PM-Jewelry: Personalized Multimodal Adaptation for Virtual Jewelry Try-On with Latent Diffusion

Yangfan He
University of Minnesota-Twin Cities
Minneapolis, USA
he000577@umn.edu

Yinghui Xia
AutoAgents.ai
Beijing, P.R.China
vix@autoagents.ai

Jinfeng Wei
Northeastern University
Shenyang, P.R.China
202119033@stu.neu.edu.cn

Yingxuan Li
AutoAgents.ai
Beijing, P.R.China
yl3932@nyu.edu

Tianyu Shi
University of Toronto
Toronto, Canada
tianyushi3@mail.mcgill.ca

Jingsong Yang *
AutoAgents.ai
Beijing, P.R.China
edward.yang@autoagents.ai

Abstract

Virtual jewelry try-on systems offer an innovative way for users to experience personalized, realistic jewelry interactions in an online setting. This paper introduces PM-Jewelry, a virtual try-on framework designed to leverage multimodal learning and personalized adaptation techniques for an enhanced user experience. By integrating data from multiple modalities—such as images, text descriptions, and user interaction data—the model generates lifelike simulations of jewelry on various users. Using a latent diffusion framework, PM-Jewelry captures the intricate details of different jewelry types (e.g., rings, earrings, necklaces), ensuring high precision in aspects like texture, shine, and fit. The model further incorporates personalized adaptation mechanisms, allowing users to tailor the virtual experience to their preferences. Extensive experiments demonstrate the system's ability to handle diverse jewelry types while preserving critical details, making PM-Jewelry a scalable and robust solution for virtual jewelry try-on. This work also explores challenges such as realistic rendering, jewelry alignment, and material texture simulation, offering insights into future developments in multimodal virtual try-on technologies.

1 Introduction

The rapid rise of e-commerce has revolutionized consumer interactions, particularly in the fashion and jewelry industries, where Virtual Try-On (VTO) systems now enable customers to visualize products without physical trials [1]. While virtual clothing try-ons have been widely researched, jewelry presents unique challenges due to its small, intricate, and highly reflective nature. This paper introduces PM-Jewelry, a novel framework leveraging multimodal learning and personalized adaptation to address these complexities [2]. By integrating images, text, and user preferences, PM-Jewelry generates high-fidelity simulations of various jewelry types—earrings, rings, necklaces, and bracelets—using a latent diffusion model to accurately render texture, color, shine, and fit.

*The corresponding author

The NeurIPS 2024 Workshop on Adaptive Foundation Models (AFM 2024), Vancouver, Canada, 2024.

The system also allows real-time personalization, enabling users to adjust fit and style, enhancing engagement. Despite the challenges of handling jewelry's reflective surfaces and fine details across diverse body types and lighting conditions, PM-Jewelry demonstrates scalability and flexibility, offering a high-quality virtual try-on experience supported by advanced multimodal learning and diffusion modeling.Our main contribution are:

- **Personalized Multimodal Adaptation**: We introduce a framework that combines image and latent diffusion models to adaptively fit jewelry on diverse user images. Using multimodal inputs such as the model image, salience map, and earring mask, we enhance personalization. Our dual-pathway network processes both image data (e.g., face and body for jewelry like necklaces or earrings) and latent embeddings that capture user-specific attributes like skin tone, facial structure, and personal style.
- **Attention-Based Latent Diffusion for 3D Jewelry from 2D Images**: Jewelry positioning and angles, such as earrings, vary with head movements. We incorporate a depth map for accurate alignment of jewelry to the user's body, ensuring realistic interactions, such as how a necklace drapes across the chest or how earrings hang and move.
- **Personalization and User Preferences**: We introduce a personalization mechanism that allows users to input style preferences (e.g., modern, traditional, minimalistic), tailoring the try-on experience to individual tastes. A feedback loop enables users to adjust features (e.g., necklace length or earring style), fine-tuning the system for improved interaction and engagement.
- **State-of-the-art Performance**: Our model delivers superior results in virtual jewelry try-on tasks, achieving high levels of realism in both visual quality and user satisfaction.

