INSHUMAN: TOWARDS NATURAL AND IDENTITY-PRESERVING HUMAN INSERTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human insertion aims to naturally place specific individuals into a target background. Although existing image editing models may have such ability, they often produce failure cases, including inappropriate human pose in new background, inconsistent number of people, and modified facial identity. Moreover, publicly available human datasets often lack full-body portraits and realistic physical interaction between humans and their background. To address these challenges, we propose **InsHuman** for natural and identity-preserving human insertion. Specifically, we propose **Human-Background Adaptive Fusion (HBAF)**, which detects foreground humans to obtain a binary mask and applies region-aware weighting to align the human regions between predicted and ground-truth latents, ensuring the person's pose, count, and overall appearance are coherently adapted to the target background. We further propose **Face-to-Face ID-Preserving (FFIP)**, which detects and matches faces between the generated image and the source image in terms of face recognition features to enforce identity consistency for each face. In addition, we propose **Bidirectional Data Pairing (BDP)** strategy to construct *BDP-InsHuman*, a high-quality dataset with realistic human-background interactions. Experiments demonstrate that InsHuman achieves significant improvements in generating plausible images while keeping human identity unchanged.

1 INTRODUCTION

In recent years, image editing technology has made progress (Rombach et al., 2022; Peebles & Xie, 2023; Brooks et al., 2023; Hertz et al., 2023), and the image editing models (Labs, 2025; Xia et al., 2025; Cao et al., 2025; Wu et al., 2025b;a) are widely used in scenarios e.g., post-production for film and television, creative ad generation, virtual try-on (Choi et al., 2021), social media content creation, and digital marketing. By simply providing reference images and text prompts, users can perform various complex edits e.g., background replacement (Yang et al., 2023), style transfer (Tumanyan et al., 2023), object rearrangement (Chen et al., 2024; Song et al., 2023), and content reconstruction. Among these tasks, naturally inserting people into a target scene is becoming a practically valuable, yet technically challenging direction (Lu et al., 2023; Kulal et al., 2023; Gao et al., 2025; Zhang et al., 2026). Whether it is inserting persons into a travel landscape to create immersive souvenir photos, seamlessly inserting actors into film scenes for low-cost reshoots, or placing models in different environments for advertising design, these *Human Insertion* tasks require a reasonable spatial relationship between the person and the background, while preserving identity information e.g., the number of people and facial features (Ruiz et al., 2023; Gal et al., 2023). Essentially, it represents a comprehensive test of image editing models' capabilities in image understanding, spatial modeling, and identity preservation.

Directly applying existing image editing models to human insertion faces three core challenges. **First**, these models lack person-scene physical interaction priors (Zhu et al., 2022), often producing poses mismatched with the background, incorrect person counts, distorted limbs, or altered clothing. **Second**, these models exhibit a strong "repainting" tendency, generating visually plausible but identity-inconsistent faces; while recent identity-preserving methods (Ye et al., 2023; Wang et al., 2024; Li et al., 2024; Xiao et al., 2023) have made progress, balancing facial fidelity with natural scene integration remains unsolved. **Third**, existing portrait datasets (e.g., FFHQ (Karras et al., 2019), VGGFace2 (Cao et al., 2018), as well as generic human datasets (Liu et al., 2016; Shao et al., 2018))

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: **Qualitative comparison of different image editing models on human insertion task.** Existing image editing models often exhibit issues e.g., poses failing to adapt to the background, deviations in the number of people or appearance from the reference image, and loss of facial features. In contrast, models fine-tuned with our method (InsHuman) more stably maintain human structure and identity features, achieving more natural human insertion. **Red** boxes highlight the errors.

mostly consist of head or half-body shots with plain backgrounds, lacking realistic real person-scene interactions, making high-quality paired training data difficult to obtain.

Regarding these challenges, we address them in three aspects. **First**, to establish a reasonable interaction between the inserted human and the background, we **strengthen the model's attention to the region of the whole human body, guiding it to prioritize learning the pose and interactive relationship with the background**. We observe that global structural information is mainly learned during the early denoising stage of diffusion models and remains relatively stable thereafter. Motivated by this, we introduce a timestep-adaptive strategy that applies stronger guidance at early stages and gradually reduces its weight as the timestep decreases, following the learning dynamics of diffusion models. **Second**, to preserve human identity, we introduce a facial constraint that minimizes the distance between face features extracted from the source and predicted images. While identity-preserving objectives are not new, our timestep-adaptive training **enables the model to jointly learn human pose and retain facial identity. Third**, to address the lack of human insertion data, we construct a dataset containing hundreds of high-quality training pairs. Since only one of the source/background and ground-truth images can generally come from the real-world distribution while the other must be synthesized, we propose exchanging their roles, **allowing both sides to access real-world data**. This bidirectional role-exchange strategy differs from existing approaches and has been shown to be effective in practice.

In this work, we propose **InsHuman** for natural and identity-preserving human insertion. Overall, our main **contributions** are summarized as follows: 1) We propose ***Human-Background Adaptive Fusion (HBAF)***, which adaptively adjusts the pose, position, and size of people from the source image to fit the target background while keeping the number of people unchanged. 2) We propose ***Face-to-Face ID-Preserving (FFIP)***, a training strategy that encourages each person's face in the predicted image to maintain the same identity as in the source image. 3) We propose ***Bidirectional Data Pairing (BDP)***, which constructs training pairs through complementary forward and reverse paths, ensuring real-world images are present on both the source and ground-truth sides of the training

data. Extensive experiments show that InsHuman significantly outperforms existing image editing models in pose adaptation, human body integrity, quantity control, and identity preservation (Figure 1). Furthermore, our approach generalizes well to existing advanced image editing frameworks, reducing person count discrepancy by 57.14% and yielding consistent improvements in facial identity consistency and overall insertion quality, as shown in Table 4.

