RETHINKING INFORMATION PROPAGATION IN DIFFUSION-BASED POSE-GUIDED PERSON IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pose-guided person image synthesis aims to generate realistic human images under arbitrary target poses while preserving appearance consistency. Existing diffusion-based methods have achieved remarkable progress, yet they still suffer from erroneous information propagation, where appearance features are indiscriminately propagated across semantically unrelated body regions under large pose variations. To address this issue, we propose a diffusion-based framework that explicitly rethinks information propagation during denoising. Specifically, selective self-attention suppresses redundant feature interactions through dual-sided channel gating, while the proposed query-gated cross-attention dynamically regulates appearance injection according to semantic relevance, effectively preventing erroneous feature propagation. Furthermore, a transformer refiner equipped with window-based attention and perceptual loss is introduced to further correct residual geometric inconsistencies and recover high-frequency texture details lost during latent-space compression. Extensive experiments demonstrate that our method achieves state-of-the-art generative quality, attaining competitive FID and LPIPS while requiring substantially fewer parameters and lower GPU memory.

1 INTRODUCTION

Pose-Guided Person Image Synthesis(PGPIS) aims to generate a person image with a specified target pose while preserving the appearance of a reference person image, which has been widely applied in virtual try-on, animation production, game content creation and et al. Early PGPIS methods were primarily built upon the generative adversarial network frameworkEsser et al. (2018); Li et al. (2019); Liu et al. (2019); Lv et al. (2021), establishing correspondences between the source image and the target pose through feature warpingDong et al. (2018), optical flow estimationRen et al. (2020), human parsingZhang et al. (2021), and attention-based alignmentRen et al. (2020). Although achieving promising progress, their one-shot forward generation paradigm makes it difficult to faithfully reconstruct realistic human structures under large pose variations.

Recently, diffusion models have substantially advanced PGPIS through the denoising mechanism. Representative methods including PIDMBhunia et al. (2023), PoCoLDHan et al. (2023), and MCLDLiu et al. (2025), have improved generation quality from different perspectives such as appearance correspondence modeling, DensePoseGüler et al. (2018) guidance, and fine-grained person detail enhancement. However, these approaches mainly focus on constructing stronger pose correspondences, while paying relatively less attention to the mechanism of information propagation throughout the denoising process.

We argue that the bottleneck of existing diffusion-based PGPIS methods lies not in insufficient information correspondence, but in erroneous information propagation during the denoising process. Existing methods generally inject reference appearance features through cross-attention, where each query interacts with all reference features. When the target pose differs substantially from the reference pose, semantically irrelevant appearance information can easily propagate to incorrect regions, leading to artifacts. as illustrated in Fig. 1. Moreover, conventional self-attention establishes fully connected dependencies among all tokens, causing a large amount of ineffective feature interactions to accumulate throughout the denoising process. In addition, latent diffusion inevitably loses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Conventional cross-attention lacks effective filtering of injected features, allowing erroneous information propagation, causing a gray sweater to be incorrectly generated as black mesh.

high-frequency information during latent-space compression, making the generated images prone to over-smoothed textures and insufficient fine-grained details. These observations suggest that the key challenge in diffusion-based pose-guided person image synthesis is not to further strengthen pose correspondence, but to fundamentally rethink how information should propagate throughout the denoising process.

To address the above issues, we propose a two-stage diffusion–refiner framework that reconstructs PGPIS from the perspective of information propagation. Specifically, in the diffusion stage, a selective self-attention(SSA) is introduced to suppress redundant feature interactions. By applying dual-sided channel gating to both query and key, SSA encourages attention to establish dependencies only between semantically consistent tokens instead of densely modeling all token relationships. Then, we further propose a query-gated cross-attention(QGCA) that adaptively controls appearance injection according to the semantic relevance between target pose and reference feature. Unlike conventional cross-attention, which uniformly propagates reference appearance features, QGCA learns a region-aware gating strategy that selectively injects appearance information into corresponding semantic regions such as hair, facial regions and et al. as illustrated in Fig. 5. Furthermore, we introduce a transformer refiner as the second stage for further refinement. Equipped with window-based self-attention, the refiner explicitly corrects residual spatial misalignments left by the diffusion stage, while perceptual loss restores high-frequency texture details lost during latent diffusion. By decoupling selective information propagation from refinement, the proposed two-stage design enables each stage to concentrate on its own objective, resulting in more higher-quality PGPIS.

