Wear-Any-Way: Manipulable Virtual Try-on via Sparse Correspondence Alignment

Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao*

Alibaba Group

Abstract. This paper introduces a novel framework for virtual try-on, termed `Wear-Any-Way`. Different from previous methods, `Wear-Any-Way` is a customizable solution. Besides generating high-fidelity results, our method supports users to precisely manipulate the wearing style. To achieve this goal, we first construct a strong pipeline for standard virtual try-on, supporting single/multiple garment try-on and model-to-model settings in complicated scenarios. To make it manipulable, we propose sparse correspondence alignment which involves point-based control to guide the generation for specific locations. With this design, Wear-Any-Way gets state-of-the-art performance for the standard setting and provides a novel interaction form for customizing the wearing style. For instance, it supports users to drag the sleeve to make it rolled up, drag the coat to make it open, and utilize clicks to control the style of tuck, *etc.* `Wear-Any-Way` enables more liberated and flexible expressions of the attires, holding profound implications in the fashion industry. Project page is mengtingchen.github.io/wear-any-way-page.

Keywords: Virtual Try-on · Customizable Generation · Diffusion Model

1 Introduction

Virtual try-on aims to synthesize an image of the specific human wearing the provided garments. It emerges as a significant technology within the realm of the fashion industry, providing consumers with an immersive and interactive means of experiencing apparel without physically trying.

Classical solutions for virtual try-on rely on Generative-Adversarial-Networks (GANs) [12, 23, 29, 35, 36] to learn the feature warping from the garment image to the person image. Recent advances in diffusion models [19] bring a huge improvement in the generation quality for various image/video synthesis tasks, including virtual try-on. Some recent works [26, 66] could generate high-fidelity results leveraging pre-trained text-to-image diffusion models [50]. However, there still exist drawbacks to these solutions.

First, most of them only support simple cases like well-arranged single garments with simple textures as input. For garments with complicated textures or

* Corresponding author

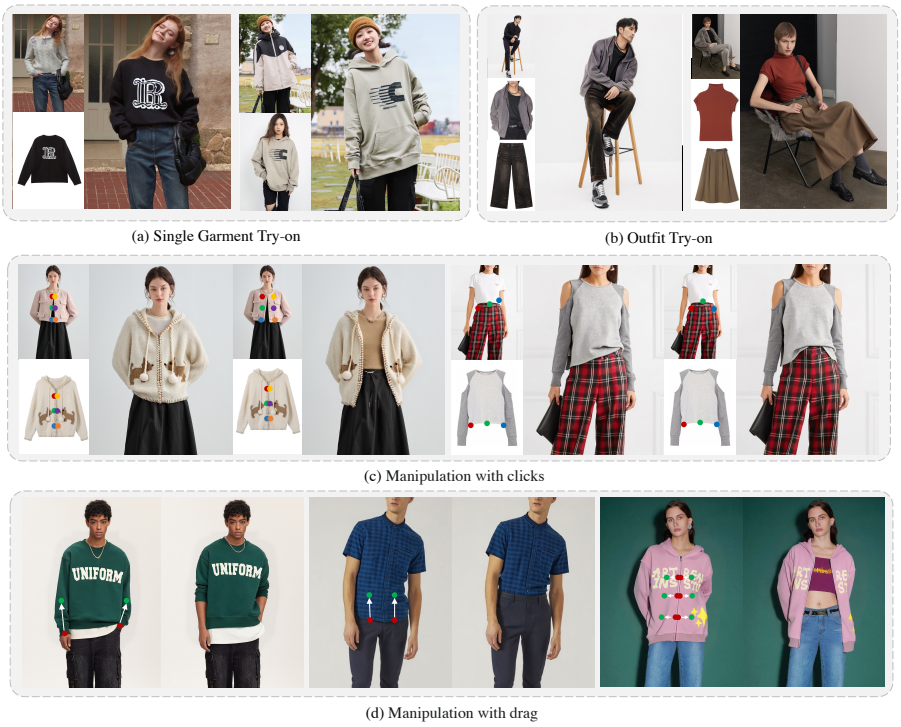


Fig. 1: Manipulable try-on with `Wear-Any-Way`. Our method achieves state-of-the-art performance for the standard setting of virtual try-on (first row), supporting diversified input formats and scenarios. An impressive feature is that our method supports users to manipulate the way of wearing using simple interactions like click (second row) and drag (third row). It should be noted that all these applications are accomplished with a single model in one pass.

patterns, existing solutions often fail to maintain the fidelity of the fine details. In addition, most of the previous solutions do not solve the challenges in real-world applications like model-to-model try-on, multi-garment try-on, complex human poses, and complicated scenarios, *etc.*

Second, previous methods are unable to exert control over the wearing style. However, the actual wearing style holds significant importance in the realm of fashion. For instance, variations in the rolling up or down of sleeves, the tucking and layering of tops and bottoms, the decision to leave a jacket open or closed, and even the exploration of different sizes for the same garment all contribute to diverse ways of wearing. These distinct styles can showcase varied states of the same piece of clothing, highlighting the pivotal role of styling options, particularly in the context of fashion applications.

