, 2024. [3](#)
- [101] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *ECCV*, pages 145–162, 2024. [3](#)
- [102] Tongchun Zuo, Zaiyu Huang, Shuliang Ning, Ente Lin, Chao Liang, Zerong Zheng, Jianwen Jiang, Yuan Zhang, Mingyuan Gao, and Xin Dong. Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. *arXiv preprint arXiv:2508.02807*, 2025. [3](#), [7](#)

Dress&Dance: Supplementary Material

This document contains additional analysis and extra experiments.

Contents.

- Sec. X: Supplementary video.
- Sec. Y: Dataset release and construction.
- Sec. Z: Code release.
- Sec. A: Additional related work.
- Sec. B: CondNet versus other conditioning mechanisms.
 - Sec. B.1: Comparison with ControlNet.
 - Sec. B.2: Comparison with token- and reference-based conditioning.
- Sec. C: Quantitative result analysis.
 - Sec. C.1: Recent released-model comparisons.
 - Sec. C.2: “Image try-on + image-to-video generation” baselines.
 - Sec. C.3: Video virtual try-on baselines.
 - Sec. C.4: User study.
- Sec. D: Prompt generation and text-only try-on.
 - Sec. D.1: Obtaining text prompts.
 - Sec. D.2: Text-only try-on.
- Sec. E: Implementation details.
 - Sec. E.1: Detailed architecture.
 - Sec. E.2: Basic designs.
 - Sec. E.3: Conditioning networks.
 - Sec. E.4: Pose condition implementation.
 - Sec. E.5: Training setting.
 - Sec. E.6: Evaluation setting.
- Sec. F: FPS upsampling and artifact removal.
 - Sec. F.1: Auto-regressive video generation.
 - Sec. F.2: Data augmentation for visual refinement training.
 - Sec. F.3: Implementation details.
- Sec. G: Efficiency optimization and fast inference.
 - Sec. G.1: UniPC multi-step scheduler.
 - Sec. G.2: CFG strategy and distillation.
 - Sec. G.3: Upsampling model 1-step generation.
 - Sec. G.4: Modes and results.
- Sec. H: Limitations.
- Sec. I: Additional results and GPT prompts.
- Sec. J: Ethics and reproducibility.

X. Supplementary Video (SV)

We provide a **Suppl. Video (SV)** as `demo.mp4` to visualize our videos and comparisons against baselines. Here is a summary of the **SV** contents:

- **SV0:00**: Introduction and try-on modes.
- **SV0:15**: Single garment try-on.
- **SV2:02**: Multiple garment try-on.
- **SV2:48**: Garment transfer.
- **SV3:09**: Text-only try-on.
- **SV3:14**: Comparison with baselines and ablation study.

- **SV3:29**: FPS upsampling and artifact removal (Sec. F).
- **SV3:34**: Fast inference (Sec. G).

In the following sections, we use **SV** to refer to this supplementary video.

Y. Dataset Release and Construction

The benchmark contains a captured subset and an Internet-derived subset, independently collected or assembled by the authors. It is distributed as one complete research release through a data vendor: after the vendor’s standard due-diligence and licensing process, approved researchers can download the full package. The two subsets retain separate consent, privacy, and redistribution checks under that process. This access model permits research use while preserving the applicable participant and source-media obligations.

Regarding the construction of our captured dataset, we adhered to strict ethical standards for all 183 participants involved. First, regarding **conditions**, data collection was conducted in a fully equipped professional studio environment. To ensure the safety and comfort of the participants, each subject was assisted by a dedicated support team of five professionals, including photographers and stylists. Second, we obtained explicit **informed consent** from all participants; each subject signed a formal service agreement prior to the session, authorizing the data capture and its usage for research purposes. Finally, regarding **compensation**, all participants received compensation for their time and effort.

Z. Code Release

Upon acceptance, we will release the complete training, inference, and evaluation implementation, including data-preparation utilities, configurations, training schedules, and evaluation scripts.

A. Related Work: Video Generation

Image-to-Video Generation. A representative method, Stable Video Diffusion (SVD) [3], generates short videos from a single image but is limited to landscape formats and short durations. Similarly, I2VGen-XL [91] takes an image and text prompt to produce short landscape videos. A more flexible solution, CogVideoX-I2V [87], supports portrait videos and integrates text guidance for controllable synthesis. Commercial systems, such as Kling Video 1.6 [37] and Ray2 [60], also enable portrait image-to-video generation with text input, offering customizable outputs. These methods highlight progress in animating still images but do not address garment try-on or motion-driven control. We note that Kling also provides a separate image try-on model, which we analyze in this work.

Video-to-Video Generation. A related body of work investigates video editing and translation for general content manipulation. For example, VideoShop [15] propagates edits from the first frame to subsequent ones using a diffusion model, and BIVDiff [65] refines frame edits to improve temporal consistency. CogVideoX-V2V [87] combines a pre-trained text-to-video diffusion model with SDEdit [49] for text-driven video translation. While these methods demonstrate the power of temporal editing, they do not address virtual try-on specifically. Our setting requires independent user, garment, and motion inputs within a VVT system.

B. CondNet versus Other Conditioning Mechanisms

B.1. Comparison with ControlNet

ControlNet [89] established parameter-copying and additive feature injection for spatial controls in a UNet. Its feature-map interface is especially natural for aligned conditions such as edges and depth. A flat product garment and its posed video appearance are spatially unaligned; CondNet represents the garment as tokens participating in non-local attention, matching the correspondence structure of VVT.

At each block, CondNet supplies condition key/value tokens to the main video/text attention (Figs. 4 and E.1). Each condition keeps its native image or sparse-video extent before fusion, avoiding a dense $F \times H \times W$ expansion for a single image or sparse skeleton.

The resulting overhead relative to the CogVideoX base model is reported in Tab. G.1.

B.2. Comparison with Token-, Adapter-, and Reference-Based Conditioning

ReferenceNet-style appearance injection [28], OminiControl [69], IC-LoRA [29], EasyControl [93], VACE [33], and Phantom [48] provide close precedents for appearance features, sequence concatenation, parameter reuse, LoRA adaptation, unified adapters, and triplet-based text-image alignment. Tab. B.1 summarizes how these interfaces relate to CondNet’s VVT-specific organization.