2 Related Work

With the rapid development of e-commerce, virtual try-on technology has become a key tool in the fashion industry, allowing users to preview the appearance of clothing and accessories without actually wearing them. The successful application of Generative Adversarial Networks (GANs)[3] in the field of image synthesis has inspired researchers to apply such models to virtual try-on[4, 5, 6, 7, 8, 9]. By learning how garments deform and fit on different body types, these methods are able to generate realistic fitting effects while maintaining the texture and style of the garment. To further enhance the realism of fitting images, the latest research trend has shifted towards utilizing diffusion models to generate more detailed and realistic images [10, 11, 12, 13, 14, 15]. For example, StableVITON [15] based on conditional generative modeling have achieved higher quality fitting effects through potential diffusion modeling techniques, while ensuring accurate rendering of garment details and image realism. Although a great deal of research has been conducted in the field of virtual clothing fitting, the complex details and diverse characteristics of jewelry present unique challenges. The PM-Jewelry framework presented in this paper is based on these state-of-the-art technologies to overcome the challenges in jewelry fitting by integrating multimodal learning and personalized adaptation mechanisms.

3 Proposed Framework for Virtual Jewelry Try-On

As illustrated in Figure 1, our approach, named **PM-Jewelry**, is inspired by garment-centric methodologies such as those introduced in the StableVITON framework [15]. The PM-Jewelry model incorporates several types of input data, including agnostic maps, dense pose, jewelry features, and depth maps, all of which work together to optimize the spatial arrangement and preservation of jewelry details. The objective function can be formalized as follows:

$$\mathcal{L}_{PMJ} = \mathbb{E}_{\zeta, I_h, M_j, D_j, S_h, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(\zeta, t, I_h, M_j, D_j, S_h)\|_2^2\right], \tag{1}$$

where we introduce the variable $\zeta = [z_t; I_h; M_j; D_j; S_h]$. Here, z_t represents the latent space, I_h refers to the head image, M_j corresponds to the jewelry mask (e.g., earrings, necklaces, rings) generated by SAM [16], D_j indicates the depth map obtained from DepthFM [17], and S_h denotes the salience map. By utilizing a zero cross-attention mechanism, similar to the method employed in STABLEVITON [15], jewelry features are seamlessly integrated with the multimodal inputs, while

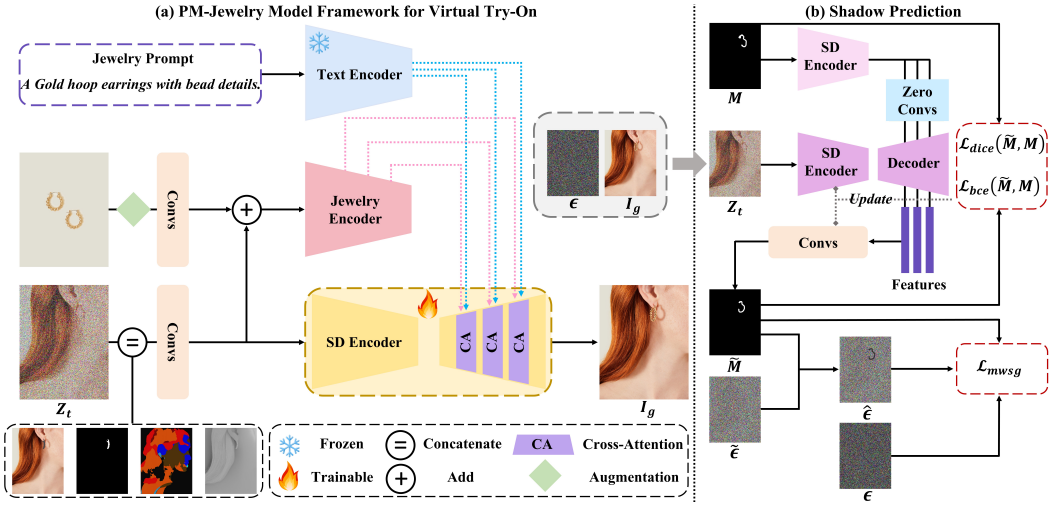

Figure 1: (a) The model integrates text and jewelry encoders with the SD encoder to generate realistic try-on images. Cross-attention ensures proper alignment and detailed rendering. (b) The shadow prediction module refines shadows using SD encoders and convolution layers, optimizing with dice loss and binary cross-entropy loss.

