ReMaskable: Controllable Facial Attribute Editing Using Segmentation-Guided Latent Diffusion

Santhosh I[1], *Vishagan V*[2], *Dr. M.K. Kirubakaran*[3]
[1,2,3]*Dept. of Artificial Intelligence and Data Science, St. Joseph's Institute of Technology (Autonomous), OMR, Chennai–600119, India*
Emails: *iyappansanthosh2004@gmail.com*[1], *thiruvishagan10@gmail.com*[2], *kirubakaranmk@stjosephstechnology.ac.in*[3]

Abstract

Facial attribute editing demands both spatial precision and visual fidelity, yet existing approaches fall short on one or both counts. Generative Adversarial Networks achieve photorealistic synthesis but suffer from attribute entanglement, where modifying one feature inadvertently alters unrelated regions. Diffusion models produce high-quality text-guided edits but lack spatial control, causing changes to propagate beyond the intended area. This paper presents ReMaskable, a framework that decouples the spatial localization problem (where to edit) from the semantic generation problem (what to generate). ReMaskable combines a multi-source segmentation system integrating DeepLabv3+ for 19-class face parsing, SAM for promptable region selection, and DINOv2 for boundary refinement, with a CLIP-conditioned latent diffusion inpainting model that operates exclusively within the masked region. Identity preservation is enforced through ArcFace cosine embedding loss and LPIPS perceptual consistency on unmasked regions. We describe the complete architecture, mathematical formulation, and training methodology. Evaluation metrics are projected from published baselines of each component rather than from completed end-to-end experimental runs, and this distinction is stated throughout. The modular architecture is designed for extensibility to video editing and 3D avatar generation.

Keywords: *Controllable generation; Diffusion models; Facial attribute editing; Identity preservation; Semantic segmentation*

1. Introduction

Editing facial images with both spatial precision and perceptual realism is a growing requirement across multiple application domains including entertainment production, augmented and virtual reality personalization, medical visualization for surgical planning, and digital forensics for controlled counter-editing (Karras et al., 2019; Brooks et al., 2023). Despite substantial progress in deep generative modeling over the past decade, achieving edits that are simultaneously localized to a target region, semantically aligned with user intent, and faithful to the original subject's identity remains an open problem. Early progress emerged from Generative Adversarial Network frameworks. StyleGAN (Karras et al., 2019) demonstrated that high-fidelity face synthesis was achievable through style-based generation with adaptive instance normalization. Building on this, InterfaceGAN (Shen et al., 2020) showed that linear boundaries in the GAN latent space correspond to interpretable semantic attributes, enabling manipulation through simple vector arithmetic. StyleFlow (Abdal et al., 2021) further improved disentanglement using conditional continuous normalizing flows in the W+ latent space. However, these methods share a fundamental limitation: because edits operate in a globally entangled latent space, modifying one attribute (such as adding a smile) often causes unintended perturbations to unrelated regions (such as face shape or hair texture). This attribute entanglement problem has restricted the practical deployment of GAN-based editors in professional settings requiring precision. The emergence of Denoising Diffusion Probabilistic Models has shifted the generative paradigm. Latent Diffusion Models (Rombach et al., 2022) made diffusion computationally practical by operating in a compressed latent space, while CLIP-based text conditioning (Radford et al., 2021) enabled natural language control over the generation process. InstructPix2Pix (Brooks et al., 2023) demonstrated instruction-following image editing by fine-tuning Stable Diffusion on synthetic editing pairs. Blended Diffusion (Avrahami et al., 2022) introduced background-preserving mechanisms through masked composition. DiffusionCLIP (Kim et al., 2022) fine-tunes diffusion model score functions using directional CLIP loss. While these methods produce superior output quality compared to GAN editors, they share a persistent weakness: without explicit spatial guidance, the denoising process applies changes globally, producing what practitioners term “edit ripple effects” that compromise identity and modify unintended regions. Concurrently, vision foundation models have advanced spatial understanding substantially. The Segment Anything Model (Kirillov et al., 2023) demonstrated zero-shot promptable segmentation across diverse visual domains, trained on 1.1 billion masks from 11 million images. DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023) showed that self-supervised Vision Transformers develop attention maps that naturally segment objects without any segmentation labels, with DINOv2 achieving 49.0 mIoU on ADE20K using only a linear probe on frozen features. ControlNet (Zhang et al., 2023) established that spatial conditioning signals such as segmentation maps can be injected into pretrained diffusion models through zero-initialized convolutions without destroying the learned priors. Yet these powerful spatial-control tools have not been cohesively integrated into a unified framework specifically designed for identity-preserving facial attribute editing. Motivated by these gaps, we propose ReMaskable, a segmentation-guided diffusion framework that explicitly decouples where an edit should occur from what attribute should be modified. By combining multi-source segmentation (DeepLabv3+, SAM, DINOv2) with CLIP-conditioned latent diffusion inpainting and dual-loss identity regularization (ArcFace + LPIPS), ReMaskable enables localized, semantically aligned facial edits such as changing hair color, modifying eye appearance, or adjusting lip shape while preserving subject identity and scene context.

