Cross-view Masked Diffusion Transformers for Person Image Synthesis

Trung X. Pham * **Zhang Kang** * **Chang D. Yoo**
Korea Advanced Institute of Science and Technology (KAIST)

Abstract

We present X-MDPT (Cross-view Masked Diffusion Prediction Transformers), a novel diffusion model designed for pose-guided human image generation. X-MDPT distinguishes itself by employing masked diffusion transformers that operate on latent patches, a departure from the commonly-used Unet structures in existing works. The model comprises three key modules: 1) a denoising diffusion Transformer, 2) an aggregation network that consolidates conditions into a single vector for the diffusion process, and 3) a mask cross-prediction module that enhances representation learning with semantic information from the reference image. X-MDPT demonstrates scalability, improving FID, SSIM, and LPIPS with larger models. Despite its simple design, our model outperforms state-of-the-art approaches on the DeepFashion dataset while exhibiting efficiency in terms of training parameters, training time, and inference speed. Our compact 33MB model achieves an FID of 7.42, surpassing a prior Unet latent diffusion approach (FID 8.07) using only $11\times$ fewer parameters. Our best model surpasses the pixel-based diffusion with $\frac{2}{3}$ of the parameters and achieves $5.43\times$ faster inference. The code is available at https://github.com/trungpx/xmdpt.

1. Introduction

The task of Pose-guided Human Image Generation (PHIG) (Ma et al., 2017) has gained considerable attention with the advent of diffusion models recently. Initially, GAN-based approaches (Ma et al., 2017; Men et al., 2020; Wu et al., 2023), showed potential in PHIG but often struggled to generate target images accurately, resulting in high

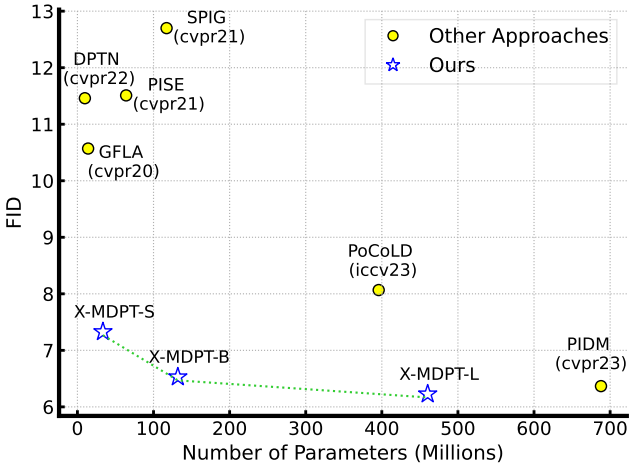

Figure 1: FID score of SOTAs approaches on the DeepFashion dataset. Our transformer-based models, X-MDPT (size of S, B, L) are marked in stars. X-MDPT-S surpasses the latent Unet-based PoCoLD with only $11\times$ fewer parameters.

Frechet Inception Distance (FID) scores due to the presence of undesired artifacts. To address these challenges, Bhunia et al. (2023) introduced PIDM, a diffusion-based methodology employing iterative generation processes. While PIDM achieved state-of-the-art results in synthesizing high-quality images from target poses and source images, it suffered from slow inference speeds and high memory consumption owing to its pixel-based operation. In response to these efficiency concerns, Han et al. (2023) proposed PoCoLD, a latent diffusion framework operating on autoencoder latent outputs. However, while PoCoLD offered improvements in efficiency over PIDM, it fell short in terms of FID and Structural Similarity Index Metric (SSIM) metrics compared to PIDM. Notably, prevailing methodologies predominantly relied on Unet-based architectures with Convolutional Neural Networks (CNNs) for denoising diffusion processes.

In contrast, our approach introduces a novel class of diffusion models based on transformers, aimed at addressing the challenges in PHIG more effectively. Diffusion models have demonstrated success in learning data distributions across various domains, particularly in generation tasks (Ho et al., 2020; Rombach et al., 2022). One significant advantage of diffusion models lies in their stable training process compared to Generative Adversarial Networks (GANs), which

*Equal contribution . Correspondence to: Trung X. Pham <trungpx@kaist.ac.kr>, Zhang Kang <zhangkang@kaist.ac.kr>, Chang D. Yoo <cd_yoo@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

are prone to mode collapse (Isola et al., 2017). While Unet-based diffusion models, including Stable Diffusion (Rombach et al., 2022) and related works (Zhang et al., 2023a; Zhao et al., 2023), have achieved notable success in various applications, DiT (Peebles & Xie, 2023) showcased the effectiveness of transformer-based designs in learning diffusion processes, effectively challenging traditional Unet-based networks in image generation. The emergence of masked diffusion transformers (Gao et al., 2023) as a state-of-the-art approach for class-conditional image generation on ImageNet further inspired our work. Motivated by these advancements, we propose a mask-based framework conditioned on both pose and style images to generate target images in the PHIG task (Ma et al., 2017; Wu et al., 2023). This novel approach aims to leverage semantic information from reference images, enhancing the model's ability to generate realistic and contextually coherent human images.

In this paper, we introduce X-MDPT, leveraging diffusion Transformer models for pose-guided human image generation. We design a specialized aggregation network to consolidate all conditions into a single vector to guide the diffusion Transformer via Adaptive Layernorm modulation. Additionally, we introduce a novel masking network to enhance the learning capabilities of transformers. By scaling up the number of layers or heads, we obtain configurations X-MDPT-S, X-MDPT-B, and X-MDPT-L, following standard transformer configurations (Dosovitskiy et al., 2010; Vaswani et al., 2017). Despite its simple design, X-MDPT achieves state-of-the-art FID, SSIM, and LPIPS scores on the DeepFashion dataset while employing significantly fewer parameters compared to existing SOTAs (Fig. 1). X-MDPT combines simplicity and effectiveness, maintaining the scalability and flexibility of the Transformer architecture, demonstrating its potential for human image generation. Moreover, Fig. 2 demonstrates that our model produces more consistent and plausible outputs than the SOTA PIDM. Our contributions are as follows:

- We propose X-MDPT, the first masked diffusion transformer framework for the PHIG task. Our model is scalable and efficient as it works on latent patches.
- We propose CANet to aggregate all conditions into a unified vector and we show that a single vector provides sufficient information to guide the diffusion process. This is conceptually simple as it neither requires any modification to the main diffusion framework nor extracts multiple-level features from different layers of conditions as did in existing approaches (CASD, PIDM, and PoCoLD).
- We propose a novel cross-view strategy to predict masked tokens across images, enhancing representation learning and improving generation quality.
- X-MDPT outperforms other state-of-the-art approaches on the DeepFashion dataset while being more efficient. For example, our compact models outperform the latent diffusion-based PoCoLD and pixel-based PIDM using only $11\times$ and $\frac{2}{3}\times$ fewer parameters, respectively. We qualitatively show that our model is robust to various difficult cases where the other approaches failed to generate the desired images.

Figure 2: **Source-View Invariant**. The 2nd and 7th columns display different views of the same individuals from the DeepFashion. PIDM yields inconsistent outputs if varying source image views, whereas ours produces consistent ones closer to the ground truth. Best view at 200% zoom.

2. Related Works

Pose-guiged Person Image Synthesis. Bhunia et al. (2023) introduced PIDM as a solution to the PHIG problem using a diffusion model in pixel space, showing significant performance compared to traditional GAN-based methods. However, PIDM requires substantial computational resources and exhibits inefficiencies in both training and inference. (Han et al., 2023) introduced PoCoLD, a latent diffusion model based on Unet design, to address these limitations. Concurrently, Karras et al. (2023) presented DreamPose, applying the diffusion model to the fashion video domain by fine-tuning a text-to-image model, specifically Stable Diffusion (Rombach et al., 2022). While these approaches enhance the diffusion model's capabilities for PHIG with CNN-based denoisers, the potential of transformer-based methods remains untapped. To bridge this gap, we propose X-MDPT, leveraging pure transformer diffusion models to generate target person images.

Diffusion Transformers. The CNN U-Net structure (Ronneberger et al., 2015) initially served as the foundation for diffusion models and remains a standard choice across various diffusion-based generation tasks (Ho et al., 2020; 2022; Song & Ermon, 2019). The introduction of DiT (Peebles & Xie, 2023) marked a significant advancement, integrating the architecture of the pure transformer ViT (Dosovitskiy et al., 2021) into latent diffusion models. DiT demonstrated exceptional scalability and outperformed Unet-like architectures. Gao et al. (2023) further advanced the diffusion transformer model, achieving state-of-the-art class image generation on ImageNet by leveraging contextual representation learning. While they explored the potential of

transformers in general generation tasks, our focus lies on the application of the diffusion transformer specifically to pose-guided human image generation (PHIG) within the Fashion domain for the first time. PHIG represents a pivotal and intricate generation task, requiring comprehensive information extraction from the source image including clothing, identity, background, and more, to faithfully generate the desired target pose (Ma et al., 2017; Bhunia et al., 2023).

Mask Prediction Modeling. Mask-based vision models, inspired by mask language models like BERT (Devlin et al., 2018), have demonstrated remarkable scalability and performance. Notable examples include MAE (He et al., 2022) in self-supervised learning (SSL), MaskGIT (Chang et al., 2022), Muse (Chang et al., 2023), and MAGVIT (Yu et al., 2023) for learning discrete token distributions for image generation. In contrast, MDT (Gao et al., 2023) introduced an asymmetric masking schedule to enhance contextual representation learning in diffusion transformers. However, MDT struggles to establish correspondence between source and target images, limiting its representation learning capabilities. To address this limitation, we propose the Mask Inter-Prediction Network (MIPNet) inspired by prior works such as SiamMAE (Gupta et al., 2023) and PatchMAE (Zhang et al., 2023b) in SSL. MIPNet focuses on inter-semantic masking prediction within diffusion models for PHIG tasks. While it shares similarities with these MAEs in utilizing different views to predict masks, MIPNet differs in its lightweight, single self-cross-attention block and asymmetric masking, as opposed to these MAEs. Additionally, MIPNet is designed for generation tasks, while SimMAE/PatchMAE is for downstream recognition tasks.

3. Method

We aim to design a simple yet scalable framework for addressing the PHIG task using Transformers. The overall architecture is depicted in Fig. 3. Our X-MDPT comprises three core modules: 1) Transformer-based Denoising Diffusion Network (TDNet), 2) Conditional Aggregation Network (CANet), and 3) Mask Inter-Prediction Network (MIPNet). Here, TDNet performs the denoising diffusion, while CANet consolidates all necessary condition inputs into a single vector for TDNet's input. Additionally, MIPNet enhances the diffusion learning process by predicting masked tokens using a novel reference-based predictor. We provide detailed insights into each component next.

