IMAGDressing-v1: Customizable Virtual Dressing

Fei Shen[1], Xin Jiang[1], Xin He[2], Hu Ye[3], Cong Wang[4], Xiaoyu Du[1], Zechao Li[1], Jinhui Tang[1]*

[1]Nanjing University of Science and Technology
[2]Wuhan University of Technology
[3]Tencent AI Lab
[4]Nanjing University

Abstract

Existing virtual try-on (VTON) methods provide only limited user control over garment attributes and generally overlook essential factors such as face, pose, and scene context. To address these limitations, we introduce the virtual dressing (VD) task, which aims to synthesize freely editable human images conditioned on fixed garments and optional user-defined inputs. We further propose a comprehensive affinity metric index (CAMI) to quantify the consistency between generated outputs and reference garments. We present IMAGDressing-v1, which leverages a garment-specific U-Net to integrate semantic features from CLIP and texture features from a VAE. To incorporate these garment features into a frozen denoising U-Net for flexible text-driven scene control, we employ a hybrid attention mechanism composed of frozen self-attention and trainable cross-attention layers. IMAGDressing-v1 seamlessly integrates with extension modules, such as ControlNet and IP-Adapter, enabling enhanced diversity and controllability. To alleviate data constraints, we introduce the Interactive Garment Pairing (IGPair) dataset, comprising over 300,000 garment–image pairs and a standardized data assembly pipeline. Extensive experiments demonstrate that IMAGDressing-v1 achieves state-of-the-art performance in controlled human image synthesis.

Code — https://github.com/muzishen/IMAGDressing

Introduction

Virtual dressing (VD) aims to achieve comprehensive and personalized clothing displays for merchants by utilizing given garments and optional faces, poses, and descriptive texts. This technology holds significant potential for practical applications in e-commerce and entertainment. However, existing works primarily focus on virtual try-on (VTON) (Han et al. 2018; Choi et al. 2021; Kim et al. 2024; Li et al. 2021b, 2022b,a; Xu et al. 2024b; Yang et al. 2024) tasks for consumers, which involve given garments and fixed human conditions, lacking flexibility and editability. While VD offers greater freedom and appeal, it also presents more significant challenges.

To enhance the shopping experience for consumers in e-commerce, VTON (Han et al. 2018; Choi et al. 2021) tasks have become increasingly popular within the community. Early methods primarily relied on generative adversarial network (GAN) (Creswell et al. 2018), typically comprising a warping module to learn the semantic correspondence between clothes and the human body, and a generator module to synthesize the warped clothes onto the person's image. However, GAN-based methods (Choi et al. 2021; Han et al. 2018) often suffer from instability due to the min-max training objective, and they have limitations in preserving texture details and handling complex backgrounds. Recently, latent diffusion models (Ramesh et al. 2022; Zhang, Rao, and Agrawala 2023; Saharia et al. 2022) have made significant advances in VTON applications. These methods (Kim et al. 2024; Zeng et al. 2024; Xu et al. 2024b) better retain the texture information of the input garments through a multi-step denoising process, ultimately generating images

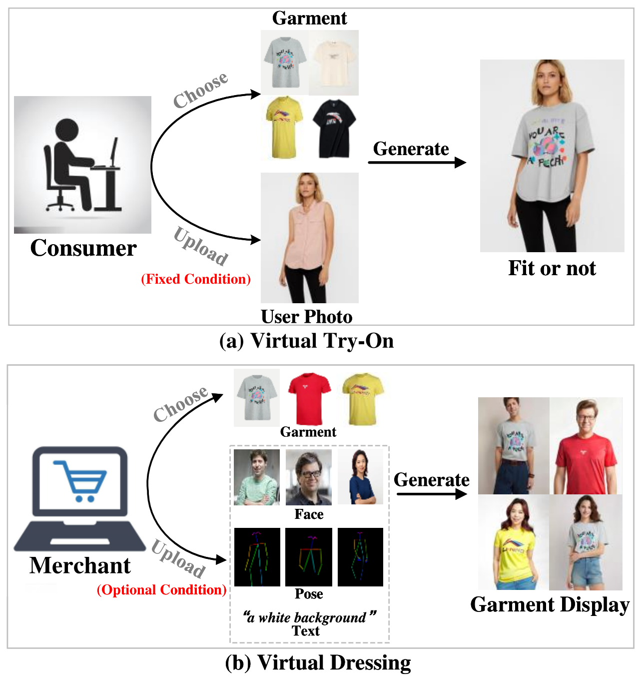

Figure 1: Differences between virtual try-on and virtual dressing tasks in conditions and applicable scenarios.

*Corresponding Author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

of specific individuals wearing the target clothing. Nevertheless, as illustrated in Figure 1 (a), VTON is essentially a local image inpainting task for consumer scenarios, requiring only the faithful preservation of given garment features. It overlooks the need for comprehensive clothing displays in merchant scenarios, lacking the ability to customize faces, poses, and scenes.