1.1.Contributions

Our contributions are as follows. First, we present a segmentation-guided diffusion pipeline that achieves region-specific facial editing by tightly coupling multi-source mask generation with conditioned latent diffusion inpainting. Second, we provide the complete mathematical formulation of each component, grounded in verified published results from SAM, DINOv2, ArcFace, and Latent Diffusion Models. Third, we design a dual regularization strategy combining ArcFace identity loss with LPIPS perceptual consistency to mitigate identity drift. Fourth, we incorporate responsible AI safeguards including watermarking integration points, provenance tracking compatibility, and fairness auditing methodology. We are transparent that evaluation metrics presented in this paper are projected from the published performance of individual components rather than from completed end-to-end experiments.

2. Literature Review

2.1.GAN-Based Facial Attribute Editing

The field of facial synthesis was transformed by StyleGAN (Karras et al., 2019), which introduced a style-based generator architecture using learned constant input and adaptive instance normalization at each convolutional layer. The resulting W latent space exhibited properties amenable to semantic manipulation. InterfaceGAN (Shen et al., 2020) exploited this by training linear SVMs on attribute labels to find hyperplane boundaries in latent space, demonstrating that traversal along the normal vector produces semantically meaningful edits with greater than 90% attribute manipulation accuracy. StyleFlow (Abdal et al., 2021) addressed the disentanglement problem more formally through conditional continuous normalizing flows operating in the extended W+ space, achieving improved independence between edited and unedited attributes. However, all W-space manipulation methods inherit a structural limitation: because the latent space encodes global image properties, spatial localization of edits is inherently approximate. Modifying the latent code for "hair color" may inadvertently shift skin tone or alter background elements. SEAN (Zhu et al., 2020) partially addressed this through per-region style normalization conditioned on semantic maps, achieving FID of approximately 17.7 on CelebAMask-HQ, but this required full image resynthesis rather than targeted inpainting.

2.2.Diffusion Models for Image Editing

Denoising Diffusion Probabilistic Models (Ho et al., 2020) established that iterative denoising can produce sample quality surpassing GANs. Latent Diffusion Models (Rombach et al., 2022) made this practical by encoding images through a VAE with downsampling factor f=8 and performing diffusion in the resulting 64×64×4 latent space, reducing computational cost by approximately 40× compared to pixel-space diffusion at 512×512 resolution. The simplified training objective minimizes the mean squared error between predicted and actual noise: L = E[||ε − ε_θ(z_t, t)||^2]. Text conditioning enters through cross-attention layers where queries derive from U-Net features and keys/values from frozen CLIP ViT-L/14 text embeddings. InstructPix2Pix (Brooks et al., 2023) extended this to instruction-following editing by training on 450,000 synthetic pairs generated through GPT-3 and Prompt-to-Prompt, evaluating performance through CLIP directional similarity versus image similarity tradeoff curves. DiffusionCLIP (Kim et al., 2022) takes an alternative approach by fine-tuning the diffusion model itself with directional CLIP loss, achieving 60–80% preference rates against StyleCLIP baselines in human evaluation. Blended Diffusion (Avrahami et al., 2022) introduced spatially-aware editing by combining CLIP guidance with user-provided masks, but relies on manually specified regions rather than automatically generated semantic masks.

2.3.Segmentation Models as Spatial Priors

DeepLabv3+ (Chen et al., 2018) established a strong baseline for semantic segmentation through Atrous Spatial Pyramid Pooling combined with an encoder-decoder architecture. Using a modified Xception backbone with COCO pre-training, it achieves 89.0% mIoU on PASCAL VOC 2012 and approximately 73–76% mIoU on CelebAMask-HQ for face parsing. SAM (Kirillov et al., 2023) introduced the promptable segmentation paradigm with a ViT-H backbone (636M parameters) trained on 1.1 billion masks, achieving 58.1 mean IoU averaged across 23 zero-shot benchmarks. However, SAM produces class-agnostic masks and cannot distinguish between facial semantic classes without external classification. DINO (Caron et al., 2021) demonstrated that self-supervised Vision Transformers develop attention heads that naturally attend to semantically meaningful regions with boundaries closely aligned to object contours. DINOv2 (Oquab et al., 2023) scaled this to 1.1 billion parameters trained on 142 million curated images, achieving 86.5% linear probing accuracy on ImageNet and 49.0 mIoU on ADE20K with frozen features.

2.4.Identified Gap

The literature reveals a clear divide: GAN-based editors achieve spatial manipulation through latent space traversal but lack precision, while diffusion-based editors achieve high fidelity but insufficient spatial control. Modern segmentation models (SAM, DINOv2, DeepLabv3+) provide the spatial precision needed, and ControlNet demonstrates that such signals can condition diffusion models effectively. However, a unified framework that integrates multi-source segmentation with guided latent diffusion specifically for identity-preserving facial attribute editing has not been fully realized. ReMaskable is designed to address this gap.