3.1. Transformer-based Denoising Diffusion Network

In our X-MDPT framework, we denote this network component as TDNet for brevity. TDNet is built on top of DiT (Peebles & Xie, 2023) to establish the diffusion process using Transformer architecture. Here's an overview of the framework for a 256×256 resolution case: Given the source image $X_s \in \mathbb{R}^{256\times256\times3}$ and the target pose y_p, the objective is to learn the model parameterized by θ to capture the target pose and the style of the source image to generate the final target image $Y \in \mathbb{R}^{256\times256\times3}$. Initially, we employ a pre-trained VAE (Rombach et al., 2022) to map the pixel images to latent representations $x_s \in \mathbb{R}^{32\times32\times4}$ and $y \in \mathbb{R}^{32\times32\times4}$ for denoising. The denoising network ϵ_θ is a transformer-based diffusion model that learns the condition distribution $p_\theta(y|x_s, y_p)$. The denoising process progressively adds Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to image y to obtain y_t at timestep $t \in [1, T]$. The conditions x_s and pose y_p are represented by c. The training objective is to predict the added noise using mean squared error:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{y,c,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\|\epsilon - \epsilon_\theta(y_t, c, t)\|^2. \quad (1)$$

Once p_θ is trained, inference proceeds by initiating with a random noise image $y_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively sampling $y_{t-1} \sim p_\theta(y_{t-1}|y_t)$ to obtain the final target image y_0. Our diffusion transformer network adheres to the structure outlined in DiT (Peebles & Xie, 2023). We transform the noisy latent $y_t \in \mathbb{R}^{32\times32\times4}$ into patches with a patch size of $p = 2$, forming the sequence $z_{y_t} = [z_y^{(1)}, z_y^{(2)}, ..., z_y^{(L_y)}] \in \mathbb{R}^{L_y\times D}$, where L_y and D denote the sequence length and dimension, respectively. The condition c is integrated into the TDNet through adaptive layer normalization (AdaLN-Zero), following the default setup of DiT.

3.2. Conditional Aggregation Network

The CANet network integrates three inputs: 1) the target pose condition feature (TPF), 2) the local source image feature (LSIF) obtained from the output of the VAE, and 3) the global source image feature (GSIF) derived from the pre-trained feature of DINOv2. CANet processes these inputs to produce a unified vector c with dimensions matching the width of the Transformer.

Local Source Image Feature. This feature ensures alignment with the noisy target image within the TDNet, enabling the transfer of information from the source image (including clothing, person, and background) to generate the target image. Specifically, the latent image $32 \times 32 \times 4$ is converted into a sequence $z_{x_s} = [z_x^{(1)}, z_x^{(2)}, ..., z_x^{(L_x)}] \in \mathbb{R}^{L_x\times D}$, where $L_x = 256$ tokens (LSIF $= z_{x_s} \in \mathbb{R}^{256\times D}$). An 1×1 conv. layer is then applied to linearly map 256 channels to a single channel, yielding the local vector $v_L \in \mathbb{R}^D$.

Pose Representation. We utilize a 3-channel RGB visualization image of the pose ($256 \times 256 \times 3$), resized to $224 \times 224 \times 3$, as the input to CANet. In contrast, PIDM (Bhunia et al., 2023) employs a more complex pose representation with $256 \times 256 \times 20$, where the 20 channels are 3 RGB channels and 17 Gaussian heatmaps. Our experiments demonstrate that the use of a simple RGB pose representation is adequate for generating satisfactory person images. We employ pre-trained DINOv2-B to extract RGB

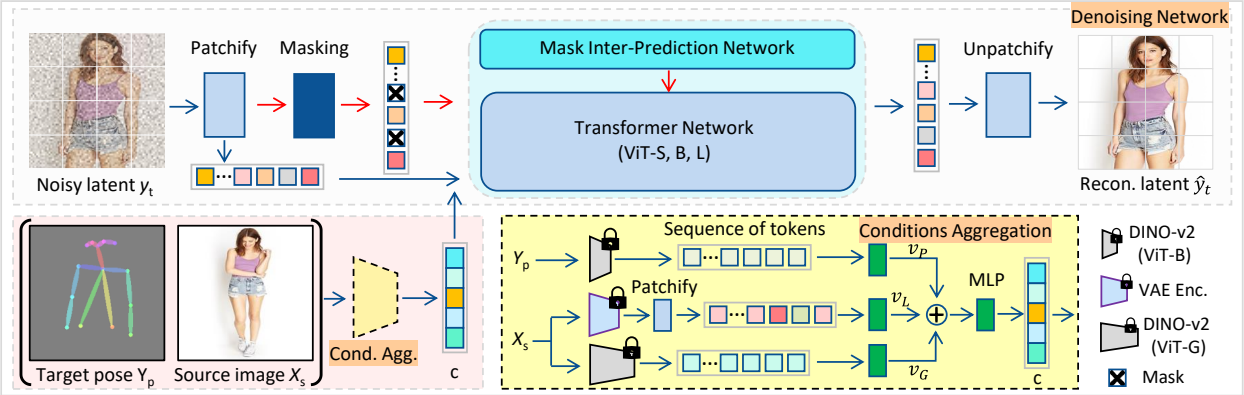

Figure 3: Overview of Our X-MDPT framework, built on transformers, facilitates pose-guided human image generation. During training, we randomly mask target image tokens at a 30% ratio. The noisy target image is then processed through the Transformer Diffusion network, conditioned on the aggregated vector (with $D = 768$ for our X-MDPT-B model) via AdaLN modulation (Peebles & Xie, 2023). Concurrently, we train a mask prediction objective alongside our novel mask inter-prediction network to capture semantics between source and target images when predicting mask tokens. The red arrow $\rightarrow$ signifies the training-only branch, discarded during inference, while the bue arrow$\rightarrow$ serves both training and inference purposes. "Cond. Agg." denotes "Conditions Aggregation" on the bottom right. VAE is omitted for simplicity.

pose features, resulting in a CLS and 256 patch tokens, concatenated into a sequence of 257 (TPF $\in \mathbb{R}^{257\times D}$), which is then processed by one 1×1 convolutional layer to map 257 channels to 1 channel, yielding vector $v_P \in \mathbb{R}^D$.

Global Source Image Feature. We observed that relying solely on local features from the source image is insufficient for the Transformer to capture both the details and the identity of the person. As demonstrated in AnyDoor (Chen et al., 2023), self-supervised models like DINO features can effectively capture identity and details. Hence, we utilize DINOv2-G to extract the CLS token and concatenate all other tokens to form the global feature (GSIF $\in \mathbb{R}^{257\times D}$). For a resolution of 256×256, we resize it to 224×224 as required by DINO (where both width and height are divisible by 14) to obtain a CLS token and 256 patch tokens. Interestingly, for a resolution of 512×512, we find that using the same resolution of 224×224 for both pose and global features when extracting DINO tokens is effective, eliminating the need for extracting pose and global features of size 448×448 (close to 512×512). This helps save memory, as the output feature sequence of the DINO transformer is larger in this case ($32 \times 32 = 1024$) compared to $16 \times 16 = 256$ tokens (DINO output for 224×224 images). This approach differs from Unet-based frameworks like PIDM and PoCoLD, which necessitate exact 512×512 size for the pose condition. Finally, we pass the resulting sequence GSIF $\in \mathbb{R}^{257\times D}$ through one 1×1 conv. layer to obtain the global vector $v_G \in \mathbb{R}^D$.

Aggregation. We obtain the final vector by either simply using an addition operation or concatenating them to have three channels and use a 1×1 convolution operation $\mathcal{H}$ to get a unified conditional vector c, *i.e.* $c = \mathcal{H}(v_L, v_P, v_G) \in \mathbb{R}^D$. We find that an MLP gives a slightly better FID as this MLP can automatically put weights on each condition and is learned during backpropagation. We show that having both local and global vectors, *i.e.* $v_L + v_G$ of the source image together with v_P is crucial to generating the high-quality target image in the ablation section.

3.3. Mask Inter-Prediction Network

Gao et al. (2023) demonstrated improvements in transformer-based diffusion models by introducing a lightweight predictor to fill masked regions within images, resulting in enhanced FID scores. While this masking strategy proved effective in general image-generation tasks like ImageNet, its application to the DeepFashion dataset remained unexplored. In the PHIG task, merely predicting masks using unmasked tokens within an image falls short of capturing the necessary correspondence between source and target images, critical for conditional generation. As a result, this approach yields suboptimal performance in person image generation. To address this limitation, we propose a novel prediction module that integrates information from the source image (cross-view) to guide the diffusion model in predicting masked tokens within the target image. This is in contrast to MDT's reliance solely on the target image. By incorporating information from the reference image, our approach, MIPNet, gains richer contextual cues for completing mask patches in the target image and learning meaningful semantic correspondence. The primary distinction between MDT and MIPNet is illustrated in Fig. 4. In the ablation section, we validate the effectiveness of our method compared to MDT's mask prediction. The objective loss for training with masked tokens remains consistent with the standard loss $\mathcal{L}_{\text{denoise}}$.

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{y,c,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\|\epsilon - \epsilon_\theta(f_\theta(x_s, y_m), c, t)\|^2, \quad (2)$$

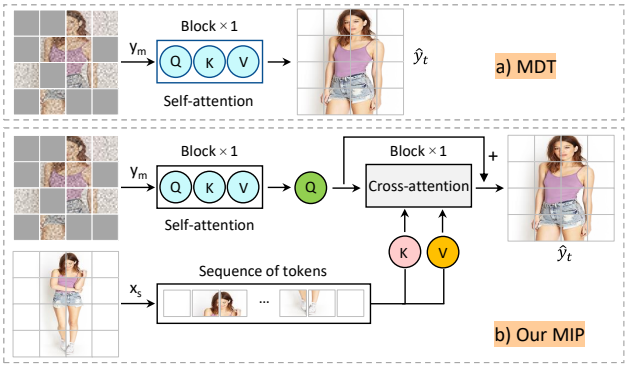

Figure 4: **MIPNet vs. MDT**. Ours MIPNet predicts masked tokens by using all tokens from the reference image x_s.

where f_θ is the network that contains MIPNet g_θ, the N_1 encoder layers and N_2 decoder layers of the TDNet. Here, N_1 and N_2, along with other settings, remain consistent with those defined in MDT (Gao et al., 2023). The output of MIPNet is computed by the following equation:

$$\begin{aligned} g_\theta(z_{x_s}, z_{y_m}) = \phi_{\text{attn}}(z_{y_m}, z_{y_m}, z_{y_m}) \\ + \phi_{\text{attn}}\left(\phi_{\text{attn}}(z_{y_m}, z_{y_m}, z_{y_m}), z_{x_s}, z_{x_s}\right), \end{aligned} \tag{3}$$

where ϕ_{attn} denotes the attention mechanism proposed by (Vaswani et al., 2017) to learn cross-view mask prediction:

$$\phi_{\text{attn}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

Final objective function. We jointly optimize two objective functions in parallel through the following equation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{mask}}. \tag{5}$$

In Eq. 5, if $\mathcal{L}_{\text{mask}}$ is removed, the framework resembles the style of DiT. Alternatively, replacing g_θ with only the self-attention transforms the model into the MDT style. During inference, MIPNet is omitted, retaining only positional embeddings, as implemented in MDT (Gao et al., 2023).