To address this, as illustrated in Figure 1 (b), we define a virtual dressing (VD) task aimed at generating freely editable human images with fixed garment and optional conditions, and then design a comprehensive affinity metric index (CAMI) to evaluate the consistency between generated images and reference garments. The differences between VTON and VD are as follows: (1) ***User Experience.*** VTON synthesizes images based on given clothing and specific person conditions, providing users with a static experience of partial inpainting. In contrast, VD centers on clothing and combines it with optional conditions to synthesize images, offering users a more interactive and flexible experience. (2) ***Application Scenarios.*** VTON is primarily used for personalized services for consumers, such as trying on clothes online to see if they suit them. In comparison, VD is mainly used by merchants on e-commerce platforms to showcase clothing, providing a comprehensive view of clothing ensembles. (3) ***Accuracy Requirements.*** VTON focuses on ensuring natural transitions and detailed handling between the clothing and the given model's body. Building on these requirements, VD further emphasizes the uniformity and aesthetics of clothing displays under given clothing conditions and optional elements.

Furthermore, this paper presents IMAGDressing-v1, a latent diffusion model specifically designed for custom virtual dressing for merchants. IMAGDressing-v1 consists primarily of a trainable garment UNet and a denoising UNet. Since the VAE can nearly losslessly reconstruct images, the garment UNet is used to simultaneously capture semantic features from CLIP and texture features from the VAE. The denoising UNet introduces a hybrid attention module, comprising a frozen self-attention and a trainable cross-attention modules, to integrate clothing features from the clothing UNet into it. This integration allows users to control different scenes through text prompts. Moreover, IMAGDressing-v1 can be combined with other extensions, such as ControlNet (Zhang, Rao, and Agrawala 2023) and IP-Adapter (Ye et al. 2023), to enhance the diversity and controllability of generated images. Lastly, to address the issue of data scarcity, we have collected and released the large-scale interactive garment pairing (IGPair) dataset, containing over 300,000 pairs of clothing and dressed images. The contributions of this paper are summarized as follows:

- This paper introduces a new virtual dressing (VD) task for merchants and designs a comprehensive affinity measurement index (CAMI) to evaluate the consistency between generated images and reference garment.
- We propose IMAGDressing-v1, which includes a garment UNet for extracting fine-grained clothing features and a denoising UNet with a hybrid attention module to balance clothing features with text prompt control.
- IMAGDressing-v1 can be combined with other extensions, such as ControlNet and IP-Adapter, to enhance the diversity and controllability of generated images.
- We collect and release a large-scale interactive garment pairing (IGPair) dataset, containing over 300,000 pairs, available for the community to explore and research.

Related Work

Virtual Try-On

Early virtual try-on (Lee et al. 2020; Liu et al. 2020; Choi et al. 2021) typically utilized generative adversarial networks (GANs) (Creswell et al. 2018) and a two-stage strategy. Initially, they would warp the clothing to the desired shape, then use a GAN-based generator to merge the warped clothing onto the human model. For instance, VITON-HD (Choi et al. 2021) addresses issues of clothing-body occlusion and mismatch by performing warping and segmentation simultaneously. GP-VTON (Xie et al. 2023) introduces local warping and global parsing to independently model the deformation of different clothing regions, aiming for a more accurate fit. To achieve precise clothing deformation, some methods (Han et al. 2019; Lee et al. 2022) estimate a dense flow map to guide the reshaping process. Additionally, some approaches (Ge et al. 2021; Zhang et al. 2020a,b) use normalization or distillation strategies to address the misalignment between the warped clothing and the human body. However, GAN-based methods face instability due to the min-max nature of their training objectives and have limitations in preserving texture details and handling complex backgrounds.

Recent research (Morelli et al. 2023; Wang et al. 2024; Wang, Guo, and Zhao 2022; Zhang et al. 2024b; Wei and Zhang 2024; Gou et al. 2023; Guan et al. 2024a; Zhang et al. 2024a; Wan et al. 2025; Zhu et al. 2023) have incorporated pre-trained diffusion models as priors for VTON tasks. For example, LADI-VTON (Morelli et al. 2023) and DCI-VTON (Gou et al. 2023) explicitly warp clothes to align them with the human body, then use diffusion models to merge the clothes with the body. TryOnDiffusion (Zhu et al. 2023) proposed an architecture with two parallel UNets and demonstrated the capability of diffusion-based virtual try-on by training on large-scale datasets. Similarly, OOTDiffusion (Xu et al. 2024b) and IDM (Choi et al. 2024) utilize parallel UNets for garment feature extraction and enhance integration through self-attention. StableVITON (Kim et al. 2024) introduces a zero-initialized cross-attention block to inject intermediate features of the spatial encoder into the UNet decoder. While diffusion-based VTON methods can combine clothing with a fixed model, producing fine-grained static images, VTON is essentially a local image inpainting task tailored for consumer scenarios, merely needing to faithfully preserve the given clothing features. As previously mentioned, VTON overlooks the need for comprehensive garment presentation in commercial contexts and cannot customize faces, poses, and scenes.