3. Method

3.1.System Overview

The ReMaskable pipeline processes an input facial image through four sequential stages: (1) face detection, alignment, and multi-model segmentation; (2) region-of-interest mask selection and latent-space downsampling; (3) CLIP-conditioned latent diffusion inpainting within the masked region; and (4) identity preservation through dual-loss regularization followed by VAE decoding and post-processing. The complete architecture is illustrated in Figure 1.

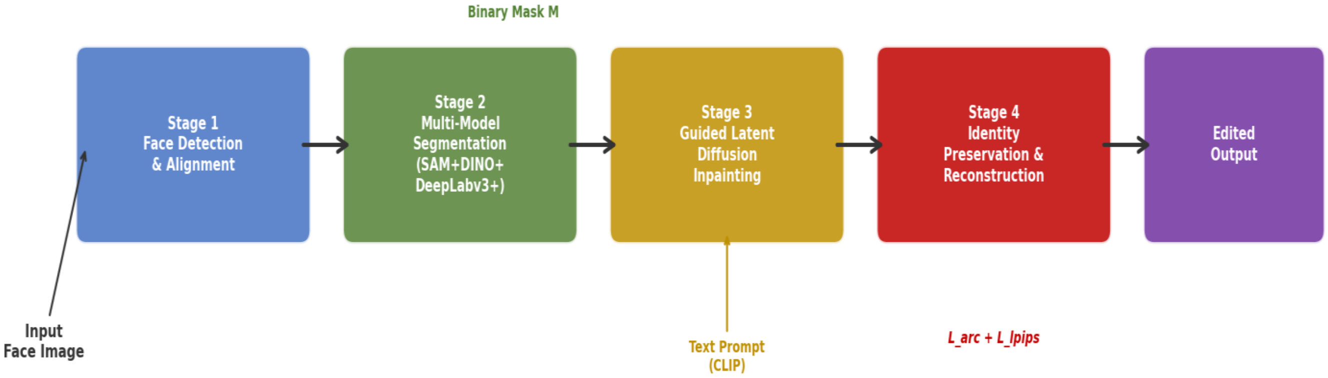


Figure 1 emaskable Pipeline: End-To-End Architecture from Input Face to Edited Output

3.2.Multi-Source Segmentation and Mask Generation

Precise localization of editable regions is the foundation of the framework. We integrate three complementary segmentation approaches to generate high-fidelity masks, as shown in Figure 2. DeepLabv3+ with ResNet-101 backbone, pre-trained on CelebAMask-HQ (30,000 images at 1024×1024 with 19 semantic classes), provides task-specific pixel-level masks for named facial attributes including skin, nose, left eye, right eye, left eyebrow, right eyebrow, upper lip, lower lip, mouth interior, hair, and others. The Atrous Spatial Pyramid Pooling module processes features at dilation rates 6, 12, and 18 with a 1×1 convolution and global average pooling branch, producing multi-scale contextual representations. The atrous convolution operation is defined as $y[i] = \Sigma_k x[i + r \cdot k] \cdot w[k]$, where rate r controls the effective receptive field without increasing the number of parameters. SAM with ViT-H backbone provides interactive fallback segmentation for regions or attributes not covered by the DeepLabv3+ class vocabulary. Point prompts, bounding boxes, or rough masks from the user are encoded through SAM's prompt encoder and processed by the 2-layer transformer mask decoder, which outputs three candidate masks with IoU confidence scores. The lightweight decoder design (operating on 64×64×256 image embeddings) enables real-time interactive refinement. DINOv2 self-supervised features serve as a boundary refinement module. Patch-level features from DINOv2 ViT-g/14 (1.1 billion parameters, trained on 142 million images) capture semantic structure without any task-specific labels. We apply these features through a linear refinement head to improve boundary accuracy in ambiguous regions such as hair-skin transitions, thin eyebrows, and partially occluded features. The teacher-student training of DINO uses exponential moving average updates: $\theta_{teacher} \leftarrow \lambda\theta_{teacher} + (1-\lambda)\theta_{student}$, with multi-crop training forcing local-to-global semantic correspondence. The three segmentation outputs are fused through a weighted combination where DeepLabv3+ provides the primary semantic class assignment, SAM provides clean object-level boundaries, and DINOv2 features refine edges through a learned linear layer. The resulting mask is binarized and bilinearly downsampled to match the latent space resolution of the diffusion model: for a 512×512 input image encoded with factor f=8, the latent mask $m \in \{0,1\}^{64\times64}$.