CondNet’s separate streams delay the mixing of semantically different conditions, while per-block text exchange coordinates their roles. A matched study of single-stream, shared-branch, no-text-exchange, and ControlNet-style variants would require separately training several high-resolution 5B VVT models; Direct Training itself required roughly 200 wall-clock hours on eight H100 GPUs. We evaluate this concrete organization through the complete Dress&Dance system, whose end-to-end results demonstrate effective heterogeneous conditioning.

C. Quantitative Result Analysis

C.1. Recent Released-Model Comparisons

Both protocols retain 24 FPS and exact frame-index correspondence without temporal resampling. The processed motion-reference, Dress&Dance, and CatV2TON videos are 480×768 ; MagicTryOn outputs follow its released 624×832 geometry. For reproducibility, Full-frame DINO/CLIP Consistency uses facebook/dinov2-base [55] and openai/clip-vit-base-patch32 [58]; for each frame after the first, we average its cosine similarities to the preceding and first frames, then average over time. AMT-S is the official VBench MotionSmoothness score [30]. For Masked CLIP-I/DINO, the supplied agnostic try-on mask is resized by nearest-neighbor interpolation, its bounding region is composited on white, and cosine similarity to the target-garment crop is averaged over all valid frames using the same encoders.

With the common motion-video input, all three baselines operate directly on the processed motion-reference clip. When the target-dressed Dress&Dance output is used as person-video input, the Dress&Dance source retains higher Masked CLIP-I/DINO than all three generated baseline outputs, especially across side/back views, while maintaining comparable full-frame temporal consistency and smoothness. MagicTryOn-Turbo and MagicTryOn-14B contain approximately 1.58B and 16.45B denoising-transformer parameters, respectively, complementing our approximately 5B backbone at two scale regimes.¹

We include additional quantitative results, user studies, and computational analyses in Tabs. C.1, C.2, C.3, C.4, C.5, and G.1.

C.2. “Image Try-On + Image-to-Video Generation” Baselines

The quantitative results of our approach comparing with “Image Try-On + Image-to-Video Generation” baselines are presented in Tab. C.1.

Within this shared task block, GPT_{Try-On} assigns higher point scores to pipelines using our try-on component. Following VQAScore [47], repeated fixed-rubric grading provides a rubric-controlled perceptual signal alongside FVD and released-baseline automatic metrics.

Rows combining our image try-on with Kling or Ray2 evaluate the same try-on output under external animation backbones. Motion scores follow each pipeline’s native interface: text-driven animation for Kling/Ray2 and skeleton control for Dress&Dance.

¹ The MagicTryOn counts exclude the VAE, T5, and CLIP components.

Method	Primary scope	Established condition interface	Relation to CondNet
ControlNet [89]	Spatial image control	Copied UNet branch with additive, spatial feature injection	Precedent for parameter reuse; CondNet uses attention tokens for unaligned/native-shape inputs.
ReferenceNet [28]	Appearance-guided animation	Reference appearance features enter attention; pose is supplied separately	Reference appearance injection with separately supplied pose; CondNet extends this organization to wearer, garment, and motion streams.
OminiControl [69]	General DiT image control	DiT VAE/blocks are reused; condition and image tokens are jointly processed with dynamic positions	Establishes unified token processing for aligned and unaligned image controls.
IC-LoRA [29]	In-context image tasks	Multiple images are concatenated and jointly captioned; task-specific LoRA activates DiT in-context generation	Establishes concatenation and LoRA adaptation as prior art.
EasyControl [93]	Efficient DiT control	An isolated Condition Injection LoRA processes a control; position-aware training and cached causal attention improve flexibility/efficiency	Closest precedent for isolated LoRA condition processing; CondNet uses multiple persistent VVT streams with per-block text exchange.
VACE [33]	Video creation/editing	Dense context channel-packs reactive/inactive latents and masks, then additively couples to noisy-video tokens through a Context Adapter	Generic reference+pose composition through dense context; CondNet assigns separate wearer, flat-garment, and sparse-motion roles.
Phantom [48]	Subject-to-video	Joint text-image injection is learned from text-image-video triplets for single/multiple subjects	Establishes cross-modal reference alignment and triplet training for subject consistency.
CondNet (ours)	Single-image + independent-motion VVT	User, flat garment, and sparse skeleton use separate parameter-shared/LoRA layer-wise streams; each exchanges with text before main attention	Task-specific organization of heterogeneous high-resolution VVT conditions with per-block text exchange.

Table B.1. Mechanism-level comparison of conditioning interfaces. CondNet organizes token conditioning, parameter reuse, LoRA adaptation, and triplet supervision as separate wearer, garment, and motion streams for high-resolution VVT.

Table C.1. Fixed-rubric GPT comparison of image-try-on-plus-animation pipelines on a shared set of tasks. “Ours” and “Commercial” identify pipeline components. Forty repeated grades reduce sampling variance; bold, underlined, and italic values indicate the within-table ranking.

Method Property Ours Commercial	Method	Evaluation of Try-On Capability and Visual Quality						
		GPT _{Try-On} ↑	GPT _{User} ↑	GPT _{Motion} ↑	GPT _{Visual} ↑	GPT _{Overall} ↑	FID _{Interact} ↓	FID _{Captured} ↓
	TPD [85]	69.67	73.98	68.15	65.45	68.64	1146	753
	OOTD [79]	70.57	76.41	68.71	70.78	70.76	1089	884
	ML-VTON	69.69	76.62	68.50	69.95	70.50	1185	739
✓	Dress&Dance Image Try-On	85.05	87.54	75.68	76.88	80.71	1109	760
	Kling Image Try-On	80.10	89.97	85.48	<i>84.70</i>	84.38	982	655
✓	Dress&Dance Image Try-On	86.85	89.77	82.59	84.94	85.85	1008	700
✓	Dress&Dance Image Try-On + Kling Video 1.6	86.79	88.99	79.31	83.48	84.18	1094	735
✓	Dress&Dance Image Try-On + Ray2	79.48	78.47	72.00	71.24	74.85	1073	788
✓	Dress&Dance , Direct Training	87.41	88.89	<i>80.35</i>	84.48	<i>84.95</i>	1055	<i>691</i>

Table C.2. Fixed-rubric GPT-based VVT comparisons. Each two-row block uses a case from the named method’s official presentation and is interpreted as a paired within-block comparison. GPT_{Motion} evaluates motion realism and smoothness.