In this work, we propose `Wear-Any-Way`, a novel framework for virtual try-on that solves the two aforementioned challenges at once. `Wear-Any-Way` could act as a strong option for the standard setting of virtual try-on, it synthesizes high-quality images and preserves the fine details of the patterns on the garment. Besides, it serves as a universal solution for real-world applications, supporting various sub-tasks like model-to-model try-on, multi-garment try-on, and complicated scenarios like street scenes. The most important feature is that the `Wear-Any-Way` supports users in customizing the wearing style. As shown in Fig. 1, users could use simple interactions like click and drag to control the rolling of sleeves, the open magnitude of the coat, and even the style of tuck.

To archive these functions, we first build a strong baseline for virtual try-on. Recently, Reference-only [64] structure has proven effective in many downstream tasks like image-to-video [22, 60] and image editing [5, 41, 42]. Inspired by these methods, we build a dual-branch pipeline, the main branch is a denoising U-Net initialized from a pre-trained inpainting Stable Diffusion [50]. The main U-Net takes the person image as input while the reference U-Net extracts the features from the garment image. The referential garment features are then injected into the main branch with self-attention. To further improve the flexibility and robustness, we also inject the guidance of the human pose. This pipeline achieves state-of-the-art performance in the standard try-on setting.

Afterward, we take a further step to make this strong baseline customizable. Specifically, we investigate a point-based control that forces the specific points on the garment image to match the target points on the person image in the generation result. To align the features of the paired points, we propose sparse correspondence alignment, which first learns a series of permutable point embeddings and injects these embeddings into both the main and reference U-Net by modifying the attention layers. To assist the network learn the feature alignment better, we design several strategies like condition dropping, zero-initialization, and point-weighted loss to ease the optimization.

Equipped with all these techniques, `Wear-Any-Way` demonstrates superior quality and controllability for virtual try-on. In general, our contributions could be summarized in three folds:

- We construct a novel framework, `Wear-Any-Way`, which generates high-quality results and supports users to precisely manipulate the way of wearing.
- We propose a strong, flexible, and robust baseline for virtual try-on, which reaches state-of-the-art with extensive comparisons with previous methods.
- We design the sparse correspondence alignment to enable the point-based control and further develop several strategies (*i.e.*, conditional dropping, zero-initialization, point-weighted loss) to enhance the controllability.

2 Related Work

GAN-based virtual try-on. Initially, numerous methods [12, 23, 29, 31, 35, 36] have utilized Generative Adversarial Networks (GANs). Some works [9, 14, 30, 58] typically adopt a two-stage strategy: first deforming the garment to fit the target

body shape, then integrating the transformed clothing onto the human model using a GAN-based try-on generator. To achieve accurate deformation, several techniques estimate a dense flow map that guides the clothing reshaping process [2,14,16,30,58]. Some works achieve simple fashion editing by dressing order [11] and warp policies [31]. However, these existing approaches have limitations, particularly when dealing with images of individuals in complex poses or against intricate backgrounds, resulting in noticeable drops in performance. Moreover, Conditional GANs (cGANs) encounter difficulties with significant spatial transformations between the clothing and the subject's posture, as highlighted by CP-VTON [56], which brings to light the need for improved methods capable of handling these challenges.

Diffusion-based virtual try-on. The exceptional generative capabilities of diffusion have inspired several approaches to incorporate diffusion models into fashion synthesis, covering tasks like visual try-on [3,6,24,39,59,66]. TryOnDiffusion [66] utilizes dual U-Nets for the try-on task. However, the requirement for extensive datasets capturing various poses presents a significant challenge. Consequently, there has been a pivot towards leveraging large-scale pre-trained diffusion models as priors in the try-on process [20,48,51,61]. Approaches like LADI-VTON [39] and DCI-VTON [15] have been introduced, which treat clothing as pseudo-words or use warping networks to integrate garments into pre-trained diffusion models. StableVITON [26] proposes a novel approach that conditions the intermediate feature maps of a spatial encoder using a zero cross-attention block. While these methods have addressed the issues of background complexity, they struggle to preserve fine details and encounter problems such as inaccurate warping. Moreover, they fail to enable flexible control over how garments are worn, producing only a rigid, static image.

Our work extends the interactive capabilities of virtual try-on, which sets new standards for performance and user interaction. Rather than just generating static images, `Wear-Any-Way` empowers users to manipulate garments dynamically, allowing for an unprecedented level of customization that signifies a significant leap in the personalization of digital fashion experiences.

Point-based image editing. Building on the achievements of diffusion models [20], various diffusion-based image editing techniques [1,4,7,18,25,38] have emerged, predominantly relying on textual instructions for editing. Techniques such as those in [25, 28, 55] apply fine-tuning to models on single images to produce alterations directed by descriptive texts. However, this text-guided approach often yields only broad-stroke modifications, lacking the precision required for detailed image editing.