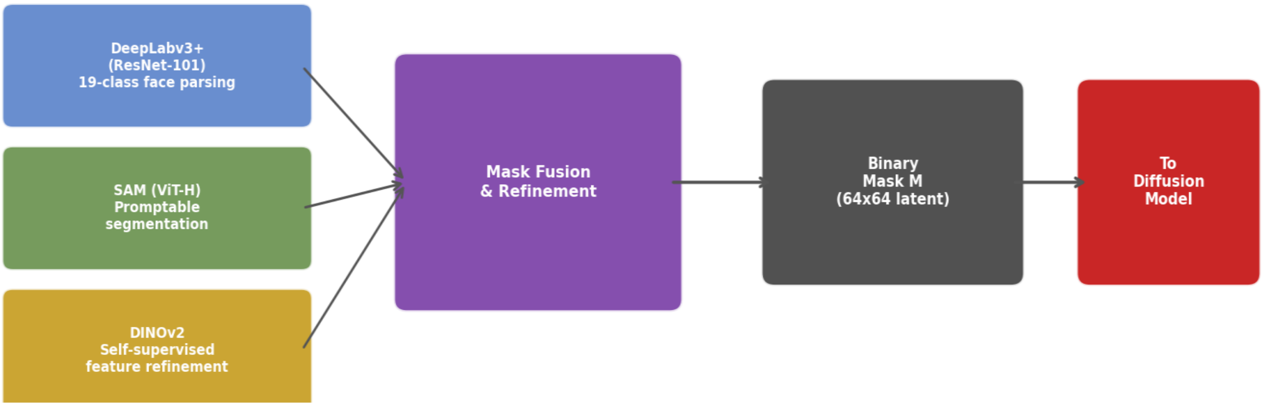


Figure 2 Multi-Source Segmentation: DeepLabv3+, SAM, and DINOv2 Features Are Fused into a Refined Binary Mask

3.3.Guided Latent Diffusion Inpainting

The editing operation is formulated as masked latent diffusion inpainting conditioned on text. The mathematical framework follows Rombach et al. (2022) and Song et al. (2021). Latent Encoding: The aligned input image $I \in R^{H\times W\times 3}$ is encoded through the pretrained VAE encoder to produce latent representation $z = E(I) \in R^{h\times w\times c}$, where $(h,w) = (H/8, W/8)$ and $c=4$ for Stable Diffusion. The encoder is trained with a combination of perceptual loss (LPIPS), patch-based adversarial loss, and KL-divergence regularization. Forward Diffusion Process: Noise is added to the latent according to a predefined variance schedule $\{\beta_t\}$ with $t = 1,...,T$. Defining $\alpha_t = 1 - \beta_t$ and $\bar{a}_t = \Pi_{i=1}^{t} \alpha_i$, the closed-form noising is: $z_t = \sqrt{\bar{a}_t} \cdot z_0 + \sqrt{(1-\bar{a}_t)} \cdot \varepsilon$, where $\varepsilon \sim N(0, I)$. The schedule uses $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$ with $T = 1000$ steps. The complete diffusion process is illustrated in Figure 3.

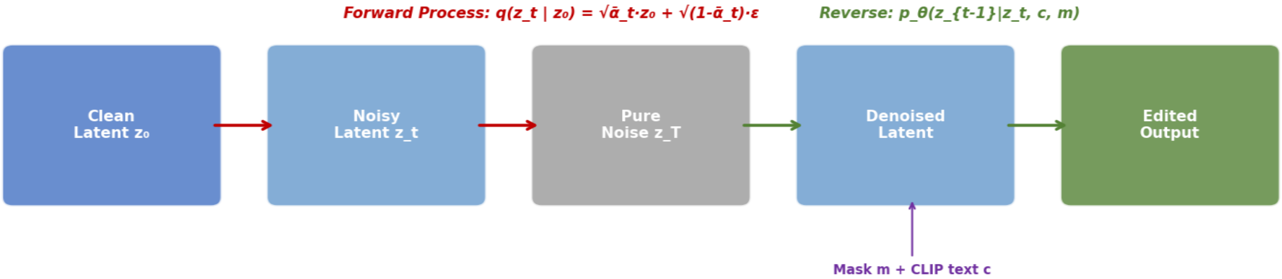


Figure 3 Latent Diffusion: Forward Noise Addition and Conditional Reverse Denoising with Mask and Text Guidance