Source case	Method	GPT _{Try-On} ↑	GPT _{User} ↑	GPT _{Motion} ↑	GPT _{Visual} ↑	GPT _{Overall} ↑
ClothFormer [32]	ClothFormer	65.0	72.4	75.2	60.6	68.8
	Dress&Dance	90.6	88.5	85.0	87.4	88.5
GPD-VVTO [74]	GPD-VVTO	70.9	75.8	78.4	65.1	72.1
	Dress&Dance	92.4	90.9	87.1	88.1	89.2
Tunnel Try-On [80]	Tunnel Try-On	74.2	86.0	82.0	69.2	74.8
	Dress&Dance	76.9	88.6	80.0	75.4	81.6
ViViD [16]	ViViD	72.2	78.9	78.1	68.6	84.0
	Dress&Dance	85.6	87.3	86.3	84.6	85.2
WildVidFit [25]	WildVidFit	82.4	83.4	77.1	75.8	81.0
	Dress&Dance	86.0	86.4	80.2	74.6	81.4
SwiftTry [53]	SwiftTry	70.8	70.9	72.7	56.0	65.2
	Dress&Dance	83.5	83.1	78.6	75.2	80.3

Table C.3. Aggregate point preferences from 40 participants. Each value weights Strongly Prefer/Prefer/Slightly Prefer as 5/3/1 and reports the preference rate of Dress&Dance over a baseline; values $> 50\%$ indicate preference for Dress&Dance. Human_{Motion} is reported only for the Kling and Ray2 animation comparisons.

Method	Human _{Try-On} $\uparrow$	Human _{User} $\uparrow$	Human _{Motion} $\uparrow$	Human _{Visual} $\uparrow$	Human _{Overall} $\uparrow$
Prefer Dress&Dance to Image Try-On (Ours) + Kling [37]	77.0	77.3	86.5	76.5	77.5
Prefer Dress&Dance to Image Try-On (Ours) + Ray2 [60]	77.0	76.5	85.5	75.0	77.0
Prefer Dress&Dance to Image Try-On (Ours) + Fashion-VDM [35]	84.8	78.8	-	78.8	78.8
Prefer Dress&Dance to ClothFormer [32]	85.0	73.5	-	78.0	78.0
Prefer Dress&Dance to GPD-VVTO [74]	77.0	76.5	-	70.0	70.0
Prefer Dress&Dance to Tunnel [80]	74.3	63.8	-	63.0	63.0
Prefer Dress&Dance to WildVidFit [25]	73.3	67.3	-	66.5	67.0
Prefer Dress&Dance to ViViD [16]	76.3	67.5	-	69.5	68.8
Prefer Dress&Dance to SwiftTry [53]	61.3	70.5	-	60.5	60.5

Table C.4. FVD [71] comparison on Internet and captured benchmarks. Lower values indicate better temporal distribution quality.

Benchmark	Method	FVD $\downarrow$
Internet	Dress&Dance Image Try-On + CogVideoX I2V	174
Internet	Kling Try-On + Kling Video	197
Internet	Dress&Dance Image Try-On + Kling Video	184
Internet	Dress&Dance Image Try-On + Ray2	207
Internet	Dress&Dance	157
Captured	Dress&Dance Image Try-On + CogVideoX I2V	42
Captured	Kling Try-On + Kling Video	49
Captured	Dress&Dance Image Try-On + Kling Video	45
Captured	Dress&Dance Image Try-On + Ray2	88
Captured	Dress&Dance	20

Input setting	Method	Full-frame DINO Cons. $\uparrow$	Full-frame CLIP Cons. $\uparrow$	AMT-S $\uparrow$	Masked CLIP-I $\uparrow$	Masked DINO $\uparrow$
Motion input (105 frames)	CatV2TON (repaint) [10]	0.774429	0.883410	0.991603	0.750302	0.297751
	MagicTryOn-Turbo (105-frame extension) [42]	0.852983	0.940497	0.992758	0.764597	0.231090
	MagicTryOn-14B (105-frame full sequence) [42]	0.817690	0.943373	0.993892	0.759915	0.236756
	Dress&Dance source	0.824177	0.904180	0.992722	0.751365	0.398013
Dress&Dance output (121 frames)	CatV2TON (repaint) [10]	0.768184	0.901619	0.991993	0.689312	0.246528
	MagicTryOn-Turbo (121-frame extension) [42]	0.825594	0.921624	0.993183	0.709624	0.251073
	MagicTryOn-14B (121-frame full sequence) [42]	0.820540	0.938463	0.993346	0.720419	0.282375
	Dress&Dance source					

Table C.5. Fixed-case automatic metrics on unmodified full-length outputs under two input settings. The first block uses the processed 105-frame motion-reference clip from the same fixed Dress&Dance example as every baseline’s person-video input. The second gives every baseline the 121-frame target-dressed Dress&Dance output as person-video input. All outputs remain at 24 FPS with one-to-one frame alignment and no temporal resampling. Results are interpreted within each block because the inputs and timelines differ. DINO/CLIP columns measure full-frame temporal consistency; AMT-S measures smoothness; masked columns measure garment-region feature similarity. CatV2TON uses its official repaint option, which restores the input video outside the try-on mask. MagicTryOn-Turbo extends its released 81-frame sequence budget without changing weights; MagicTryOn-14B uses its released full-sequence path at the non-default 105- and 121-frame lengths.

C.3. Video Virtual Try-On (VVT) Baselines

In Tab. C.2, each pair follows a task presented by the respective baseline authors, preserving its intended conditions and supporting a paired within-block comparison.

All methods are graded repeatedly with the same fixed rubric, which reports try-on fidelity and visual quality as separate criteria.

C.4. User Study

In Tab. C.3, we report the results of a user study comparing our Dress&Dance against each baseline. For each comparison, participants were presented with the input user image,

garment image, and motion reference, and asked to indicate their preference based on five criteria: Human_{Try-On} (faithfulness to the garment), Human_{User} (preservation of user identity), Human_{Motion} (fidelity to the reference motion), Human_{Visual} (overall visual quality), and Human_{Overall} (comprehensive assessment). Human_{Motion} is reported for the Kling and Ray2 animation comparisons, where participants compare each generated video with the displayed reference motion. For person-video VVT methods, whose input videos determine motion, we report the applicable Try-On, User, Visual, and Overall criteria. We adopted a weighted scoring system where users selected a degree of

preference from {Strongly Prefer, Prefer, Slightly Prefer}, corresponding to weights of 5, 3, and 1, respectively. The final preference rate is calculated as the ratio of the weighted sum favoring our method to the total weighted sum.

We conducted the study with 40 participants. Aggregate point preferences exceed 50% in every reported comparison under the stated criteria.