To overcome the limitations of text-guided editing, studies explored point-based editing [13, 44, 57]. DragGAN, notable for its intuitive drag-and-drop manipulation, optimizes latent codes for handle points and incorporates point tracking. However, GANs' inherent limitations constrain DragGAN. FreeDrag [34] refines DragGAN by eliminating point tracking, while [53] extends DragGAN's framework to diffusion models, showcasing versatility. Simultaneously,

[41] utilizes diffusion models for drag-based editing, employing classifier guidance to convert editing intentions into actionable gradients.

To address these shortcomings and enhance the granularity and adaptability of image editing, our research leverages diffusion models' exceptional generative capabilities. We propose a novel editing paradigm, `Wear-Any-Way`, combining point-based precision with diffusion models' rich generative potential. `Wear-Any-Way` enhances fine-grained, context-aware image alterations, offering unprecedented control over the editing process and outcomes.

3 Method

We first introduce the basic knowledge required for diffusion models in Sec. 3.1. Afterward, we present our strong baseline for virtual try-on in Sec. 3.2. Next, we dive into the details of our proposed sparse correspondence alignment in Sec. 3.3 and the training strategies in Sec. 3.4. In addition, in Sec. 3.5 we also elaborate on the details for collecting the training point pairs and finally summarize the the inference pipeline in Sec. 3.6.

3.1 Preliminaries

Text-to-image diffusion model. Diffusion models [19] exhibit promising capabilities in both image and video generation. In this study, we select the widely adopted Stable Diffusion [50] as our foundational model, leveraging its efficient denoising procedure in the latent space. The model initially employs a latent encoder [27] to project an input image $\mathbf{x}_0$ into the latent space: $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. Throughout training, Stable Diffusion transforms the latent representation into Gaussian noise using the formula:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

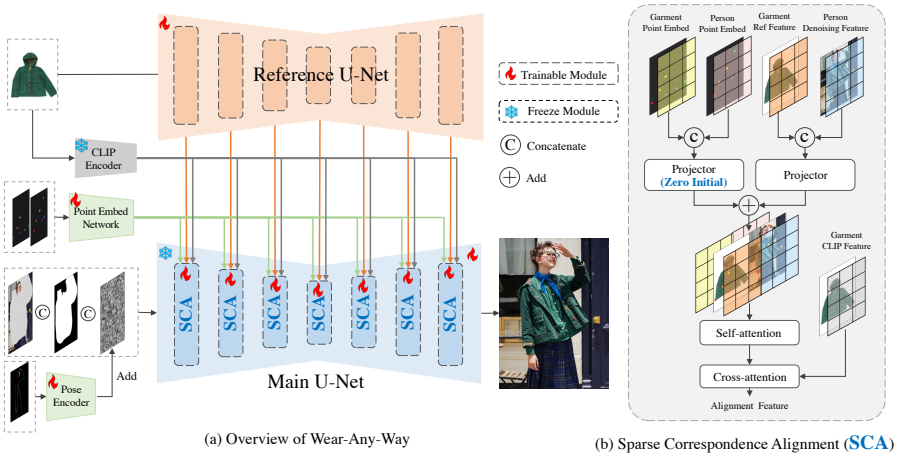
where $\epsilon \sim \mathcal{U}([0,1])$, and $\bar{\alpha}_t$ is the cumulative product of the noise coefficient α_t at each step. Subsequently, the model learns to predict the added noise as:

$$\mathbb{E}_{\mathbf{z},\mathbf{c},\epsilon,t}(|\epsilon_\theta(\mathbf{z}_t,\mathbf{c},t) - \epsilon|_2^2). \tag{2}$$

Here, t represents the diffusion timestep, and $\mathbf{c}$ denotes the conditioning text prompts. During the inference phase, Stable Diffusion effectively reconstructs an image from Gaussian noise step by step, predicting the noise added at each stage. The denoised results are then fed into a latent decoder to regenerate colored images from the latent representations, denoted as $\hat{\mathbf{x}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0)$.

3.2 Virtual Try-on Pipeline

As demonstrated in Fig. 2, our pipeline consists of two branches. The main branch is the inpainting model initialized with the pre-trained weight of Stable Diffusion [50]. It takes in a 9-channel tensor as input, with 4 channels of latent



(a) Overview of Wear-Any-Way (b) Sparse Correspondence Alignment (**SCA**)

Fig. 2: The pipeline of Wear-Any-Way. The overall framework consists of two U-Nets. The reference U-Net takes the garment image as input to extract fine-grained features. The main U-Net takes charge of generating the try-on results. It takes the person image (masked), the garment mask, and the latent noise as input. We exert the pose control via an additional pose encoder. The point-based control is realized by a point embedding network and a sparse correspondence alignment module. The detailed structures are demonstrated on the right part. Symbols of flames and snowflakes denote trainable and frozen parameters respectively.

noise, 4 channels of the latent for the inpainting background (*i,e.,* person image with the masked clothes region), and 1 channel for the binary mask (representing for inpainting region). The original SD receives text embedding as conditions to guide the diffusion procedure. Instead, we replace the text embedding with an image embedding of the garment image extracted by a CLIP [49] image encoder.