Extensive experimental results demonstrate that the proposed method achieves superior performance over existing SOTA approaches. More importantly, through visual analyses, we further show that our method more effectively preserves the semantic boundaries of clothing and human structures under complex pose variations, providing an intuitive and convincing interpretation of selective information propagation in diffusion-based PGPIS. Our main contributions are summarized as follows:

- We revisit PGPIS from a perspective of information propagation, revealing that erroneous information propagation is a primary cause of appearance inconsistency under complex pose variations.
- We propose a two-stage diffusion-refiner framework that decomposes the generation process into two consecutive stages: selective propagation-based generation and end-to-end detail refinement.
- We propose selective self-attention and query-gated cross-attention that effectively suppressing erroneous appearance information injection.

2 Related Work

2.1 Pose-Guided Person Image Synthesis

Early methods were primarily built upon the GAN framework. PG2Ma et al. (2017) was the first to introduce a two-stage framework for PGPIS. Deformable GANSiarohin et al. (2018) alleviated spatial misalignment under large pose variations through deformable skip connections. Subsequently, PISEZhang et al. (2021) incorporated human parsing priors to further improve semantic consistency during the generation process. Despite the remarkable progress achieved by GAN-based methods, they remain prone to structural distortions and texture artifacts under complex pose variations due to their one-shot forward generation paradigm and the inherent instability of adversarial training.

With the development of diffusion models, PSPIS has gradually shifted toward a progressive denoising paradigm. PIDMBhunia et al. (2023) was the first to adopt diffusion models for person appearance transfer. PoCoLDHan et al. (2023) further strengthened pose correspondence by incorporating DensePoseGüler et al. (2018) guidance, while MCLDLiu et al. (2025) introduced multi-focus conditional constraints to improve detail restoration. Although these methods have substantially improved generation quality, their efforts remain primarily focused on strengthening the correspondence between appearance and pose. In contrast, our work revisits how appearance information should propagate throughout the denoising process.

2.2 Information Propagation in Diffusion

Cross-AttentionVaswani et al. (2017) has become the core mechanism for conditional information propagation in diffusion models. Representative PGPIS methods, including Stable Diffusion, PIDMBhunia et al. (2023), and PoCoLDHan et al. (2023), inject reference appearance features into the diffusion network through cross-attention, allowing each query to interact with all reference features. However, dense cross-attention tends to propagate semantically irrelevant appearance information into incorrect regions, leading to clothing texture interference1 and mixing. PoCoLDHan et al. (2023) reduces redundant interactions through low-rank compression, but imposes no constraint on the value features, and still fails to prevent the injection of irrelevant information.

Meanwhile, self-attentionVaswani et al. (2017) establishes fully connected dependencies among all tokens. While it effectively captures long-range relationships, it also introduces substantial redundant dependencies, causing irrelevant features to be propagated indiscriminately. Rather than continually strengthening correspondence modeling as in existing methods, our approach addresses this issue from the perspective of selective information propagation. Through a gating mechanism and low-rank filtering, appearance features are propagated only to semantically relevant regions, thereby effectively suppressing erroneous feature injection during the diffusion denoising process.