3.4. Classifier-Free Guidance

We use the common technique of classifier-free guidance (Ho & Salimans, 2022) to predict noise via the linear combination of the unconditional model $\epsilon_\theta(y_t, t)$ and conditional model $\epsilon_\theta(y_t, c, t)$ as follows:

$$\hat{\epsilon}_\theta(y_t, x, t) = \gamma_t \epsilon_\theta(y_t, c, t) + (1 - \gamma_t)\epsilon_\theta(y_t, t). \tag{6}$$

The guidance scale γ_t is determined at timestep t. During training, we randomly assign the unified conditional vector $c \in \mathbb{R}^D$, obtained by CANet, to the zero vector $\emptyset \in \mathbb{R}^D$ with a probability of $\eta = 10\%$. Another parameter that facilitates dynamic scale guidance is $\gamma_t = \frac{1-\cos\pi(\frac{t}{T})^\alpha}{2} \times \gamma$. This sets the power-cosine schedule (α) for the guidance scale during the sampling procedure, as used in MDT (Gao et al., 2023). By default, we set $\gamma = 2.0$. Our experiments indicate that using $\alpha = 1.0$ yields the best FID, albeit slightly lower values for other metrics such as SSIM and LPIPS. Conversely, $\alpha = 0.01$ results in slightly higher FID but improved SSIM and LPIPS.

4. Experiments

4.1. Implementation Details

Dataset. We evaluate our method against the state-of-the-art (SOTA) using high-resolution images from the DeepFashion In-shop Clothes Retrieval Benchmark dataset (Liu et al., 2016) at resolutions of 256×256 and 512×512. The dataset comprises non-overlapping train and test subsets, containing 101,966 and 8,570 pairs, respectively. We adopt preprocessing steps consistent with prior works (Bhunia et al., 2023; Han et al., 2023). We employ OpenPose (Cao et al., 2017) to extract 18 key points from each person's image and then utilize OpenCV to generate RGB visualizations, which are resized to 224×224. This fixed size is applied for pose images in both the 256×256 and 512×512 resolution cases.
Metrics. We use the common measurements as utilized in prior studies (Bhunia et al., 2023; Han et al., 2023) including FID, SSIM, LPIPS, and optionally PSNR for ablations.

Training. We finetune the pre-trained VAE ft-MSE of Stable Diffusion on the DeepFashion training set. For 256×256 images, training was conducted on a single A100 GPU (80GB RAM) with a batch size of 32, spanning 800k steps. Meanwhile, for 512×512 images, we employed two A100 GPUs with a batch size of 10 (5 images per GPU), trained for 1M steps. For ablations, we trained X-MDPT-B with 300k steps on a 256×256 resolution. The learning rate was set to 1e-4, the model's EMA rate to 0.9999, and other settings aligned with DiT (Peebles & Xie, 2023). Note that, the original images in the DeepFashion dataset have resolutions of 256×176 and 512×352, which we resized to 256×256 and 512×512, respectively, using bicubic interpolation before inputting them into the models.

4.2. Main Results

Results are reported in Tab. 1, with the comparative performance of different approaches. Our X-MDPT demonstrates consistent superiority across FID, SSIM, and LPIPS metrics at a resolution of 256×256. We find that our model achieves its best FID score, around 6.25, during the mid-training phase (350-400k steps). Subsequently, with extended training (800k-1M steps), the FID stabilizes at around 7.28, while SSIM and LPIPS metrics exhibit continual enhancement. Here, the evaluation process performs the comparison between the synthetic test set and the real training data. Ground truth images yield an FID of 7.86, indicating a substantial distribution gap in DeepFashion's

test and training sets, as confirmed by PoCoLD (Han et al., 2023). As the model fully converges, we observe a narrowing of the FID, LPIPS, and SSIM as the model learned generates images closer to the ground truth. At a higher resolution of 512×512, X-MDPT exhibits a slight lag behind PIDM in FID, but better in other key metrics. Its resource efficiency-batch sizes of 32 and 10 with one and two GPUs for 256 and 512 resolutions, suggests room for improvement compared to PIDM's 8 GPUs with a batch size of 128.

Table 1: Comparison of X-MDPT with SOTA approaches. † we reproduced with the public checkpoint. **Bold** and underline denotes the best and second-best, respectively.

Dataset	Method	FID ↓	SSIM ↑	LPIPS ↓	Type
DeepFashion (256×176)	Def-GAN (Siarohin et al., 2018)	18.457	0.6786	0.2330	Non-Diffusion
	PATN (Zhu et al., 2019)	20.751	0.6709	0.2562	
	ADGAN (Men et al., 2020)	14.458	0.6721	0.2283	
	PISE (Zhang et al., 2021)	13.610	0.6629	0.2059	
	GFLA (Ren et al., 2020)	10.573	0.7074	0.22341	
	DPTN (Zhang et al., 2022c)	11.387	0.7112	0.1931	
	CASD (Zhou et al., 2022b)	11.373	0.7248	0.1936	
	NTED (Ren et al., 2022)	8.6838	0.7182	0.1752	
	PoCoLD (Han et al., 2023)	8.0667	0.7310	0.1642	Unet
	PIDM (Bhunia et al., 2023)	6.3671	0.7312	0.1678	
	PIDM (Bhunia et al., 2023)†	6.6182	0.7294	0.1715	
	X-MDPT-S (300k), $\alpha = 0.01$	7.4282	0.7128	0.1961	Transformer
	X-MDPT-S (800k), $\alpha = 0.01$	7.6724	0.7194	0.1875	
	X-MDPT-B (300k), $\alpha = 0.01$	6.7288	0.7215	0.1814	
	X-MDPT-B (800k), $\alpha = 0.01$	7.3293	0.7284	0.1734	
	X-MDPT-L (350k), $\alpha = 1.00$	**6.2512**	0.7298	0.1671	
	X-MDPT-L (800k), $\alpha = 0.01$	7.2865	**0.7405**	**0.1589**	
	VAE reconstruction	8.0126	0.9168	0.0142	-
	Ground Truth	7.8610	1.0000	0.0000	
DeepFashion (512×352)	CocosNet-v2 (Zhou et al., 2021)	13.325	0.7236	0.2265	Unet GAN
	NTED (Ren et al., 2022)	7.7821	0.7376	0.1980	
	PoCoLD (Han et al., 2023)	8.4163	0.7430	0.1920	
	PIDM (Bhunia et al., 2023)	**5.8365**	0.7419	0.1768	
	X-MDPT-L (500k), $\alpha = 1.0$	5.9264	0.7416	0.1788	Trans
	X-MDPT-L (1M), $\alpha = 0.01$	7.1615	**0.7522**	**0.1645**	
	VAE reconstruction	8.1815	0.9122	0.0266	-
	Ground Truth	7.9150	1.0000	0.0000	

Qualitative Results. We compared the output generated by X-MDPT and other methods in Fig. 7. The generated person images exhibit high quality across various scenarios. Notably, pixel-based diffusion models like PIDM and other CNN-based methods often struggle to faithfully capture intricate style details, resulting in noticeable artifacts, as observed in specific cases. In contrast, X-MDPT operates on latent space with a transformer equipped with a semantic understanding scheme, enabling a more accurate depiction of clothing elements such as shirts and trousers. This results in more complete and satisfactory images that align well with the intended pose and source image.

Visual comparisons between ours and existing approaches in Fig. 7 highlight X-MDPT's consistent generation of plausible and complete target images that closely resemble ground truth. Furthermore, compared to the previous best-performing model, PIDM, our model demonstrates superior alignment when the image viewpoint is changed, as shown in Fig. 2. We analyze this property in Fig. 5. More qualitative results can be found in the **Appendix**.

4.3. Ablation Studies

Learning Invariant Views of Source Images. For the task of PHIG, with different views of a person's image and the same target pose, we expect to generate the same target image. However, as shown in Fig. 2, PIDM does not capture the specific structure of the clothes and produces inconsistent target images when the person's view is changed. By contrast, X-MDPT gives consistent target images and it is closer to the ground truth. This consistent generation can be elaborated by measuring the cosine similarity of unified conditional vectors when varying views of the source image. We find that existing works such as PoCoLD and PIDM utilize conditions in different places and multiple levels in their networks, which makes it challenging to produce the unified vectors of all conditions (*i.e.* pose and source image).

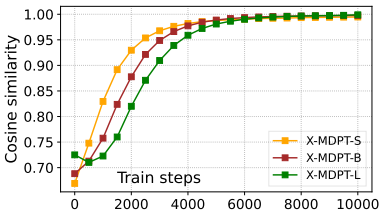

Figure 5: **CANet.** Views of same person have 99.99% similarity.

By contrast, our CANet module supports generating such a unified conditional vector. These vectors can easily support monitoring the cosine similarity score of different views of the same person. We conducted the whole test set of DeepFashion and took the average. Fig. 5 shows that the similarity reached above 99.99% after 10k training steps, indicating that CANet learns to capture the invariant features of the same person. This explains why X-MDPT can generate a consistent target given the same pose with different views of a person, which is helpful for the task.

Scalability. We discover the scalability of X-MDPT with sizes S, B, and L in Fig. 6. We observe FID, SSIM, LPIPS, and PSNR improved as scaling up model size, demonstrating the potential of transformers for the PHIG problem.

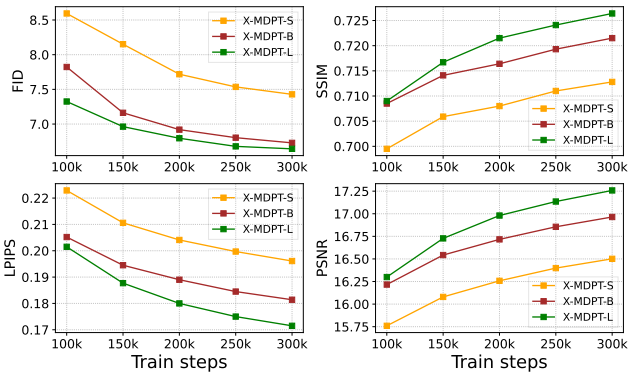

Figure 6: **Scalability.** X-MDPT (model sizes S, B, and L) is scalable for four metrics FID, SSIM, LPIPS, and PSNR.