Dataset	Public	Caption	#Garments	#Pairs	Resolution
TryOnGAN	✗	✗	52,000	52,000	512×512
Revery AI	✗	✗	321,000	321,000	512×512
VITON-HD	✓	✗	13,679	13,679	1024×768
Dress Code	✓	✗	53,792	53,792	1024×768
IGPair (Ours)	✓	✓	86,873	324,857	$2K \times 2K$

Table 1: Comparison between IGPair and the widely used datasets.

Latent Diffusion Model

While latent diffusion models (LDMs) (Rombach et al. 2022a; Ma et al. 2024c,b; Shen et al. 2023, 2024; Shen and Tang 2024; Guan et al. 2024b; Zhang et al. 2025; Ma et al. 2024a) have been widely used for text-to-image (T2I) generation and editing tasks, the inaccuracy of natural language limits fine-grained image control. Various methods have been proposed to add conditional control to T2I diffusion models to address this. For example, ControlNet (Zhang, Rao, and Agrawala 2023) and T2I Adapter (Mou et al. 2024) introduce additional conditional encoding modules, such as edges, depth, and human poses, to control diffusion models and text prompts. IP-Adapter (Ye et al. 2023) conditions T2I diffusion models on high-level semantics of reference images, using both text and visual cues to guide image generation. Uni-ControlNet (Zhao et al. 2024) proposes a unified framework that flexibly and composably handles different conditional controls within a single model to reduce computational costs. MasaCtrl (Cao et al. 2023) achieves consistent image generation and complex non-rigid image editing through self-attention transformation without additional training costs. Similarly, InstructPix2Pix (Brooks, Holynski, and Efros 2023) retrains LDMs by adding extra input channels to the first convolutional layer to follow editing instructions. PCDMs (Shen et al. 2023) proposes proposes a multi-stage conditional diffusion model for pose guided character image generation. In this paper, we leverage the capabilities of frozen LDMs in text-to-image generation to achieve garment-centric image generation and editing.

IGPair Dataset

➤ **Q1:** ***What kind of data is suitable for VD task?*** We identify three critical requirements for an ideal virtual dressing dataset: (1) it should be publicly accessible for research purposes; (2) it should include high-resolution images of both garment and models wearing the clothing; (3) it should encompass a variety of scenes and styles, with detailed textual descriptions. As shown in Table 1, the proposed IGPair dataset not only meets all the aforementioned requirements but also provides six times the number of image pairs compared to the largest publicly available dataset, VITON-HD (Choi et al. 2021), surpassing TryOnGAN (Lewis, Varadharajan, and Kemelmacher-Shlizerman 2021), Revery AI (Li et al. 2021a), VITON-HD (Choi et al. 2021), and Dress Code (Morelli et al. 2022) datasets. Notably, IGPair includes multiple models for each clothing item. It is also the only dataset with a resolution exceeding

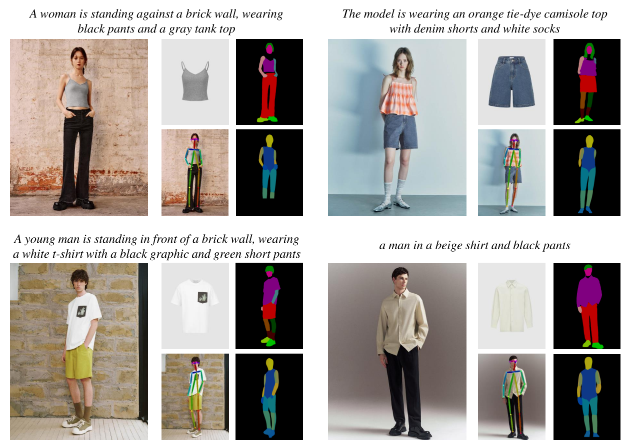

Figure 2: Sample pairs from the IGPair dataset, including pose keypoints, dense poses, and human body segmentation masks. More sample refer to supplementary material.

$2K \times 2K$. Additionally, IGPair is the only publicly available dataset that includes textual descriptions, diverse scenes, and various styles.

➤ **Q2:** ***How is the IGPair dataset collected and annotated?*** All images are sourced from the internet and encompass a variety of clothing styles, including casual, formal, athletic, fashionable, and vintage, etc. Initially, we collected 500,000 garment images, each accompanied by 2 to 5 images of people wearing the clothing from different perspectives. We then use classifiers to differentiate between clothing and human and employ a human pose estimator to select complete and usable images of clothing on models. After this automated stage, we manually verified all images. We categorize the garment into 18 types, and the dataset consists of 324,857 image pairs. To further enrich our dataset, we use OpenPose (Cao et al. 2017) to extract 18 key points for each human figure, DensePose (Güler, Neverova, and Kokkinos 2018) to compute dense poses for each reference model, and SCHP (Li et al. 2020) to generate segmentation masks for body parts and clothing items. We utilize BLIP2-OPT-6.7B (Li et al. 2023), INTERNLM-XCOMPOSER2-VL-7B (Dong et al. 2024), LLaVA-V1.5-13B (Liu et al. 2024), and Qwen-VL-Chat (Bai et al. 2023) to generate captions for the images. All model images are anonymized. Samples of human models and clothing pairs from our dataset, along with the corresponding additional information, are shown in Figure 2. More detail refer to supplementary material.