3 Methods

3.1 Overview

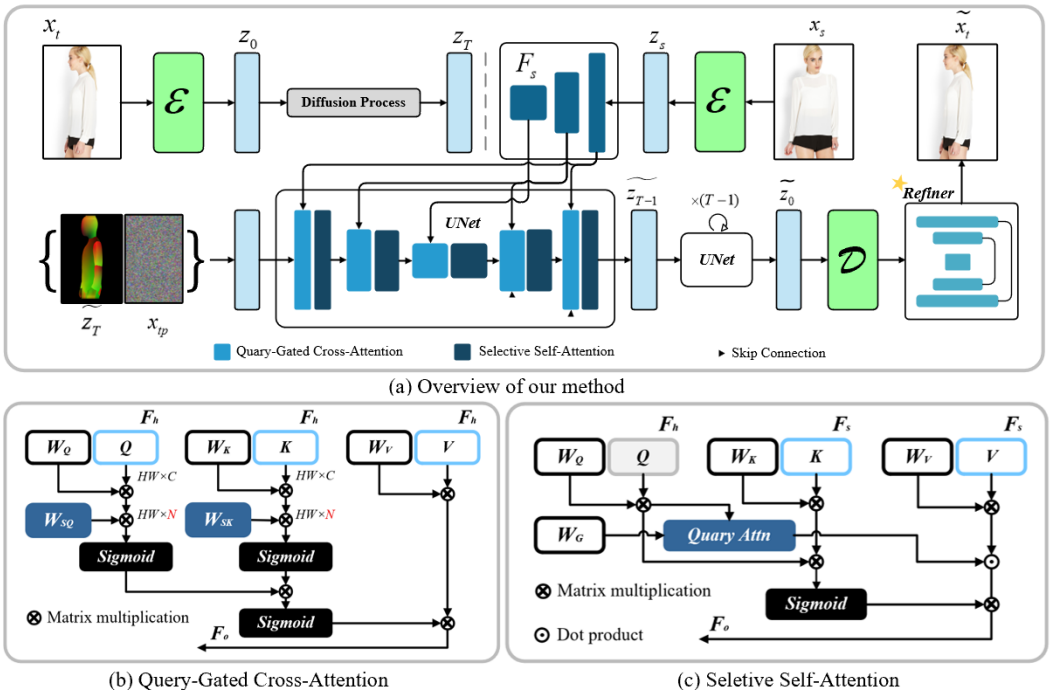
As illustrated in Fig. 2, the proposed framework consists of a diffusion generator and a transformer refiner, which form a two-stage information propagation pipeline. Given a reference image I_s and a target pose P_t, the first stage performs selective appearance propagation in the latent diffusion space. Specifically, we first employ query-gated cross-attention to adaptively control the injection of appearance features based on the semantic relevance between target pose and reference appearance. Then, we introduce selective self-attention to reduce redundant interactions among appearance features, thereby suppressing the erroneous propagation of appearance information into semantically irrelevant regions.

Unlike existing methods that rely on a single diffusion network to simultaneously perform generation and refinement, the second stage is explicitly designed for information refinement. The vision transformer leverages window-based self-attention to explicitly correct the residual spatial misalignments caused by large pose variations, while perceptual loss further restores the high-frequency texture details lost during latent diffusion. By explicitly decoupling information generation from information refinement, the proposed framework achieves higher-quality generation.

3.2 Selective Information Propagation

3.2.1 Query-Gated Cross-Attention

Traditional cross-attention allows each query to interact with all reference features. However, this dense information propagation implicitly assumes that all reference regions are equally important for every target location. When there is a large pose discrepancy between the target and reference poses, semantically irrelevant appearance information can easily propagate to incorrect human regions, resulting in typical artifacts such as clothing texture mixing and blurred semantic boundaries, as illustrated in Fig. 1. Therefore, we argue that the key challenge is not how to inject more appearance



(a) Overview of our method

(b) Query-Gated Cross-Attention

(c) Seletive Self-Attention

Figure 2: Overview of our method, which is mainly composed of two stages, diffusion generator and detail refinement.

information, but how to selectively propagate only the appearance information that is semantically relevant to the current target region.