an ATV (Attention Total Variation) Loss is applied to reduce attention dispersion, thus improving placement accuracy:

$$L_{ATV} = \sum_{i,j} \|F_{ij} - G_{ij}\|_1 , \tag{2}$$

where F_{ij} denotes the attention-driven coordinate for the i, j-th point, and G_{ij} is the target coordinate representing the correct positioning of the jewelry feature. For further refinement of the PM-Jewelry model, the final loss function is a weighted combination of losses:

$$L_{\text{finetune}} = L_{\text{PMJ}} + \lambda_{ATV} L_{ATV}, \tag{3}$$

Here, the coefficient λ_{ATV} adjusts the influence of the ATV Loss. Additionally, the model incorporates Binary Cross-Entropy (BCE) Loss, Dice Loss, and a Shadow-Weighted Noise Loss (L_{mwsq}). The predicted shadow mask $\tilde{M}$ is progressively refined through U-Net's upsampling layers, leveraging modified VGG blocks [18], while a mask alignment loss ensures precise matching with the ground truth mask M:

$$L_{\text{mask}} = L_{\text{bce}}(\tilde{M}, M) + L_{\text{dice}}(\tilde{M}, M), \tag{4}$$

This loss ensures alignment between predicted and actual shadow regions. To control the shadow intensity, noise values $\tilde{\epsilon}$ are adjusted using channel-specific scaling factors s and biases b:

$$\hat{\epsilon} = (s \odot \tilde{\epsilon} + b) \odot \tilde{M} + \tilde{\epsilon} \odot (1 - \tilde{M}), \tag{5}$$

where $\tilde{M}$ controls the noise distribution within the shadow regions. The shadow-weighted noise loss is formulated as:

$$L_{\text{shadow}} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,1)} \left[\|W_{\text{fs}} \odot (\epsilon - \hat{\epsilon})\|_2^2 \right], \tag{6}$$

where W_{fs} represents the weights extracted from VGG extracted features related to the shadow foreground. The total shadow loss integrates both mask prediction and noise adjustment losses:

$$L_{\text{shadow}} = L_{\text{mask}} + \lambda L_{\text{shadow}}, \tag{7}$$

Finally, the PM-Jewelry model is trained using a comprehensive loss function that combines the main model loss L_{PMJ}, Attention Total Variation Loss L_{ATV}, shadow noise loss L_{shadow}, a depth consistency loss L_{depth} to maintain consistency between the generated and original images, and a perceptual loss L_{adv} based on minimizing the cosine similarity in VGG feature space. The overall loss is formulated as:

$$L_{\text{total}} = L_{\text{PMJ}} + \lambda_{ATV} L_{ATV} + \lambda_{\text{shadow}} L_{\text{shadow}} + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{perception}} L_{\text{perception}}, \tag{8}$$

where each component is scaled by its respective hyperparameter λ, allowing for balanced optimization of the various model aspects.

