

A. Related Work

Existing Image Editing Methods Recent advancements in fashion image editing [5, 27, 42, 54] have enabled transformative applications such as virtual try-on [8–10, 49], dynamic garment replacement [19, 52]. Diffusion-based approaches, exemplified by VITON-HD [7] (trained on 11,647 front-facing static poses), demonstrate high-fidelity garment synthesis but remain limited to controlled environments with simplified poses, hindering deployment in real-world scenarios. Predominant frameworks [1, 14, 25, 29] adopt a two-stage paradigm: anatomical mask generation followed by diffusion-driven [38] editing. However, mask generation strategies—whether coarse human parsers like SCHP [26] or garment-specific segmenters such as TexFit [43] (encoders and decoders trained with the fashion-specific dataset DeepFashion)—struggle to reconcile pixel-level anatomical accuracy with open-vocabulary editing flexibility.

Coarse-grained human masks, exemplified by IDM-VTON [8] (DensePose [4]-driven torso segmentation) and OOTDiffusion [49] (SCHP-based [26] full-body parsing), restrict edits to predefined zones (*e.g.*, *upper torso*) through rigid anatomical priors, preventing dynamic length customization (*e.g.*, *waist-to-hip transformations*). Recent methods like CatVTON [9] and Cat2VTON [10] expand editable regions via global attention mechanisms but retain lower-body constraints from ankle-length dress training data in VITON-HD [7]. Conversely, Fine-grained clothes masks (DPDEdit [44] with Grounded-SAM [24] architecture, FICE [33] via CLIP-guided grounding) achieve pixel-aligned boundaries but rigidly adhere to dataset-specific categories, making open-vocabulary style changes (*e.g.*, *"dress→pants"*) infeasible. Even state-of-the-art approaches like Leffa [56], which introduces flow-guided attention for lower-body edits, cannot modify garment lengths due to fixed mask topologies. General-purpose editors [12, 22, 28] (*e.g.*, Instructpix2pix [2], Null-text [31]) suffer from color bleeding and detail loss (*e.g.*, *distorted hands*) due to unconstrained attention maps. Emerging solutions like PromptDresser [23] explore adjustable masks for wrinkle editing but lack anatomical length control, underscoring a persistent gap: no existing method enables dynamic, anatomy-aware masking for open-world fashion editing.

Existing Detection/Segmentation Methods The quest for anatomically precise open-world editing reveals critical flaws in existing methods. Open-vocabulary detectors like T-Rex2 [20] (noun-centric) struggle to localize rare regions (*e.g.*, *chest*), while DINOv1.6 Pro [36] misclassifies ambiguous areas (*e.g.*, *shorts→waist in twisted poses*), and YOLO-World [6] erroneously edits non-target regions (*e.g.*, *jeans*). Segmentation tools like SAM [24] suffer uncontrolled over-segmentation via pixel-similarity expansion (*e.g.*, *facial edits during torso adjustments*), distorting identity. Human parsers (20-class of SCHP [26], DeepFashions [47]-trained garment constraints of TexFit [43]) lack flexibility, omitting regions like waist or under-segmenting anatomy.

To achieve cross-modal alignment [3, 18, 51, 53], diffusion-based [38] approaches (DiffSeg [40]) generate noisy, anatomy-agnostic edges (*e.g.*, *distorted hands*) via coarse pixel classification, prioritizing texture over structure. Foreground-background segmenters (DIS [35]/ U^2 -Net [34]) misclassify subtle boundaries (hip-length→background), retaining only misaligned clothing edges. Early attempts to extract anatomy-aware masks (*e.g.*, U^2 -Net [34]/DIS [35] for belly-length regions) collapsed under foreground-background dichotomies, while DiffSeg’s text-aligned masks via inversion introduced unstable attention noise. Subsequent open-vocabulary detectors (GLEE [45]/T-Rex2 [20]), trained on noun-centric tags, failed catastrophically on rare anatomical prompts (*e.g.*, *waist/belly*) and articulated poses due to unconstrained region proposals. Human parsers (M2FP [50]/AIParsing [55]) further highlighted field-wide rigidity, omitting critical regions (*e.g.*, *chest/belly*). Collectively, these efforts expose a persistent gap: no method achieves user-defined anatomical segmentation with pose robustness and in-the-wild generalization, demanding dynamic mask generation that fuses diffusion’s open-vocabulary capacity with biomechanical priors.