To address this issue, we propose a query-gated cross-attention, which introduces a query-dependent gating mechanism before information aggregation in cross-attention. Specifically, QGCA adaptively modulates the propagation strength of different appearance features according to the semantic relevance between target pose and reference appearance that firstly predicts a gating weight based on the current query to measure the degree to which the current target position requires reference appearance information. The gating weight can be formulated as:

$$g = \sigma(W_G \cdot Q) \tag{1}$$

where $\cdot$ denotes the dot product and σ denotes the sigmoid activation function. The gating weights are then used to selectively modulate the reference appearance features:

$$V' = g \cdot V \tag{2}$$

Finally, the modulated value features are incorporated into the cross-attention aggregation process:

$$Y = softmax(\frac{QK^T}{\sqrt{d}})V' \tag{3}$$

Unlike conventional cross-attention, which applies a unified value aggregation across all spatial locations, QGCA inherently functions as a query-conditioned dynamic filter. Specifically, each query independently generates its own gating vector, enabling every location to adopt an independent appearance injection policy before cross-attention aggregation. Consequently, QGCA possesses an intrinsic capability for part-specific attention, allowing the model to selectively inject appearance information into individual body regions rather than uniformly propagating reference features5.

3.2.2 Selective Self-Attention

Conventional self-attention establishes fully connected dependencies among all tokens, assuming that all interactions are equally important. Under large pose variations, such dense dependencies

introduce a substantial amount of information that is irrelevant to the current generation region, allowing redundant feature interactions to accumulate and ultimately interfere with appearance representation in critical human regions. Therefore, we argue that, in PGPIS, not all feature dependencies should be preserved. Instead, the key is to establish discriminative feature propagation relationships.

To address this issue, we propose selective self-attention, which adaptively filters feature dependencies to suppress redundant information propagation. Specifically, we remove ineffective interactions between query and key through a low-rank approximation and sigmoid activation:

$$Q' = \sigma(W_{SQ} * Q), \quad K' = \sigma(W_{SK} * Q) \tag{4}$$

$$Q, K \in R^{C*H*W}; \quad Q', K' \in R^{N*H*W} \tag{5}$$

where σ denotes the sigmoid activation function, and N is a relatively small value, which is set to 64 in our implementation. The filtered interactions are then used for feature modulation:

$$Y = softmax(\frac{QK^T}{\sqrt{d}})V \tag{6}$$

The effectiveness of SSA originates from its dual-sided gating mechanism for query–key interactions. A channel contributes significantly if and only if it is simultaneously preserved by both the query and key masks, whereas channels suppressed on either branch are naturally attenuated through multiplicative gating and the subsequent sigmoid operation. This "NOR" behavior fundamentally distinguishes SSA from conventional self-attention("NAND"). As a result, SSA effectively reduces redundant feature interactions, enabling attention to focus on semantically consistent token dependencies rather than densely modeling relationships among all tokens6.

3.3 Information Refinement

After SSA and QGCA, the diffusion stage is able to establish a more reliable appearance propagation while effectively suppressing erroneous information propagation across semantically irrelevant regions. However, we further observe that even when semantic propagation is correctly established, the generated results may still retain minor local geometric errors, such as slight facial misalignment and asymmetric human structures. Therefore, to obtain more realistic details, a refiner is introduced to perform fine-grained adjustments in the spatial domain. Specifically, we incorporate a window-based self-attentionLiu et al. (2021) to refine human structures by modeling semantic relationships in the spatial domain.

However, due to the inherent diversity of diffusion-generated results, directly learning a mapping to the ground truth tends to produce over-smoothed outputs. Therefore, we introduce a scaling factor into the refiner, enabling it to perform genuinely fine-grained adjustments:

$$Y = X + s \cdot tanh(Refiner(X)) \tag{7}$$

where X denotes the output of the diffusion stage, $Refiner(\cdot)$ represents the mapping function of the transformer composed of window-based self-attention, $\tanh(\cdot)$ denotes the hyperbolic tangent activation function, and s is a scaling factor that constrains the refinement magnitude of the transformer. In our implementation, s is set to 0.005.

3.4 Training Objectives

In the first-stage diffusion generation, we supervise the training using the standard MSE loss:

$$\mathcal{L}_{diff} = \mathbb{E}_{Z_t, Z_s, P_t, t, \epsilon} ||\epsilon - \epsilon_\theta(Z_t, Z_s, P_t, t)|| \tag{8}$$

where Z_t denotes the latent noisy sample at diffusion timestep t, Z_s denotes the VAE-encoded source image, and P_t denotes the target pose.