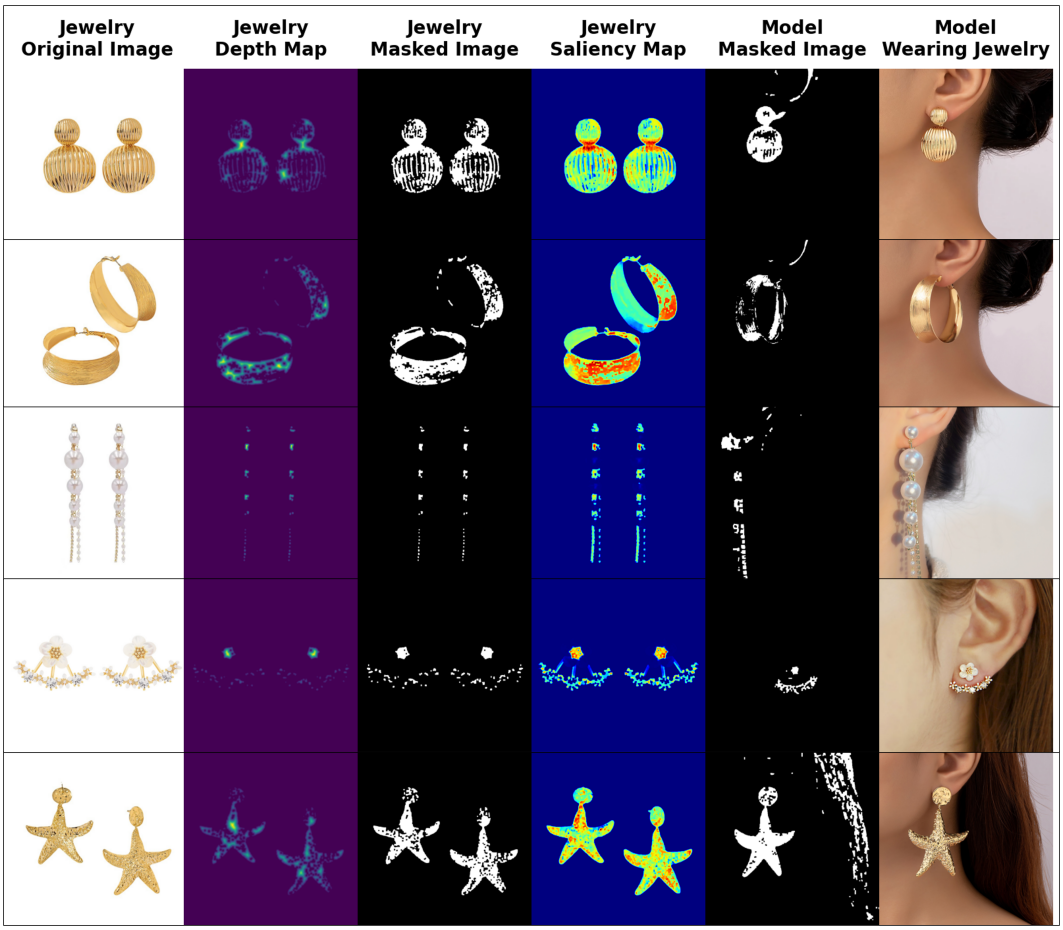

Figure 2: PM-Jewelry demonstrates superior alignment and detail preservation compared to baseline models, accurately aligning jewelry with the user's facial feature

4 Experimental Results

We evaluate the PM-Jewelry framework using SSIM [19], LPIPS [20], and FID (Fréchet Inception Distance). Our evaluation involves comparisons against several baselines as well as ablation studies. For this purpose, we introduce a novel jewelry try-on dataset consisting of 6,157 paired images, including 4,248 earrings, 3,332 necklaces, and 6,766 rings, showcasing diverse jewelry styles, wear effects, and adaptations on the human body.This dataset, split into 90% training and 10% testing, provides a comprehensive foundation for training diffusion models tailored to jewelry try-on tasks.

4.1 Quantitative Evaluation

Table 1 provides a comprehensive comparison of different models across Earrings, Necklaces, and Rings datasets using SSIM, LPIPS, and FID metrics. SSIM indicates structural similarity, LPIPS reflects perceptual distance, and FID measures the quality of generated images relative to real ones. Across all categories, our model achieves the highest SSIM values, indicating superior structural preservation, and the lowest LPIPS scores, showing minimal perceptual distortion. Additionally, our model consistently attains the lowest FID values, indicating a high degree of similarity to real data and producing visually coherent images. These results demonstrate that our model outperforms all others, particularly in preserving intricate jewelry details and achieving realistic rendering quality.

Figure 3: Ablation study on the model modules, highlighting the importance of data augmentation, zero cross-attention mechanism, jewelry encoder, and text encoder in integrating features and maintaining style consistency.