2 Related Work

Human Insertion. Existing image editing models (Labs, 2025; Wu et al., 2025a; Xia et al., 2025; Cao et al., 2025; Wu et al., 2025b) achieved significant progress in general image editing and layout control (Li et al., 2023; Avrahami et al., 2022; Couairon et al., 2023), but often struggle with human insertion in two key aspects: pose-scene mismatch (often stemming from a lack of explicit human spatial priors (Xu et al., 2024)) and loss of facial identity (Ruiz et al., 2023; Kumari et al., 2023). Research dedicated to this task remains limited. Kulal et al. (2023) learn scene affordances and person-scene interaction relationships to predict reasonable positions and poses, but lack high-fidelity facial feature control. Text2Place (Parihar et al., 2024) further leverages scene priors from text-to-image models to optimize semantic masks for plausible human placement. However, its affordance prediction and human generation are handled as separate stages and require test-time optimization for each scene. TelePortraits (Gao et al., 2025) improves identity preservation via attention manipulation, while Insert Anyone (Zhang et al., 2026) employs a dual-branch structure to separately model identity and appearance, but depends on additional modules and shows limited robustness. In contrast, our method jointly achieves pose-scene adaptation, structural integrity, person count control, and facial identity preservation within a unified training strategy.

Human-Centric Datasets. Full-body datasets (Liu et al., 2016; Fu et al., 2022; Jiang et al., 2022) generally use studio-like backgrounds, while natural-scene datasets (Lin et al., 2014; Shao et al., 2018) suffer from occlusions or limited image quality. Meanwhile, recent compositional editing benchmarks (Huang et al., 2023; Wang et al., 2023b) primarily evaluate general object relations, lacking a specific focus on high-fidelity person-scene physical interactions. HumanSD (Ju et al., 2023) improve structural diversity but still lack paired input-output images required for scene insertion. Our *BDP-InsHuman*, constructed via a bidirectional pipeline, provides realistic input-output distributions with high-fidelity person-scene interactions, offering more reliable training support.

3 Natural and Identity-Preserving Human Insertion

We focus on the human insertion task, which is challenging due to three key difficulties: keeping the overall structure of the human, preserving facial identity consistency, and the lack of high-quality paired training data. To address these, we propose **HBAF** for overall human structural coherence (Section 3.1), **FFIP** for facial identity alignment (Section 3.2), and **BDP** for high-quality data construction (Section 3.3). The overall framework is illustrated in Figure 2

3.1 Human-Background Adaptive Fusion (HBAF)

To address pose-scene mismatch, person count mismatches, and clothing inconsistencies with the reference image in human insertion, HBAF applies a region-aware weighted mask on the human area, guiding the model to prioritize structural coherence between inserted person and target background.

Given a ground-truth image $I_{gt} \in \mathbb{R}^{H\times W\times 3}$, we use YOLOv12-m (Tian et al., 2025) to detect and merge all foreground persons bounding boxes, generating a binary mask $M \in \{0,1\}^{H\times W}$, where pixels within the foreground person region are set to 1. As shown in Figure 4, the model establishes overall human structure at large timesteps and refines details at small timesteps. We therefore design a dynamic weight function λ_t that assigns a higher intensity of guidance to the human region at large timesteps and gradually relaxes it as the denoising progresses (Figure 3):

$$\lambda_t = \begin{cases} \lambda_{\max}, & \text{if } t > T_{start} \\ \lambda_{\min} + \dfrac{t - T_{end}}{T_{start} - T_{end}} \cdot (\lambda_{\max} - \lambda_{\min}), & \text{if } T_{end} < t \le T_{start} \\ \lambda_{\min}, & \text{if } t \le T_{end} \end{cases} \tag{1}$$

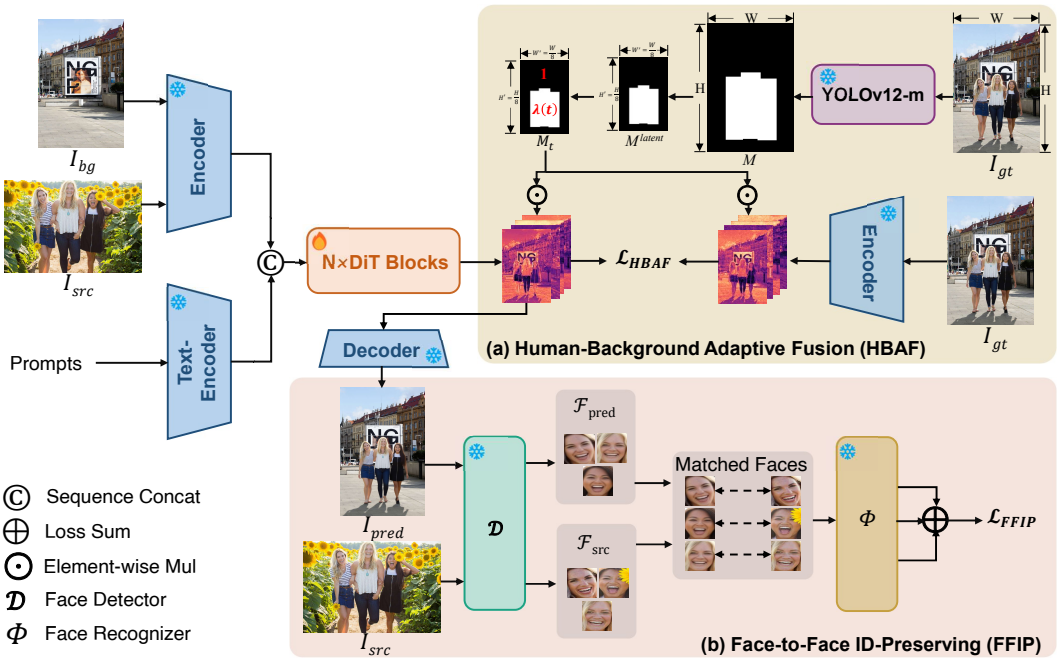


Figure 2: **Overview of our proposed InsHuman.** (a) Human-Background Adaptive Fusion (HBAF): Utilizing the human region mask to dynamically adjust the weight of the human region. (b) Face-to-Face ID-Preserving (FFIP): Activated when $t \leq T_{end}$, it employs a pretrained facial feature extraction network and matching algorithm to perform facial feature alignment without interfering with the overall person structure, thereby preserving person's facial identity.