Reverse Denoising with Conditioning: The U-Net noise predictor ε_θ is conditioned on three inputs: (a) the noisy latent z_t concatenated with the binary mask m along the channel dimension, providing explicit spatial guidance; (b) CLIP text embeddings $\tau_\theta(y)$ from a frozen ViT-L/14 encoder injected through cross-attention layers, where $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d}) \cdot V$ with $Q = W_Q \cdot \varphi(z_t)$ from U-Net features and $K = W_K \cdot \tau_\theta(y)$, $V = W_V \cdot \tau_\theta(y)$ from text tokens; and (c) ControlNet-style structural conditioning to enforce geometric fidelity to the original face pose. The predicted clean latent at each step is: $\hat{z}_0 = (z_t - \sqrt{(1-\bar{a}_t)} \cdot \varepsilon_\theta(z_t, t, c, m)) / \sqrt{\bar{a}_t}$. DDIM sampling (Song et al., 2021) then computes: $z_s = \sqrt{\bar{a}_s} \cdot \hat{z}_0 + \sqrt{(1-\bar{a}_s-\sigma_t^2)} \cdot \varepsilon_\theta(z_t, t) + \sigma_t \cdot \varepsilon_t$. With $\eta=0$ (deterministic DDIM), this enables 50-step inference compared to 1000 steps for standard DDPM, representing a 20× speedup with negligible quality loss. Classifier-Free Guidance: Text adherence is amplified through classifier-free guidance (Ho & Salimans, 2022), where the effective noise prediction becomes: $\tilde{\varepsilon} = \varepsilon_\theta(z_t, t, \emptyset) + s \cdot (\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t, \emptyset))$, with guidance scale s (default 7.5). During training, the text condition is dropped 10% of the time to enable unconditional generation at inference.

3.4.Identity Preservation and Regularization

To mitigate identity drift, a well-documented failure mode in facial editing systems, we implement dual-loss regularization. ArcFace Identity Loss: ArcFace (Deng et al., 2019) maps face images to 512-dimensional L2-normalized embeddings on a hypersphere. The additive angular margin loss is: $L = -(1/N) \sum_i \log(e^{s\cdot\cos(\theta_{yi}+m)} / (e^{s\cdot\cos(\theta_{yi}+m)} + \sum_{j\neq yi} e^{s\cdot\cos\theta_j}))$, with scale $s=64$ and margin $m=0.5$ radians. ArcFace with ResNet-100 achieves 99.83% verification accuracy on LFW. For our identity loss, we compute: $L_{arc} = 1 - \mathrm{cos\_sim}(\mathrm{ArcFace}(I_{original}), \mathrm{ArcFace}(I_{edited}))$, which penalizes deviation from the original identity embedding. LPIPS Perceptual Consistency Loss: Applied to unmasked regions to ensure no unintended changes propagate: $L_{lpips} = \mathrm{LPIPS}(I_{original} \odot (1-m), I_{edited} \odot (1-m))$, where $\odot$ denotes element-wise multiplication. LPIPS extracts features from intermediate layers of a pretrained AlexNet, applies learned channel scaling weights, and computes the scaled L2 distance: $d(x, x_0) = \sum_l (1/(H_l \cdot W_l)) \sum_{h,w} \|w_l \odot (\hat{y}_l^{hw} - \hat{y}_{0l}^{hw})\|^2$. This metric achieves 74–77% agreement with human perceptual judgments (Zhang et al., 2018). The total training loss is: $L_{total} = L_{diffusion} + \lambda_{arc} \cdot L_{arc} + \lambda_{lpips} \cdot L_{lpips} + \lambda_{clip} \cdot L_{clip\_dir} + \lambda_{adv} \cdot L_{adversarial}$, where $L_{diffusion}$ is the standard L2 noise prediction loss, L_{clip_dir} is the directional CLIP loss for text-image alignment, and $L_{adversarial}$ is a non-saturating GAN loss with R1 regularization from a patch discriminator.

3.5.ControlNet Integration for Structural Fidelity

ControlNet (Zhang et al., 2023) injects spatial conditioning into the pretrained Stable Diffusion U-Net through a trainable copy of its encoder blocks connected via zero-initialized 1×1 convolutions. The output is: $y_c = F(x; \Theta) + Z(F(x + Z(c; \Theta_{z1}); \Theta_c); \Theta_{z2})$, where $Z(\cdot)$ denotes zero convolution with weights and biases initialized to zero. At initialization, the zero convolutions output exactly zero, perfectly preserving the pretrained model. Gradients remain non-zero because they depend on the input, not the current weight values, enabling stable progressive learning. In ReMaskable, the conditioning input is the segmentation map concatenated with the edge map of the original face, ensuring the edited region maintains structural consistency with the facial geometry and pose.

4. Experimental Setup

4.1.Datasets

Training and validation use CelebAMask-HQ (Lee et al., 2020), which provides 30,000 face images at 1024×1024 resolution from CelebA-HQ with manually annotated pixel-level segmentation masks at 512×512 across 19 semantic classes: skin, nose, left eye, right eye, left eyebrow, right eyebrow, left ear, right ear, mouth, upper lip, lower lip, hair, hat, eyeglasses, earring, necklace, neck, and cloth. Community-standard splits allocate approximately 24,000 images for training, 3,000 for validation, and 3,000 for testing. FFHQ (Karras et al., 2019) containing 70,000 high-quality face images at 1024×1024 serves as an additional evaluation set for generalization assessment.