➤ **Q3:** ***How to evaluate the consistency between the generated image and the garment?*** We propose a comprehensive affinity metric index (CAMI) for evaluating VD task, which includes the unspecified score (CAMI-U) and the specified score (CAMI-S). CAMI-U represents the score for image generation without specified pose, face, and text scenarios. In contrast, CAMI-S represents the score for image generation with the specified pose, face, and text scenarios. CAMI-U focuses on the clothing images' structure S_s, texture S_t, and keypoints S_k. CAMI-S builds upon CAMI-U by adding pose matching degree S_p, facial similarity S_f, and

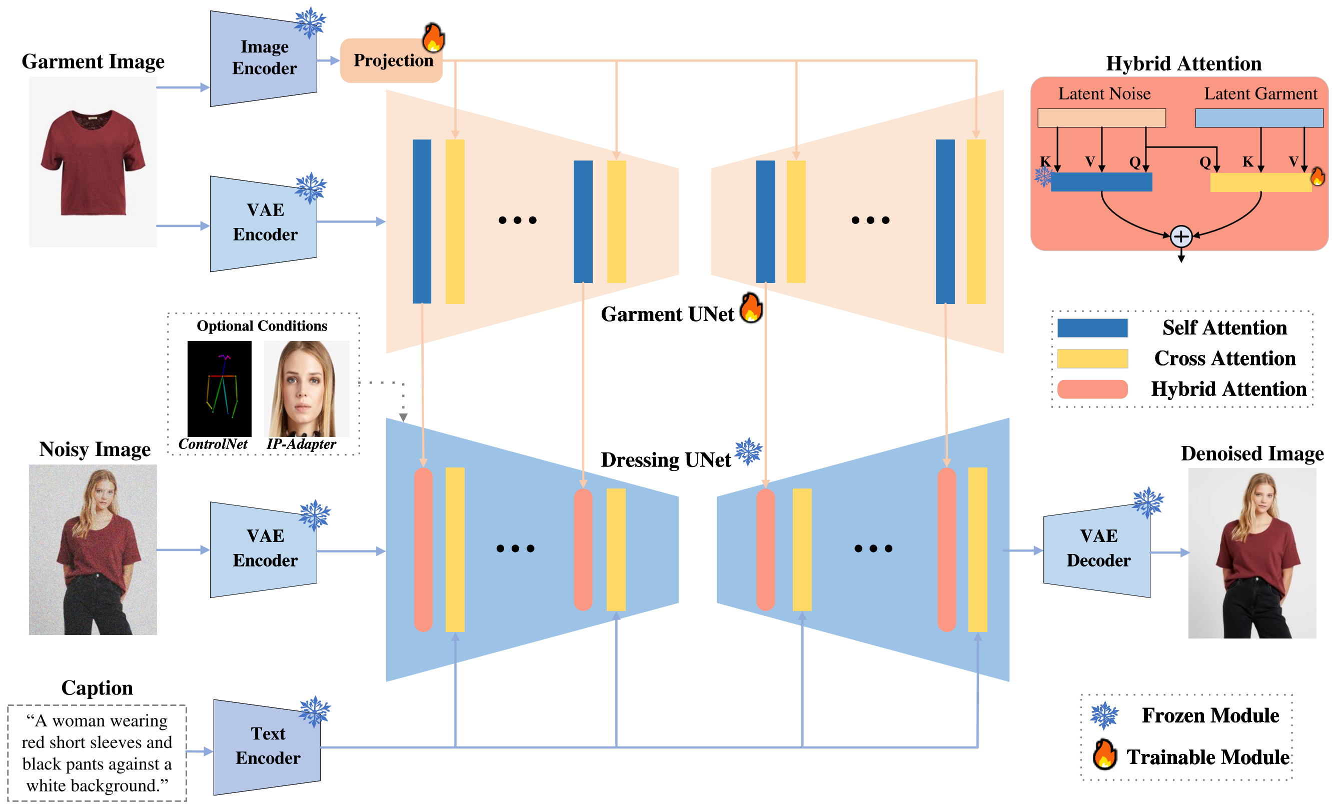

Figure 3: Illustration of the proposed IMAGDressing-v1 framework. It mainly consists of a trainable garment UNet and a frozen denoising UNet. The former extracts fine-grained garment features, while the latter balances these features with text prompts. IMAGDressing-v1 is compatible with other community modules, such as ControlNet and IP-Adapter.

text-image matching degree S_c.

$$S_{\text{CAMI-U}} = S_s + S_t + S_k, \quad (1)$$
$$S_{\text{CAMI-S}} = S_{\text{CAMI-U}} + S_p + S_f + S_c. \quad (2)$$

More detailed settings are to be provided in the supplementary material. Additionally, we also utilize MP-LPIPS (Chen et al. 2024), and ImageReward (Xu et al. 2024a) to evaluate the quality of the generated images.