B. DeepSeek Anatomy Parser

Category	Included Tokens	Coverage Principle
Star Tokens	Neck, Shoulder, Elbow, Wrist, Hip, Knee, Ankle	Skeletal joints for pose calibration
Fleshy Tokens	Forehead, Chest, Waist, Belly, Arms, Hip, Hand, Thigh, etc	Soft-tissue regions for volume editing

Table 1. Star tokens (skeletal joints) and fleshy tokens (volumetric anatomy) classification rules.

The Pose-Star framework employs DeepSeek-V3-Base [11], which is based on the MoE architecture, as the anatomical instruction parser. This module interprets user-provided semantic commands (*e.g.*, *"belly-length blouse"*) into structured

Garment	Start Point	End Point	Coverage Zone
Blouse/Shirt	Neck/User-defined	User-defined	Upper body (arms included)
Dress	Neck/User-defined	User-defined	Full torso + legs
Pants/Skirt	Waist	User-defined	Lower body

Table 2. **Partial Instruction Mapping Protocol.** Length Anchor: Explicit endpoint (e.g., *belly* in "*belly-length*"), Implicit Start: Auto-derived from garment type.

Example: "belly-length blouse"

- **Start:** Neck (implicit from blouse)
- **End:** Belly (explicit length anchor)
- **Output:**

```
{
  "star_tokens": ["Neck", "Shoulder", "Elbow", "Wrist"],
  "fleshy_tokens": ["Chest", "Belly", "Arms", "Waist"],
  "coverage_zone": "Upper body (neck→belly)"
}
```

Figure 10. Instruction mapping protocol example and DeepSeek target output in .json form.

body representations that define garment coverage zones. As shown in Tab. 2, the parser first identifies explicit length anchors (e.g., "*belly*") and implicit coverage ranges derived from garment semantics: blouses/shirts default to neck-to-hip coverage including arms; dresses extend from neck to user-specified endpoints; pants/skirts initiate from waist landmarks. As shown in Tab. 1, key to this process is the decomposition into two complementary anatomical token sets: Star Tokens encode skeletal joints (Neck, Shoulder, Elbow, Wrist, Hip, Knee, Ankle) for pose calibration. Fleshy Region Tokens represent volumetric anatomy (Chest, Waist, Belly, Arms, Thighs, Shanks, Torso) for style editing. A example of an anatomy-aware instruction parsing and output results is shown in Fig. 10.

C. Attention Mechanisms in U-Net Architecture

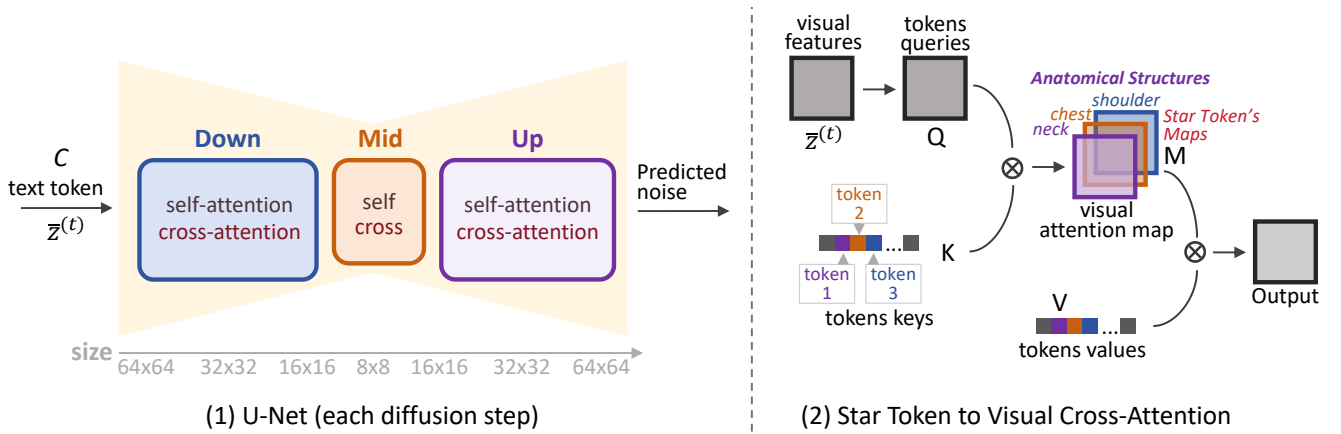


Figure 11. **Background and Preliminaries.** Diffusion-based token initialization.