The CLIP image embedding could guarantee the overall colors and textures of the garment but fails to preserve the fine details. Recently, reference-only [64] has proven effective in keeping the fine details of the reference image in many fields of applications [5, 22, 42, 60]. TryonDiffusion [66] also leverages two U-Net to make generations with high fidelity. Inspired by those explorations, we also use a reference U-Net to extract the detailed features of the garment. Our reference U-Net is a standard text-to-image diffusion model with 4-channel input. We conduct feature fusion after each block by concatenating the "key" and "value" of the reference U-Net after the main U-Net.

To further enhance the generation, we add the pose map as an additional control. We construct a tiny convolution network to extract the features of the pose map, and directly add it to the latent noise of the main U-Net. The pose map is extracted from the provided person image with DW-Pose [62].

3.3 Sparse Correspondence Alignment

To make our `Wear-Any-Way` customizable, we introduce a sparse correspondence alignment mechanism into the diffusion procedure. Specifically, for the point pair, one point is marked on the garment image and another is given on the person image. We utilize the correspondence between the two points to control the generation results: the marked position of the garment would match the targeted position on the person image. In this way, users could precisely control the wearing style by manipulating multiple point pairs.

As shown in Fig. 2, we first learn a series of point embedding to represent the pair of control points. Afterward, we inject this control signal into both the main U-Net and referential U-Net. To assist the model in learning the correspondence relationships, we propose several strategies like condition dropping, zero-initialization, and point-guided loss.

Point embedding. Assuming we sample N pairs of points, we use disk maps $\mathbf{D}_{g/p}^{1\times H\times W}$ to represent the points on the garment and person images respectively. The background values are zeros. Points on the garment image are filled with values from 1 to K randomly without repeat, while the points on the person image are filled with the corresponding values. K denotes the maximum number of control points enabled in one image with $N \leq K$. In this work, we set K = 24. This random assignment decouples the semantics and the points thus making the point representation permutable. It serves as the basis to support arbitrary numbers of point controls at arbitrary locations. Afterward, we design a point embedding network with stacked convolution layers to project the disk maps into high-dimension embeddings: $\mathbf{E}_{g/p}^{C\times H\times W}$. This network is optimized alone with the diffusion model in the end-to-end training.

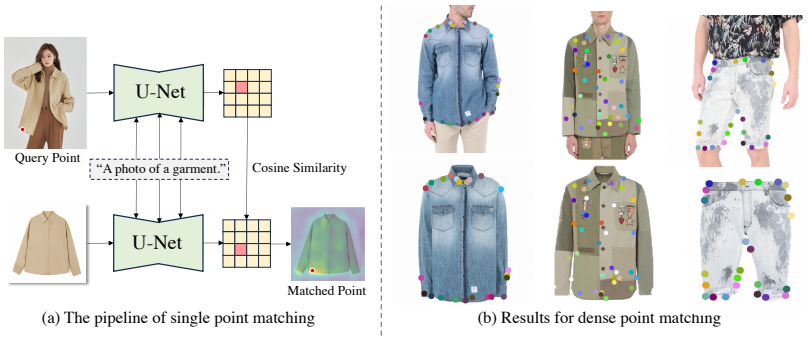
Embedding injection. We excel in controls by injecting the point embedding into the attention layer. In our baseline, the features from the reference U-Net are concatenated on the "key" and "value" of the self-attention as illustrated in Eq. (3), where the subscripts of m,r represents the main and reference U-Net.

$$\text{Attention} = \text{softmax}(\frac{Q_m \cdot \text{cat}(K_m, K_r)^T}{\sqrt{d_k}}) \cdot \text{cat}(V_m, V_r) \tag{3}$$

This attention layer enables garment features extracted by the reference U-Net to be integrated into the main U-Net. To enable the correspondence control with point guidance, we modify this attention layer via adding the point embedding of the person and garment with the "query" and "key" as in Eq. (4).

$$\text{Attention} = \text{softmax}(\frac{(Q_m + E_p) \cdot \text{cat}(K_m + E_p, K_r + E_g)^T}{\sqrt{d_k}}) \cdot \text{cat}(V_m, V_r) \tag{4}$$

In this way, when integrating the garment feature into the main U-Net, the feature aggregation would consider the correspondence of the point pairs. The feature located by the point of the garment could be aligned to the position of the point on the person image. Thus, users could assign control points to manipulate the wearing style by clicking and dragging.



(a) The pipeline of single point matching (b) Results for dense point matching

Fig. 3: Pipeline of collecting the training point-pairs. As shown on the left, the person and garment images are sent into the same Stable Diffusion to extract the feature. We calculate the cosine similarity between the two feature maps to get the point pairs. Some densely sampled point pairs are demonstrated on the right.