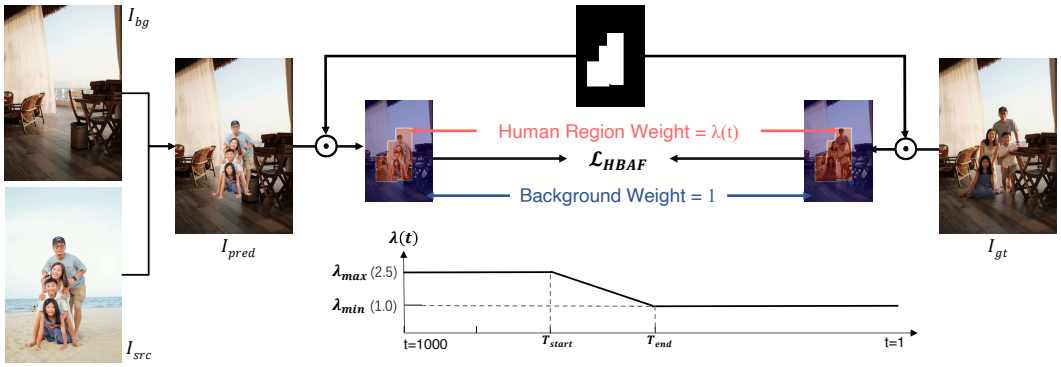


Figure 3: **Weighted supervision for human region.** This mechanism uses a human region mask to guide the model in prioritizing the optimization of human structure and overall appearance.

M^{latent} is a binary human region mask in the latent space, obtained by downsampling M to the latent resolution. We then construct a timestep-adaptive mask M_t by scaling M^{latent} with λ_t:

$$M_t = \mathbf{1} + \big(\lambda_t - 1\big) \cdot M^{latent} \tag{2}$$

where $\mathbf{1}$ is an all-one matrix. M_t assigns the weight $\lambda_t > 1$ to the human region and the weight 1 to the background, amplifying the supervision signal in the human area. The HBAF training loss is:

$$\mathcal{L}_{\text{HBAF}} = \frac{1}{N} \sum_{i,j} \| M_t[i,j] \cdot (\epsilon_\theta(z_t, t, c)[:, i, j] - \epsilon[:, i, j]) \|^2 \tag{3}$$

$\epsilon_\theta(z_t, t, c)$ denotes the noise predicted by the model and ϵ is the ground-truth noise. $[:, i, j]$ indexes the tensor at spatial position (i, j) across all channels. $M_t[i, j]$ is the scalar weight assigned to spatial position (i, j) by M_t. N is the total number of spatial elements in the latent feature map. With this design, the model can better achieve adaptation between the person and the background during

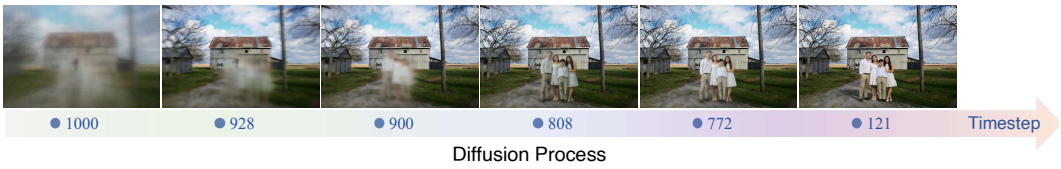


Figure 4: **Variation of model-predicted images over timesteps.** In the early stages of denoising, the model primarily determines the number of people and the overall composition. In the later stages of denoising, it mainly refines details and textures.

human insertion, while ensuring that both the number of persons and the overall appearance remain consistent with the input reference image (see Figure 7).

3.2 Face-to-Face ID-Preserving (FFIP)

Although HBAF effectively stabilizes the number of humans and the overall structure, the model still has facial distortions, resulting in inconsistencies of the facial ID between the generated image and the source image. To mitigate this, we propose Face-to-Face ID-Preserving (FFIP), which explicitly matching each face in the predicted image to its counterpart in the source image and minimizing the distance between their identity features.

We first utilize a pretrained face detection model (Deng et al., 2019) to detect faces in both the currently predicted image I_{pred} and the reference image I_{src}, obtaining two sets of facial regions: $\mathcal{F}_{\text{pred}} = \{f_i^{\text{pred}}\}_{i=1}^{N_p}$, $\mathcal{F}_{\text{src}} = \{f_j^{\text{src}}\}_{j=1}^{N_s}$, where N_p and N_s denote the face counts in I_{pred} and I_{src}, respectively. Each f contains the corresponding facial bounding box information. To establish one-to-one faces correspondences, we compute a face similarity matrix $S_{ij} = \langle \mathbf{e}_i^{\text{pred}}, \mathbf{e}_j^{\text{src}} \rangle$, where $\mathbf{e}$ is the normalized feature vector from the face recognition network and $\langle \cdot, \cdot \rangle$ denotes cosine similarity. We then apply the Hungarian algorithm to obtain the optimal face pairs $\mathcal{M}$:

$$\mathcal{M} = \{(i,j) \mid j = \pi^*(i)\}, \pi^* = \arg\max_{\pi} \sum_{i=1}^{N_p} S_{i,\pi(i)} \tag{4}$$

where $\pi(i)$ denotes the index of the face in I_{src} that matches the i-th face in I_{pred}. For each matched pair, we feed the cropped facial regions into a pretrained face recognition network $\Phi(\cdot)$ with fixed parameters to extract identity features. Then, we obtain: $\mathbf{e}_i^{\text{pred}} = \Phi(f_i^{\text{pred}})$, $\mathbf{e}_j^{\text{src}} = \Phi(f_j^{\text{src}})$. The face alignment loss is then defined as:

$$\mathcal{L}_{FFIP} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} (1 - \langle \mathbf{e}_i^{\text{pred}}, \mathbf{e}_j^{\text{src}} \rangle) \tag{5}$$

The total loss is defined as $\mathcal{L}_{total} = \mathcal{L}_{HBAF} + \lambda_{face}\mathcal{L}_{FFIP}$, where λ_{face} is a balancing coefficient that keeps the two loss terms at a similar order of magnitude. This design guides the model to preserve facial identity consistent with the source image (Figure 8).

3.3 Bidirectional Data Pairing (BDP)

Most existing human portrait datasets consist of the upper body or heads, often with blurred or even unrecognizable background information, making it difficult for the model to learn reasonable human pose and position information during training. To address this, we propose Bidirectional Data Pairing (BDP) and construct *BDP-InsHuman*, a high-quality human-scene interaction dataset where each training sample consists of a source image I_{src}, a target background I_{bg}, and a ground-truth I_{gt}.

During training, the model takes (I_{src}, I_{bg}) as input and generates an output image under fixed text prompts (provided in Appendix F). The objective is to make the generated image as close as possible to I_{gt} in terms of human pose, scale, and spatial relationship with the environment. As shown in Figure 5 (a), in the forward pairing path, we use real-world human images I_{real} (collected from Open-source image library, discussed in supplementary) as I_{src} and web-collected scene images I_{bg}^{web} as I_{bg}. High-quality composite results I_{syn}, generated by Nano Banana Pro, are selected as I_{gt} under strict quality criteria covering pose plausibility, lighting consistency, and perspective correctness.