To obtain the attention maps from the diffusion process of real images for inference in Pose-Star, we need to perform diffusion (i.e., inversion-reconstruction) on these real images. Specifically, we invert (add noise to) the real image into the initial latent space using DDIM [41] and then reconstruct (denoise) it back to the real image. This work primarily focuses on the reconstruction process, with the goal of acquiring attention maps that capture the text token-image mapping during this reconstruction. Additionally, due to the approximation inherent in diffusion models—where the U-Net cannot perfectly learn the theoretical reverse conditional distribution $q(z_{t-1} | z_t, z_0)$ and instead provides an approximate $p_\theta(z_{t-1} | z_t)$ —the

inversion-reconstruction process lacks perfect symmetry, leading to errors where the reconstruction struggles to fully restore the original image. To address this, we consider introducing null-text optimization [31]. Its core idea is to align the reconstruction process with the inversion process by optimizing the MSE loss between the latent vector \bar{z}_t from the reconstruction process and the corresponding latent vector z_t from the inversion process at the same diffusion timestep:

$$\min_{\emptyset_t} \|z_{t-1} - z_{t-1}(\bar{z}_t, \emptyset_t, C)\|_2^2, \quad \bar{z}_{t-1} = z_{t-1}(\bar{z}_t, \emptyset_t, C). \quad (6)$$

Here, \emptyset denotes the null text embedding, whose value is adjusted during optimization to minimize the loss. C represents the text condition input, corresponding to target tokens (Star Tokens/Fleshy Tokens/Clothes Tokens), and $z_{t-1}(\cdot)$ signifies the latent vector update for a single diffusion step. Through this process, we obtain the latent diffusion trajectory of the real input image, from which we then extract the attention maps corresponding to the target tokens.

The U-Net structure used for noise prediction at each diffusion step is illustrated in Fig. 11(1), with our focus on its attention layer. This layer comprises a downsampling (down) block, a middle (mid) connection block, and an upsampling (up) block. Each block contains cross-attention and self-attention maps at different spatial resolutions: the down and up blocks include resolutions of 256, 1024, and 4096, while the mid block has a resolution of 64. Higher resolutions typically store higher-dimensional feature information. Crucially, as demonstrated in Prompt-to-Prompt [15] and shown in Figure Fig. 11(2), the cross-attention layer primarily maps the text condition C to a set of visual attention maps M , where each map in M corresponds to a specific text token (e.g., 'neck'). This finding is essential to our method. Specifically, the deep spatial features of the noisy image \bar{z}_t are projected into the query matrix Q , while the text embeddings (including: Star Tokens/Fleshy Tokens/Clothes Tokens) are projected into the key matrix K and value matrix V . Learned linear projections then yield the visual attention map corresponding to each token within the text condition C :

$$M = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right). \quad (7)$$

Through this method, the text condition C is mapped to the output image z_0 , ensuring it aligns with the given conditional prompt. We observe that the intermediate attention map set M serves as a visual identifier for the spatial extent of each token, exhibiting a one-to-one correspondence with every text token. In essence, M functions similarly to a detection result capturing the regions associated with all text tokens. By extracting the corresponding attention map from M , we initialize the Maps for Star Tokens, Fleshy Tokens, and Clothes Tokens.

D. Post-Processing of Edge-Aware Selector.

To convert the refined edge map \hat{E} into a valid segmentation mask, we perform a series of post-processing operations including Edge Discontinuity Optimization, Edge-to-Mask Conversion, and Mask Edge Smoothing.

Edge Discontinuity Optimization We enforce spatial continuity in \hat{E} to establish well-defined boundaries. Although the initial Canny edges E suffer from fragmentation and missing segments due to illumination variations and interfering objects, we preserve relevant edges from R_k while addressing discontinuities. Specifically, we bridge discontinuous endpoints through direct linear interpolation, as empirical evaluations indicate this approach introduces boundary deviations within 5% of the total target area - an acceptable tolerance for practical applications. This optimization process ultimately yields a topologically continuous edge representation suitable for mask generation.

Edge-to-Mask Conversion To generate a binary mask \hat{M} from a closed curve represented in \hat{E} (where 1 indicates edge pixels and 0 indicates non-edge pixels), we propose an efficient boundary propagation algorithm that fills the interior region while preserving the curve topology. First, initialize \hat{M} by copying \hat{E} such that edge pixels retain value 1 and non-edge pixels are set to 0. Next, identify external regions by propagating from image boundaries: Enqueue all boundary-adjacent pixels (i, j) where $\hat{M}_{i,j} = 0$ (non-edge), temporarily marking them as external (value 2). Perform breadth-first search using 4-connectivity (up/down/left/right neighbors), iteratively marking connected non-edge pixels as external (2). Upon queue exhaustion, finalize \hat{M} by assigning 0 to external pixels (value 2), 1 to all remaining non-edge pixels (interior), and preserving original edge pixels (1). This approach robustly distinguishes interior/exterior regions in $O(HW)$ time while handling complex curve topologies through boundary-connected propagation, ensuring closed curves yield watertight masks.