For the second-stage refiner, we observe that the outputs of the first stage inevitably suffer from missing texture details and over-smoothing due to the reconstruction loss introduced by VAE encoding and decoding. Therefore, in addition to the standard L1 loss, we incorporate a perceptual loss to supervise texture restoration:

$$\mathcal{L}_{ref} = |Y - \hat{Y}| + w_l \cdot |lpips(Y) - lpips(\hat{Y})| \tag{9}$$

where Y denotes the ground truth, $\hat{Y}$ denotes the output predicted by the refiner, and $lpips(\cdot)$ denotes the LPIPS perceptual metric computed from multi-layer feature representations of AlexNet.



Figure 3: Results of SOTA methods on Deepfashion dataset.

4 Experiment

4.1 Experiment Setting

Dataset All experiments are conducted on the DeepFashion dataset Liu et al. (2016), which is the most widely used benchmark for PGPIS. We follow the same dataset split and inference resolution as MCLD Liu et al. (2025) and other previous methodsLu et al. (2024); Ren et al. (2022).

Evaluation Metrics We conduct a comprehensive evaluation using metrics including PSNR, SSIM, LPIPS Zhang et al. (2018), and FID Heusel et al. (2017). Specifically, PSNR and SSIM are used to evaluate the fidelity between generated images and ground-truth images. LPIPS measures perceptual similarity in the deep feature space, while FID evaluates the realism of generated images and their consistency with the overall data distribution.

Details Our method is trained on two NVIDIA A100 GPUs. We adopt the pre-trained VAE from Stable Diffusion 1.5. During diffusion training, the batch size is set to 6, and the model is trained for 100 epochs. The diffusion outputs on the training set are then used as the inputs to train the refiner, with a batch size of 4 for 10 epochs. During sampling, a classifier-free guidance(CFG) strategyHo & Salimans (2022) is adopted to improve generation quality, and the CFG scale is set to 3.0.

4.2 Comparison with State-of-the-Art

Quantitative Comparison We quantitatively compare our method with recent PGPIS approaches in Table 1. At both 256×176 and 512×352 resolutions, our method achieves the lowest FID and LPIPS. Although MCLD attains slightly higher SSIM and PSNR, our method exhibits a clear advantage in LPIPS while using only 855M parameters, which is substantially fewer than MCLD's 1717M. These results demonstrate that our method achieves a more favorable efficiency-quality trade-off.

Qualitative Comparison As shown in Fig. 3, the most significant advantage of our method lies in its stronger structural consistency under large pose variations. For example, in the left example of the second row, all competing methods exhibit erroneous feature injection caused by the pose transformation, where black mesh from the back is incorrectly transferred to the front garment. In contrast, our gating-based feature injection effectively suppresses this artifact.

Table 1: Comparison with SOTA methods of PGPIS. Params indicates the total parameters of the model including VAE and et al. Memory indicates the usage of memory when evaluating images in a resolution of 512×512. †We strictly follow the evaluation implementation in CFLDLu et al. (2024) and reproduce these results. ††Results are obtained using the images released by the authors.

Methods	FID	LPIPS	SSIM	PSNR	Params	Memory
*Evaluationon*256×176*resolution*						
NTED(CVPR22)	8.51	0.1770	0.7156	17.74	—	—
PIDM†(CVPR23)	6.81	0.2006	0.6621	15.63	180M	—
PIDM††(CVPR23)	6.44	0.1686	0.7109	17.40	180M	—
PoCoLD(ICCV23)	8.06	0.1642	0.7310	—	207M	—
CFLD(CVPR24)	6.80	0.1519	0.7378	18.23	1184M	—
MCLD(CVPR25)	<u>6.69</u>	<u>0.1482</u>	**0.7511**	**18.84**	1717M	—
Ours(B4)	**6.47**	0.1543	0.7409	18.44	767M	—
Ours	**6.47**	**0.1455**	<u>0.7419</u>	<u>18.52</u>	772M	—
*Evaluationon*512×352*resolution*						
NTED(CVPR22)	7.64	0.1999	0.7359	17.38	—	—
PoCoLD(ICCV23)	8.41	0.1920	0.7430	—	207M	—
CFLD(CVPR24)	7.14	0.1819	0.7478	17.64	1184M	6.3G
MCLD(CVPR25)	<u>7.07</u>	0.1757	**0.7557**	**18.21**	1717M	10.3G
Ours(B4)	**6.57**	<u>0.1756</u>	0.7480	17.86	850M	3.8G
Ours	**6.57**	**0.1704**	<u>0.7488</u>	<u>17.92</u>	855M	3.8G