3.4 Training strategies

Besides the design of the model structure, we also develop several training strategies to assist `Wear-Any-Way` to learn the correspondence alignment.

Condition dropping. We observe that, even without the guidance of point pairs, the generation results on the training set are already very close to the ground truth. We analyze that, the inpainting mask and the human pose map could indicate the wearing style of the training sample to some extent. To enforce the model to learn from the control point, we increase the possibility of dropping the pose map and degrading the inpainting mask to a box around the mask.

Zero-initialization. Adding the point embedding to the attention "key" and "value" causes the unstability of the training optimization. To achieve a progressive integration, we get the inspiration of ControlNet [65] to add a zero-initialized convolutional layer at the output of the point embedding network. This zero-initialization brings better convergence.

Point-weighted loss. To enhance the controllability of the paired points. We increase the loss weight around the sampled points on the person image. The supervision of `Wear-Any-Way` is an MSE loss for predicting the noise.

3.5 Training Points Collection

It is crucial to get precisely matched point pairs for training. Considering that there are no densely annotated point data between the garment and person image, we make extensive explorations to collect point pairs.

The challenge lies in the fact that garments are not rigid like steel. When worn on the human body, garments could undergo deformation. Previous virtual matching/correspondence learning methods [33,46,52] could only deal with rigid objects like buildings. Fashion/human key points detection methods [37,62,67] could localize point pairs. However, they could only detect a few predefined key

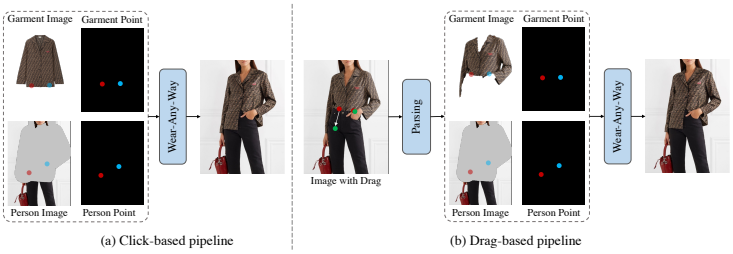


Fig. 4: The inference pipeline of `Wear-Any-Way`. For click-based control, users provide garment images, person images, and point pairs to customize the generation. When the user drags the image, the starting and end points are translated as the garment and person points. While the parsed clothes are regarded as the garment image. Thus, the drag could be transformed into the click-based setting.

points, which fail to generate arbitrary sampled points for truly flexible control. Recently, some works [17, 54, 63] find that the pre-trained diffusion models are naturally image matchers. In this work, we also explore correspondence learning by leveraging pre-trained text-to-image diffusion models.

As illustrated in Fig. 3, we leverage the siamese text-to-video diffusion model to extract features from the person and garment image respectively. We take the feature map of the last later and ensemble the prediction results at multiple time steps to get a robust matching. Given a point on the person image, we select the corresponding point on the garment image with the maximum cosine similarity. The dense point-matching results are demonstrated on the right of Fig. 3. Given a person image, we first extract the mask for the wearing clothes. Afterward, we randomly sample points in the internal and boundary regions of the mask as queries and leverage the matching pipeline to extract the corresponding points on the garment image. We chose the mapping direction from the person image to the garment image because some points on the garment image could not be matched on the on-body image when the poses were complex.

3.6 Inference with manipulation

Equipped with the sparse correspondence alignment, `Wear-Any-Way` supports users to customize the try-on results using control points. The inference pipeline is illustrated in Fig. 4. For the click-based setting, besides providing the garment image and person images, users could assign multiple point pairs on these two images as control signals. The coordinates on the garment image indicated by the garment points could be aligned to the corresponding position of the person points in the generation result. For the drag-based control, the starting and end points are processed as garment points and person points, while the parsed clothes are regarded as the garment image. In this way, the drag-based manipulation could be transformed into click-based controls.

4 Experiments

4.1 Implementation Details

Detailed configurations. The main U-Net and the reference U-Net both leverage the pre-trained weights from Stable Diffusion-1.5 [50]. We collect 0.3 million high-quality try-on data with "person image, up-clothes, down-clothes" triplets to train our model. However, for fair comparisons with other works, we also train `Wear-Any-Way` on VITON-HD [10] and Dresscode [40] respectively to report the quantitative results and make qualitative comparisons. To make our model try the upper and down clothes in one pass. We concatenate two input garment image together from H × W × 3 to H × 2W × 3. We randomly drop a garment image to an all-zeros image to preserve the ability of single-garment generation.
Training hyper-parameters. During training, we set the initial learning rate 5e-5 with a batch size of 64. The models are trained using 8× A100 GPUs. For the main U-Net, we train the parameters of the decoder, and the self-attention layers of the encoder; All parameters of the reference U-Net are trained. For the data augmentation, we conduct random crop on the person image, and exert random flip simultaneously on the garment and person images. In addition, we also use random color jitter to improve the robustness. We train our model with the resolution of 768 × 576 on the self-collected data for better visual quality. We also train a 512 × 384 version for fair comparisons with previous works.