Figure 7: **Qualitative comparison.** Person images are generated by state-of-the-art approaches on the DeepFashion dataset. Given two inputs: the target pose and source image, our transformer-based X-MDPT-L (460M) handles various difficult cases and creates high-quality, more realistic, and closer to the ground truth (GT) compared to other CNN-based methods.

Impact of Mask Prediction. Tab. 2a shows that naive applying transformer (DiT) (Peebles & Xie, 2023) to PHIG task does not give a satisfactory FID. MDT (Gao et al., 2023) can help improve FID but is not beneficial for SSIM and LPIPS. By contrast, our method improves all metrics.

Qualitatively, images generated by our method are more realistic and closer to ground truth images as shown in Fig. 8. This can be attributed to the fact that correspondence is the key to achieving the best performance in the posed-guide person image generation task, for which our X-MDPT equipped with the proposed Inter-Semantic module (MIP-Net) can capture strong clues between the source and target images.

Inference Time. As presented in Tab. 2f, all variants of our models demonstrate significantly faster inference speeds and require fewer parameters compared to the runner-up,

Table 2: **Ablation experiments** on the DeepFashion dataset at 256 × 176 resolution. Default settings are marked in gray.

(a) **Transformer baselines**. Compared different diffusion transformers. MIPNet improves all metrics.

Method	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Transformer	7.1150	0.7199	0.1841	16.8882
+ MDT	6.8616	0.7182	0.1851	16.8227
+ MIPNet	**6.7288**	**0.7215**	**0.1814**	**16.9642**

(b) **Masking ratio effect**. Model X-MDPT-B trained on the DeepFashion dataset for 300k iterations.

Mask Ratio	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
30%	**6.7288**	0.7215	0.1814	16.9642
50%	6.8081	0.7231	0.1792	17.0609
70%	6.9533	**0.7246**	**0.1759**	**17.1442**

(c) **Global & pose feature**. (G), (B), and (S) mean DINOv2-G, DINOv2-B, and DINOv2-S, respectively.

Features	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Global (S) + Pose (S)	6.9239	0.7187	0.1894	16.7466
Global (G) + Pose (S)	6.9237	**0.7228**	**0.1798**	**17.011**
Global (G) + Pose (B)	**6.7288**	0.7215	0.1814	16.9642

(d) **Attention for MIPNet**. Compared performance of different designs, **self-cross** attention gives the best FID.

Method	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
self-att	6.8616	0.7182	0.1851	16.8227
cross-att	6.9067	0.7222	0.1791	17.010
cross-self att	6.8938	**0.7244**	**0.1766**	**17.1215**
self-cross att	**6.7288**	0.7215	0.1814	16.9642

(e) **Conditions Aggregation**. Combining pose feature v_P with both local v_L and global feature v_G gives the best FID.

CANet	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
$v_L + v_P$	11.2661	0.6679	0.3131	12.8694
$v_P + v_G$	6.9241	**0.7230**	**0.1805**	**17.0456**
$v_L + v_P + v_G$	**6.7288**	0.7215	0.1814	16.9642

(f) **Inference speed**. It is averaged over 10 times for each 8 image generation with 50 DDIM steps (256 × 176), using one A100 GPU.

Method	Param (M) ↓	Time (s) ↓	Speed up ↑	FID ↓
PIDM	688.00	16.975±0.055	1.0×	6.36
X-MDPT-S	**33.52**	**1.191±0.021**	**14.25×**	7.42
X-MDPT-B	**131.92**	**1.299±0.022**	**13.07×**	6.72
X-MDPT-L	**460.24**	**3.124±0.026**	**5.43×**	**6.25**

PIDM. Specifically, models X-MDPT-S, X-MDPT-B, and X-MDPT-L speed up PIDM by $14.25\times$, $13.07\times$, and $5.43\times$, respectively. This speed advantage can be attributed to several factors. Firstly, X-MDPT operates on latent patches of 32×32, while PIDM works directly on pixel space of 256×256. Here, the VAE in X-MDPT works a single forward, whereas PIDM necessitates 50 forward passes in DDIM. Secondly, PIDM employs disentangled classifier-free guidance for both pose and source, requiring two forwards and resulting in $50\times$ more evaluation steps. Conversely, our X-MDPT utilizes CFG for a unified condition and needs only one forward. The training time can be found in the **Appendix** due to space constraints, where our method proves significantly more efficient than PIDM.

Figure 8: **Generated images with different training steps**. Our X-MDPT performs much better than MDT. It is best viewed with zoom in 200%.

MIPNet Components. (a) **Mask Ratio**: In Tab. 2b, we observe that a lower mask ratio of the target image in MIPNet yields the best FID score, consistent with findings from Gao et al. (2023) on MDT applied to ImageNet. Conversely, a higher ratio, such as 70%, results in better SSIM and LPIPS scores. This higher ratio compels the model to prioritize reconstruction over semantic representation learning, which is crucial for effective generation.

(b) **Attention for MIPNet**: Tab. 2d illustrates that the self-cross attention design produces the best FID score while utilizing only self-attention (as MDT) yields the worst FID. This discrepancy highlights the significance of reference-based mask prediction in enhancing the performance of person image generation.

CANet Components. (a) **Conditions Aggregation**: Tab. 2e illustrates that solely utilizing the local feature v_L, *i.e.* the VAE's output of the source image, yields the poorest performance (the $v_L + v_P$ case). Conversely, relying solely on the global feature v_G (the $v_G + v_P$ case) leads to significant improvement. However, this approach lacks local information crucial for the generation, as the noisy target latent of TDNet operates at the VAE level. The optimal choice arises from combining both local and global features, as demonstrated in the $v_L + v_P + v_G$ case.

(b) **Global and Pose Representations**: AnyDoor (Chen et al., 2023) demonstrates that DINOv2-G is good at capturing object details. In Tab. 2c we observe that the quality of global features significantly impacts performance, with higher-capacity DINOv2 models yielding better FID scores.

For the pose, we find that DINOv2-B outperforms DINOv2-S in FID but slightly lags in SSIM and LPIPS. We opt for DINOv2-G for the global feature of x_s and DINOv2-B for the pose y_p as default. The simplicity of the pose image allows a smaller model like DINOv2-B to effectively guide TDNet. Conversely, the complexity of the human image necessitates a more powerful variant.

5. More Discussions

During the discussion phase, we delved into additional properties of the proposed method, uncovering several findings.

Compared with Prior Efficient Method: PoCoLD introduced a latent diffusion more efficient than PIDM but falls behind FID. Since PoCoLD's code is not complete, we conducted our own implementation. Tab. 3 shows that our variants run faster PoCoLD on A100. Notably, X-MDPT-S, with $11\times$ fewer parameters, surpasses PoCoLD on FID, generating 8 images in just 1 second, $3\times$ faster than PoCoLD.

Our method's enhanced speed over PoCoLD arises from

Table 3: **Compare Efficiency.** Results are presented for every 8-image generation (batch size=8) for 256×176 image, 50 denoising steps, using one NVIDIA A100 GPU. We conduct 10 runs and take the average.

Method	Infer. Time (s)	Mem. Use (M)	Param. (M)	Guidance	#Forward
PIDM	16.975 ± 0.055	9572	688.00	D-CFG	150
PoCOLD	3.215 ± 0.028	5686	395.89	D-CFG	150
X-MDPT-L (Ours)	$\mathbf{3.124 \pm 0.026}$	6438	460.24	CFG	**100**
X-MDPT-B (Ours)	$\mathbf{1.299 \pm 0.022}$	**5680**	**131.92**	CFG	**100**
X-MDPT-S (Ours)	$\mathbf{1.191 \pm 0.021}$	**5485**	**33.52**	CFG	**100**

PoCoLD's use of cumulative CFG, a variant of Disentangled Classifier-Free Guidance (D-CFG, similar to PIDM), which significantly slows down its operation. Specifically, each generation in PoCoLD requires a total of 150 forwards. In 50 denoising steps, one step necessitates three forwards (unconditional, pose-condition, source-image condition), resulting in a total of $50 \times 3 = 150$.

In contrast, X-MDPT employs standard CFG, requiring only 100 forwards, where each denoising step needs two forwards (unconditional, unified condition), totaling $50 \times 2 = 100$, saving 50 forwards compared to PoCoLD and PIDM.

Performance for Higher Resolution. In Tab. 1, for the DeepFashion dataset, X-MDPT shows a slightly lower FID score for higher-resolution images compared to lower-resolution ones, consistent with PIDM and NTED but not with PoCoLD. We suggest that our Transformer-based framework can better capture finer details and contextual information in higher-resolution images compared to PoCoLD, which relies on Unet. In contrast, on the ImageNet dataset, the behavior of Diffusion Transformer (DiT paper) is opposite to its performance on DeepFashion. We believe these disparities warrant a systematic exploration of dataset diversity and model architectures (CNN, Transformer, etc.) for a conclusive understanding.

Insights of Adding MIPNet. We integrate MIPNet into the Transformer backbone during training to facilitate the model in leveraging contextual information among patches within an image (via self-attention) and learning the correspondence between the source and target image (via cross-attention). This enhances the Transformer's learning capacity more rapidly. This insight is corroborated by Tab. 2(a), where MIPNet significantly improves performance compared to using the Transformer alone. The efficacy of masking loss, trained concurrently with the main loss, has been observed in previous works such as iBOT (Zhou et al., 2022a) and PatchMAE (Zhang et al., 2023b) in the self-supervised representation learning domain.

Generalizations. We have conducted an evaluation of in-the-wild images to prove the generalization of our method. We use our ready-to-use X-MDPT-L model trained on the DeepFashion dataset. We compare with the second-best method PIDM using their published checkpoint. We test with more challenging cases such as people with darker skin (as DeepFashion most data are collected with white people, a few for black) and complex background images. Fig. 9 demonstrates that X-MDPT is much more stable than the runner-up method PIDM, highlighting the distinguishing properties of diffusion transformers on this task. We believe one of the reasons PIDM suffers from overfitting more seriously is because it uses the features of pose and source image that are trained from scratch, while our method employs a pre-trained model DINO that gives much better features for both pose and source image. Training on more in-the-wild data can be a solution for mitigating the overfitting (toward background, identity) of ours and PIDM.