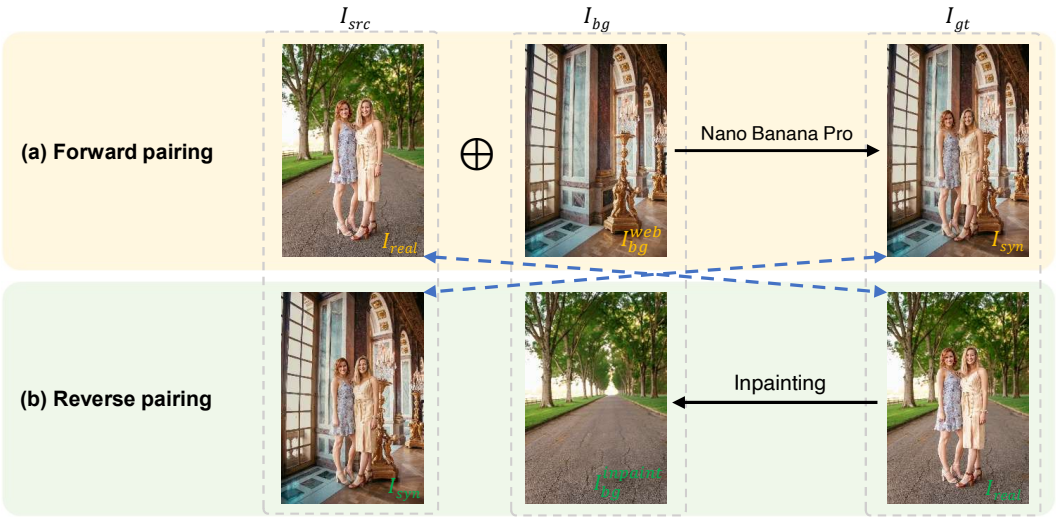


Figure 5: **The construction process of *BDP-InsHuman*.** Through (a) forward pairing and (b) reverse pairing, real-world images can serve as both ground truth and input to the model.

However, relying solely on generated images as ground truth risks introducing generative style biases into the model, causing it to learn synthesis artifacts rather than true physical priors. To address this, we further design a reverse pairing path, as shown in Figure 5 (b). We directly select high-quality real-world human-scene images as I_{gt}, and remove the person via inpainting (Qwen-Image-Edit-2509 (Wu et al., 2025a)) to obtain $I_{bg}^{inpaint}$. The synthetic images from the forward path are then used as I_{src}, forming the corresponding training pair. This ensures that both I_{src} and I_{gt} are grounded in real-world data, enabling the model to learn more natural human insertion.

3.4 Advantages and Differences from Existing Methods

InsHuman differs from existing methods in three aspects. **First**, for region supervision, existing methods apply fixed weights throughout training (Tu et al., 2025), without distinguishing structural learning from detail refinement. In contrast, HBAF uses a dynamic weight function λ_t, concentrating supervision at large timesteps when global human structure is learned and gradually relaxing it at smaller timesteps. **Second**, for identity preservation, existing methods (Ye et al., 2023; Wang et al., 2024; Li et al., 2024) inject reference identity features during generation, requiring additional adapters or inference-time modules. FFIP instead achieves identity consistency through training-time supervision by matching corresponding faces with the Hungarian algorithm and minimizing their identity-feature distances, without additional inference modules. **Third**, for data construction, existing methods mainly use generative outputs as ground truth, which may introduce synthesis biases. BDP addresses this by exchanging the roles of source and ground-truth images, allowing both to be grounded in real-world data and enabling more natural human insertion.

4 Experiment

4.1 Implementation Details and Metrics

Implementation details. We train our model on BDP-InsHuman (Section 3.3), which consists of 529 high-quality data pairs covering diverse identities and real-world scene backgrounds, with each identity–pose combination appearing at most twice. Using the proposed *HBAF* and *FFIP* strategies, we perform LoRA fine-tuning on Qwen-Image-Edit-2509 (Wu et al., 2025a). Detailed training hyperparameters and inference configurations are provided in Appendix A.

Evaluation and Metrics. We construct a dedicated test set of 131 image pairs, each consisting of a reference image containing humans and a target background image. We adopt established evaluation metrics from prior work to assess different aspects of human insertion. Following MultiHuman-Testbench (Borse et al., 2026), we use **IDS to measure facial identity consistency**. Following

Table 1: **Quantitative comparisons among InsHuman and other image editing models.** InsHuman achieved best performance in the following metrics.

Method	IDS↑	LPIPS↓	CA↓	CLIP-IQA↑	BRISQUE↓	UNIQUE↑	VLM-T↑
DreamOmni2	0.26	0.5834	71.00	0.6200	19.30	1.7205	4.7405
HunyuanImage-3.0-instruct	0.21	0.2503	74.05	0.6038	11.39	1.6541	6.7481
OmniGen2	0.28	0.4786	68.70	0.5093	27.10	1.5245	6.4733
Insert-Anything	0.28	0.2630	90.08	0.4010	25.26	0.7376	6.7328
Nano-Banana-Pro	0.54	0.2644	87.02	0.5196	10.47	1.0436	7.2290
Qwen-Image-Edit-2509	0.50	0.4540	70.99	0.6173	10.14	1.6453	5.9160
InsHuman (Ours)	**0.55**	**0.2220**	**96.18**	**0.6207**	**9.50**	**1.7712**	**7.8626**

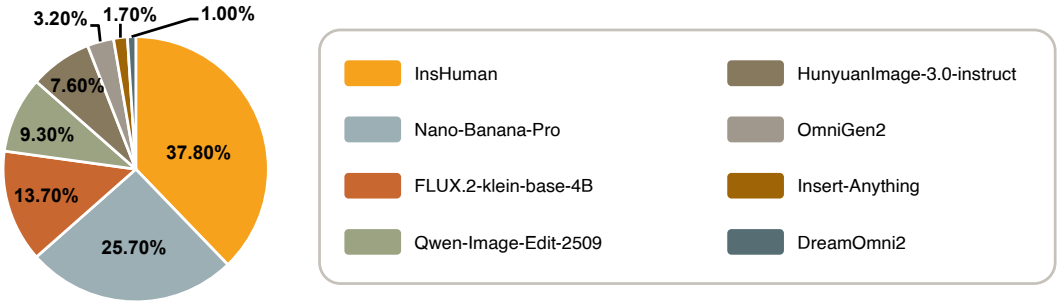


Figure 6: **Human preference distribution from a blind best-pick study with 10 annotators on our test set.** InsHuman achieves the highest preference results compared to all the considered methods

.