4.2 Evaluation protocols.

Standard virtual try-on. We first evaluate the performance of `Wear-Any-Way` on standard virtual try-on benchmarks (*i.e.,*VITON-HD [10] Dresscode [40]) to report qualitative results. Meanwhile, we give qualitative comparisons with state-of-the-art methods to prove the effectiveness of our design.
Evaluation for point control. To evaluate the ability of point-based control, we get inspiration from previous drag-based image editing methods [41,45,53] to calculate the landmark distance. Specifically, we detect the fashion landmarks using FashionAI [67] detector on the pair of garment image and the person image. Afterward, we use the paired landmarks $\mathcal{L}_{garment}$ and $\mathcal{L}_{person}$ as the control points to generate a try-on image (the person image could be viewed as the ground truth). Next, we use the same detector to localize the landmarks on the newly generated image, noted as $\mathcal{L}_{gen}$. We calculate the Euclidean distance between $\mathcal{L}_{person}$ and $\mathcal{L}_{gen}$ to evaluate the control ability. Ideally, the landmark distance should be small if the generation is well-controlled by the points. We construct a benchmark covering the upper-, down-, and coat-clothes with 1000 samples in total for a comprehensive evaluation.

4.3 Ablations Studies

We first conduct experiments for our strong baseline of standard virtual try-on. Afterward, we provide a detailed analysis of the sparse correspondence alignment

Fig. 5: Ablation studies on feature extractors. Our design shows notable superiority compared with CLIP image encoder [49], DINOv2 [43], ControlNet [65].

module. In addition, we also discuss the correspondence matching methods used for collecting the paired points.

Strong baseline. We first investigate the design of our strong baseline for virtual try-on. We claim that the reference U-Net is crucial for preserving the fine details of garments. Previous works like AnyDoor [8] uses a image encoder (*e.g.*, DINOv2 [43]) to extract the garment features. StableVITON [26] leverages a ControlNet-like [65] structure to extract finer representations. We organize the qualitative comparisons in Fig. 5. We leverage the CLIP image encoder, DINOv2, and ControlNet as feature extractors and apply the same training settings for fair comparison. We observe that CLIP, DINOv2, and ControlNet could only encode the global appearances of the garments, but fail to preserve the identity of the detailed patterns/texts/logos. In contrast, the reference U-Net provides fine-grained details and is able to preserve the high-fidelity details of the garments.

Sparse correspondence alignment. It is the core component of our point-based control. We first conduct ablation studies for the control injection methods in Tab. 1. We follow the evaluation protocol introduced in Sec. 4.2 to calculate the landmark distance. Without the point embedding, our try-on baseline (row 1) gets a high landmark distance. In the second row, we first explore injecting the point embedding at the input noise of the main and reference U-Net. In the third row, we report the results of injecting the control signal in the attention layer as introduced in Sec. 3.3. In Tab. 2, we add the enhancement strategies presented in Sec. 3.4 step-by-step to verify the effectiveness of our designs.

Training point pair collection. As introduced in Sec. 3.5, we collect the control point by using a siamese U-Net structure. In this section, we make extensive experiments by comparing different correspondence-matching methods. We also utilize the benchmarks of fashion landmarks to evaluate the matching accuracy. Concretely, given a pair of landmarks on the person and garment image, we leverage different matching methods to map the landmarks from the

Fig. 6: Qualitative comparison for classical virtual try-on. We make comparisons on VITON-HD [10] test split with DCI-VTON [15], LaDI-VTON [39], KGI [32], AnyDoor [8], and StableVITON [26]. Our solution demonstrates notable superiority in detail preservation and generation quality.

Table 1: Point embedding injection. We compare different injection methods and report the landmark distance.

	$\text{Dist}_{\text{upper}}$	$\text{Dist}_{\text{down}}$	$\text{Dist}_{\text{coat}}$
None	35.65	21.13	43.34
Latent Noise	27.32	16.34	30.38
Attention q,k	24.35	15.79	27.27

Table 2: Enhancing strategies for correspondence alignments are added step by step to verify the effectiveness.

	$\text{Dist}_{\text{upper}}$	$\text{Dist}_{\text{down}}$	$\text{Dist}_{\text{coat}}$
Base (Attention k,q)	24.35	15.79	27.27
+ Zero-init	22.65	15.33	25.56
+ Condition-dropping	18.39	12.04	20.44
+ Point-weighted loss	17.65	10.32	20.32

person image to the garment image. Then, we calculate the distance between the mapped results with the ground truth landmarks. We observe that he pre-trained diffusion models demonstrated superior abilities compared with other feature extractors like CLIP [49] and DINOv2 [43]. We also include several specific correspondence-matching methods, and our pre-trained reference U-Net. Among them, the diffusion matcher demonstrates the best performance.

4.4 Comparisons with Existing Alternatives

In this section, we conduct intensive comparisons with existing alternatives. We first compete for previous arts in the standard setting of virtual try-on. Afterward, we compare interactive image editing methods that support click/drag.