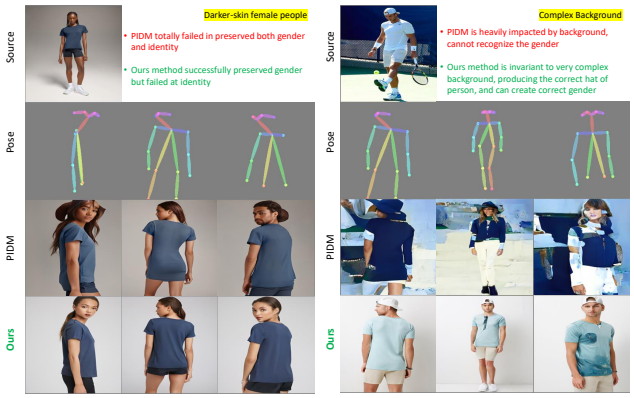

Figure 9: **In-the-wild testing.** Our X-MDPT-L consistently generates meaningful images with accurate gender representation, whereas PIDM (Bhunia et al., 2023) fails.

Overlapping Cases. Overlaps occur rarely, but in DeepFashion, we discovered several overlapping scenes, notably, a duplicated data pair with IDs `00005136` and `00005188`. Further details can be found in the Appendix. When testing on these duplicate samples, other approaches failed to reconstruct the target images. Why can X-MDPT fully reproduce previously seen samples while other models cannot? We attribute this capability to the Diffusion Transformer, which generates based on its training sample observations more effectively than previous CNN Unet-based algorithms.

6. Conclusion

In this paper, we present X-MDPT, a novel masked diffusion generative model for pose-guided human image generation (PHIG). Unlike previous methods using Unet for denoising diffusion, X-MDPT employs a Transformer on latent patches. Our analysis shows that X-MDPT achieves 99.99% similarity in generating view-invariant vectors, ensuring consistent target images across poses. Extensive experiments demonstrate X-MDPT's efficiency in producing high-quality, high-resolution images, surpassing existing approaches in inference speed and setting a new state-of-the-art for PHIG on the DeepFashion benchmark.

Impact Statement

Our method proficiently produces high-quality images featuring individuals in diverse poses, utilizing any person's image as a reference point. While it offers numerous advantages, including swift image generation, there's a risk of misuse, such as the creation of deceptive content for fraudulent purposes, a well-documented issue in image synthesis. We are committed to implementing measures to regulate access, thereby preventing misuse and ensuring that the technology contributes to the community's welfare safely.

Acknowledgements

This work was supported by the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01381, *Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments*) and partly supported by IITP grant funded by the Korea government (MSIT) (No. 2022-0-00184, *Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics*).

References

Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Laaksonen, J., Shah, M., and 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.

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

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., and Zhao, H. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Gao, S., Zhou, P., Cheng, M.-M., and Yan, S. Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023.

Gupta, A., Wu, J., Deng, J., and Fei-Fei, L. Siamese masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=yC3q7vInux.

Han, X., Zhu, X., Deng, J., Song, Y.-Z., and Xiang, T. Controllable person image synthesis with pose-constrained latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22768–22777, 2023.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Jung, Y. H., Hong, S. K., Wang, H. S., Han, J. H., Pham, T. X., Park, H., Kim, J., Kang, S., Yoo, C. D., and Lee, K. J. Flexible piezoelectric acoustic sensors and machine learning for speech processing. *Advanced Materials*, 32 (35):1904020, 2020.

Jung, Y. H., Pham, T. X., Issa, D., Wang, H. S., Lee, J. H., Chung, M., Lee, B.-Y., Kim, G., Yoo, C. D., and Lee, K. J. Deep learning-based noise robust flexible piezoelectric

acoustic sensors for speech processing. *Nano Energy*, 101:107610, 2022.

Karras, J., Holynski, A., Wang, T.-C., and Kemelmacher-Shlizerman, I. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023.

Kim, J., Ma, M., Pham, T., Kim, K., and Yoo, C. D. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10106–10115, 2020.

Lee, D., Park, H., Pham, T., and Yoo, C. D. Learning augmentation network via influence functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10961–10970, 2020.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

Men, Y., Mao, Y., Jiang, Y., Ma, W.-Y., and Lian, Z. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5084–5093, 2020.

Niu, A., Zhang, K., Pham, T. X., Sun, J., Zhu, Y., Kweon, I. S., and Zhang, Y. Cdpmsr: Conditional diffusion probabilistic models for single image super-resolution. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 615–619. IEEE, 2023.

Niu, A., Pham, T. X., Zhang, K., Sun, J., Zhu, Y., Yan, Q., Kweon, I. S., and Zhang, Y. Acdmsr: Accelerated conditional diffusion models for single image super-resolution. *IEEE Transactions on Broadcasting*, 2024a.

Niu, A., Zhang, K., Pham, T. X., Wang, P., Sun, J., Kweon, I. S., and Zhang, Y. Learning from multi-perception features for real-word image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024b.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Pham, T., Zhang, C., Niu, A., Zhang, K., and Yoo, C. D. On the pros and cons of momentum encoder in self-supervised visual representation learning. *arXiv preprint arXiv:2208.05744*, 2022a.

Pham, T. X., Mina, R. J. L., Issa, D., and Yoo, C. D. Self-supervised learning with local attention-aware feature. *arXiv preprint arXiv:2108.00475*, 2021.

Pham, T. X., Mina, R. J. L., Nguyen, T., Madjid, S. R., Choi, J., and Yoo, C. D. Lad: A hybrid deep learning system for benign paroxysmal positional vertigo disorders diagnostic. *IEEE Access*, 2022b.

Pham, T. X., Niu, A., Zhang, K., Jin, T. J. T., Hong, J. W., and Yoo, C. D. Self-supervised visual representation learning via residual momentum. *IEEE Access*, 2023.

Ren, Y., Yu, X., Chen, J., Li, T. H., and Li, G. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7690–7699, 2020.

Ren, Y., Fan, X., Li, G., Liu, S., and Li, T. H. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13535–13544, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

Siarohin, A., Sangineto, E., Lathuiliere, S., and Sebe, N. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3408–3416, 2018.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Trung, P. X. and Yoo, C. D. Short convolutional neural network and mfccs for accurate speaker recognition systems. *International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Vu, T., Jang, H., Pham, T. X., and Yoo, C. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. *Advances in neural information processing systems*, 32, 2019.

Wu, J., Si, S., Wang, J., Qu, X., and Jing, X. Pose guided human image synthesis with partially decoupled gan. In *Asian Conference on Machine Learning*, pp. 1133–1148. PMLR, 2023.

Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A. G., Yang, M.-H., Hao, Y., Essa, I., et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14441–14450, 2022a.

Zhang, C., Zhang, K., Zhang, C., Pham, T. X., Yoo, C. D., and Kweon, I. S. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=bwq6O4Cwdl.

Zhang, J., Li, K., Lai, Y.-K., and Yang, J. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7982–7990, 2021.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023a.

Zhang, P., Yang, L., Lai, J.-H., and Xie, X. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7713–7722, 2022c.

Zhang, S., Zhou, Q., Wang, Z., Wang, F., and Yan, J. Patch-level contrastive learning via positional query for visual pre-training. In *International Conference on Machine Learning*, pp. 41990–41999. PMLR, 2023b.

Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., and Wong, K.-Y. K. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=ydopy-e6Dg.

Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., and Wen, F. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11465–11475, 2021.

Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., and Li, Q. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pp. 161–178. Springer, 2022b.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., and Bai, X. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2347–2356, 2019.

A. Appendix

A.1. Compare with the previous best-reported images

To be more complete, in Fig. 10 we also compare the best-generated images reported in the prior papers for reference. Our model consistently produces comparable or better outputs.

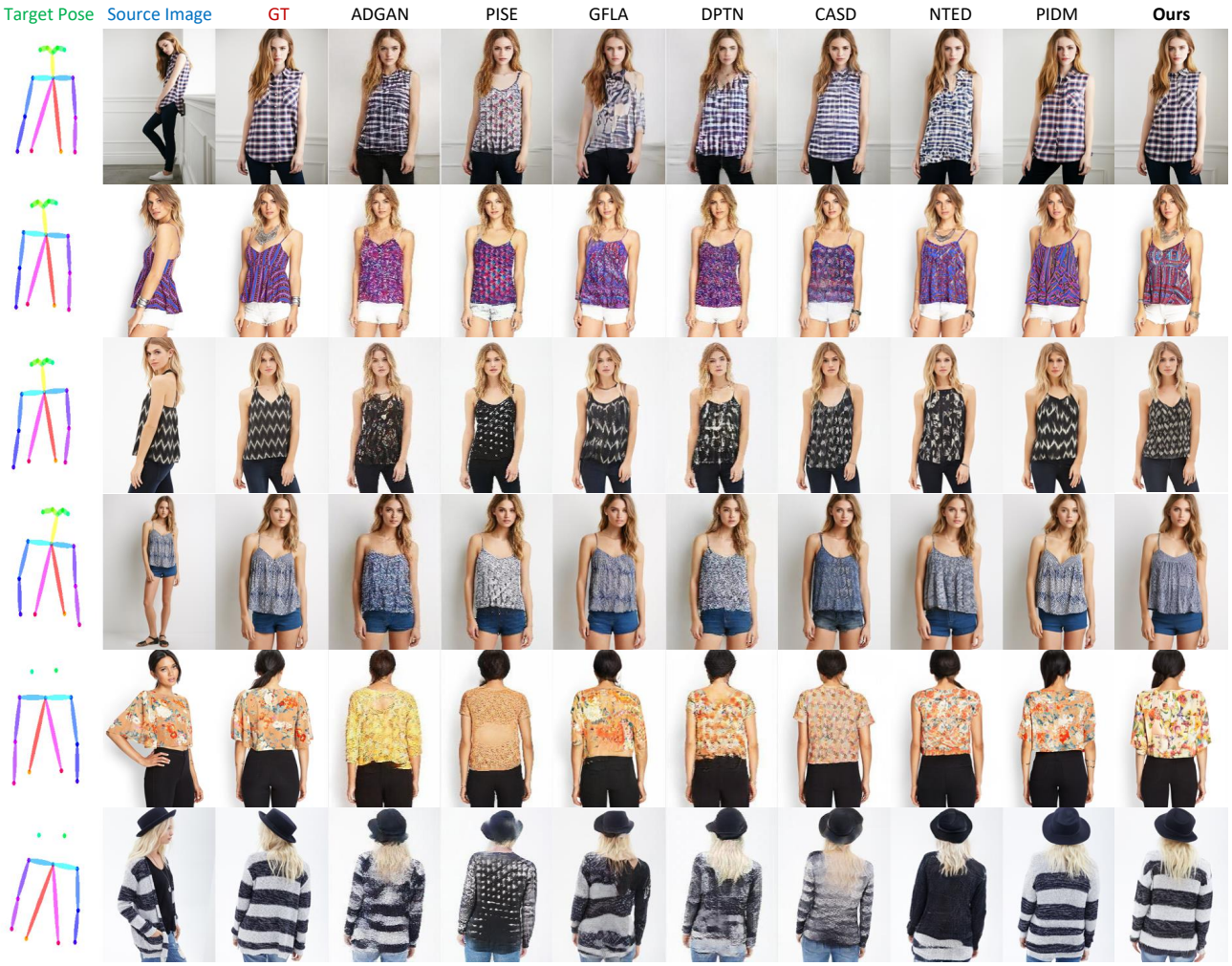

Figure 10: Compare some reported images in the prior paper and **ours**. Images generated by our method are also comparable or even better as it is closer to the ground truth images (*e.g.* samples in rows 1, 2, 6). Images at 256×256 resolution are resized to 256×176.