D. Prompt Generation and Text-Only Try-On

D.1. Obtaining Text Prompts

When other image and video conditions are present, this text prompt can be easily generated using a vision-language model (VLM) like GPT [54]. We follow a simple multi-step process: first, we send sampled frames of the reference video to the VLM and request a detailed description of the motion. Then, we send the user and garment images, asking the VLM to generate a detailed prompt of the form: “the person in the user image, changing the garment to the one in the garment image, performing the motion as described in the motion description.” This composite approach generates a plausible and comprehensive text prompt. In practice, since most of the critical information is contained in the image and video conditions, our Dress&Dance remains highly effective even with a very simple text description.

D.2. Text-Only Try-On

In our Dress&Dance, the text prompt can serve as a crucial function as a fallback or refinement mechanism: if certain required information is not fully described in other conditions, the details provided in the text will be utilized. This capability allows our framework to achieve text-only try-on without any garment images. As shown in Fig. D.1, our Dress&Dance supports text-guided virtual try-on, enabling users to specify a garment and even add new accessories purely through the prompt.

E. Implementation Details

E.1. Detailed Architecture

Fig. E.1 visualizes the detailed architecture of our diffusion transformer integrated with CondNets for multi-modal conditioning. The CondNet for a specific condition, *e.g.*, the garment image, contains LoRA-augmented [27] parameter-shared copies of each transformer block. At every transformer block, the main sequence is constructed by concatenating the video sequence, the text sequence, and the condition sequences (specifically, the key and value sequences from the attention). The model then computes only the output video and text sequences. The output condition sequences for each transformer block are computed in their respective CondNet branches through cross-attention with the text sequence.

E.2. Basic Designs

Our main video diffusion model generates a 41-frame, 8-FPS video at 1152×720 (resulting in a video length of approximately 5 seconds, consistent with prior work like CogVideoX [87] which generates a 49-frame, 8-FPS video of 6 seconds). The subsequent upsampling model operates on overlapping 25-frame segments to produce a 121-frame, 24-FPS video at 1152×720 . All image conditions are maintained at a resolution of 1152×720 . The skeleton videos are ViTPose [78] outputs at the same resolution and length as the target video, featuring a gray background and a customized skeleton coloring to clearly distinguish left and right arms and legs.

E.3. Conditioning Networks

We utilize LoRA [27] adapters to enable the heterogeneous behavior required for different conditions. Specifically, we apply LoRA adapters in the `to_query`, `to_key`, `to_value`, and `to_out` linear layers of the attention modules, as well as in the feed-forward and convolutional layers. The rank of LoRA is set to 64 to ensure sufficient capacity.

E.4. Pose Condition Implementation

To provide precise motion guidance and resolve geometric ambiguities (*e.g.*, distinguishing frontal from dorsal views), we employ a modified skeleton representation based on ViTPose [78]. We introduce three specific enhancements to the standard visualization scheme to enforce structural consistency:

Auxiliary Midpoints and Central Axis. We introduce three additional auxiliary keypoints representing the midpoints of the eyes, shoulders, and hips. We connect these midpoints to their respective left and right anatomical keypoints (*e.g.*, connecting the left shoulder to the shoulder midpoint) using edges with distinct colors. This explicitly defines the body’s central axis and symmetry plane.

Lateral Color Differentiation. We implement a strict color separation strategy for the body topology. The edges representing the left limbs (arms, legs) and the left side of the torso are rendered in contrasting colors to their counterparts on the right side. This provides the model with unambiguous strong signals regarding the subject’s orientation.

Robust Hand Representation. We use one consistent edge color for all fingers of each hand, matching its corresponding arm (*e.g.*, all left fingers share the left arm’s color). This simplification reduces the visual complexity of the condition map and improves the model’s robustness against high-frequency jitter or errors in finger pose estimation.

Compressing Skeleton Videos Into Sparse Sequences.

A naive approach for converting a F -frame, $H \times W$ -resolution skeleton video into a sequence is to treat it as a standard $F \cdot H \cdot W$ sequence, similar to ordinary videos. However, this dramatically increases computational costs, doubling the main sequence length and quadrupling the self-attention computation. We observe that skeleton videos contain numerous “empty” pixels (background regions without any part of the skeleton), accounting for roughly 80% of each frame. Since our primary concern is the human pose, we convert the skeleton video into a sparse sequence of variable length by masking and preserving only the sequence elements that correspond to non-background regions. With positional embeddings tracking the location of each sparse element, our Dress&Dance is able to learn the correct motion from this sparse sequence. This strategy significantly reduces computational costs by 70% while preserving the non-background pose tokens.

E.5. Training Setting

We build our Dress&Dance architecture based on CogVideoX [87]-5b. We reuse the parameters of the pre-trained CogVideoX model and initialize the new parameters related to CondNets. We set separate learning rates for non-attention parameters (4×10^{-5}), non-LoRA attention parameters (10^{-5}), and LoRA parameters (2.5×10^{-6}). We use an AdamW optimizer with default hyperparameters and no scheduler.

We train our Dress&Dance on a single machine with 8 NVIDIA H100 GPUs. During garment-aware target steering, we train Dress&Dance’s main video diffusion model for 500 iterations for each of the three phases. Following this, we execute resolution scaling in stages: we train the main video diffusion for 24,000 iterations in stage 1 at 768×480 resolution, and then 11,000 iterations in stage 2; afterward, we add the skeleton video condition and CondNet, and continue to train the main video diffusion for 6,000 iterations in stage 3 at 768×480 resolution, and then 6,000 iterations in stage 4 at full resolution to obtain our initial main video diffusion model. Then, initialized from this main video diffusion, we train our subsequent upsampling model for 12,000 iterations at full resolution, and continue to train the main diffusion model for another 12,000 iterations in stage 4. The 1–2 hours per GATS phase in Sec. 3.1 and the roughly 200-hour Direct Training baseline in Sec. 4 are wall-clock times measured on this 8-GPU setup; the corresponding GPU-hours are $8\times$ larger.

E.6. Evaluation Setting

Evaluation uses held-out real user images, garment images, and motion references; synthetic image-try-on triplets are confined to training-data construction. For the captured dataset with ground truth, we randomly sample 32 virtual

try-on tasks from the test subset as our evaluation benchmarks. For the Internet dataset, we sample tasks by randomly pairing user images and garment images from the test subset, and construct 32 single garment try-on tasks, 32 multiple garment try-on tasks, and 16 garment transfer tasks. We compute the Fréchet Inception Distance (FID) of a generated video with respect to a randomly sampled subset of 8,192 videos from the corresponding dataset.