Text2Place (Parihar et al., 2024), we use **LPIPS to evaluate background preservation**. Following TISE (Dinh et al., 2022), we adopt **CA (Counting Alignment) to measure whether the generated image preserves the person count of the reference image**. We further use **CLIP-IQA** (Wang et al., 2023a), **BRISQUE** (Mittal et al., 2012), and **UNIQUE** (Temel et al., 2016) **for perceptual quality**, and follow TelePortraits (Gao et al., 2025) to employ **VLM-T for evaluating overall human insertion quality**. Detailed metric definitions and evaluation protocols are provided in Appendix B.

4.2 Comparison with State-of-the-Arts

Quantitative Results. InsHuman achieves the best performance across all seven metrics. Notably, our method outperforms all compared approaches on IDS (0.55), LPIPS (0.2220), CA (96.18), three IQA metrics and VLM-T (7.8626), including both open-source models and the closed-source commercial model Nano-Banana-Pro. The human preference study (Figure 6) further validates our results, with InsHuman obtaining the highest preference rate of 37.80%.

Qualitative Results. As shown in Figure 1, when I_{src} contains multiple people or complex poses, other models often suffer from missing persons or severe limb distortions. Comparative models also struggle to adapt the person's pose, position, and size to the target background, making the inserted person appear unnatural. Additional qualitative results can be found in the Appendix E.

4.3 Ablation Study

Impact of HBAF and FFIP. As shown in Table 2, combining both HBAF and FFIP achieves the best performance. Removing HBAF leads to a significant increase in CA, accompanied by spatial mismatches and incorrect person counts (Figure 7). Without FFIP, facial features deviate noticeably from the reference image (Figure 8), confirming that both components are necessary.

Impact of Bidirectional Construction in BDP-InsHuman. As shown in Figure 9, Forward-only results lack real-world physical logic between the person and background, while Reverse-only results show unnatural person size and position. In contrast, BDP achieves more natural human insertion.

Impact of $\lambda_{\max}$ and λ_t in HBAF. As shown in Table 3, setting $\lambda_{\max}$ at 5 causes severe structural distortions (CA=0.2824), with visual evidence provided in Appendix C.1 (Figure 11). Fixing λ to 2.5 without the dynamic schedule disrupts facial feature generation (as illustrated in Figure 12

Table 2: **Ablation study on HBAF and FFIP.** Their combination achieves the best.

HBAF	FFIP	CA↓	VLM-T↑
✗	✓	0.1985	7.7405
✓	✗	0.1445	7.7926
✗	✗	0.2137	7.2115
✓	✓	**0.1298**	**7.8626**

Table 3: **Ablation study on $\lambda_{\max}$ and λ_t in HBAF.** Our choice of $\lambda_{\max} = 2.5$ with dynamic λ_t achieves the best performance on both FR and IDS.

Method	CA↓	VLM-T↑
$\lambda_{\max} = 5$, w/ λ_t	0.2824	7.7099
Fix $\lambda = 2.5$, w/o λ_t	0.1384	7.8550
$\lambda_{\max} = 2.5$, w/ λ_t (Ours)	**0.1298**	**7.8626**



Figure 7: **Visual comparison without HBAF.** Removing HBAF causes the character's pose does not match the target environment and limb distortions.



Figure 8: **Visual comparison after removing FFIP.** Rremoving FFIP, facial features differ greatly from the reference image.



Figure 9: **Ablation study results of the bidirectional construction strategy.** A comparison among forward-only construction, reverse-only construction, and bidirectional construction validates the superiority of bidirectional construction strategy.



Figure 10: **Visual comparison of FLUX.2-klein-base-4B (FLUX.2) before and after applying InsHuman.** After fine-tuning with InsHuman, the model can adaptively adjust the person's position and scale, while keeping the number of people consistent with the reference image.

Table 4: **Generalization of InsHuman across different image editing models.** It demonstrats that the proposed approach generalizes effectively beyond its original model and remains effective across different image editing models.

Method	IDS↑	CA↓	CLIP-IQA↑	VLM-T↑
Qwen-Image-Edit-2509	0.50	1.0229	0.6173	5.9160
Qwen-Image-Edit-2509 + Ours	**0.55**	**0.1298**	**0.6207**	**7.8626**
FLUX.2-klein-base-4B	0.61	0.3740	0.4717	7.3435
FLUX.2-klein-base-4B + Ours	**0.63**	**0.1603**	**0.5897**	**7.6947**

in Appendix C.2), which confirms that both the weight magnitude and the dynamic schedule are necessary for balanced performance.

Impact of HBAF mask design. We compare our bounding box-based mask with a finer SAM-segmented mask within the same bounding boxes. As shown in Figure 13 (in Appendix C.3), the box-based mask produces more natural transitions between the inserted person and the background, since the broader supervision region encourages the model to learn perceptually important transition areas such as contact boundaries, cast shadows, and local illumination changes.

Cross Model Generalization of InsHuman. To validate the generalization of InsHuman, we apply InsHuman to FLUX.2-klein-base-4B under the same settings. As shown in Table 4, consistent improvements are observed across all metrics. As illustrated in Figure 10, the fine-tuned model adapts well to diverse scenes, demonstrating that our approach is not limited to a single base model.

5 Conclusion

We present InsHuman, a framework for natural and identity-preserving human insertion, comprising three components: HBAF, which applies a region-aware adaptive mask to ensure structural coherence between the inserted person and target background; FFIP, which enforces facial identity consistency by minimizing recognition-feature distances between matched face pairs; and BDP, which constructs *BDP-InsHuman*, a high-quality dataset with realistic human-background interactions. Experiments demonstrate that InsHuman significantly reduces person count errors and achieves consistent improvements in pose adaptation, structural integrity, and facial identity preservation.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and editing of the manuscript. We used generative model for dataset construction and the evaluation of VLM-T metric. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Shubhankar Borse, Seokeon Choi, Sunghyun Park, Jeongho Kim, Shreya Kadambi, Risheek Garrepalli, Sungrack Yun, Durga Malladi, and Fatih Porikli. Multihuman-testbench: Benchmarking image generation for multiple humans. *Advances in Neural Information Processing Systems*, 38, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6418–6427, 2024.