A.2. Generation given the same pose and different source views and vice versa

In Fig. 11, we provide more comprehensive examples for comparison when all state-of-the-art approaches perform the generation with the given the same target pose + different views of the source image of the same person and vice versa. As shown in Fig. 11, in the case of given the same target pose, when varying the source views of the person, all existing methods failed to capture the consistent target image, while our proposed X-MDPT handles it very well. Our model's superiority is also demonstrated when we keep the source image unchanged and generate different target images. Fig. 12 shows more generated by our X-MDPT-L for different persons with different views when generating the same target pose.

A.3. More details of experimental setups

We use 50 steps DDIM (Song et al., 2020) for inference which is the same as PIDM. The details for three variants of X-MDPT-S, B, and L are provided in Tab. 4 and Fig.13. For VAE, we fine-tuned only the decoder using VAE of Stable Diffusion (Rombach et al., 2022) on the training data of DeepFashion for 77 epochs. The face is distorted if not fine-tuning.

Figure 11: **Different views to the same target pose and vice versa**. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176.

Table 4: **Parameters and Configs**. We follow ViT (Dosovitskiy et al., 2010) to name models for Small (S), Base (B), and Large (L). Our X-MDPT has slightly more parameters compared to DiT (Peebles & Xie, 2023) and MDT (Gao et al., 2023) as it needs one more cross-attention block, but overall, the total parameters are almost similar.

Method	Layers	Dim.	Heads	Param. (M)	Method	Layers	Dim.	Heads	Param. (M)	Method	Layers	Dim.	Heads	Param. (M)
DiT-S	12	384	6	32.9	MDT-S	12	384	6	33.1	X-MDPT-S	12	384	6	33.52
DiT-B	12	768	12	130.3	MDT-B	12	768	12	130.8	X-MDPT-B	12	768	12	131.92
DiT-L	24	1024	16	458.0	MDT-L	24	1024	16	459.1	X-MDPT-L	24	1024	16	460.24

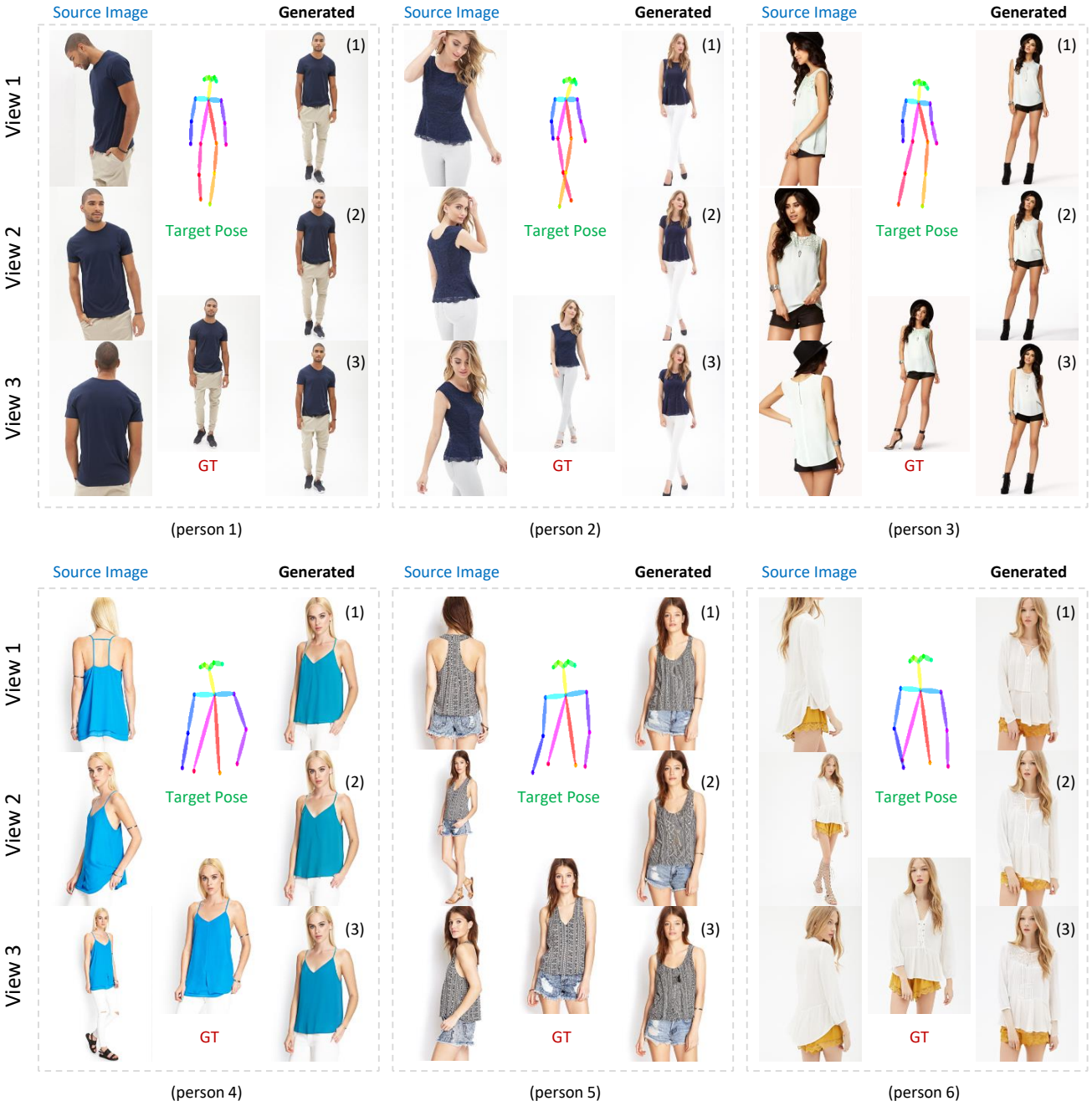

Figure 12: **Different views of different persons**. X-MDPT-L generates consistent target images for the same target pose with three corresponding views, GT denotes ground truth. Images at 256×256 resolution are resized to 256×176.

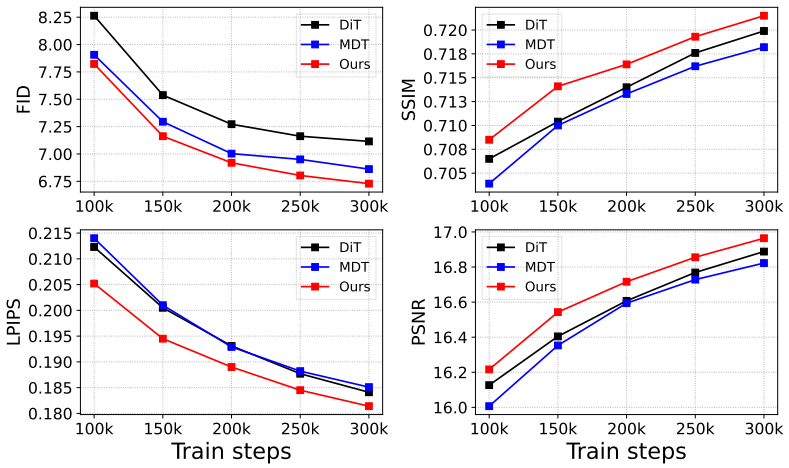

Figure 13: **Transformer baselines.** We compare three frameworks: Applying to the PHIG problem with 1) DiT, 2) MDT, and 3) Ours. All methods use the size ViT-B (131M). MDT can improve DiT on FID but NOT on other metrics, while Ours improves on both metrics.

A.4. Failure cases

We note that our methods failed in certain cases that potentially mitigate it to achieve the best performance on the test set with SSIM, and LPIPS, as shown in Fig. 14. First, changing in clothing of the target images can lead to worse quantitative scores. Second, the pose is under-represented, for example, the left hand of the woman is missing in the pose image making the model predict without that part. Third, the wrong-predicted pose that came from the OpenPose (Cao et al., 2017) can make the model predict wrong targets. Increasing more training pairs of clothes changing and improving detected pose algorithms may resolve these failures.

Figure 14: **Failure cases.** Several use cases can result in worse metrics such as SSIM and LPIPS when compared to GT.

A.5. Different random seeds

In Fig. 15 we show some samples generated by X-MDPT-L for six different seeds: 0, 100, 200, 300, 400, 500. As shown in the figure, our method generated the target images quite stable, demonstrating that it is not sensitive to the random seeds.

Figure 15: **Different random seeds**. X-MDPT-L generates stable samples for 6 different random seeds, GT denotes ground truth. Images at 256×256 resolution are resized to 256×176.

A.6. Training time

We compare the training times of our method and the runner-up PIDM (Bhunia et al., 2023). We refer to DreamPose (Karras et al., 2023) showed that PIDM uses 4 A100 GPUs trained for 26 days (section 2.3 of Dreampose), which results in if the use of the same resource as ours, *i.e.* only 1 GPU A100, it would expect to need 104 days for PIDM. While our method trains for 800k steps (with a single GPU) takes 15.7 days. However, training our method using a single GPU with only 300k steps (5.9 days for X-MDPT-L, and 4.1 days for X-MDPT-B) already produces very good-quality images, while PIDM (needs 2/3 of their training to get a model can create good enough image, *i.e.* would take approximately 69.3 days).

Table 5: **Training time**. Compare PIDM and our method at 256×256 resolution. The results show that our method is much more efficient in training time compared to the pixel-based PIDM when using the same computation resource (1 GPU). The inference time taken from our main paper is for generating 8 images using a single A100 GPU.