Methodology

Preliminaries

Unlike other pixel-based diffusion models, latent diffusion models (LDMs) (Rombach et al. 2022a) aim to perform the denoising process in the latent space to reduce computational costs. An LDM typically comprises a variational autoencoder (VAE) (Kingma and Welling 2013), a CLIP text encoder (Radford et al. 2021), and a denoising UNet. The VAE transforms images into latent space representations and vice versa. Specifically, the VAE encoder $\mathcal{E}$ compresses the original image x into a latent representation z, i.e., $z = \mathcal{E}(x)$, while the VAE decoder $\mathcal{D}$ reconstructs the image x from the latent representation z, i.e., $x = \mathcal{D}(z)$. The CLIP text encoder converts text prompts into token embeddings c. During the diffusion process, Gaussian noise ϵ is added to the latent representation z over timestep t to produce z_t, where $t \in [0, T]$. The denoising UNet then iteratively denoises z_t during the denoising process. To learn such a denoising UNet ϵ_θ parameterized by θ, for each timestep t, the training objective usually adopts a mean square error loss L_{LDM}, as follows,

$$L_{LDM} = \mathbb{E}_{\mathbf{z}_t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{c}, t} \left\| \epsilon_\theta \left(\mathbf{z}_t, \mathbf{c}, t \right) - \epsilon_t \right\|^2, \quad (3)$$

where $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon_t$ is the noisy latent at timestep t and ϵ_t is the added noise. Here, $\mathbf{z}_0 = \mathcal{E}\left(\mathbf{x}_0\right)$ and x_0 represents the real data with a text condition c.

During the sampling stage, the predicted noise is calculated using both the conditional model $\epsilon_\theta(x_t, c, t)$ and the unconditional model $\epsilon_\theta(x_t, t)$ via classifier-free guidance (Ho and Salimans 2022).

$$\hat{\epsilon}_\theta(x_t, c, t) = w\epsilon_\theta(x_t, c, t) + (1-w)\epsilon_\theta(x_t, t). \quad (4)$$

Here, w is the guidance scale used to adjust the influence of the condition c.

IMAGDressing-v1

As shown in Figure 3, the proposed IMAGDressing-v1 mainly consists of a trainable garment UNet, architecturally same Stable Diffusion V1.5 (SD v1.5) [1]. The difference lies in the garment UNet's ability to simultaneously capture garment semantic features from CLIP and texture features from VAE, since the VAE can nearly losslessly reconstruct images. The lower part is a frozen denoising UNet, similar to SD v1.5, used for denoising the latent image under conditions. Unlike SD v1.5, we replace all self-attention modules with hybrid attention modules to more easily integrate garment features from the garment UNet and leverage the existing text-to-image capabilities for scene control via text prompts. Additionally, IMAGDressing-v1 includes an image encoder and projection layer for encoding garment features, as well as a text encoder for encoding textual features.

Garment UNet. Extracting fine-grained garment features is crucial for maintaining the consistency of garment details in VD task. To achieve this, the proposed garment UNet simultaneously extracts semantic information and texture features as garment characteristics. Specifically, given a garment image $\mathcal{X} \in \mathbb{R}^{3\times H\times W}$, we first convert it into a latent space representation $\mathbf{Z}_g \in \mathbb{R}^{4\times \frac{H}{8}\times \frac{W}{8}}$ using a frozen VAE Encoder [2]. Simultaneously, token embeddings are extracted from $\mathcal{X}$ using a frozen CLIP image encoder [3] and a trainable projection layer, where we utilize a Q-Former (Li et al. 2023) as the projection layer. Subsequently, the garment features from garment UNet interact thoroughly in the cross-attention mechanism, similar to the interaction between text and image in the original T2I model. Finally, the garment UNet aligns parallelly with the denoising UNet, injecting fine-grained features into the denoising UNet through hybrid attention. It is important to note that the garment UNet is only used to encode the reference image. Therefore, no noise is added to the reference image, and only a single forward pass is performed during the diffusion process.