Table 1: Performance metrics (SSIM, LPIPS, FID) for various models across Earrings, Necklaces, and Rings datasets

Model	**Earrings**			**Necklaces**			**Rings**		
	SSIM (↑)	**LPIPS (↓)**	**FID (↓)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**
Stable Diffusion v1.5 [21]	0.7135	0.1278	24.732	0.7018	0.1351	25.823	0.7092	0.1234	25.578
ControlNet [22]	0.7123	0.1292	25.176	0.7068	0.1313	26.512	0.7074	0.1247	25.981
PICTURE [23]	0.7809	0.1223	19.057	0.7607	0.1268	19.984	0.7729	0.1195	19.674
Gal4way/TPD [24]	0.6752	0.1307	28.245	0.6546	0.1419	29.887	0.6629	0.1368	28.582
StableVITON [15]	0.6704	0.1015	30.298	0.6699	0.1116	30.834	0.6718	0.1073	29.621
ComfyUI [25]	0.7942	0.1214	29.769	0.7783	0.1227	30.165	0.7847	0.1192	30.456
Ladi-VTON [12]	0.7134	0.1273	24.612	0.7064	0.1298	26.055	0.7087	0.1246	25.735
GP-VTON [8]	0.7119	0.1264	25.411	0.7046	0.1322	26.213	0.7072	0.1238	25.619
DCI-VTON [10]	0.7295	0.1259	24.831	0.7081	0.1345	26.356	0.7084	0.1211	25.945
Ours	**0.8127**	**0.0972**	**13.529**	**0.9065**	**0.1003**	**9.805**	**0.8913**	**0.1094**	**10.854**

4.2 Qualitative Evaluation

In addition to the quantitative evaluation, we provide visual comparisons of the generated jewelry images. As illustrated in Figure 2, our model demonstrates its ability to produce visually realistic jewelry try-ons. Each intermediate stage, such as the depth map, masked image, and saliency map, contributes to the overall quality of the final rendered jewelry. Specifically, our approach effectively preserves intricate details like color, gloss, and texture while ensuring accurate placement and alignment with the user's facial features. The model's output images, shown in the "Model Wearing Jewelry" column, highlight its effectiveness in maintaining the realism of the jewelry's physical characteristics. For necklace and ring examples, please refer to Figure 4 and Figure 5 in the appendix for more detailed visual results.

4.3 Ablation Study

Ablation of Loss Design: Depth consistency, perception, Attention Total Variation (ATV), and shadow-weighted noise losses are crucial for precise jewelry alignment, perceptual similarity, accurate placement, and realistic shadows. Figure 6 in the appendix illustrates how removing these losses degrades visual quality, with results summarized in Table 2 in appendix A.1
Ablation of Modules: Data augmentation, zero cross-attention, the jewelry encoder, and the text encoder are essential for generalization, multimodal integration, detail capture, and style consistency. Figure 3 demonstrates the performance impact of removing any of these components, with results presented in Table 3 in appendix A.1

5 Conclusion and Future work

In this paper, we introduced the PM-Jewelry framework, a personalized multimodal adaptation system for virtual jewelry try-on, which effectively integrates latent diffusion techniques with user preferences to achieve realistic and tailored jewelry simulations. Our experimental results demonstrate significant improvements in both visual quality and user satisfaction across various jewelry types. In the future, we will focus on enhancing real-time interaction capabilities and expanding the framework's adaptability to a wider range of jewelry designs and body types. Additionally, we plan to incorporate advanced texture rendering techniques to further refine visual details and explore the integration of augmented reality features for an immersive virtual try-on experience.

References

[1] Tasin Islam, Alina Miron, Xiaohui Liu, and Yongmin Li. Deep learning in virtual try-on: A comprehensive survey. *IEEE Access*, 2024.

[2] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. Multimodal conditioned diffusion model for recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1733–1740, 2024.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[4] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022.

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021.

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

[7] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022.

[8] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23550–23559, 2023.

[9] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020.

[10] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607, 2023.

[11] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023.

[12] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8580–8589, 2023.

[13] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[15] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[17] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024.

[18] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019.

[19] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.

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

[21] 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.