Standard virtual try-on. We report the qualitative results in Tab. 3 on VITON-HD [10] and Dresscode [40] datasets. Our model gets the best result for the FID and KID and competitive performance for the SSIM and LPIPS.

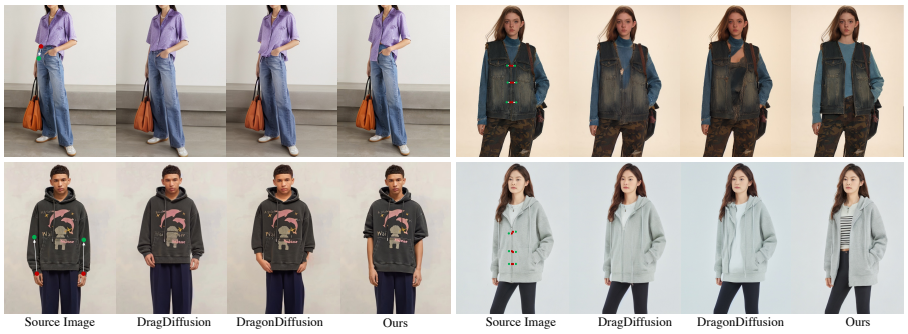


Fig. 7: Qualitative comparison with drag-based image editing methods. We make comparisons with DragDiffusion [53] and DragonDiffusion [41]

Table 3: Quantitative comparisons with existing state-of-the-art try-on solutions on VITON-HD and DressCode upper-body (D.C. Upper) datasets. **Bold** and underline denote the best and the second best result, respectively.

Train / Test	VITON-HD / VITON-HD				D.C. Upper / D.C. Upper			
Method	SSIM	LPIPS	FID	KID	SSIM	LPIPS	FID	KID
VITON-HD [21]	0.862	0.117	12.117	3.23	-	-	-	-
HR-VITON [30]	0.878	0.1045	11.265	2.73	0.936	0.0652	13.820	2.71
LADI-VTON [39]	0.864	0.0964	9.480	1.99	0.915	0.0634	14.262	3.33
Paint-by-Example [61]	0.802	0.1428	11.939	3.85	0.897	0.0775	15.332	4.64
DCI-VTON [15]	0.880	0.0804	8.754	1.10	**0.937**	0.0421	11.920	1.89
GP-VTON [58]	**0.884**	0.0814	9.072	0.88	0.769	0.2679	20.110	8.17
AnyDoor [8]	0.821	0.099	10.846	2.46	0.899	0.119	14.834	3.05
StableVITON [26]	0.852	0.0842	8.698	0.88	0.911	0.0500	**11.266**	0.72
`Wear-Any-Way`	0.877	**0.078**	**8.155**	**0.78**	0.934	**0.0409**	11.72	**0.33**

Considering that the quantitative results could not perfectly align with the real generation quality. We make qualitative comparisons with previous state-of-the-art solutions in Fig. 6. `Wear-Any-Way` demonstrates obvious advantages over other works for the generation quality and detail preservation.

Controllable generation. We prove the controllability of our model by comparing it with drag-based image editing methods like DragDiffusion [53] and DragonDiffusion [41]. We illustrate the comparison results in Fig. 7. We observe that DragDiffusion [44] could not precisely follow the instructions of drag, while DragonDiffusion [41] usually destroys the structure of humans and garments.

4.5 Qualitative Analysis

We illustrate more examples in Fig. 8. It is demonstrated that `Wear-Any-Way` supports the manipulation for various types of garments including coats, T-shirts, pants, hoodies, *etc.* Besides, users could assign arbitrary numbers of control points to get the customized generation results. As shown in the first row, assisted by the precise point control, `Wear-Any-Way` can conduct "continuous"

Fig. 8: Manipulable virtual try-on with click. `Wear-Any-Way` supports users to assign arbitrary numbers of control points on the garment and person image to customize the generation, bringing diverse potentials for real-world applications.

editing for splitting the coat gradually. The high controllability of `Wear-Any-Way` enables it to realize many fantastic styles of wearing, like rolled-up sleeves or paints, and the different types of tuck.

5 Conclusion

We propose `Wear-Any-Way`, a novel framework for manipulable virtual try-on. Besides reaching state-of-the-art performance on the classical setting, it enables users to customize the style of wearing by assigning control points. To achieve this, we introduce a sparse correspondence align module to make our model customizable. `Wear-Any-Way` serves as a practical tool for e-commerce and provides novel inspirations for the future research of virtual try-on.

Limitations and potential effect. Our methods could still generate some artifacts for fine details like human hands, especially when the hands only occupy a small region in the full image. This could be improved by using higher resolutions and larger diffusion models like SD-XL [47].