Furthermore, our method demonstrates superior fine-grained appearance consistency by preserving clear semantic separation across different clothing regions. For example, in the left example of the third row, NTED incorrectly renders the shoulder in a skin tone due to feature interference from the skin, CFLD completely loses the garment's specular appearance, and MCLD incorrectly transfers the garment color to the arm, causing it to appear blue. In contrast, our selective self-attention effectively prevents these errors by selectively preserving discriminative feature interactions.

Model Size and Memory Use Recently, the computational cost of AIGC has attracted increasing attention. Therefore, we further compare the parameter scale and memory usage of existing SOTA methods, as summarized in Table 1. Despite introducing an additional transformer refiner, our two-stage framework maintains high computational efficiency. The final model contains only 885M parameters, representing reductions of approximately 25.3% and 48.5% compared with CFLD(1184M) and MCLD(1717M), respectively.

In terms of GPU memory consumption, our method also demonstrates a clear advantage. Under the 512×352 evaluation setting, our method requires only 3.8 GB of GPU memory, making it deployable even on older GPUs with 4 GB memory, such as the GTX 1050 Ti. In contrast, CFLD and MCLD require 6.3 GB and 10.3 GB, respectively, corresponding to memory reductions of approximately 39.7% and 63.1%.

4.3 Ablation Study

To validate the contribution of each component, we conduct ablation studies on the DeepFashion dataset at a resolution of 512×352. The model configurations are defined as follows. B1 is the baseline, which employs only standard cross-attention(CA). B2 replaces part of the convolution blocks in B1 with conventional self-attention(SA). B3 replaces the SA in B2 with the proposed selective self-attention(SSA). B4 further replaces the CA in B3 with the proposed query-gated cross-attention(QGCA). Ours denotes the complete two-stage diffusion-refiner framework.

As shown in Table 2, conventional SA performs even worse than the baseline due to its redundant feature dependencies. Replacing SA with SSA substantially improves perceptual quality, reducing FID from 7.41 to 6.63, demonstrating that SSA effectively suppresses redundant feature dependencies. Further replacing CA with QGCA yields the best FID(6.57), verifying that QGCA enables

Table 2: Ablation study. Results are obtained by evaluating images in a resolution of 512×352.

Variants	SA	SSA	CA	QGCA	Refiner	FID	LPIPS	SSIM	PSNR	Param
B1			✓			6.61	0.1784	0.7465	17.81	841M
B2	✓		✓			7.41	0.2015	0.7369	16.96	747M
B3		✓	✓			6.63	0.1766	0.7481	17.82	749M
B4		✓		✓		**6.57**	0.1756	0.7480	17.86	767M
Ours		✓		✓	✓	**6.57**	**0.1704**	**0.7488**	**17.92**	772M



Figure 4: Result of different variants of our methods.

more effective selective appearance information propagation. Finally, introducing the refiner increases the model size by only 5M parameters while achieving a substantial reduction in LPIPS.

As shown in Fig. 4, conventional CA(B1) and SA(B2) fail to prevent texture confusion between the skin-colored jacket and the black inner garment. Although SSA(B3) largely resolves this issue, erroneous feature injection still remains around the arm region. By introducing the gating mechanism, QGCA(B4) effectively suppresses this artifact. Finally, with the addition of the refiner, the generated textures are further refined, producing more realistic appearance details.