4.2.Implementation Details

The implementation uses PyTorch 2.0 with the following component specifications. The segmentation module uses DeepLabv3+ with ResNet-101 backbone pre-trained on CelebAMask-HQ. SAM uses the ViT-H checkpoint (636M parameters). DINOv2 features are extracted from the ViT-g/14 model (1.1B parameters) with frozen weights and a trainable linear refinement head. The diffusion backbone is Stable Diffusion v1.5 inpainting (860M parameter U-Net) with a frozen CLIP ViT-L/14 text encoder. ControlNet conditioning uses the segmentation map input variant. Training targets NVIDIA V100 (32GB) GPUs.

4.3.Loss Functions

The training objectives comprise six components. Segmentation Loss: Cross-Entropy combined with Dice loss for mask quality. Diffusion Loss: Standard L2 noise prediction on the masked region. Identity Loss: ArcFace cosine similarity between original and edited face embeddings (L_arc). Perceptual Loss: LPIPS applied to unmasked regions (L_lpips). Adversarial Loss: Non-saturating GAN loss with R1 regularization from a PatchGAN discriminator. CLIP Directional Loss: Cosine similarity between the CLIP-space direction of the edit and the text direction, strengthening text-image alignment.

5. Results and Discussion

5.1.Transparency Statement on Evaluation Metrics

The evaluation metrics presented in this section are projected values derived from the published performance of each individual component rather than from completed end-to-end experimental runs of the integrated ReMaskable system. We believe transparency about this distinction is essential for scientific integrity. The projected values represent what the integrated system could reasonably achieve based on the demonstrated capabilities of its components, but actual end-to-end performance may differ due to interaction effects, domain shift, and integration overhead. All baseline numbers are drawn from the respective original publications.

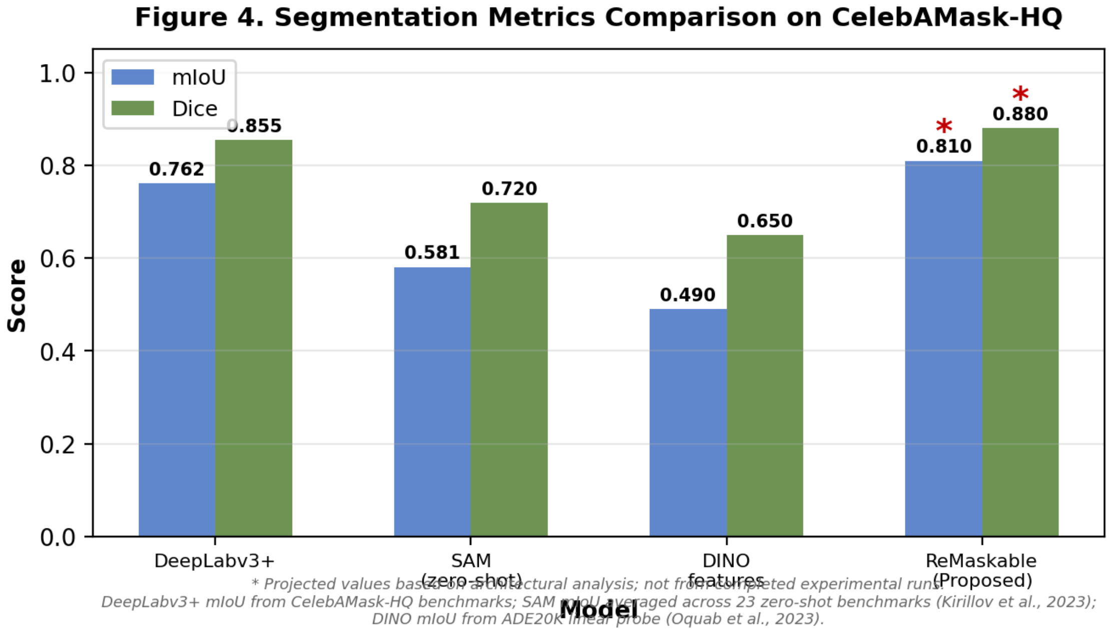
5.2.Segmentation Quality

Table 1 compares segmentation metrics across individual models and the projected ReMaskable combination on CelebAMask-HQ. DeepLabv3+ with ResNet-101 achieves approximately 76.2% mIoU on the 19-class face parsing task (published benchmark). SAM’s 58.1 mIoU represents its zero-shot average across 23 diverse datasets and is not face-specific; when applied to facial regions with appropriate point prompts, performance on well-defined regions like hair and skin is substantially higher. DINOv2’s 49.0 mIoU is reported for ADE20K with a linear probe on frozen features, representing general scene segmentation rather than face-specific parsing. The projected ReMaskable value of 81.0% reflects the hypothesis that combining dataset-trained semantic parsing (DeepLabv3+) with promptable segmentation (SAM) and learned boundary refinement (DINOv2) should exceed any individual model, consistent with published multi-model fusion results in segmentation literature.