F. FPS Upsampling and Artifact Removal

The CogVideoX backbone produces 41-frame videos at 8 FPS per forward pass. Our subsequent refinement stage converts them into 121-frame, 24-FPS outputs while improving contrast, color, sharpness, and motion quality. We initialize this refinement stage from the final main-model checkpoint after progressive training. The upsampling model operates in an auto-regressive manner, refining the video segment by segment with overlaps. To train the upsampling model, we construct training data by temporally downsampling raw 24-FPS ground truth videos to 8 FPS and applying various data augmentations to the downsampled videos to simulate potential artifacts. This specialized training equips the upsampling model with the ability to remove artifacts and further enhance the final visual quality.

F.1. Auto-Regressive Video Generation

For memory-efficient high-resolution refinement, we train the upsampling model on overlapping 25-frame segments generated from 9-frame, 8-FPS source clips. The CondNet for this source video condition repeats each 8-FPS frame three times to form a coarse 24-FPS sequence and then a standard $F \cdot H \cdot W$ sequence. To generate a 121-frame video, we divide it into seven 25-frame segments with 9-frame overlaps and refine them progressively, leveraging DiffEdit [11] for consistent overlap refinement.

F.2. Data Augmentation for Visual Refinement Training

The upsampling model is trained within a standard diffusion training process, where we construct 8-FPS video inputs by temporally downsampling the ground truth videos and train the model to restore the original video. We observe that, beyond temporal upsampling, the model also has the potential to refine visual quality by removing artifacts. Therefore, we retain all input conditions used to generate the original video, and then apply extensive data augmentations to the 8-FPS video inputs during training. These augmentations include color jitter, Gaussian blur, VAE compression (encoding and decoding the video with VAE, which may lose details), noising, local pixel shuffling [97], and temporal blurring, to simulate various artifacts that may occur in the generated source video. These intentional degradations make the refiner learn both temporal interpolation and restoration,

Table G.1. Computational analysis of the main 41-frame video diffusion model for one denoising forward pass at 1152×720 resolution. Both models are measured under the same inference setup.

Metric	Dress&Dance	CogVideoX Base
Resolution	1152×720	1152×720
Frames	41	41
VRAM Peak	19 GB	16 GB
FLOPs/Forward	3.18P	1.01P
Latency/Forward	20 sec	5 sec

completing the progressive training regimen. With this robust training strategy, our upsampling model significantly enhances the visual quality of the final video, especially in addressing blurriness, contrast, and brightness issues.

F.3. Implementation Details

The full refined video is a 121-frame, 24-FPS video, refined from a 41-frame, 8-FPS video. We first repeat each frame three times and remove the last two frames to convert the 41-frame video into a coarse 121-frame version. Then, we refine this coarse video segment by segment. In the first step, we take frames $[0, 25)$ as the input to the upsampling model to generate the refined segment. In the subsequent steps, given the last refined index i where frames $[0, i)$ are already refined, we take frames $[i - 9, i + 16)$ as the input to the upsampling model, where frames $[i - 9, i)$ are preserved, and frames $[i, i + 16)$ are newly refined. We apply DiffEdit [11] on the first 9 frames to ensure they preserve the previously refined frames $[i - 9, i)$, and the model generates smooth frames $[i, i + 16)$ as the continuation. The entire refinement process is completed in 7 steps through 7 segments: $[0, 25)$, $[16, 41)$, $[32, 57)$, $[48, 73)$, $[64, 89)$, $[80, 105)$, and $[96, 121)$.

To grant the upsampling model the capability to regenerate blurred parts, it also takes all the conditions used by the main video diffusion model. More specifically, we directly pass the same garment image(s) and user image, and the corresponding segment of the skeleton video. The refinement source frames are also input to the upsampling diffusion model through an individual CondNet. This design lets the model preserve the first 9 input frames or refine the whole segment, while DiffEdit maintains overlap consistency.

G. Efficiency Optimization and Fast Inference

Our acceleration reduces the 100-step two-stage pipeline from roughly four hours to 4–10 minutes while maintaining visual quality comparable to the unaccelerated pipeline. As shown in Tab. G.1, the main overhead over the CogVideoX base model comes from jointly handling multiple high-resolution conditions, such as multiple garment images,

user appearance, and motion guidance. We combine sampler acceleration with two complementary techniques: CFG distillation [50, 61] to remove the extra guidance forward pass, and one-step denoising for the upsampling model.

G.1. UniPC Multi-Step Scheduler

Our standard generation process uses the DDIM [68] scheduler, which relies only on the first-order derivative and requires numerous denoising steps. We replace this with the UniPC [94] multi-step scheduler, which leverages higher-order derivatives to perform denoising based on the predicted noise in multiple steps. With UniPC, we are able to reduce the denoising steps from 100 to 10 for the main model and from 100 to 2 for the upsampling model while maintaining comparable quality. Notably, this replacement can be applied to a pre-trained model in an out-of-the-box manner without requiring specific training with the UniPC scheduler.

G.2. CFG Strategy and Distillation

CFG Strategy. To handle multiple heterogeneous conditions effectively, we categorize them based on their guidance strength and conditioning difficulty into two sets:

- **Easy Conditions** ($\mathcal{C}_{\text{easy}}$): Conditions that provide strong, unambiguous signals which the model follows easily. In our framework, this includes *motion*, *user image*, and the *source video* (in the refiner module).
- **Hard Conditions** ($\mathcal{C}_{\text{hard}}$): Conditions requiring complex semantic interpretation or mapping, specifically *text prompts* and *garment images*.

We apply guidance scaling specifically to the hard conditions while treating the easy conditions as the unconditioned baseline. Let $f(\mathbf{z}_t, t, \mathcal{C})$ denote our DiT-based noise predictor network, where $\mathbf{z}_t$ is the noisy video latents at timestep t . We compute the joint noise prediction ϵ_{all} using all conditions $\mathcal{C}_{\text{all}} = \mathcal{C}_{\text{easy}} \cup \mathcal{C}_{\text{hard}}$, and the baseline prediction ϵ_{easy} using only easy conditions:

$$\epsilon_{\text{all}} = f(\mathbf{z}_t, t, \mathcal{C}_{\text{all}}), \quad \epsilon_{\text{easy}} = f(\mathbf{z}_t, t, \mathcal{C}_{\text{easy}}). \quad (1)$$

The final classifier-free guided noise ϵ_{cfg} is computed as:

$$\epsilon_{\text{cfg}} = w \cdot (\epsilon_{\text{all}} - \epsilon_{\text{easy}}) + \epsilon_{\text{easy}}, \quad (2)$$

where w is the guidance scale (set to $w = 5.0$ in our experiments).