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[23] Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. Picture: Photorealistic virtual try-on from unconstrained designs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6985, 2024.

[24] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024.

[25] ComfyUI. Comfyui: A powerful and modular stable-diffusion gui and backend. `https://github.com/comfyanonymous/ComfyUI`, 2024. Accessed: [Insert Access Date].

A Appendix / supplemental material

A.1 Ablation Study

Ablation of Loss design

Table 2 shows that the PM-Jewelry framework achieves the highest SSIM and lowest LPIPS and FID scores across all jewelry categories, with the ablation study confirming that omitting any key loss

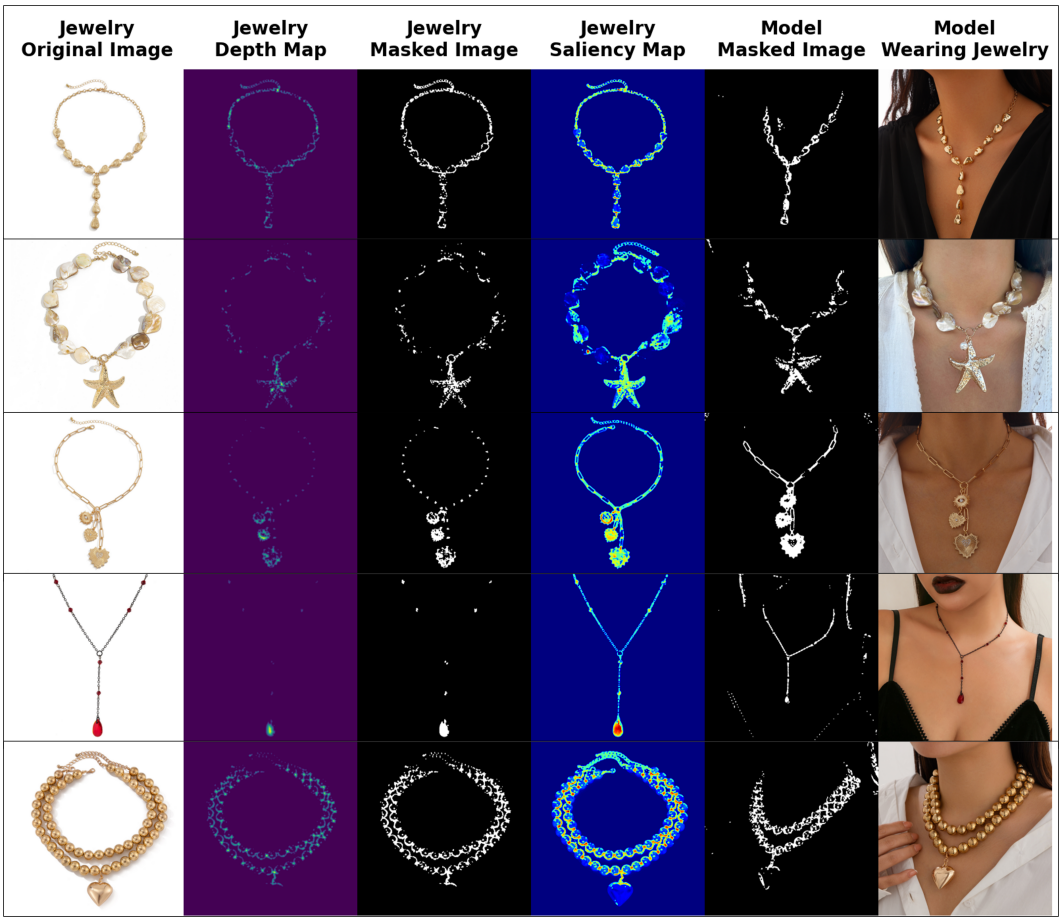

Figure 4: Qualitative comparison of PM-Jewelry with baseline models. PM-Jewelry preserves jewelry details and aligns with the user's face more accurately.

Table 2: Ablation study results for PM-Jewelry across different jewelry categories. The effect of including or excluding key loss components on performance metrics.