Mask Edge Smoothing To address potential artifacts such as dark seams along mask boundaries that may arise from excessively precise segmentation during editing, we implement a Mask Edge Smoothing procedure to enhance coherence between edited and unedited regions. This process incorporates two complementary operations: First, we apply morphological dilation using a small circular kernel (radius=2 pixels) to slightly expand the mask boundary, ensuring sufficient coverage of transitional edge areas. Second, we employ Gaussian smoothing ($\sigma=1.5$) followed by curvature-constrained B-spline fitting to maintain geometric continuity while eliminating irregular jagged artifacts along the contour. These operations collectively preserve topological integrity while generating perceptually natural transitions, ultimately producing the optimized mask \hat{M} that robustly supports seamless content generation in practical editing scenarios.

E. Implementation

We systematically evaluate Pose-Star against state-of-the-art fashion-specific and general-purpose image editors: 1) Fashion-Specific Editors: IDM-VTON [8] and CatVTON [9] for virtual try-on based on coarse-grained human masks, Leffa [56] for full-body garment replacement using fixed anatomical priors, and TexFit [43] for text-driven editing with fine-grained clothes masks. 2) General-Purpose Editors: text-guided diffusion editors Null-text [31] and InstructPix2Pix [2] without masks, PowerPoint [57] + TexFit [43]/OOTDiffusion [49] for inpainting with clothes/human masks. To validate the effectiveness of Pose-Star’s plug-in to the existing editor, our framework is configured for virtual try-on: IDM-VTON [8]+Pose-Star, text-driven editing: PowerPoint [57]+Pose-Star. Pose-Star generates fine-grained human masks. All experiments utilize the stable-diffusion-v1-4 model¹ and are conducted on NVIDIA GeForce RTX 4090 GPUs with 24GB VRAM.



Figure 12. Representative Samples from the Challenging Real-World Evaluation Dataset.

As our approach is training-free, it eliminates the need for collecting large-scale, labor-intensive training datasets. To evaluate its practical utility in real-world, challenging scenarios, we curated 16,800 captured data samples from authentic online platforms and applications (Facebook, Xiaohongshu, YouTube frames). This dataset (as shown in Fig. 12) was meticulously selected to include challenging cases across multiple dimensions: hinged poses, dynamic snapshots, wide-angle overhead shots, layer occlusions, and diverse body types/ages. Additionally, our test suite incorporates 5,136 hinged pose samples filtered from DeepFashion-MultiModal. We will publicly release this challenging benchmark to better assess the limits of current model capabilities and enable targeted optimization.

F. Robustness

This section addresses **Q3** by analyzing the performance of Pose-Star under varying hyperparameters. We focus on three critical parameters: threshold $\beta \in (0, 1)$ for thresholded mask averaging (Sec.2.2), where higher β enforces stricter attention filtering; threshold $\alpha \in (0, 1)$ for cross-self attention merge (Sec.2.3), with larger α tightening boundary alignment; and selection range $\mu \in [0, 1]$ for the edge-aware selector, where l denotes the shortest distance from point $\mathcal{R}_k(m, n)$ to boundary of \mathcal{R}_k , and smaller μ imposes stricter edge constraints. Our suggested setting range is: $\beta \in (0.2, 0.6)$, $\alpha \in (0.3, 0.7)$, and $\mu \in (0, 0.2l)$, as shown in Fig. 13.

The analysis demonstrates that parameters within our recommended ranges effectively balance noise suppression and precision. Threshold β provides first-stage noise filtering during region aggregation, while stricter thresholds α enhance