Distillation. Standard CFG necessitates two network evaluations per denoising step, doubling the inference latency. To eliminate this overhead, we employ a distillation approach building upon the framework introduced by [50]. We train a distilled model g to directly predict the guided noise ϵ_{cfg} derived from our strategy above. The distillation objective is to minimize:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{\mathbf{z}_t, t, \mathcal{C}} [\|g(\mathbf{z}_t, t, \mathcal{C}_{\text{all}}) - \epsilon_{\text{cfg}}\|^2]. \quad (3)$$

By applying this distillation training to both the main video diffusion and the upsampling model, we enable the model to output the guided result in a single forward pass instead of two passes for ϵ_{all} and ϵ_{easy} , effectively reducing the inference time by half.

G.3. Upsampling Model 1-Step Generation

As the upsampling model’s task is relatively simple and can already support 2-step generation in a training-free manner, we further train it for one-step generation. This training involves training the distilled model to output a noise prediction such that using it to denoise the noisy video obtains the target original video. Through this training, we achieve 1-step generation for the upsampling model.

G.4. Modes and Results

Finally, we support three different modes to trade off between efficiency and quality. Each mode is marked as “ $a\{, D\} + b\{, D\}$ ”, e.g., “25D+1D” or “10+1D”, representing a mode with a main video denoising steps and b upsampling model denoising steps, with (marked as a ‘D’ after a and/or b) or without (no ‘D’) CFG distillation. The “25+1D” mode (25-step non-distilled generation with 1-step distilled refinement) generates videos at very similar visual quality in 8–10 minutes. The “25D+1D” and “10+1D” modes provide faster 4–5 minute operating points; the former favors complicated garments and high-frequency detail, and the latter favors user and garment fidelity. The results are visualized in Fig. G.1 and in SV3:34.

H. Limitations

Performance can degrade for very fast or unusual motion because the 8-FPS main stage undersamples rapid transitions and such poses may fall outside the training distribution.

I. Additional Results and GPT Prompts

This section provides additional qualitative results and the GPT evaluation prompts.

- Fig. I.1: Try-on with dancing motion.
- Fig. I.2: Garment transfer from another given user image.
- Fig. I.3: Simultaneous try-on of multiple garments.
- Figs. I.4 and I.5: Prompts for GPT evaluator.

J. Ethics and Reproducibility

Intended use and misuse. Dress&Dance is intended for consent-based, offline clothing visualization and research. As a face-, identity-, garment-, and motion-conditioned human-video generator, it can also be misused for impersonation, deceptive media, harassment, or biased/stereotyped portrayals. Any deployment should require authorization to use a person’s likeness, apply access

controls and provenance/watermarking, retain auditable generation records, and present outputs as synthetic previews rather than evidence of real events or physical garment fit. Dataset and model audits should cover identities, body shapes, garments, and motions.

Participant data. The captured collection contains 183 consenting participants recorded in a professional studio. Each participant signed a service agreement authorizing capture and research use and received compensation. Studio staff, including photographers and stylists, supported safe and comfortable sessions. Distribution is mediated by the vendor access process in Sec. Y, which preserves consent, privacy, and licensing obligations.

Reproducibility. Upon acceptance, we will release the complete training, inference, and evaluation implementation, including preprocessing, configurations, and schedules. The benchmark is distributed under the vendor-managed terms in Sec. Y. Hyperparameters, training stages, pose representation, refiner, and efficiency settings are documented in Secs. E–G.



Figure C.1. Motion-video try-on using the processed 105-frame motion-reference clip; columns are source frames 0, 17, 34, 51, 68, 85, 104. Rows show the input motion video, CatV2TON (repaint), MagicTryOn-Turbo 105-frame output, MagicTryOn-Turbo 81-frame suffix (frames 24–104), MagicTryOn-Turbo 65-frame suffix (40–104), MagicTryOn-14B 105-frame output, and MagicTryOn-14B 65-frame suffix (40–104). All rows remain at 24 FPS and map one-to-one to the labeled source frames without temporal resampling; gray cells mark frames preceding each generated suffix. The full-length MagicTryOn rows use the original zip-front garment caption; the suffix rows use the corrected pullover-hoodie caption and are presented as separate qualitative settings. The input and CatV2TON rows are 480×768 ; MagicTryOn rows are generated at 624×832 and corrected only for display, while metrics use unmodified full-length outputs.



Figure C.2. Baselines given the target-dressed Dress&Dance output as person-video input on the 121-frame timeline; columns are frames 0, 20, 40, 60, 80, 100, 120. Rows show the Dress&Dance input, CatV2TON (repaint), MagicTryOn-Turbo 121-frame output, MagicTryOn-Turbo 81-frame suffix (40–120), MagicTryOn-Turbo 65-frame suffix (56–120), MagicTryOn-14B 121-frame output, and MagicTryOn-14B 65-frame suffix (56–120). Every baseline therefore receives direct access to the target appearance. All rows remain at 24 FPS and map one-to-one to the labeled source frames without temporal resampling; gray cells mark frames preceding each generated suffix. The full-length MagicTryOn rows use the original zip-front garment caption; the suffix rows use the corrected pullover-hoodie caption and are presented as separate qualitative settings. The Dress&Dance input and CatV2TON rows are 480×768 ; MagicTryOn rows are generated at 624×832 and corrected only for display, while metrics use unmodified full-length outputs.

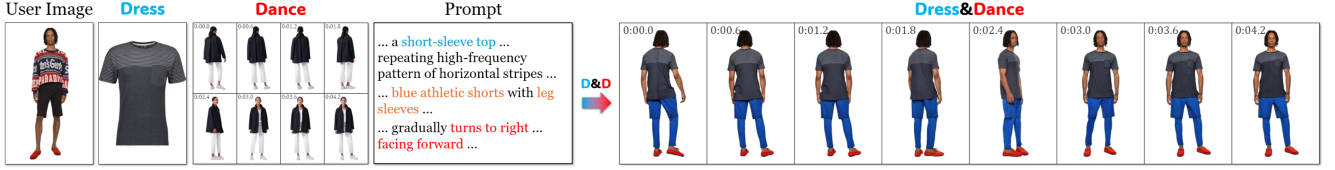


Figure D.1. Our Dress&Dance supports text-only try-on. By introducing a different garment and accessories in the text prompt, our Dress&Dance generates try-on results without corresponding images.