Hybrid Attention. For VD task, an ideal denoising UNet should possess two key capabilities: (1) maintaining the original editing and generation abilities, and (2) incorporating additional garment features. The former can be achieved by freezing the modules of the denoising UNet, while the latter is accomplished through the proposed hybrid attention modules. Consequently, the architecture of the denoising UNet in IMAGDressing-v1 is similar to that of the original text-to-image SD v1.5 model, with the main difference being that we replace all self-attention modules in the denoising UNet with hybrid attention modules. As shown in Figure 3, the hybrid attention module consists of a frozen self-attention module and a learnable cross-attention module. Here, the weights of the self-attention of hybrid attention module are initialized using the self-attention's weights from SD v1.5. Assuming $\mathbf{Z}_d$ and $\mathbf{C}_g$ represent the query features and the garment features output by the garment UNet at corresponding positions, the output of hybrid attention $\mathbf{Z}_h$ can be defined as follows:

$$\mathbf{Z}_h = \underbrace{\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}}_{\text{Self Attention}} + \lambda \underbrace{\mathrm{Softmax}\left(\frac{\mathbf{Q}\left(\mathbf{K}'\right)^\top}{\sqrt{d}}\right)V'}_{\text{Cross Attention}}, \tag{5}$$

where $\lambda \in [0, 1.5]$ is a hyperparameter used to regulate the strength of garment conditions. $\mathbf{Q} = \mathbf{Z}_d\mathbf{W}_q$, $\mathbf{K} = \mathbf{Z}_d\mathbf{W}_k$, $\mathbf{V} = \mathbf{Z}_d\mathbf{W}_v$, $\mathbf{K}' = \mathbf{C}_g\mathbf{W}'_k$, and $\mathbf{V}' = \mathbf{C}_g\mathbf{W}'_v$. Here, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$, $\mathbf{W}'_k$, and $\mathbf{W}'_v$ are the weight matrices of the trainable linear projection layers. Noted that we share a query matrix $\mathbf{Q}$ for self attention and cross attention. In the hybrid attention module, the self-attention is frozen while the cross-attention is trainable. In other words, in Eq.5, only $\mathbf{W}'_k$ and $\mathbf{W}'_v$ are learnable. This approach allows us to retain the generative capabilities of the original T2I model, such as scene generation.

Training and Inference. During training stage, we keep the parameters of the basic modules in the denoising UNet unchanged and only optimize the remaining modules. Let $\mathbf{C}_t$ represent the text condition, then the loss function L_{LDM} is as follows,

$$L_{LDM} = \mathbb{E}_{\mathbf{z}_t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\mathbf{C}_t,\mathbf{C}_g,t}\left\|\epsilon_\theta\left(\mathbf{z}_t,\mathbf{C}_t,\mathbf{C}_g,t\right)-\epsilon_t\right\|^2, \tag{6}$$

In the inference stage, we also use classifier-free guidance according to Eq. 7.

$$\hat{\epsilon}_\theta(x_t,\mathbf{C}_t,\mathbf{C}_g,t) = w\epsilon_\theta(x_t,\mathbf{C}_t,\mathbf{C}_g,t) + (1-w)\epsilon_\theta(x_t,t). \tag{7}$$

➤ **Q4:** ***How support customized generation?*** As shown in Figure 3, the weights of the basic modules are frozen in the denoising UNet, making the garment UNet essentially an adapter module compatible with other community adapters for customized face and pose generation. For instance, to generate images of people in a given outfit and consistent pose, IMAGDressing-v1 can be combined with ControlNet-Openpose. To generate specific individuals wearing specified clothing, IMAGDressing-v1 can be integrated with the IP-Adapter. Furthermore, if both pose and face need to be specified simultaneously, IMAGDressing-v1 can be used in conjunction with both ControlNet-Openpose and IP-Adapter. Additionally, for VTON task, IMAGDressing-v1 also can be combined with ControlNet-Inpaint.

Method	ImageReward (↑)	MP-LPIPS (↓)	CAMI-U (↑)	CAMI-S (↑)
Blip-Diffusion	-2.224	0.1824	1.051	-
Versatile Diffusion	-2.055	0.4321	1.253	-
IP-Adapter	-2.267	0.4093	1.381	-
MagicClothing	-0.164	0.1499	1.655	2.692
Ours	**-0.095**	**0.1466**	**1.753**	**2.719**

Table 2: Quantitative comparison of the IMAGDressing-v1 with several state-of-the-art methods.

Experiments

Implementation Details

In our experiments, we initialize the weights of our garment UNet by inheriting the pre-trained weights of the UNet in

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

[2] https://huggingface.co/stabilityai/sd-vae-ft-mse

[3] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

Figure 4: Qualitative comparison with other SOTA methods under both unspecific and specific conditions, including BLIP-Diffusion (Li, Li, and Hoi 2023), Versatile Diffusion (Xu et al. 2023), IP-Adapter (Ye et al. 2023), and MagicClothing (Chen et al. 2024).

Method	ImageReward (↑)	MP-LPIPS (↓)	CAMI-U (↑)	CAMI-S (↑)
A0 (Base)	-0.245	0.1537	1.575	2.578
A1 (Base + IEB)	-0.178	0.1504	1.637	2.625
A2 (Base + IEB + HA)	**-0.095**	**0.1466**	**1.753**	**2.719**

Table 3: Quantitative results for different Settings. IEB and HA denote the image encoder branch and hybrid attention.

