Disentangled Pose and Appearance Guidance for Multi-Pose Generation

Supplementary Material

A. Formulation of Diffusion Model

The Latent Diffusion Model (LDM) [5] is a probabilistic diffusion framework [1] designed to operate in a latent space rather than the pixel space. By shifting the diffusion process to the latent domain, LDM significantly reduces computational complexity while maintaining perceptual fidelity. This is achieved through an autoencoding model that learns a compact latent representation perceptually equivalent to the original image space. The model additionally incorporates a compression learning stage to further optimize training efficiency.

Diffusion models generally start from a data distribution z_0 and define a forward Markovian process q, wherein Gaussian noise is iteratively added to z_0 over t time step. The forward process is formalized as:

$$\begin{aligned} q\left(z_{1:T}|z_0\right) &= \prod_{t=1}^{T} q\left(z_t|z_{t-1}\right), \\ q\left(z_t|z_{t-1}\right) &= \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I\right), \end{aligned} \tag{1}$$

where $\beta_t \in (0,1)$ determines the noise variance schedule. Using the closed-form expression derived in [1], data at any time step t can be sampled directly:

$$\begin{aligned} q\left(z_t|z_0\right) &= \mathcal{N}\left(z_t; \sqrt{\bar{a}_t}z_0, (1-\bar{a})I\right), \\ &= \sqrt{\bar{a}_t}z_0 + \epsilon\sqrt{1-\bar{a}_t}, \epsilon \in \mathcal{N}\left(0, I\right), \end{aligned} \tag{2}$$

where $\bar{a}_t = \prod_{s=0}^{t} a_s$ and $a_t = 1 - \beta_t$. This allows for efficient noise schedule definition through $\bar{a}_t$. The posterior $q\left(z_{t-1}|z_t, z_0\right)$ is derived using Bayes' theorem and is also Gaussian:

$$q\left(z_t|z_{t-1}\right) = \mathcal{N}\left(z_{t-1}; \tilde{\mu}\left(z_t, z_0\right), \tilde{\beta}_t I\right), \tag{3}$$

where $\tilde{\beta}_t = \frac{1-\bar{a}_{t-1}}{1-\bar{a}_t}\beta_t$ and $\tilde{\mu}\left(z_t, z_0\right)$ is expressed as:

$$\tilde{\mu}\left(z_t, z_0\right) = \frac{\sqrt{\bar{a}_{t-1}}\beta_t}{1-\bar{a}_t}z_0 + \frac{\sqrt{a_t}\left(1-\bar{a}_{t-1}\right)}{1-\bar{a}_t}z_t. \tag{4}$$

To reverse the forward process and sample from the learned distribution, the conditional distribution $p(z_{t-1}|z_t)$ must be estimated. Since $p(z_{t-1}|z_t)$ is intractable, a neural network is employed to approximate it:

$$p_\theta\left(z_{t-1}|z_t\right) = \mathcal{N}\left(z_{t-1}; \mu_\theta\left(z_t, t\right), \sum_\theta\left(z_t, t\right)\right), \tag{5}$$

where μ_θ and $\sum_\theta$ are parameterized by the network. To simplify training, [1] fixes the variance $\sum_\theta$ and focuses on learning the mean μ_θ, as the variance is inherently defined by the noise schedule β_t in the forward process. This simplification accelerates convergence and improves model efficiency. Finally, the denoising step is expressed as:

$$p_\theta\left(z_{t-1}|z_t\right) = \mathcal{N}\left(z_{t-1}; \mu_\theta\left(z_t, t, c\right), \sigma^2 I\right), \tag{6}$$

with μ_θ further defined as:

$$\mu_\theta\left(z_t, t, c\right) = \frac{1}{\sqrt{a_t}}\left(z_t - \frac{1-a_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta\left(z_t, t, c\right)\right), \tag{7}$$

where $\epsilon_\theta\left(z_t, t, c\right)$ represents the network's prediction of the noise component.

B. Experiments

Additional Visulizations. We provide additional visualizations of the UBC Fashion dataset [6] and the TikTok dataset [2] in Fig. 1 and 2.