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*, pp. 14131–14140, 2021.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Tan M Dinh, Rang Nguyen, and Binh-Son Hua. Tise: Bag of metrics for text-to-image synthesis evaluation. In *European Conference on Computer Vision*, pp. 594–609. Springer, 2022.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Jialu Gao, KJ Joseph, and Fernando De La Torre. Teleportraits: Training-free people insertion into any scene. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18866–18875, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December*

10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets_and_Benchmarks.html.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11, 2022.

Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15942–15952. IEEE, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17089–17099, 2023.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Black Forest Labs. FLUX.2. Online. https://blackforestlabs.ai, 2025. Accessed: 2025-05-07.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2294–2305. IEEE, 2023.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R Venkatesh Babu. Text2place: Affordance-aware text guided human placement. In *European Conference on Computer Vision*, pp. 57–77. Springer, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4199–4209, 2023.

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*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Wensong Song, Hong Jiang, Zongxin Yang, Zheqiao Cheng, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9097–9105, 2026.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Seung Yong Kim, and Daniel Ali. Objectstitch: Object compositing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18310–18319, 2023.

Dogancan Temel, Mohit Prabhushankar, and Ghassan AlRegib. UNIQUE: unsupervised image quality estimation. *IEEE Signal Process. Lett.*, 23(10):1414–1418, 2016. doi: 10.1109/LSP.2016.2601119. URL https://doi.org/10.1109/LSP.2016.2601119.

Yunjie Tian, Qixiang Ye, and DAVID DOERMANN. Yolov12: Attention-centric real-time object detectors. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 78433–78457. Curran Associates, Inc., 2025. doi: 10.52202/085713-2627. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7103444259031cc58051f8c9a4868533-Paper-Conference.pdf.

Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *2023 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 18359–18369. IEEE, 2023b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1481–1490. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00147. URL https://doi.org/10.1109/CVPR52733.2024.00147.

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*, pp. 18381–18391, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Yifan Zhang, Jianguo Wang, Zhongliang Tang, and Wenmin Wang. Insert anyone: High-fidelity full-body photo insertion via dual-branch adapters. *Expert Systems with Applications*, pp. 131013, 2026.

Hao Zhu, Wu Wayne, Wentao Qiu, Chenxia Zhu, et al. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

Training Configuration. We fine-tune Qwen-Image-Edit-2509 (Wu et al., 2025a) using LoRA with the proposed *HBAF* and *FFIP* strategies. Training is conducted on 2 NVIDIA A100 GPUs for a total of 2,645 iterations. The batch size is set to 2 and the learning rate is 1×10^{-4}. We use the AdamW optimizer with a weight decay of 0.01. For HBAF, we set T_{start} to 900 and T_{end} to 808. For FFIP, the face loss balancing coefficient λ_{face} is set to 0.02.

Inference Configuration. All inference experiments are conducted on a single NVIDIA A100 GPU. Generating one image at a resolution of 1312×864 requires 40 denoising steps, with a total runtime of approximately 2.5 minutes, corresponding to an average of 3.75 seconds per step.

B EVALUATION DETAILS

IDS (Identity Similarity). Following MultiHuman-Testbench (Borse et al., 2026), we evaluate facial identity consistency using the IDS metric. We employ the InsightFace library to detect all faces in the generated image and the reference image, extracting normalized feature vectors for each. We then compute the cosine similarity matrix between all detected faces across the two images and solve for the optimal one-to-one face matching via the Hungarian algorithm (Munkres, 1957). The IDS score is defined as: $IDS = S_{total} / \max(n_{gen}, n_{src})$. S_{total} is the sum of cosine similarities over all matched face pairs, and n_{gen} and n_{src} denote the number of detected faces in the generated and reference images, respectively.

LPIPS. Following Text2Place (Parihar et al., 2024), we use LPIPS to measure how well the generated image preserves the target background. LPIPS computes a perceptual distance between the background regions of the synthesized and target images in feature space; a lower value indicates better background fidelity.

CA (Counting Alignment). Following the CA (Counting Alignment) metric proposed in TISE (Dinh et al., 2022), we adopt the number of persons in the reference image as the ground truth and compute the mean absolute error between the detected person count in the synthesized image and the reference, measuring whether the correct number of people is preserved after portrait insertion. A lower CA indicates better counting accuracy.

CLIP-IQA, BRISQUE, and UNIQUE. We employ CLIP-IQA (Wang et al., 2023a), BRISQUE (Mittal et al., 2012), and UNIQUE (Temel et al., 2016) to evaluate the perceptual quality of the synthesized images, which can indirectly reflect the quality of the portrait insertion task's completion.

VLM-T. Following TelePortraits (Gao et al., 2025), we use the VLM-T metric to evaluate the overall quality of human insertion. The full prompt used for evaluation is provided in List 3 (Appendix F).

Compared Methods. We compare against accessible state-of-the-art image editing models and dedicated insertion model, including DreamOmni2 (Xia et al., 2025), HunyuanImage-3.0-Instruct (Cao et al., 2025), OmniGen2 (Wu et al., 2025b), Insert-Anything (Song et al., 2026), Nano-Banana-Pro and Qwen-Image-Edit-2509 (Wu et al., 2025a).

C ADDITIONAL ABLATION RESULTS

We present supplementary ablation experiments result to further justify the design choices of InsHuman, including the effect of the $\lambda_{\max}$ and λ_t in HBAF and the mask design strategy.

C.1 EFFECT OF $\lambda_{\max}$ IN HBAF

Setting $\lambda_{\max} = 5$ imposes overly strong supervision on the human region during the early denoising stages, where the model is still forming global composition. As shown in Figure 11, this results in severe structural artifacts, including phantom limbs, duplicated heads, and distorted body proportions. The CA metric correspondingly degrades to 0.2824, significantly worse than our default ($\lambda_{\max} = 2.5$,



Figure 11: **Visual comparison with $\lambda_{\max} = 5$.** An excessively large value of $\lambda_{\max}$ will disrupt the generative process and produce anatomically implausible outputs.



Figure 12: **Visual comparison with a fixed λ.** Removal of the dynamic weight function λ_t disrupts facial feature generation, leading to identity deviations from the reference image.