Stable Diffusion v1.5 (Rombach et al. 2022b), and finetune its weight. Our model is trained on the paired images from the IGPair dataset at the resolution of 512 $\times$ 640. We adopt the AdamW optimizer with a fixed learning rate of 5e-5. The model is trained for 200,000 steps on 10 NVIDIA RTX3090 GPUs with a batch size of 5. At the inference stage, the images are generated with the UniPC sampler for 50 sampling steps and set the guidance scale w to 7.0. Please refer to the supplementary material for more details.

Main Comparisons

We compare our IMAGDressing-v1 with four state of- the-art (SOTA) methods: Blip-Diffusion (Li, Li, and Hoi 2023), Versatile Diffusion (Xu et al. 2023), Versatile Diffusion (Xu et al. 2023), and MagicClothing (Chen et al. 2024).

Quantitative Results. As shown in Table 2, since Blip-Diffusion (Li, Li, and Hoi 2023), Versatile Diffusion (Xu et al. 2023), and IP-Adapter (Ye et al. 2023) are not specif-

Figure 5: Ablation study of each component.

Figure 6: Example results with different garment strength λ.

Figure 7: Examples of plug-in results of our IMAGDressing-v1 combined with ControlNet-Inpaint for virtual try-on.

ically designed VD models, they struggle to extract fine-grained garment features and generate character images that precisely match the text, pose, and garment attributes. This results in suboptimal performance across multiple metrics. Additionally, these models are incompatible with several plugins, making it impossible to compute the CAMI-S metric. Compared to MagicClothing (Chen et al. 2024), our IMAGDressing-v1 captures more detailed garment features through its image encoder branch and employs a hybrid attention mechanism. This mechanism integrates additional garment features while retaining the original text editing and generation capabilities. As a result, IMAGDressing-v1 demonstrates superior performance, outperforming other SOTA methods across all evaluation metrics.

Qualitative Results. Figure 4 illustrates the qualitative results of IMAGDressing-v1 compared to SOTA methods, including unspecific and specific condition generations. In Figure 4(a), under unspecific conditions, BLIP-Diffusion (Li, Li, and Hoi 2023) and Versatile Diffusion (Xu et al. 2023) fail to faithfully reproduce garment textures. Although IP-Adapter maintains the overall appearance of the garments, it does not preserve the details well and, more importantly, does not follow the text prompts accurately. MagicClothing aligns closely with the text conditions; however, it struggles to retain the overall appearance and details of the garments, such as printed text or colors. In contrast, IMAGDressing-v1 not only adheres to the text prompts but also preserves fine-grained garment details, demonstrating superior performance in VD tasks. Additionally, our method supports customized text prompt scenarios, as shown in the last three rows of Figure 4 (a). Furthermore, Figure 4 (b) illustrates the qualitative results under specific conditions. We observe that IMAGDressing-v1 significantly outperforms MagicClothing in scenarios involving given poses, faces, or both. The results generated by IMAGDressing-v1 exhibit superior texture details and a more realistic appearance. This demonstrates the enhanced compatibility of IMAGDressing-v1 with community adapters, which enhances the generated images' diversity and controllability.

Ablation Studies

Effectiveness of each component. Table 3 presents an ablation study to validate the effectiveness of the proposed image encoder branch (IEB) and hybrid attention (HA) module. Here, A0 (Base) denotes the setting without IEB and HA. We observe that A1, which uses IEB, shows improvements across all metrics, indicating that IEB effectively captures semantic garment features. Furthermore, A2 surpasses A1, demonstrating that the combination of IEB and HA further enhances quantitative results. Additionally, Figure 5 provides qualitative comparisons. We notice that A0 fails to adequately capture garment features in images with complex textures (2nd row). Although IEB (A1) partially addresses this issue, directly injecting IEB into the denoising UNet can lead to conflicts with the main model's features, resulting in obscured garment details (3rd). Therefore, the HA module (A2) improves image fidelity by adjusting the intensity of garment details within the garment UNet (4th row), aligning with our quantitative results.

Hyper-parameter λ. In Figure 6 demonstrates the effects of the hyper-parameter λ on generated samples with a fixed random seed. As λ increases to 1.0, the garment in the generated character becomes more similar to the input garment. A smaller λ ensures the generated results adhere more closely to the text prompts, while a larger λ biases the results towards the input garment. This indicates that λ effectively balances original editing and generation capabilities with additional garment features. Consequently, we empirically set λ to 1.0 in our experiments.

Potential application. Figure 7 illustrates a potential application of IMAGDressing-v1 in virtual try-on (VTON). By combining IMAGDressing-v1 with ControlNet-Inpaint and masking the garment area, we achieve VTON. The results demonstrate that IMAGDressing-v1 can achieve high-fidelity VTON, showcasing significant potential.