Details of Ablation Experiments. In this section, we present detailed ablation experiments and provide a comprehensive analysis of the results.

- *w/o Global-aware Pose Generation (GPG)*: In this setup, the GPG module was removed, and reference image features were directly used as conditional embeddings for the diffusion model. For pose control, multi-pose features from the pose encoder were added to the U-Net input layer. Results indicate that the GPG module plays a key role in performance improvement. By pre-completing the spatial transformation of the pose and effectively integrating reference image features with global pose features, the GPG module alleviates the modeling burden on the diffusion model, enabling it to focus on content generation tasks.
- *w/o Multi-stage Image Encoder (MIE)*: The MIE module was replaced with the CLIP image encoder [4], which lacks the ability to produce multi-scale features. Consequently, the Appearance Adapter (AD) module could not be utilized. The experiments show that relying solely on single-scale reference image features fails to preserve the original appearance characteristics, resulting in degraded pose generation quality.
- *w/o Appearance Adapter (AD)*: To assess the AD module's effectiveness, it was removed, severing the conditional link between the MIE module and the U-Net's upsampling layers. The results highlight that the absence of the AD module prevents the diffusion model from leveraging multi-scale appearance features, thereby reducing generation quality. The AD module plays a crucial role

Figure 1. Additional qualitative results of our method on the UBC Fashion dataset [6].