CA=0.1298). This demonstrates that while amplifying the human-region weight can in principle strengthen structural constraints, an excessively large value disrupts the generative process and produces anatomically implausible outputs. A moderate $\lambda_{\max} = 2.5$ strikes the right balance between structural guidance and natural generation.

C.2 Effect of the Dynamic Weight Schedule λ_t in HBAF

When the dynamic schedule λ_t is replaced by a constant $\lambda = 2.5$ applied uniformly across all timesteps, as shown in Figure 12, this causes noticeable facial feature deviations from the source identity. Specifically, applying a constant strong constraint throughout all denoising stages limits the model's ability to refine identity-related details in the later denoising steps, resulting in noticeable facial feature deviations from the reference image. In contrast, our dynamic weighting strategy assigns stronger guidance at early timesteps, when the model mainly determines global human structure and spatial arrangement, and gradually reduces the constraint during later timesteps to allow fine-grained facial features refinement. This adaptive schedule enables the model to simultaneously maintain reasonable human-background interaction and preserve facial identity consistency.

C.3 Comparison of HBAF Mask Designs



Figure 13: **Comparison of different HBAF mask designs.** The box-based mask (Ours) provides a broader supervision region around the human and yields more natural human-background transitions than the fine-grained mask, which excluding the boundary areas.

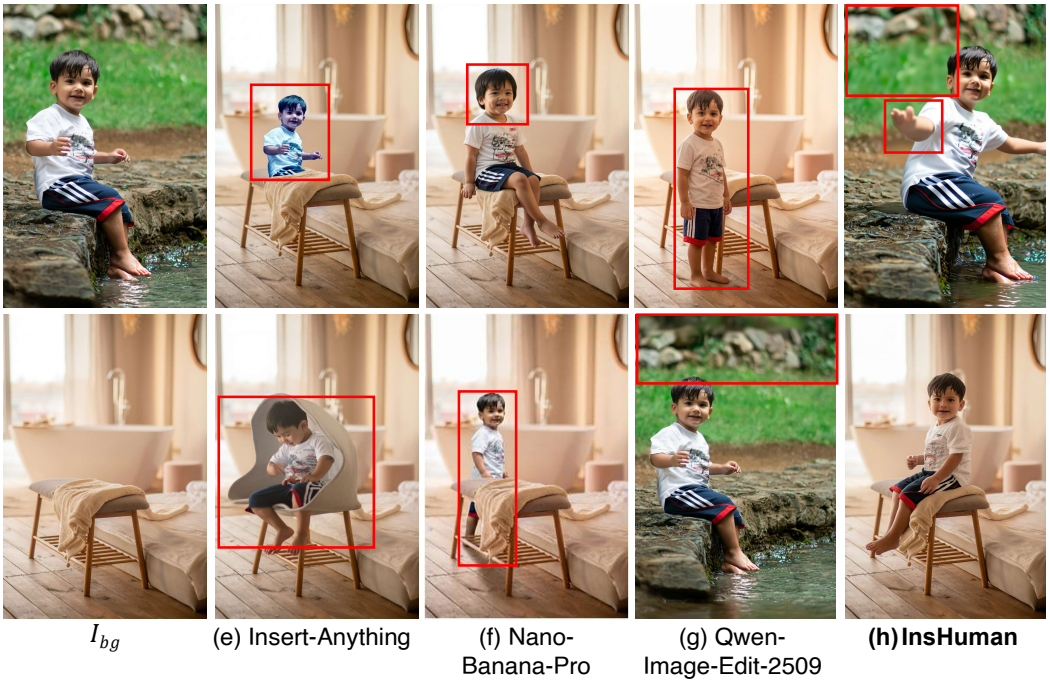
We further investigate the effect of different mask designs on human insertion performance in HBAF. In our original implementation, HBAF first employs YOLOv12-m (Tian et al., 2025) to detect foreground persons and merges their bounding boxes into a binary human-region mask. To construct a more fine-grained alternative, we retain the same YOLOv12-m bounding boxes and further apply SAM within each bounding box to segment the precise human foreground.

As shown in Figure 13, the box-based mask used in our method produces more natural transitions between the inserted person and the background. This is because natural visual blending in human insertion depends not only on the appearance of the person itself, but also on the transition and interaction regions surrounding the person, such as contact boundaries with the ground, cast shadows, and local illumination transitions. Therefore, a broader supervision region helps the model better learn these transition regions.

D Image Sources

For the real images used in constructing our dataset, we obtained the images directly from Unsplash (`https://unsplash.com/`) and used them in accordance with the applicable Unsplash License and Terms of Service.

E More qualitative comparisons with image editing models


I_{src}
(a) FLUX.2-klein-base-4B
(b) DreamOmni2
(c) HunyuanImage-3.0-instruct
(d) OmniGen2
I_{bg}
(e) Insert-Anything
(f) Nano-Banana-Pro
(g) Qwen-Image-Edit-2509
(h) InsHuman


I_{src}
(a) FLUX.2-klein-base-4B
(b) DreamOmni2
(c) HunyuanImage-3.0-instruct
(d) OmniGen2
I_{bg}
(e) Insert-Anything
(f) Nano-Banana-Pro
(g) Qwen-Image-Edit-2509
(h) InsHuman