Method	Training time to full converged ↓	Inference Time (s) ↓	Param. (M)↓
PIDM (Bhunia et al., 2023)	104 days	16.975	688.0
X-MDPT-L	**15 days**	**3.124**	**460.24**

We have not taken PoCoLD (Han et al., 2023) to compare qualitative results because we find that the published code of PoCoLD is not complete, no published checkpoint, and we are unable to reproduce it. Instead, we show some references excerpted from their paper as follows (Table 2 in PoCoLD paper). Inference time of PIDM 9.25s vs. PoCoLD 4.99s where they measured on 256×256 generation for a single GPU Tesla V100. We can see that PoCoLD only speeds up $1.85\times$ over PIDM, but our method speeds up PIDM with $14.25\times$, $13.07\times$, and $5.43\times$ for three of our model variants in the same settings. Note that, our smallest model X-MDPT-S already outperforms PoCoLD on FID score with $11\times$ fewer parameters and its inference speed is $14.25\times$ faster than PIDM.

We notice that the same PIDM, PoCoLD also uses the disentangled classifier free guidance form, so it will need one forward for unconditional, one forward for the pose condition, and one for the source image condition and results in 50 times more forwards, this will significantly mitigate their inference speed. By contrast, our method used normal CFG that required only one forward for condition and one for unconditional, saving 50 times of forward.

A.7. Self-Supervised Learning Models

There are various SSL models have been explored to learn the representations without labels (He et al., 2022; Pham et al., 2021; 2023; Oquab et al., 2023; Zhang et al., 2022a;b). These models serve as a good extractor for various applications (Pham et al., 2022b; Chen et al., 2023). DINOv2 (Oquab et al., 2023) demonstrated an excellent pre-trained model for various diffusion-based frameworks. We mainly use DINOv2, but the other options may be worth trying. With the potential of diffusion transformers for conditional learning, it is expected to have more discovery of its capability in various domains and applications such as speech processing (Jung et al., 2022; 2020; Trung & Yoo, 2019), data augmentation (Lee et al., 2020), VQA (Kim et al., 2020), visual detection learning (Vu et al., 2019), super-resolution (Niu et al., 2023; 2024a;b).

A.8. Overlapping in DeepFashion Dataset

We illustrate several cases of overlapping in Fig. 16 and Fig. 17, with a rare duplication case shown in Fig. 18.

A.9. High-resolution images 512×512

We also report some high-resolution images 512×512 generated by our method in Fig. 19, Fig. 20, and Fig. 21, etc... Model X-MDPT-L is trained on the DeepFashion dataset for high resolution and generates excellent target images.

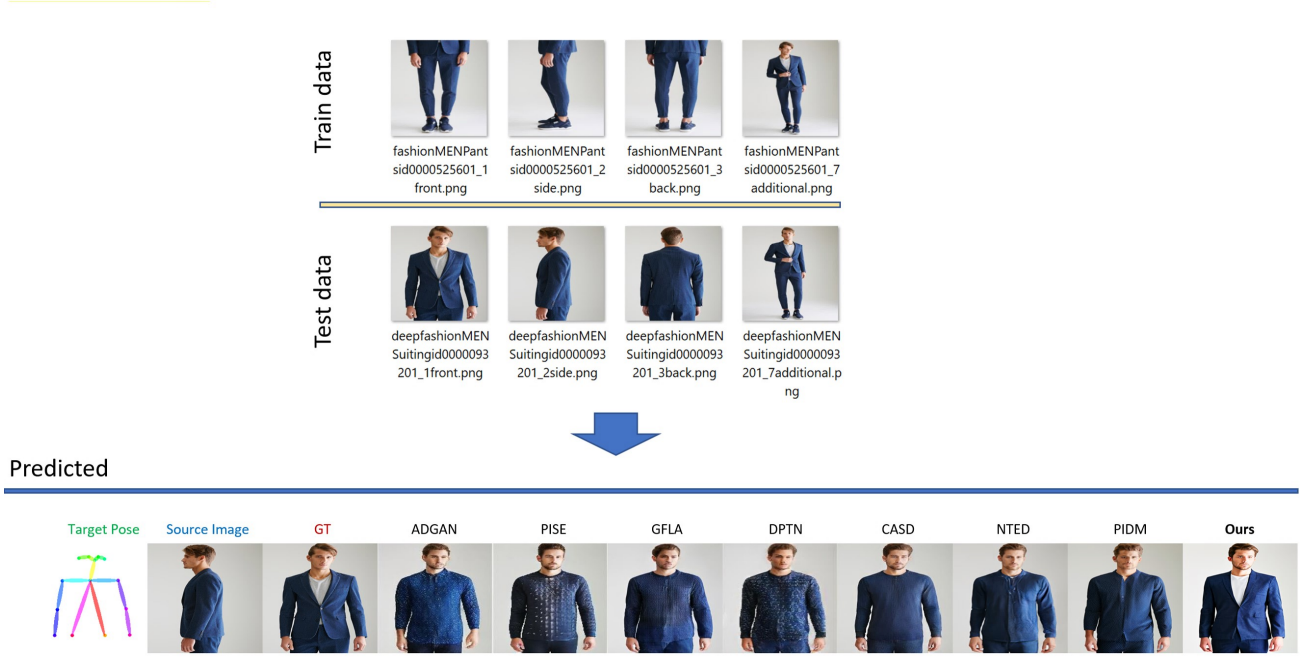

Figure 16: Overlapping case on the DeepFashion dataset (1).

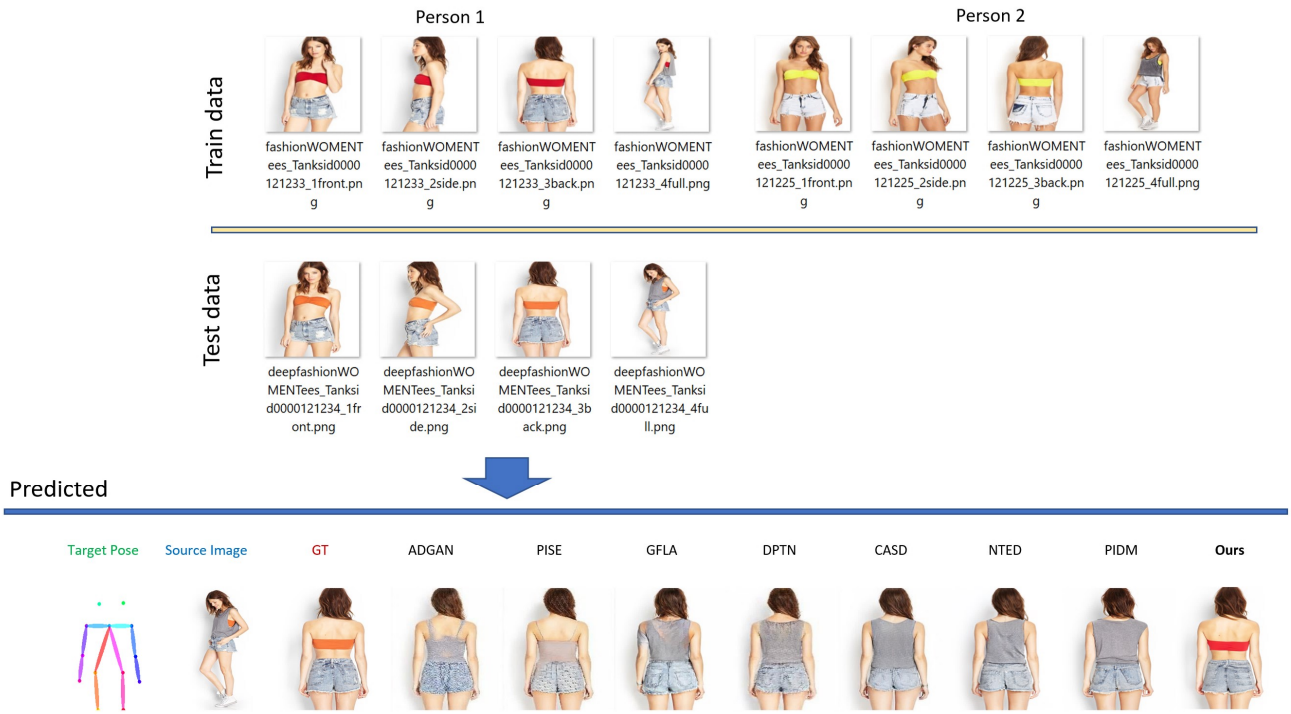

Figure 17: Overlapping case on the DeepFashion dataset (2).