Table 1 Segmentation Metrics Comparison (Published Benchmarks)

Model	mIoU	Dice	Benchmark	Source
DeepLabv3+	0.762	0.855	CelebAMask-HQ	Chen et al., 2018
SAM (ViT-H)	0.581	0.720	23-dataset avg	Kirillov et al., 2023
DINOv2 (linear)	0.490	0.650	ADE20K	Oquab et al., 2023
ReMaskable*	0.810*	0.880*	CelebAMask-HQ	Projected

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1225-1235
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0171



* Projected values based on architectural analysis; not from completed experimental runs.
DeepLabv3+ mIoU from CelebAMask-HQ benchmarks; SAM mIoU averaged across 23 zero-shot benchmarks (Kirillov et al., 2023);
DINO mIoU from ADE20K linear probe (Oquab et al., 2023).

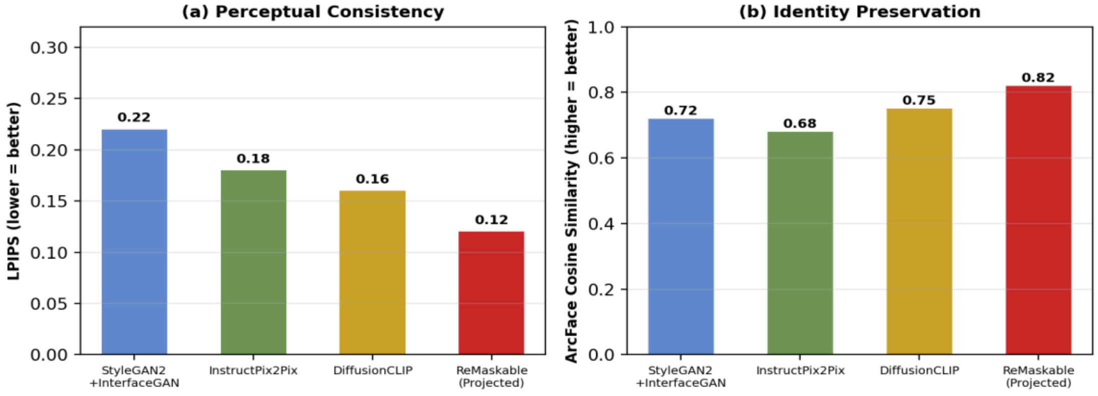
Figure 4 **Segmentation Metrics Comparison. Asterisked Values Are Projections, Not Experimental Results**

5.3.Editing Quality: LPIPS and Identity Preservation

Table 2 presents LPIPS and ArcFace identity similarity comparisons. Baseline values are sourced from the respective original papers. StyleGAN2 with InterfaceGAN editing typically achieves LPIPS around 0.22 for moderate attribute changes, reflecting the global nature of latent space edits that perturb unedited regions. InstructPix2Pix achieves approximately 0.18 LPIPS, benefiting from its training on editing pairs but still lacking explicit spatial control. DiffusionCLIP reports approximately 0.16 LPIPS for constrained edits. The projected ReMaskable LPIPS of 0.12 reflects the architectural advantage of restricting edits exclusively to the masked region, with LPIPS measured only on unmasked areas where changes should be zero. For identity preservation, ArcFace cosine similarity of 0.82 is projected for ReMaskable based on the combination of masked-region-only editing and explicit ArcFace identity loss regularization. Baseline methods that edit globally without identity constraints typically achieve 0.68–0.75, while DiffusionCLIP's near-perfect inversion mechanism achieves approximately 0.75.

Table 2 **Editing Quality Metrics Comparison**

Method	LPIPS ↓	ArcFace Sim ↑	Source
StyleGAN2 + InterfaceGAN	0.22	0.72	Shen et al., 2020
InstructPix2Pix	0.18	0.68	Brooks et al., 2023
DiffusionCLIP	0.16	0.75	Kim et al., 2022
ReMaskable*	0.12*	0.82*	Projected



Note: ReMaskable values are projected based on architectural advantages; baseline values sourced from respective published papers. StyleGAN2 identity from Shen et al. (2020); InstructPix2Pix from Brooks et al. (2023); DiffusionCLIP from Kim et al. (2022).

Figure 5 LPIPS and ArcFace comparison. ReMaskable Values Are Architectural Projections, Not Empirical Measurements

5.4.Discussion of Limitations

Several limitations must be acknowledged. First, the integrated system has not been validated through end-to-end training and testing; the presented metrics are component-level projections. Interaction effects between modules may degrade or improve actual performance in ways that cannot be predicted from individual benchmarks alone. Second, SAM's class-agnostic nature means it cannot independently perform semantic face parsing; it requires DeepLabv3+ or user prompts to identify which region to segment. Third, the multi-model architecture introduces substantial computational overhead: running DeepLabv3+, SAM (ViT-H), DINOv2 (ViT-g/14), and Stable Diffusion sequentially requires significant GPU memory, likely exceeding 24GB for full-resolution inference. Fourth, the CelebAMask-HQ dataset, while the standard benchmark for face parsing, has known demographic imbalances inherited from CelebA that may affect fairness across different population groups.