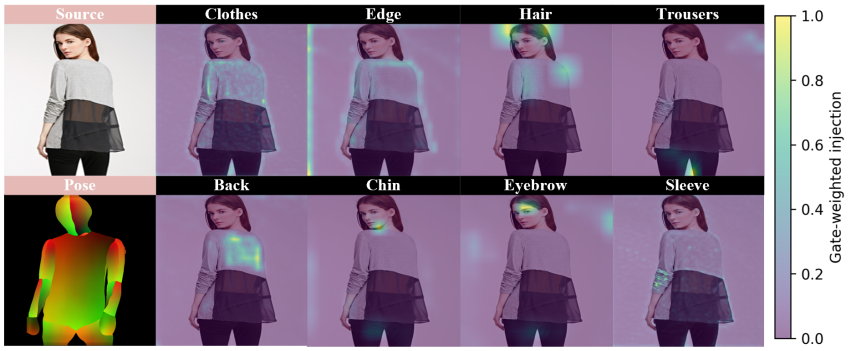


Figure 5: Feature visualization of query-gated cross-attention. Query-gates at different spatial locations selectively filter injected appearance information according to the current query features.

QGCA We visualize the query-gates from different layers of the network. As shown in Fig. 5, query-gates at different spatial locations attend to distinct semantic regions, including the back, hair, and et al. By selectively gating appearance feature injection according to the current query, QGCA suppresses erroneous feature injection at its source, thereby preventing texture confusion caused by semantically irrelevant information propagation.

SSA As illustrated in Fig. 6, SSA establishes attention primarily among semantically consistent tokens rather than densely connecting all tokens. Queries located on the hat, white stripes and et al. predominantly respond to tokens within the corresponding semantic regions, while background

queries exhibit limited activation on foreground structures. This demonstrates that SSA effectively suppresses redundant feature interactions and encourages localized semantic consistency.

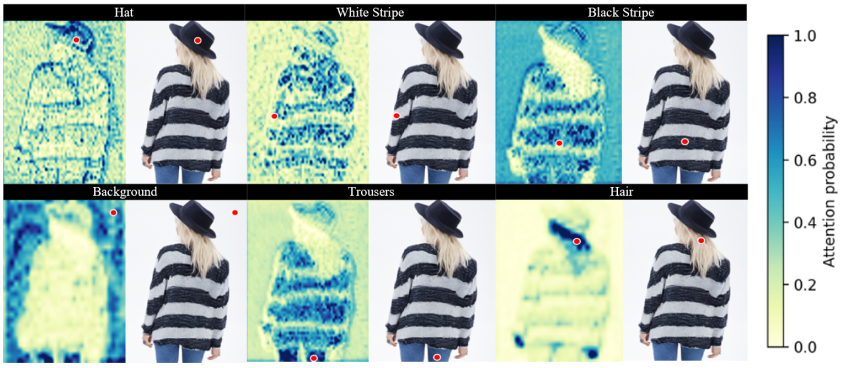


Figure 6: Feature visualization of selective self-attention. SSA only establishes associations between tokens that have similar semantic relationships.

Refiner The response maps of refiner are shown in Fig. 7. It can be observed that the refiner exhibits high responses around clothes texture, garment boundaries and regions containing high-frequency texture details, while its responses remain relatively weak in large homogeneous regions. These observations indicate that the refiner primarily fine-grained high-frequency details loss left by the diffusion stage, thereby complementing the preceding stage of selective information propagation.



Figure 7: Responses of the refiner. The refiner primarily focuses on regions containing high-frequency texture details like edges, clothes texture and et al.

5 Conclusion

In this paper, we revisit PGPIS from the perspective of information propagation. We argue that the primary limitation of existing diffusion-based methods lies not in insufficient appearance modeling, but in erroneous appearance propagation. To address this challenge, we propose a two-stage framework that decouples selective information propagation from information refinement. Specifically, the proposed selective self-attention learns cleaner appearance representations by suppressing redundant feature dependencies, while query-gated cross-attention selectively injects appearance features according to semantic relevance. Furthermore, the transformer refiner, combined with perceptual loss, corrects residual geometric errors and restores high-frequency texture details. Extensive experiments demonstrate that our method outperforms existing diffusion-based methods in perceptual generation quality while requiring a smaller model size and lower GPU memory consumption.