¹<https://huggingface.co/CompVis/stable-diffusion-v1-4>

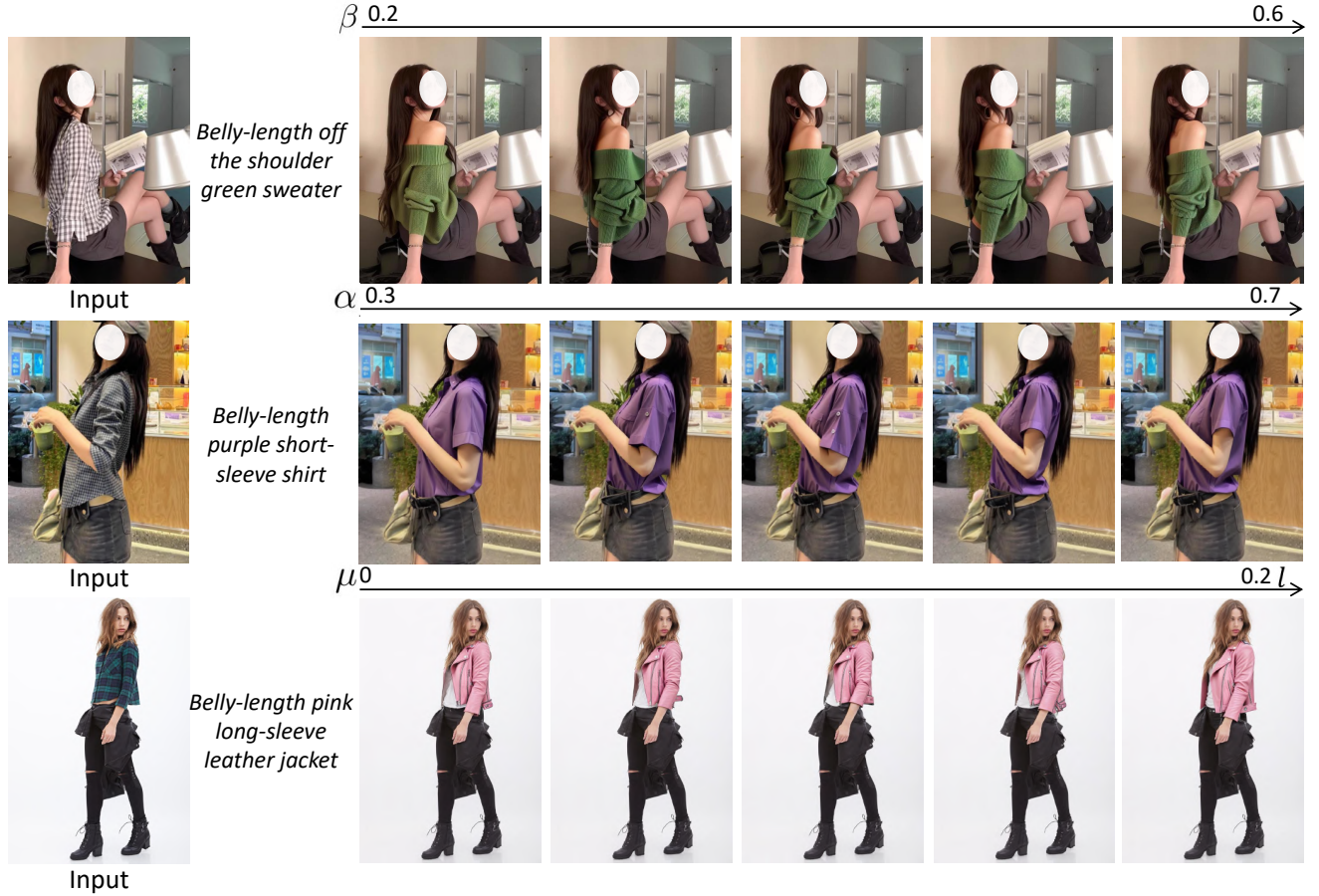


Figure 13. **Hyperparametric Evaluation.**

second-stage refinement, retaining only anatomically coherent regions. For edge optimization, tighter μ values—enabled by the localization and fusion module’s inherent precision—yield sharper boundaries aligned with anatomical contours. Collectively, Pose-Star exhibits stable performance across these ranges, demonstrating robustness against parameter variations in open-world editing scenarios.

We further evaluate the impact of two key parameters: the radius r for filtering attention noise around keypoints and the sliding window size for merging multi-stage token attention maps. For star-region constraints, r is set to the minimum, maximum, and average. The sliding window size is evaluated across four configurations: 1×1 , 2×2 , 3×3 , 4×4 . Performance is measured using mean Intersection over Union (IoU), with quantitative results summarized in Tab. 3.

Parameter Types	Radius r			Window Size			
Settings	$\min r$	$\text{ave } r$	$\max r$	1×1	2×2	3×3	4×4
Average IoU [37]	0.69	0.73	0.67	0.51	0.67	0.73	0.69

Table 3. Parametric evaluation of Radius r and Window Size.