Model Variation	Earrings			Necklaces			Rings		
	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)
w/o Depth Consistency Loss	0.851	0.153	12.07	0.843	0.164	12.58	0.835	0.176	13.03
w/o Perception Loss	0.861	0.144	11.57	0.853	0.153	12.06	0.844	0.162	12.54
w/o ATV Loss	0.872	0.135	11.11	0.864	0.142	11.51	0.852	0.154	12.13
w/o L_{mwsq} Loss	0.883	0.126	10.53	0.873	0.133	11.07	0.862	0.145	11.54
w All Losses (Full Loss)	**0.912**	**0.098**	**9.75**	**0.905**	**0.102**	**10.21**	**0.894**	**0.113**	**11.12**

(Depth Consistency, Perception, ATV, or L_{mwsq}) consistently degrades performance, validating the integrated loss approach.

Ablation of Modules The ablation study demonstrates that removing augmentation, zero cross-attention, jewelry encoder, or text encoder results in degraded performance across all jewelry categories, highlighting the importance of each module in maintaining high SSIM, low LPIPS, and FID scores.

B Architecture and Training Details

B.1 Architecture Details

Our model builds upon the foundation of Stable Diffusion v1.5 [21] . The core of our system is the latent diffusion model, which consists of a denoising U-Net and a VAE-based autoencoder.

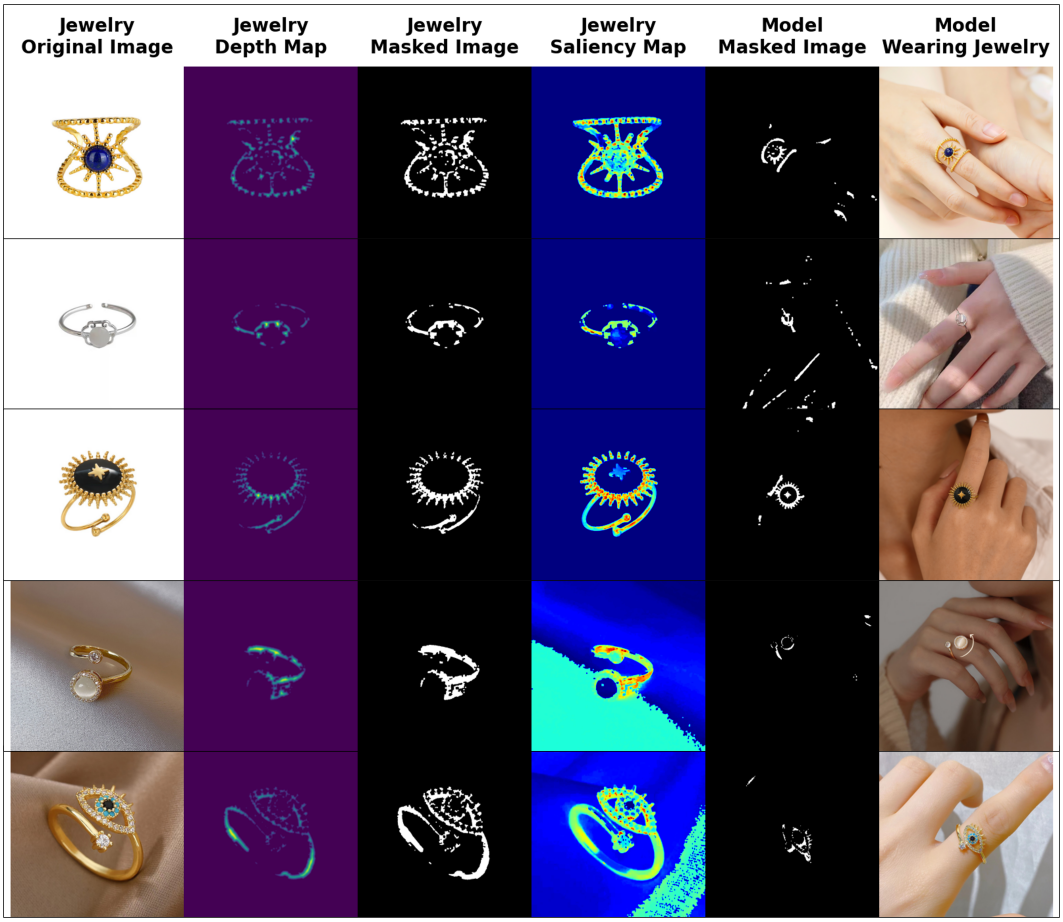

Figure 5: Qualitative comparison of PM-Jewelry with baseline models. PM-Jewelry preserves jewelry details and aligns with the user's face more accurately.