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022) 4
2. Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision. pp. 409–425. Springer (2022) 4
3. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5968–5976 (2023) 4
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 4
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023) 3, 6
6. Cao, S., Chai, W., Hao, S., Zhang, Y., Chen, H., Wang, G.: Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. arXiv preprint arXiv:2302.06826 (2023) 4
7. Chen, X., Feng, Y., Chen, M., Wang, Y., Zhang, S., Liu, Y., Shen, Y., Zhao, H.: Zero-shot image editing with reference imitation. arXiv preprint arXiv:2406.07547 (2024) 4
8. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 11, 12, 13
9. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: CVPR. pp. 14131–14140 (2021) 3
10. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: CVPR (2021) 10, 12
11. Cui, A., McKee, D., Lazebnik, S.: Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14638–14647 (2021) 4
12. Dong, H., Liang, X., Zhang, Y., Zhang, X., Shen, X., Xie, Z., Wu, B., Yin, J.: Fashion editing with adversarial parsing learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8120–8128 (2020) 1, 3
13. Endo, Y.: User-controllable latent transformer for stylegan image layout editing. arXiv preprint arXiv:2208.12408 (2022) 4
14. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: CVPR. pp. 8485–8493 (2021) 3, 4
15. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. arXiv preprint arXiv:2308.06101 (2023) 4, 12, 13
16. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: ICCV. pp. 10471–10480 (2019) 4
17. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. Advances in Neural Information Processing Systems **36** (2024) 9

18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 4
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020) 1, 5
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4
21. Honda, S.: Viton-gan: Virtual try-on image generator trained with adversarial loss. arXiv preprint arXiv:1911.07926 (2019) 13
22. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023) 3, 6
23. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1745–1753 (2019) 1, 3
24. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv preprint arXiv:2304.06025 (2023) 4
25. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) 4
26. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. arXiv preprint arXiv:2312.01725 (2023) 1, 4, 11, 12, 13
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013) 5
28. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264 (2022) 4
29. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5549–5558 (2020) 1, 3
30. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: ECCV. pp. 204–219. Springer (2022) 3, 4, 13
31. Li, K., Zhang, J., Chang, S.Y., Forsyth, D.: Controlling virtual try-on pipeline through rendering policies. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5866–5875 (2024) 3, 4
32. Li, Z., Wei, P., Yin, X., Ma, Z., Kot, A.C.: Virtual try-on with pose-garment keypoints guided inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22788–22797 (2023) 12
33. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. arXiv preprint arXiv:2306.13643 (2023) 8
34. Ling, P., Chen, L., Zhang, P., Chen, H., Jin, Y.: Freedrag: Point tracking is not you need for interactive point-based image editing. arXiv preprint arXiv:2307.04684 (2023) 4
35. Liu, J., Song, X., Chen, Z., Ma, J.: Mgcm: Multi-modal generative compatibility modeling for clothing matching. Neurocomputing **414**, 215–224 (2020) 1, 3
36. Liu, L., Zhang, H., Ji, Y., Wu, Q.J.: Toward ai fashion design: An attribute-gan model for clothing match. Neurocomputing **341**, 156–167 (2019) 1, 3

37. Liu, X., Li, J., Wang, J., Liu, Z.: Mmfashion: An open-source toolbox for visual fashion analysis. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3755–3758 (2021) 8
38. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021) 4
39. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501 (2023) 4, 12, 13
40. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: ECCV. pp. 2231–2235 (2022) 10, 12
41. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023) 3, 5, 10, 13
42. Nam, J., Kim, H., Lee, D., Jin, S., Kim, S., Chang, S.: Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. arXiv preprint arXiv:2402.09812 (2024) 3, 6
43. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv:2304.07193 (2023) 11, 12
44. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. arXiv preprint arXiv:2305.10973 (2023) 4, 13
45. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 10
46. Pautrat, R., Suárez, I., Yu, Y., Pollefeys, M., Larsson, V.: Gluestick: Robust image matching by sticking points and lines together. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9706–9716 (2023) 8
47. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952 (2023) 14
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 6, 11, 12
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 3, 5, 10
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
52. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020) 8
53. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435 (2023) 4, 10, 13

54. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36** (2024) 9
55. Valevski, D., Kalman, M., Molad, E., Segalis, E., Matias, Y., Leviathan, Y.: Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023) 4
56. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: ECCV. pp. 589–604 (2018) 4
57. Wang, S.Y., Bau, D., Zhu, J.Y.: Rewriting geometric rules of a gan. ACM Transactions on Graphics (TOG) **41**(4), 1–16 (2022) 4
58. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: CVPR. pp. 23550–23559 (2023) 3, 4, 13
59. Xu, Z., Chen, M., Wang, Z., Xing, L., Zhai, Z., Sang, N., Lan, J., Xiao, S., Gao, C.: Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. arXiv preprint arXiv:2404.17571 (2024) 4
60. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023) 3, 6
61. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023) 4, 13
62. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 6, 8
63. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. Advances in Neural Information Processing Systems **36** (2024) 9
64. Zhang, L.: Reference-only controlnet. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236 (20235) 3, 6
65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 8, 11
66. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023) 1, 4, 6
67. Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: Fashionai: A hierarchical dataset for fashion understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019) 8, 10