Based on the evaluation results, we observe that extremely small or large radius r values lead to either loss of valid regions or excessive noise inclusion. Consequently, setting r to the average distance achieves an optimal trade-off in stability. For sliding window size, results align with findings in Sec.2.2: multi-stage attention map fusion (3×3) outperforms single-stage configurations (1×1 , 2×2) by better balancing effective region coverage, while oversized windows (4×4) induce performance degradation due to over-coarsened fusion.

Additionally, to evaluate our method’s dependency on the pre-trained OpenPose model, we assess samples containing

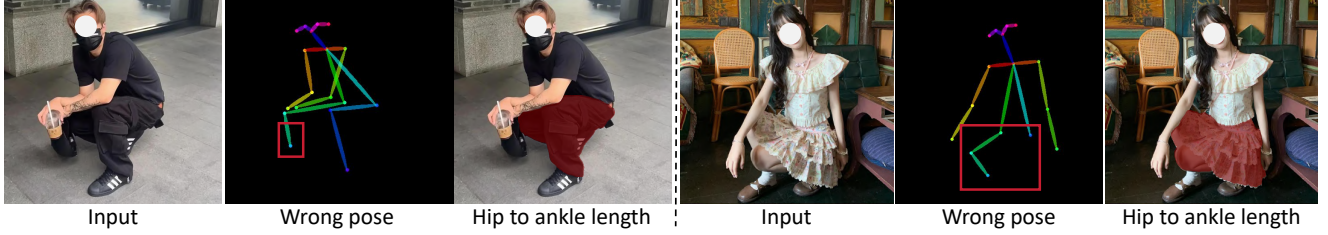


Figure 14. **Pose-Star Robustness Evaluation on OpenPose Partial Keypoints Errors.**

OpenPose keypoint detection errors. As illustrated in Fig. 14, OpenPose frequently exhibits local keypoint misalignment or loss when the human body is heavily occluded by clothing (*e.g.*, *skirts*) or in the presence of similarly colored adjacent regions (*e.g.*, *black hats and trousers*). Nevertheless, Pose-Star reliably generates anatomically consistent masks despite keypoint inaccuracies. This robustness stems from OpenPose keypoints solely filtering attention pixels rather than directly generating attentions (Sec.2.1), thus avoiding fundamental localization interference. Furthermore, clothing tokens mitigate the impact of erroneous keypoints, while calibration through aggregation and refinement modules ensures stable performance. Consequently, Pose-Star achieves resilience by synergizing existing components rather than relying on any single module. The framework’s modular design also supports alternative pose estimators, which we will explore in future work.

G. User Study Protocol

To comprehensively evaluate the performance of our proposed image editing method, we conducted a user study involving 200 participants. The participant pool was recruited to represent diverse perspectives and expertise levels relevant to image manipulation tasks, comprising 60 professional graphic designers (30%), 70 engineers with computer vision/ML experience (35%), and 70 general users without technical backgrounds (35%). This stratified sampling ensured balanced assessment across usability and technical robustness dimensions. Participants were presented with a randomized sequence of original and edited image pairs across varied scenarios and asked to evaluate results using a structured survey titled Image Editing Method User Evaluation Survey (as shown in Fig. 15).


The survey employed a 0-5 rating scale (0=Worst, 5=Best) across three critical dimensions: User-defined Region Flexibility assessed whether edits were confined strictly to user-specified areas (*e.g.*, “*hip-length*→*knee-length*” transformations), with ratings ranging from 5 for perfect localization to 0 for complete failure; Pose Robustness measured the naturalness of edits under challenging poses (*e.g.*, *extreme stances*), where 5 indicated flawless limb/joint preservation and 0 denoted severe disconnections; In-the-Wild Generalizability evaluated performance in real-world conditions (*e.g.*, *crowds*, *occlusions*, *night scenes*), with 5 representing seamless integration in cluttered backgrounds and 0 indicating catastrophic background/identity corruption. Display hardware and lighting conditions were standardized across all testing sessions to ensure consistent visual assessment.

H. Evaluation on Wild Images

Qualitative As illustrated in Fig. 16, we present additional evaluations across diverse scenarios: varying body types, multi-task editing, specific age groups, regional occlusions, wide-angle captures, and hinged poses. Existing methods, constrained by rigid masks, can only edit upper-body regions while failing to modify lower-body areas flexibly. Our approach not only enables precise editing of user-specified regions but also supports concurrent multi-task edits (*e.g.*, *simultaneous upper and lower garment replacement*). It faithfully preserves original body proportions across diverse physiques. In occlusion scenarios where non-editable foreground objects overlay target regions, our method retains foreground elements while plausibly editing underlying targets. Under extreme conditions—including acute hinged poses and wide-angle overhead shots—it robustly accomplishes virtual try-on tasks. These qualitative results demonstrate the robust and powerful editing capabilities of our method across diverse challenging scenarios.