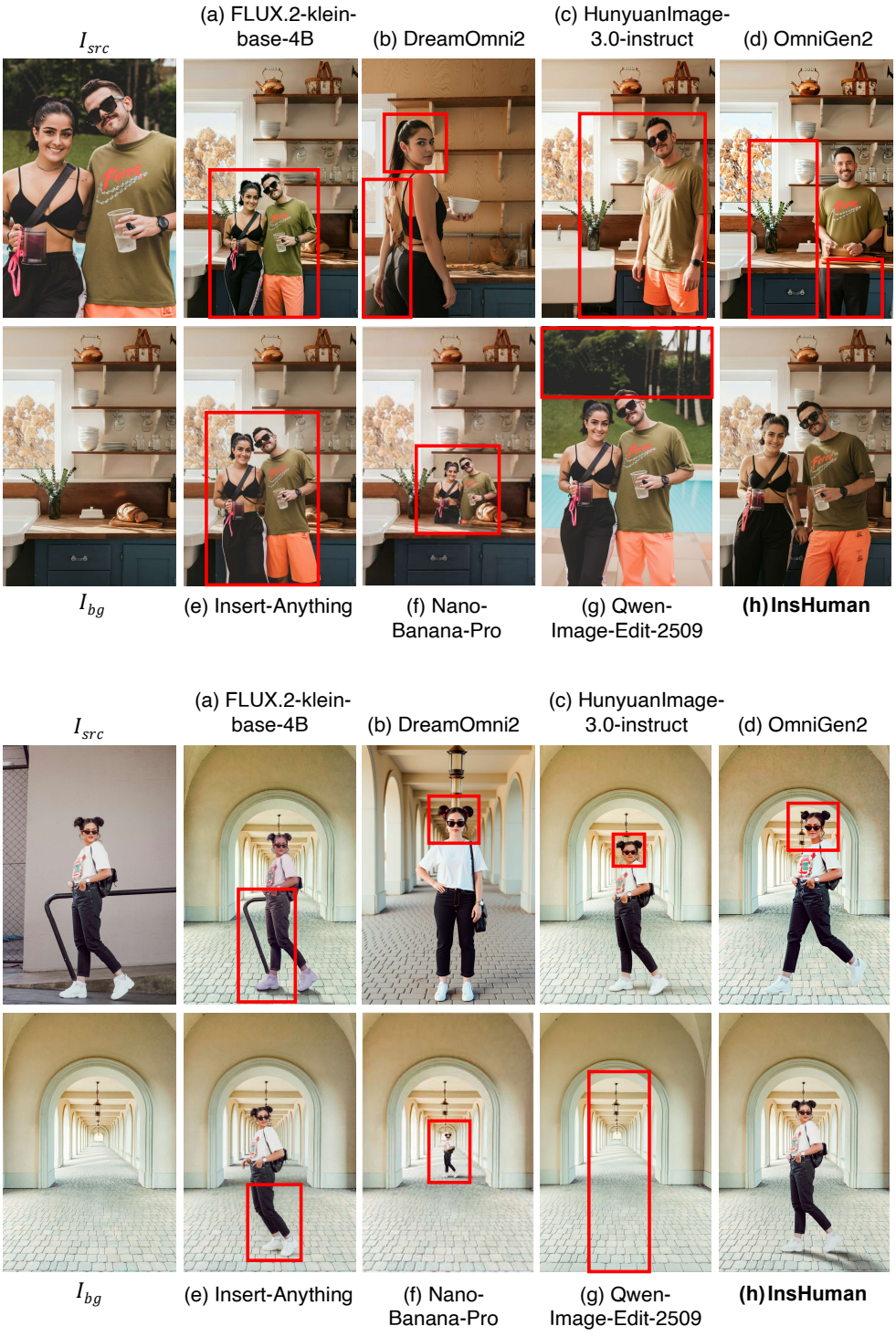


Figure 14: **More qualitative comparison of different image editing models on human insertion.** Existing image editing models often exhibit issues on human insertion, e.g., poses failing to adapt to the background, deviations in the number of people or overall appearance from the reference image, and loss of facial features. In contrast, our method (InsHuman) more stably maintain human structure and identity features, achieving more natural human insertion. The **red** boxes highlight the errors.

F PROMPTS USED

Listing 1: Nano Banana Pro prompts used for dataset synthesis.

```
Task: Precisely place the figure in the foreground of [the first image]
     into the background scene of [the second image].
Requirements:
1. Ensure the correct size and perspective. You may adjust the pose and
     position of the figure as appropriate to ensure a perfect match
    with the ground and objects in the background.
2. Identity Constraint: Strictly maintain the facial features, clothing
     (including color and shape), and identity of the figure. You may
    generate additional information such as arms or legs, but must not
    alter the identity or micro-expressions of the figure. Also, ensure
    overall clarity, especially of the face.
3. Environmental Integration: Achieve perfect geometric consistency and
     natural integration of lighting and shadows with the background
    scene.
```

Listing 2: Qwen-Image-Edit-2509 prompts used for inpainting.

```
Remove the person from the scene.
Fill the area naturally with the surrounding background, keeping
    lighting, shadows, textures, and perspective consistent. Seamless,
    photorealistic result with no traces of the removed subject.
```

Listing 3: Qwen2.5-VL-72B-Instruct prompts used for evaluating VLM-T.

```
### Task Definition
You will be provided with three images: the first is a composite image
    where one or more persons from a foreground image have been inserted
     into a background scene; the second is the foreground image
    containing the original person(s); the third is the original
    background scene image.

As an experienced evaluator, your task is to evaluate how well the
    insertion task has been completed, focusing on person count accuracy
    , scale, perspective, natural placement, and environmental
    integration.

### Scoring Criteria
Evaluate how well the composite image fulfills the following five
    requirements:

1. Person Count Accuracy: Check whether the number of persons visible
    in the composite image matches the number of persons in the
    foreground image. Missing or extra persons should significantly
    lower the score.

2. Scale and Perspective: Assess whether the size of the inserted
    person(s) is consistent with the background scene's perspective and
    depth, appearing proportionally correct and natural relative to
    surrounding objects.

3. Pose and Position Naturalness: Evaluate whether the pose and
    position of the inserted person(s) are natural and appropriate for
    the background context (e.g., feet properly grounded, body correctly
     oriented relative to objects in the scene).

4. Environmental Integration: Assess whether lighting direction, shadow
     quality, and geometric consistency between the inserted person(s)
    and the background are realistic and natural. Look for signs of
    unnatural edges, color mismatches, or lighting inconsistencies.
```

```
5. Overall Realism: Evaluate the overall visual coherence and
    photorealism of the composite image as a whole.

### Scoring Range
Give a specific integer score from 0 to 9:
- Very Poor (0): Task completely failed. No person detected in the
    composite image, or the number of persons is completely incorrect.
- Poor (1-2): Major failures. Significant issues in person count, scale,
     placement, or integration. Result looks clearly and unnaturally
    composited.
- Fair (3-4): Partial success. Some requirements met but with notable
    issues in at least two of the five criteria.
- Good (5-6): Most requirements met. Only minor issues in scale,
    perspective, or environmental integration.
- Very Good (7-8): Nearly all requirements met. Only minor
    imperfections remain, such as slight lighting inconsistency or minor
     scale deviation.
- Excellent (9): All requirements perfectly met. The composite is
    photorealistic, natural, and indistinguishable from a real
    photograph.

### Input Format
You will receive three images: the first is the composite (generated)
    image, the second is the foreground reference image, the third is
    the original background scene image.

### Output Format
[Your Score]
Output only the integer score. Do not include any analysis or
    explanation.
```