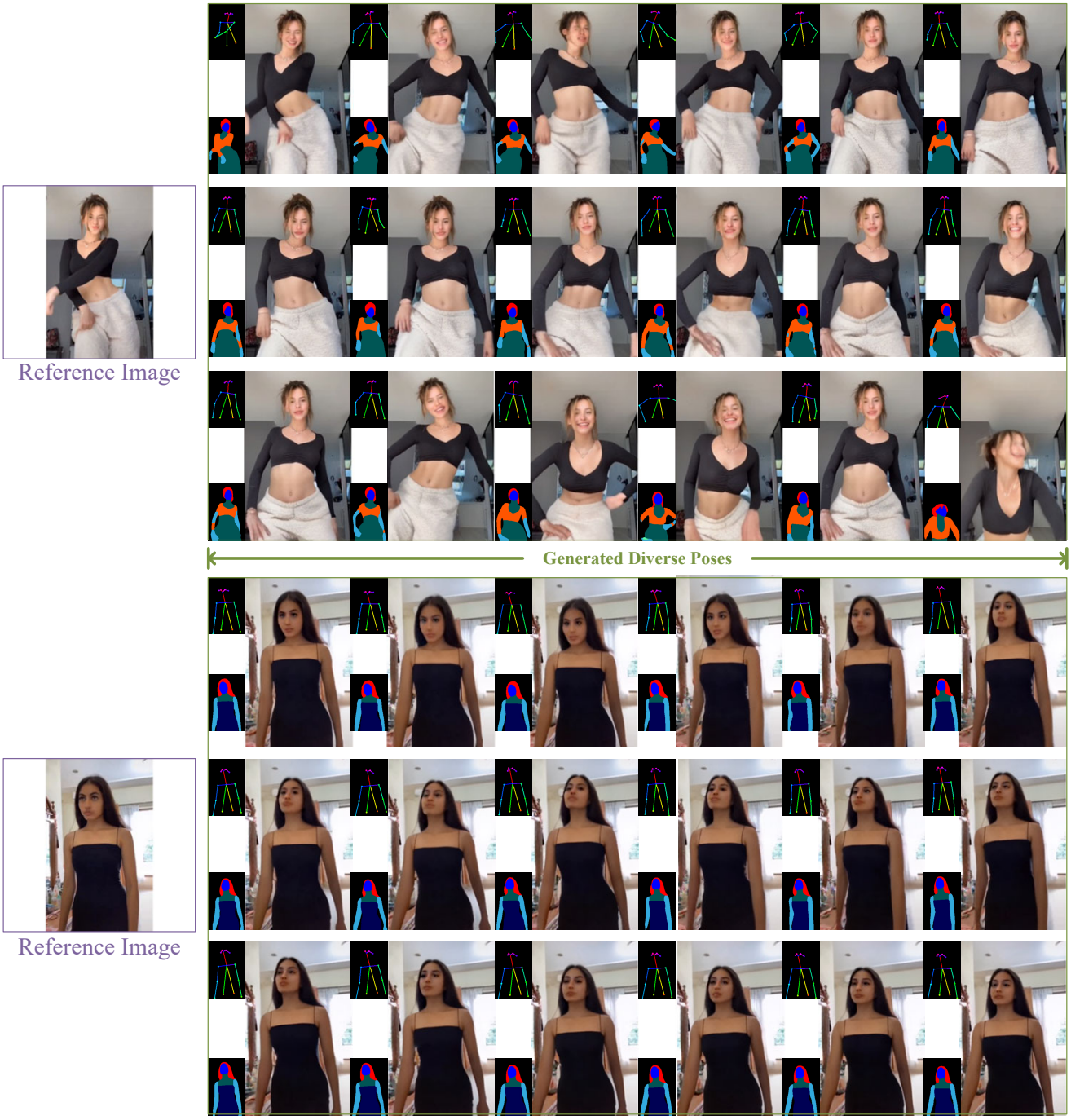

Figure 2. Additional qualitative results of our method on the TikTok dataset [2].

in learning high-level semantic information about human appearance and providing the model with effective appearance priors.

- *w/o Fine-tune VAE (FVAE)*: Consistent with conclusions in [6], our experiments confirm that sample-specific fine-tuning is essential for maintaining identity and clothing characteristics while achieving cross-pose consistency. Moreover, FVAE significantly enhances the clarity and realism of the generated images.
- *w/o* $\mathcal{L}_{rec}$: The $\mathcal{L}_{rec}$ term was removed to evaluate its contribution. Inspired by [7], the reconstruction of the person from the reference image was incorporated during pose generation, encouraging the model to generate robust and consistent human poses. Results demonstrate the critical role of $\mathcal{L}_{rec}$ in improving pose quality and structural integrity.

Figure 3. Qualitative results of video generation at 16 FPS using our method on the UBC Fashion dataset [6] are presented. The figure demonstrates our method's ability to generate coherent and realistic human pose videos in a video format.

- *Alternative GTD Variants*: To validate the effectiveness of the Global-aware Transformer Decoder (GTD) within the GPG module, we compared it against two alternative designs:
  - Ours+(a): An additional cross-attention layer was added to a standard Transformer decoder for global pose interaction.
  - Ours+(b): A cross-stacked Transformer decoder was employed, where the Nth layer performs spatial transformation, and the $(N+1)$th layer handles global pose interaction.

In contrast, our proposed GTD adopts the reverse order: it first enhances target pose features using global pose features through similarity-based querying, followed by interaction with reference image features to complete spatial transformation. Experimental results demonstrate that GTD outperforms both alternatives, achieving superior performance and generation quality.

Model	SSIM ↑	PSNR ↑	FID ↓	LPIPS ↓	L_1 ↓	FVD(16f) ↓
DreamPose [3] (ICCV'23)	0.885	-	13.04	0.068	0.025	238.75
Ours	0.939	25.617	11.74	0.052	0.015	66.410

Table 1. Quantitative metrics for generated videos (16 frames) on the UBC Fashion dataset [6].

Discussion on Video Generation. Our method disentangles pose control from appearance guidance, enabling high-quality multiple human pose generation. Pose control is implemented through the iterative application of the Global-aware Transformer Decoder (GTD), a design that seems to naturally extends to video generation tasks. To assess the performance of our method on video generation, we conducted experiments using the dataset from [6], with quantitative results presented in Tab. 1. The iterative generation mechanism of the GTD, combined with the Global-aware block, allows our approach to achieve superior performance in video generation. Future work will focus on enhancing temporal consistency by incorporating an optional motion module into the primary U-Net architecture, further strengthening the model's capability to generate temporally coherent video sequences.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[2] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 1, 3

[3] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22623–22633. IEEE, 2023. 5

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[6] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 1, 2, 3, 4, 5

[7] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 3