Quantitative To further quantify the performance of our method, we employ ImageReward [48] as the evaluation metric. Editing results generated by different models on our collected test dataset are assessed, with mean ImageReward scores reported. As shown in Tab. 4, Pose-Star achieves superior human feedback ratings compared to fixed-mask baselines, attributable to its anatomy-aware masks that faithfully adhere to user-specific editing instructions.

Anatomy-Aware Instruction: Belly-length

Virtual Try-On Image :


User-defined Region Flexibility *(All ratings: 0=Worst, 5=Best)*

0.05.00.0

Does the edited image accurately modify ONLY your specified area?
5.0: Perfectly edits custom regions (e.g., "hip-length→knee-length")
3.0: Partially edits custom regions with minor errors
0.0: Fails to edit specified regions

Pose Robustness


0.05.00.0

Does the editing maintain natural appearance under complex poses?
5.0: Flawless for dancing/yoga/extreme poses
3.0: Minor distortions in limbs/joints
0.0: Severe body part disconnections

In-the-Wild Generalizability

0.05.00.0

Does it work in real scenarios (occlusions/cluttered backgrounds)?
5.0: Seamless in crowds/night/reflections
3.0: Minor background artifacts
0.0: Destroyed background/identity

Original Image:



Edited Image:


Figure 15. **Participant Evaluation Interface.** Participants rate edited images across three dimensions: User-defined Region Flexibility, Pose Robustness, In-the-Wild Generalizability. Scores range 0–5 (5=Best).

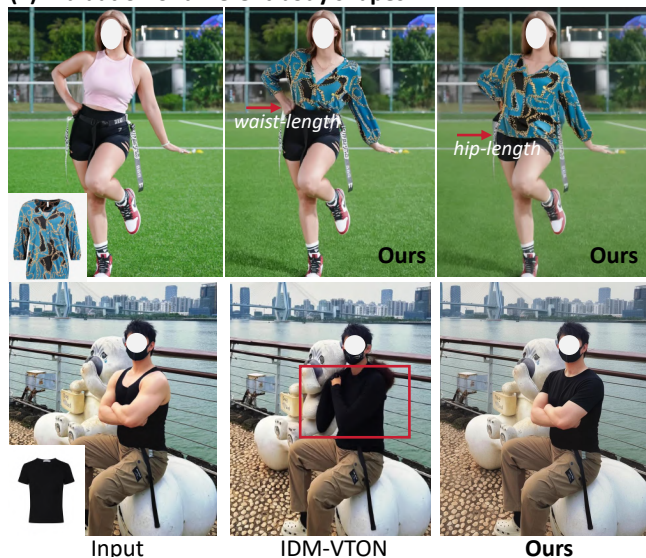
Methods	TexFit [43]	Leffa [56]	CatVTON [9]	IDM-VTON [8]	Ours
ImageReward [48]	1.17	1.35	1.19	1.16	1.61

Table 4. Quantitative evaluation of human feedback.