Our Dress&Dance includes a text prompt as part of the model input. Built upon CogVideoX [87], Dress&Dance inherits a strong text-conditioned video prior, making text a natural semantic interface for controllable generation. Beyond this architectural compatibility, text provides complementary cues that are often unavailable from visual inputs alone. In particular, when some regions are occluded or absent in the input images, text can specify the desired appearance of non-try-on regions and thereby reduce ambiguity. This is especially useful in practical try-on scenarios where the target outfit is only partially determined by the garment image and the user appearance image. Text further enables lightweight semantic control over the final outfit, such as adjusting garment attributes or harmonizing non-try-on regions.

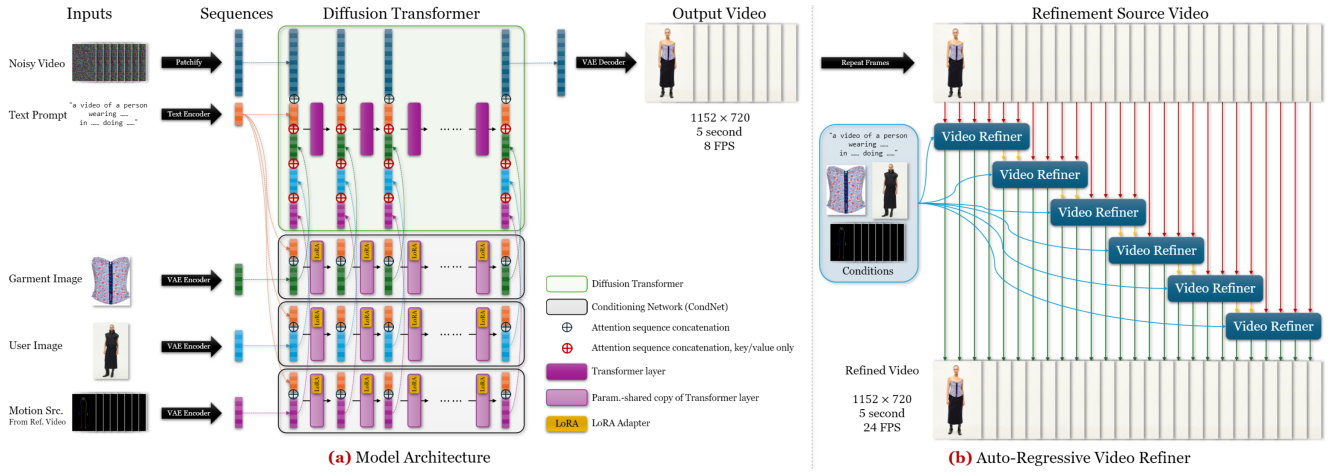


Figure E.1. Our Dress&Dance designs. (a) Dress&Dance supports multi-modal conditioning through our CondNet mechanism, a unified conditioning framework augmented with LoRA-parameter-shared [27] transformers. (b) The two core models in our Dress&Dance framework: a main model for 8-FPS video generation, and an upsampling model to auto-regressively increase the frame rate to 24-FPS while removing artifacts.



Figure G.1. In mode $a\{,D\} + b\{,D\}$, a and b denote main-model and refiner denoising steps, and D denotes classifier-free-guidance (CFG) distillation. The “25+1D” mode retains quality comparable to the 100-step two-stage pipeline in 8–10 minutes; “25D+1D” and “10+1D” reduce generation time to 4–5 minutes. More results are available in [SV3:34](#).

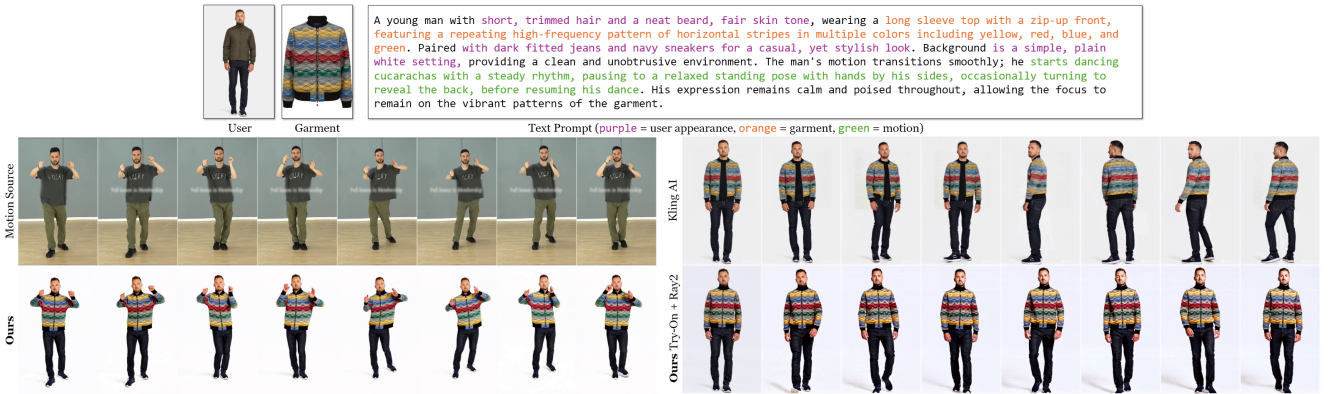


Figure I.1. Dress&Dance allows a user to **dress** themselves in a desired garment and perform the desired **dance**. These complex movements are challenging to express with text alone, which makes motion generation difficult for models like Kling [37] and Ray2 [60].