Table 3: Ablation study results for PM-Jewelry across different jewelry categories. The effect of augmentation, zero cross-attention, and encoder inclusion on performance metrics.

Variation	Earrings			Necklaces			Rings		
	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)
Full Model	0.884	0.123	10.53	0.872	0.132	11.07	0.863	0.147	11.37
w/o Aug	0.853	0.154	12.08	0.843	0.162	12.59	0.835	0.171	13.03
w/o Zero Cross-Attn	0.842	0.165	13.06	0.832	0.173	13.53	0.824	0.183	14.01
w/o Jewelry Enc	0.862	0.145	11.57	0.851	0.153	12.08	0.841	0.164	12.54
w/o Text Enc	0.879	0.136	11.02	0.862	0.144	11.51	0.857	0.155	12.12

The encoder and decoder of the U-Net feature 12 residual blocks, with three downsampling and upsampling stages, producing feature maps at multiple resolutions (8×6, 16×12, 32×24, and 64×48). We initialize the U-Net's weights using pretrained models from Paint-by-Example [24], and utilize a dual-branch pathway in the latent space for enhanced multimodal alignment. Specifically, the U-Net leverages cross-attention at all resolutions except 8×6, improving feature fusion between jewelry-specific attributes and input image features. The spatial encoder adopts a similar structure, with attention mechanisms fine-tuned for capturing fine details like jewelry positioning, alignment, and reflective textures. This enables the model to handle diverse jewelry types and variations across user body types and poses.

Figure 6: Ablation study on the loss design, illustrating the compromised rendering quality and jewelry alignment when any of the key losses—Depth Consistency, Perception, ATV, or L_{mwsq} are removed from the PM-Jewelry framework.

B.2 Training and Inference Details

We train our model using the AdamW optimizer with an initial learning rate of $1e^{-4}$, over 400k iterations, using a batch size of 24. The model is fine-tuned with a total variation weight (λ_{ATV}) of 0.001 for an additional 40K iterations, maintaining the same learning rate and batch size. Training is conducted on four NVIDIA A100 GPUs, taking approximately 120 hours.

For data augmentation, we apply a series of transformations to the input images, including horizontal flips (p=0.5), random shifts (limit=0.2, p=0.5), random scaling (limit=0.2, p=0.5), as well as contrast and HSV adjustments. These augmentations are crucial for enhancing the model's robustness in handling diverse lighting conditions and jewelry styles.

To address the challenge of jewelry alignment and facial distortion, we fine-tune the decoder separately using the VITON-HD [5] and DressCode [23] datasets, applying the AdamW optimizer with a learning rate of $5e^{-5}$ over 12k iterations. This step ensures that the final output preserves key jewelry attributes like shine, texture, and fit while maintaining realism in facial regions.

During inference, we utilize the pseudo linear multi-step (PLMS) sampler [21], running with 50 steps to ensure high-fidelity results with minimized noise and optimal detail preservation.

B.3 Augmentation and Regularization

Augmentation techniques, including contrast and HSV adjustments, were applied to both the clothing and the jewelry regions to maintain consistency in the training process. We also incorporated an attention-based loss function (λ_{ATV}) to reduce attention dispersion, particularly during the cross-attention process in jewelry-specific regions. A shadow-weighted noise loss was introduced to enhance the realistic rendering of shadows and reflective surfaces.

B.4 Inference Strategy

For inference, we adopt the PLMS sampler with a 50-step sampling strategy, similar to the approach used in Stable Diffusion [21]. This allows us to balance computational efficiency with high-quality output, enabling our model to generate detailed, lifelike jewelry try-on results efficiently.