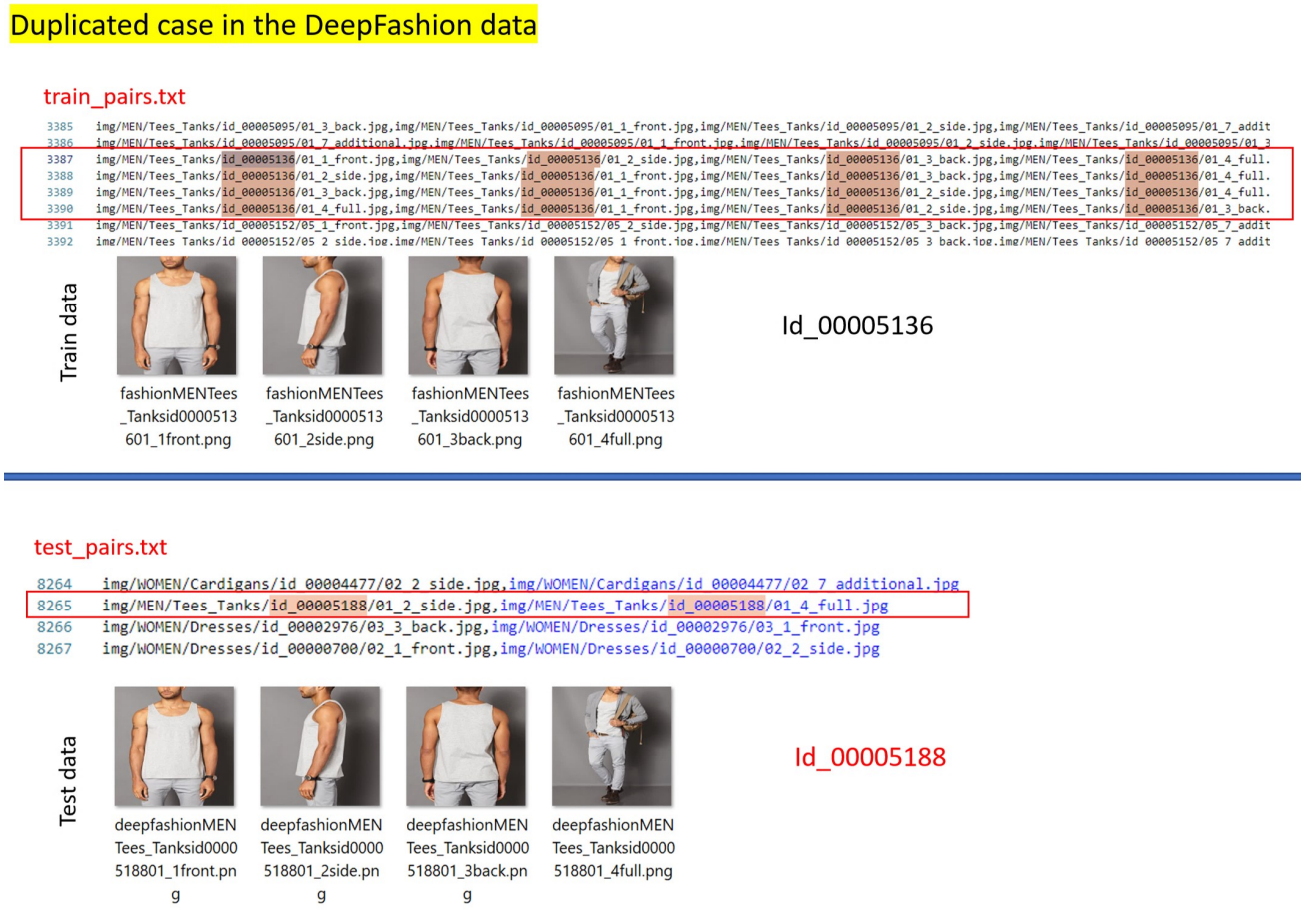

Figure 18: Duplicated case on the DeepFashion dataset.

Figure 19: 512×512 Images generated by our X-MDPT-L model training on the DeepFashion dataset (1).

Figure 20: 512×512 Images generated by our X-MDPT-L model training on the DeepFashion dataset (2).

Figure 21: 512×512 Images generated by our X-MDPT-L model training on the DeepFashion dataset (3).

Figure 22: 512×512 Images generated by our X-MDPT-L model training on the DeepFashion dataset (4).

Figure 23: 512×512 Images generated by our X-MDPT-L model training on the DeepFashion dataset (5).

A.10. More qualitative comparisons on 256×176

More visualization results in resolution 256×176 are provided below from Fig. 24 below. For various difficult cases such as rare poses and a close look at the source image, other methods failed to generate the correct target, while X-MDPT can handle them adequately.

Figure 24: More comparison images with state-of-the-art approaches. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176. (1)

Figure 25: More comparison images with state-of-the-art approaches. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176. (2)

Figure 26: More comparison images with state-of-the-art approaches. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176. (3)

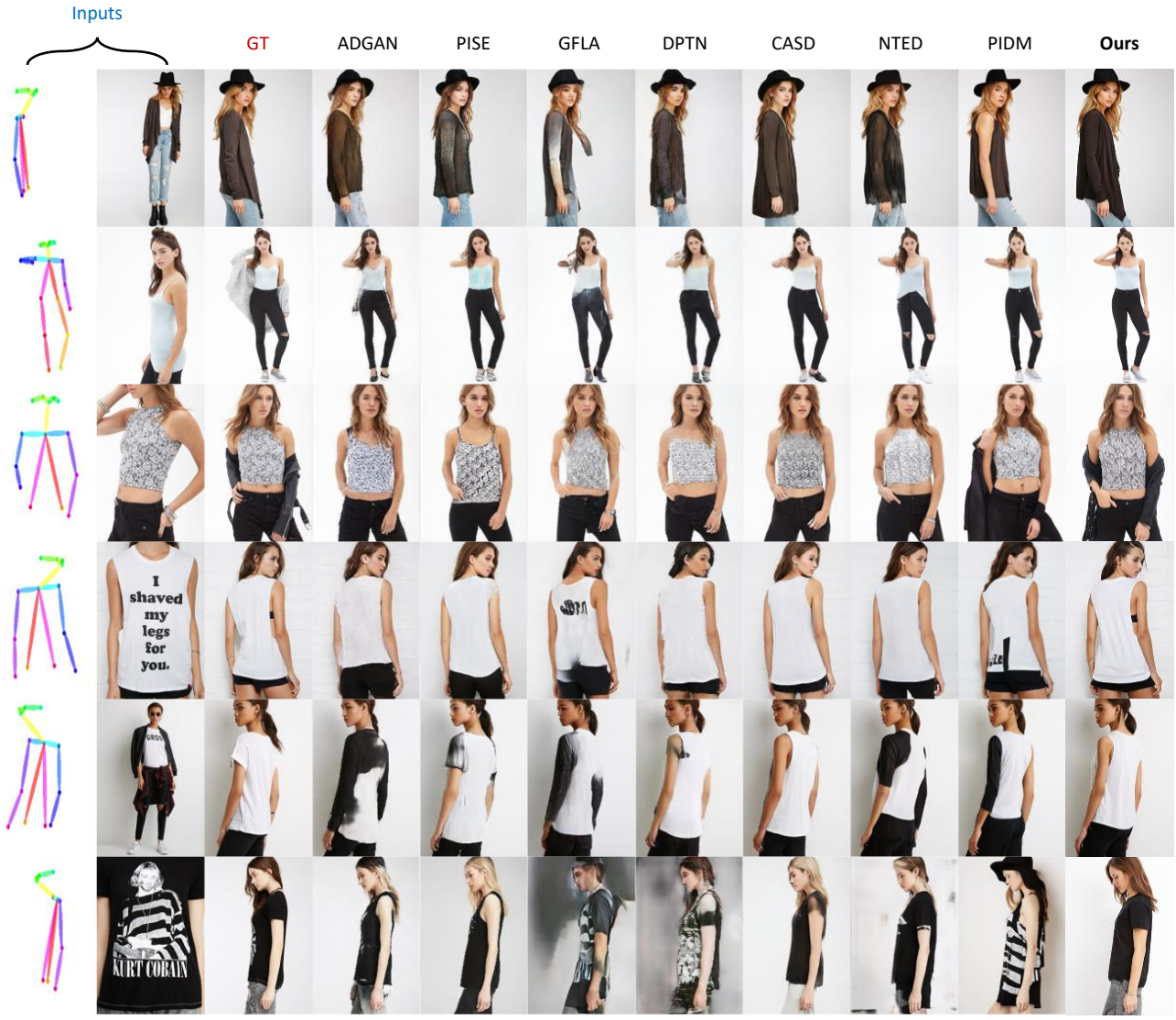

Figure 27: More comparison images with state-of-the-art approaches. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176. (4)

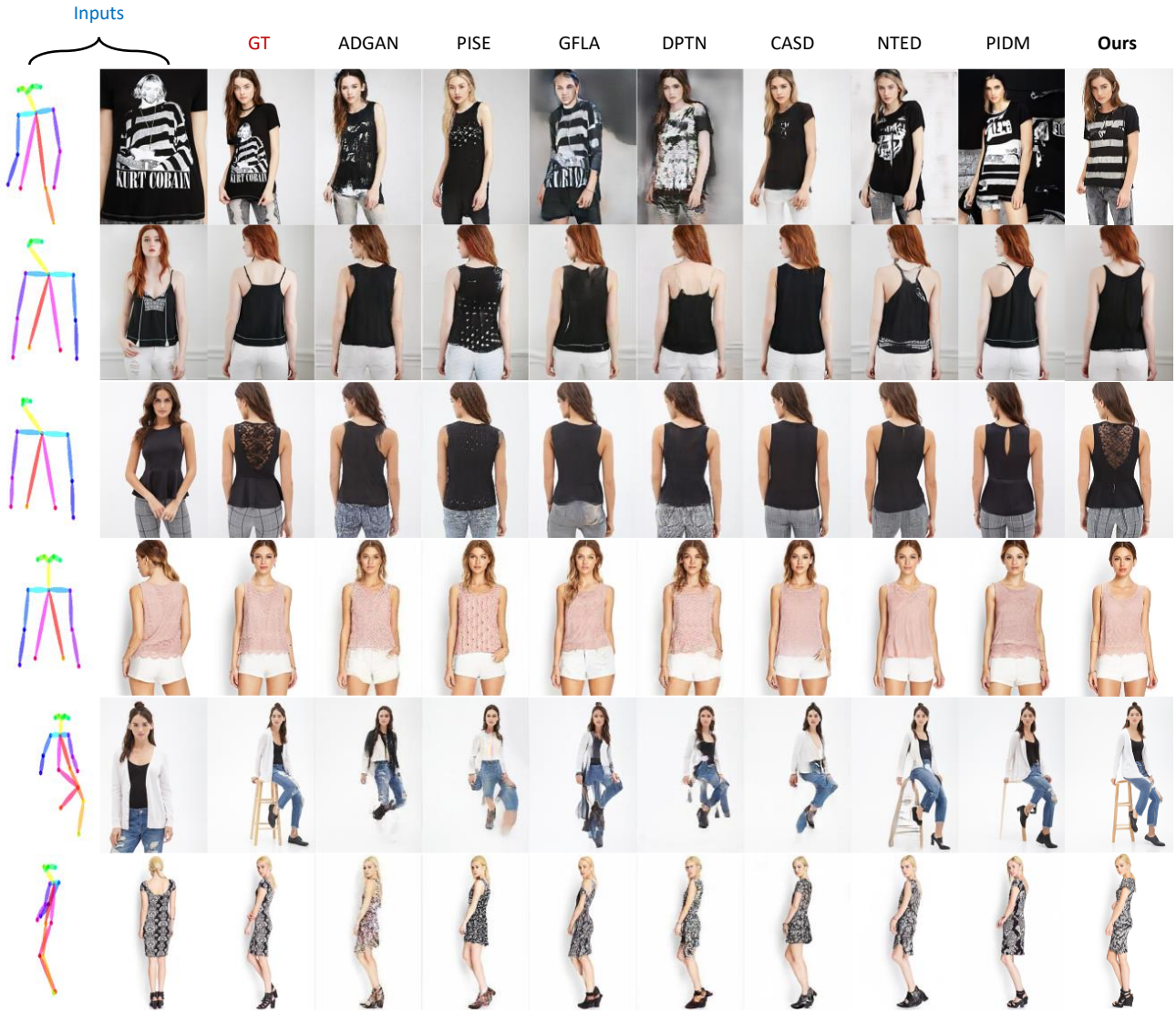

Figure 28: More comparison images with state-of-the-art approaches. X-MDPT generates a target image that is more comprehensive and closely aligns with the ground truth image. Images at 256×256 resolution are resized to 256×176. (5)