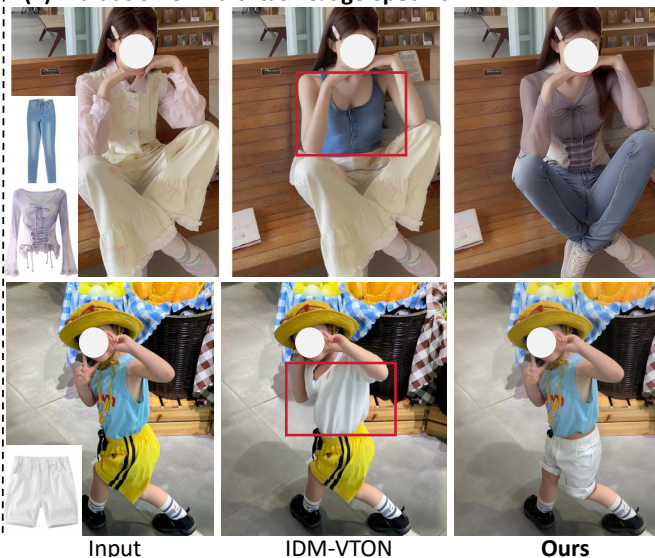
I. Discussion and Limitations

Pose-Star exhibits limitations on specific samples: when editing transparent occlusions (as shown in Fig. 17), such as cases where a human arm is partially obscured by glass and the occluded area falls within the target editing region, the method struggles to recognize body parts behind the glass and fails to generate the edited target with consistent occlusion. Similar challenges arise in scenarios like wire-mesh occlusion editing or veil occlusion editing. This limitation indicates that neither Pose-Star nor existing editing models possess layer-awareness, necessitating more advanced layer-perceptive editing capabilities. Another limitation involves cross-garment category edits (e.g., "hoodie→T-shirt") when the original facial region is occluded: such edits cannot preserve the subject’s original facial identity, instead producing randomly generated facial identities—an undesirable outcome. This specific issue remains unaddressed in prior works; we propose that future research could incorporate additional facial identity constraints to mitigate identity loss and random generation during editing.

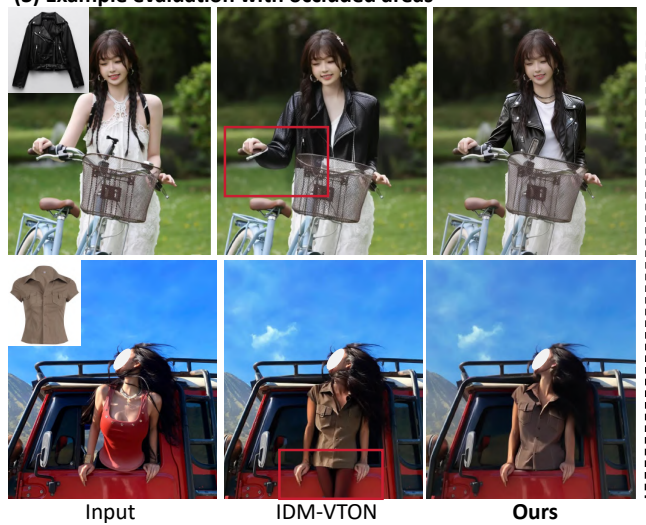
(1) Evaluation of different body shapes



(2) Evaluation of multi-task & age-specific



(3) Example evaluation with occluded areas



(4) Complex poses: wide-angle overhead shot & articulated pose

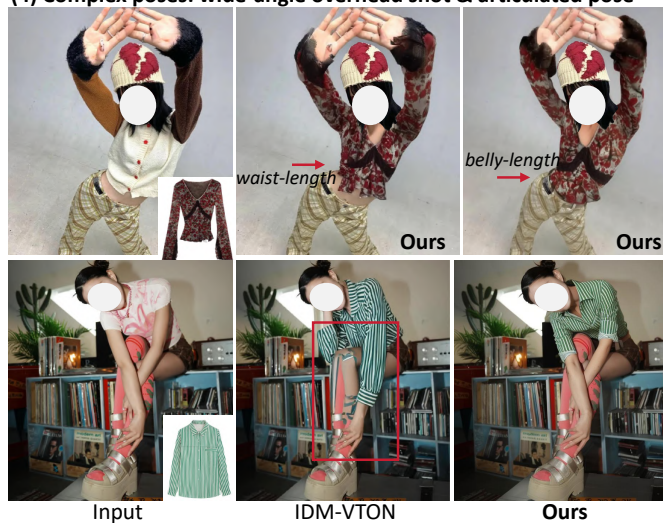
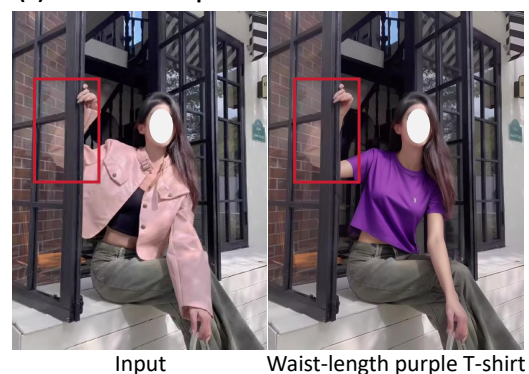


Figure 16. Multi-scenario Evaluation: Diverse Body Types, Multi-task, Age-specific, Partial Occlusion, Wide-angle Shots, Articulated Poses

(1) Localized transparent occlusion



(2) Unclear edges & Cross-category clothing translation



Figure 17. **Limitations.** Foreground occlusion scene and random face ID problem.