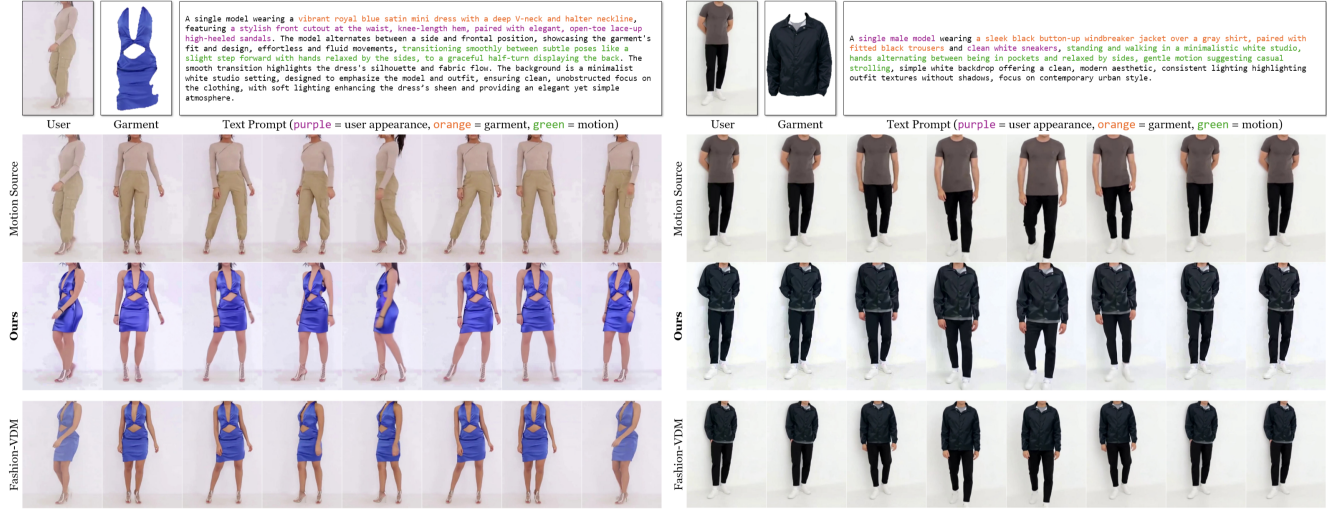


Figure I.2. Our **Dress&Dance** supports transferring a garment from another given image via segmentation, regardless of the pose of that image. Notably, our **Dress&Dance** generates high-resolution videos at 1152×720 with clear appearance and more details, while the baseline Fashion-VDM [35] exhibits color fading and generates low-resolution 512×384 videos.



Figure I.3. Dress&Dance supports the simultaneous try-on of a top and bottom garment, correctly understanding and representing both without explicit labeling. In contrast, the Kling AI baseline [37] incorrectly misrepresents the trousers as a skirt. More results in [SV2:02](#).

You are a professional fashion expert and an evaluator of human video generation. You task is to evaluate the overall quality of the generated human fashion video.

In each task, You will be given:

- (1) a user image, which is a photo of the model person;
- (2) one or two garment frontal images, which are photos of the garments to be worn by the model person (these images may be different from the user image, as they are the garments to be try-on) along with several image or video as conditions, with each of their descriptions;
- (3) a generation prompt, with is a detailed description of a human fashion video, you can assume that the prompt will be consistent with the model person and the garment images;

Then, you will be given one candidate video to evaluate, which is expected to generate the human fashion video of the model person wearing the garment(s) in the garment images, following the generation prompt.

To evaluate the quality of the human video, you should consider the following aspects:

- "user": user appearance fidelity
Whether the appearance of the person in the generated video matches the user image.
Also, for all the other garments that are not indicated to change in the prompt and garment images, the person should wear the same garments as in the user image.
- "garment": garment appearance fidelity
Whether the appearance of the garments in the generated video matches the garment images, where the frontal of the garment should match the garment image from all the viewpoints at any human motion, and the back and sides of the garment, which are not shown in the garment images, should be reasonably
↳ generated
to maintain a consistent appearance with the frontal.
- "motion": video motion quality
Whether the motion of the person, garments, and possible accessories in the generated video is realistic,
↳ smooth, and natural,
with clear appearance with reasonable speed, acceleration, and deceleration, without sudden or unnatural movements or morphing.
- "visual": visual quality
Whether the generated video is visually appealing, with reasonable brightness, contrast, and color saturation, with smooth and natural motion, and without visual artifacts.
The person, garments, and background should be visually clear and sharp, with high resolution and without
↳ blurring.
The video should be visually appealing and engaging to watch.
- "overall": overall quality
This is an "overall" evaluation, which should consider all the aspects of the video and try-on quality. Which means it should evaluate
the fidelity of user image and garment images, the motion quality, and the visual quality, and give an overall score based on these aspects.
To receive a high "overall" evaluation score, the video should satisfy all the aspects' requirements, and any failure in any aspect will result in a lower score.
The overall score should be a comprehensive evaluation of the quality of such video try-on generation.
You will be paid tips based on the quality of your evaluation.

Figure I.4. GPT evaluator prompt: different aspects.

```

{{ For: each frame }}
This is the frame at {timestep} of the generated video.
{{ End For }}

Now, please evaluate the quality of the generated human video.
You should consider the four aspects: "user", "garment", "motion", and "visual", and finally provide an "overall"
↪ evaluation.
You should consider all the frames from the video, and have a comprehensive evaluation.
In your evaluation, for each scoring item of "user", "garment", "motion", "visual", and "overall",
you should give an integer score from 0 to 100,
where higher is better.
To better align your score:
    "0" means the worst, failure, or completely wrong;
    "60" means a passable quality, the generation is not contradict with the inputs, but with multiple noticeable
    ↪ issues;
    "80" means a good quality, with minor issues that can be ignored;
    "90" means a very good quality, with only very minor issues that are hard to notice;
    "100" means perfect and flawless quality, with no issues at all.
For "user" and "garment",
    a score of 80 means that the appearance of user or the frontal of garment is almost the same as the image.
For "motion",
    a score of 80 means that the motion is at least be able to be performed by a human,
    with no obvious artifacts like morphing, or the head turns 360 degrees.
For "visual",
    a score of 80 means that the video is visually clear and sharp, with only a few blurring,
    and the brightness, contrast, and color saturation are somehow reasonable.
For "overall",
    a score of 80 means that the video can roughly score 80 in all the other aspects above.
You can provide intermediate scores like 73, 86, 92, etc., to indicate the quality more precisely.
In your reply, you should provide the evaluation scores for each aspect in json format.
For example, the evaluation can be:
    {
        "user": 93,
        "garment": 87,
        "motion": 81,
        "visual": 62,
        "overall": 75,
    }
or
    {
        "user": 92,
        "garment": 64,
        "motion": 89,
        "visual": 55,
        "overall": 71,
    }
These are just examples. You should provide the actual evaluation for the actual video.
Please do not include any other contents, like quotes, or markdown code blocks like ```json".
You can just directly reply with the json string, starting with "{" and ending with "}".
All other information or incorrect format will be considered as invalid,
which will result in a penalty in your payment.
Please provide your evaluation as a json string now:

```

Figure I.5. GPT evaluator prompt: rubrics.