ACKNOWLEDGE

The author expresses his deep gratitude to his mentor and senior colleagues, who have been providing guidance and support for a long time.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance and code development, including improving the organization and clarity of the manuscript, refining language expression, and assisting with the implementation and analysis of experimental code. We have not used generative AI tools to generate or fabricate experimental results, research data, or evaluation results, and other required disclosure tasks are not applicable to this work. All AI-assisted text, ideas, code, and analyses were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including all text, claims, methods, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, human experiments, or the collection of personally identifiable information. We use publicly available benchmark datasets for research and evaluation, following their intended research purposes and applicable usage licenses. The proposed method is developed for academic research in pose-guided person image synthesis and does not target harmful or discriminatory applications. We have no conflicts of interest or undisclosed sponsorship related to this work.

REPRODUCIBILITY STATEMENT

We provide sufficient details to facilitate the reproduction of our results. The proposed framework and its implementation details are described in Sec. 3, while the training and evaluation settings are provided in Sec. 4.1. Detailed experimental configurations, additional implementation details, and further qualitative and quantitative results are provided in the supplementary materials. The source code, model configurations, and pretrained model information will also be released as supplementary materials to facilitate independent verification and reproduction of our results.

REFERENCES

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5968–5976. IEEE, 2023.

Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *Advances in neural information processing systems*, 31, 2018.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8857–8866, 2018.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22711–22720. IEEE, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3688–3697. IEEE, 2019.

Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16019–16028. IEEE, 2025.

Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5903–5912. IEEE, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 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.

Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6420–6429. IEEE, 2024.

Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. *arXiv preprint arXiv:2104.06650*, 2021.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7687–7696. IEEE, 2020.

Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13525–13534. IEEE, 2022.

Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3408–3416. IEEE, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. *arXiv preprint arXiv:2103.04023*, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A Appendix

You may include other additional sections here.



Figure 8: Results of appearance editing. Our method consistently performs localized appearance transfer across different clothing styles, including complex appearances such as printed, striped, and solid-color garments, while preserving the person's identity, human structure, and clear semantic boundaries in unedited regions.

A.1 Appearance Editing

Beyond pose-guided person image synthesis, our framework naturally supports controllable appearance editing during inference. Given a reference image, a donor appearance image, and a target DensePose, our method regenerates only the user-specified editing regions while strictly preserving the remaining identity characteristics, background, and unedited regions.

Let I_r denote the reference image, I_d the donor appearance image, P_t the target DensePose, and $M \in \{0,1\}^{H\times W}$ a binary editing mask, where $M = 1$ indicates the region to be edited. We first encode the donor image into the latent space to obtain its appearance representation:

$$z_d = \mathcal{E}(I_d) \tag{10}$$

The donor appearance representation is then injected into the unet as an appearance condition through the Reference Encoder and the proposed QGCA, while the target DensePose provides pose constraints via the pose mapper. To preserve the unedited regions throughout the diffusion process, we impose region-wise constraints on the latent representation at every sampling step:

$$z_t = M \odot \hat{z}_t + (1 - M) \odot {z_t}^{ref} \tag{11}$$

where z_t denotes the latent feature predicted at the current diffusion timestep, and ${z_t}^{ref}$ denotes the corresponding noisy latent representation of the reference image. Through this step-wise region constraint, only the masked region is allowed to be regenerated according to the donor appearance and the target pose, while the content outside the mask consistently preserves the information from the reference image. After diffusion sampling is completed, the proposed Transformer Refiner is further applied to improve local geometric consistency and restore high-frequency texture details.

As shown in Fig. 8, our method consistently performs localized appearance transfer across different clothing styles, including complex appearances such as printed, striped, and solid-color garments, while preserving the person's identity, human structure, and clear semantic boundaries in unedited regions. These results further demonstrate that the proposed selective information propagation mechanism learns transferable appearance representations rather than fixed pose correspondences.