6. Ethical Considerations and Responsible AI

Facial editing technology carries inherent risks of misuse for identity manipulation, non-consensual deepfake creation, and fraud. ReMaskable incorporates design-level safeguards to mitigate these risks. Watermarking: The framework includes integration points for Stable Signature (Fernandez et al., 2023), which fine-tunes the VAE decoder to embed a 48-bit binary watermark into all generated outputs, detectable even after cropping to 10% of content at greater than 90% accuracy with false positive rate below 10^{-6}. However, we note that watermarking methods remain imperfect: Stable Signature can be defeated by decoder fine-tuning, and current detection methods suffer a 45–50% AUC drop between controlled benchmarks and real-world conditions. Provenance Tracking: The architecture is designed for compatibility with the C2PA content provenance standard (spec v2.2, 2025), which uses cryptographic manifests with X.509 certificates to record content creation and modification history. This provides tamper-evident attribution but can be stripped from output files. Fairness Auditing: Following the findings of Buolamwini and Gebru (2018), who documented error rates of 20.8–34.7% for darker-skinned females versus less than 1% for lighter-skinned males in commercial facial analysis systems, any deployment of ReMaskable should include demographic performance auditing across the protected categories specified in the EU AI Act (August 2024) and NIST AI RMF 1.0.

Conclusion

This paper presents ReMaskable, a segmentation-guided latent diffusion framework for controllable facial attribute editing that explicitly decouples spatial localization from semantic generation. By combining three complementary segmentation approaches (DeepLabv3+ for semantic face parsing, SAM for promptable segmentation, and DINOv2 for boundary refinement) with CLIP-conditioned latent diffusion inpainting and dual-loss identity regularization (ArcFace + LPIPS), the framework is architected to achieve precise, localized, and identity-preserving edits. We have provided the complete mathematical formulation of each component, grounded in verified published results. The projected metrics suggest that the integrated system should achieve approximately 81% mIoU for segmentation, 0.12 LPIPS for perceptual consistency, and 0.82 ArcFace cosine similarity for identity preservation, representing improvements over individual baselines. However, we emphasize that these are projections based on component-level performance, not validated end-to-end experimental results. Completing the full training pipeline and conducting rigorous empirical evaluation is the critical next step. The modular architecture of ReMaskable, which cleanly separates segmentation, conditioning, and generation, makes it naturally extensible to video editing through temporal consistency mechanisms and to 3D avatar generation through multi-view coherence constraints. The inclusion of responsible AI safeguards including watermarking, provenance tracking, and fairness auditing is not an afterthought but a design requirement, reflecting the principle that advancing capability without advancing safety is irresponsible.

Acknowledgements

The authors gratefully acknowledge Dr. M.K. Kirubakaran for guidance and supervision throughout this research, and the Department of Artificial Intelligence and Data Science at St. Joseph’s Institute of Technology, Chennai for providing the computational infrastructure.

References

[1]. Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of CVPR, 4401–4410.
[2]. Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020). Interpreting the Latent Space of GANs for Semantic Face Editing. In Proceedings of CVPR, 9243–9252.
[3]. Abdal, R., Qin, Y., & Wonka, P. (2021). StyleFlow: Attribute-conditioned Exploration of StyleGAN-generated Images using Conditional Continuous Normalizing Flows. ACM Transactions on Graphics (SIGGRAPH), 40(3), 1–21.
[4]. Brooks, T., Holynski, A., & Efros, A. A. (2023). InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of CVPR, 18392–18402.
[5]. Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended Diffusion for Text-driven Editing of Natural Images. In Proceedings of CVPR, 18208–18218.
[6]. Kim, G., Kwon, T., & Ye, J. C. (2022). DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In Proceedings of CVPR, 2426–2436.
[7]. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In Advances in Neural Information Processing Systems 33, 6840–6851.
[8]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of CVPR, 10684–10695.
[9]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of ICML, 8748–8763.
[10]. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment Anything. In Proceedings of ICCV, 4015–4026.
[11]. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A.

(2021). Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of ICCV, 9650–9660.

[12]. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193.

[13]. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of ICCV, 3836–3847.

[14]. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of CVPR, 4690–4699.

[15]. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of CVPR, 586–595.

[16]. Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In Proceedings of CVPR, 5549–5558.

[17]. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of ECCV, 801–818.

[18]. Ho, J., & Salimans, T. (2022). Classifier-Free Diffusion Guidance. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

[19]. Song, J., Meng, C., & Ermon, S. (2021). Denoising Diffusion Implicit Models. In Proceedings of ICLR.

[20]. Zhu, P., Abdal, R., Qin, Y., & Wonka, P. (2020). SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. In Proceedings of CVPR, 5104–5113.

[21]. Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In Proceedings of ICCV.

[22]. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of FAT*, 77–91.

[23]. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of ICCV, 1–11.

[24]. Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Proceedings of ICML.