Conclusion

While recent advancements in VTON using latent diffusion models have enhanced the online shopping experience, they fall short of allowing merchants to showcase garments comprehensively with flexible control over faces, poses, and scenes. To bridge this gap, we introduce the virtual dressing (VD) task, designed to generate editable human images with fixed garments under optional conditions. Our proposed IMAGDressing-v1 employs a garment UNet and a hybrid attention module to integrate garment features, enabling scene control through text. It supports plugins like ControlNet and IP-Adapter for greater diversity and controllability. Additionally, we release the IGPair dataset with over 300,000 pairs of clothing and dressed images, providing a robust data assembly pipeline. Extensive experiments validate that IMAGDressing-v1 achieves state-of-the-art performance in controlled human image synthesis.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61925204) and the China Scholarship Council program.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chen, W.; Gu, T.; Xu, Y.; and Chen, C. 2024. Magic Clothing: Controllable Garment-Driven Image Synthesis. *CoRR*, abs/2404.09512.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420*.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7599–7607.

Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; and Feng, R. 2024a. Contrastive Multiview Subspace Clustering of Hyperspectral Images Based on Graph Convolutional Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Guan, R.; Tu, W.; Li, Z.; Yu, H.; Hu, D.; Chen, Y.; Tang, C.; Yuan, Q.; and Liu, X. 2024b. Spatial-Spectral Graph Contrastive Clustering with Hard Sample Mining for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7297–7306.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10471–10480.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7543–7552.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8176–8185.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5549–5558.

Lee, S.; Gu, G.; Park, S.; Choi, S.; and Choo, J. 2022. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, 204–219. Springer.

Lewis, K. M.; Varadharajan, S.; and Kemelmacher-Shlizerman, I. 2021. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–10.

Li, D.; Li, J.; and Hoi, S. C. H. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *NeurIPS*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, K.; Chong, M. J.; Zhang, J.; and Liu, J. 2021a. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15546–15555.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3260–3271.

Li, Y.; Wang, X.; Xiao, J.; and Chua, T.-S. 2022a. Equivariant and invariant grounding for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4714–4722.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022b. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

Li, Y.; Yang, X.; Shang, X.; and Chua, T.-S. 2021b. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4091–4099.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, J.; Song, X.; Chen, Z.; and Ma, J. 2020. MGCM: Multi-modal generative compatibility modeling for clothing matching. *Neurocomputing*, 414: 215–224.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024c. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. LaDI-VTON: latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2231–2235.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2024. Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models. *arXiv preprint arXiv:2407.02482*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2023. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Wan, S.; Li, Y.; Chen, J.; Pan, Y.; Yao, T.; Cao, Y.; and Mei, T. 2025. Improving Virtual Try-On with Garment-focused Diffusion Models. In *European Conference on Computer Vision*, 184–199. Springer.

Wang, S.; Feng, Y.; Lan, T.; Yu, N.; Bai, Y.; Xu, R.; Wang, H.; Xiong, C.; and Savarese, S. 2024. Text2Data: Low-Resource Data Generation with Textual Control. *arXiv preprint arXiv:2402.10941*.

Wang, S.; Guo, X.; and Zhao, L. 2022. Deep generative model for periodic graphs. *Advances in Neural Information Processing Systems*, 35: 35797–35810.

Wei, J.; and Zhang, X. 2024. DOPRA: Decoding Over-accumulation Penalization and Re-allocation in Specific Weighting Layer. *arXiv preprint arXiv:2407.15130*.

Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; and Liang, X. 2023. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23550–23559.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024a. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Xu, X.; Wang, Z.; Zhang, E. J.; Wang, K.; and Shi, H. 2023. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In *ICCV*, 7720–7731. IEEE.

Xu, Y.; Gu, T.; Chen, W.; and Chen, C. 2024b. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.

Yang, X.; Ding, C.; Hong, Z.; Huang, J.; Tao, J.; and Xu, X. 2024. Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7017–7026.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR*, abs/2308.06721.

Zeng, J.; Song, D.; Nie, W.; Tian, H.; Wang, T.; and Liu, A.-A. 2024. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8372–8382.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhang, L.; Meng, W.; Zhong, Y.; Kong, B.; Xu, M.; Du, J.; Wang, X.; Wang, R.; and Liu, L. 2025. U-COPE: Taking a Further Step to Universal 9D Category-Level Object Pose Estimation. In *European Conference on Computer Vision*, 254–270. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, L.; Zhong, Y.; Wang, J.; Min, Z.; Liu, L.; et al. 2024a. Rethinking 3D Convolution in ℓ_p-norm Space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, X.; Shen, C.; Yuan, X.; Yan, S.; Xie, L.; Wang, W.; Gu, C.; Tang, H.; and Ye, J. 2024b. From Redundancy to Relevance: Enhancing Explainability in Multimodal Large Language Models. *arXiv preprint arXiv:2406.06579*.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.