JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 6, pp. 67-80, June 2026
https://doi.org/10.9708/jksci.2026.31.06.067

High-Fidelity Face Swap via Prompt-Driven Inpainting and Pixel-Level Background Preservation

Moonsung Kang*, Jihoon Lee*, Seungwon Jang**, Suin Kim*, Doheun Cha*, Sangtae Ahn***

*Student, School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Korea
**Student, School of Electronics Engineering, Kyungpook National University, Daegu, Korea
***Associate Professor, School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Korea

[Abstract]

In this paper, we propose a novel pipeline that integrates mask-weighted loss and a face-aware text adapter to address the unrealistic painterly textures and unstable text guidance inherent in existing latent diffusion models during high-resolution face swapping. To restore fine facial details and photorealistic textures, we first employ a fine-tuning strategy for the U-Net using a mask-weighted loss. While this optimization enhances visual fidelity, it often leads to a degradation of semantic information or unintended background distortions. To mitigate these issues, we introduce a face-aware text adapter that dynamically calibrates the intensity of text embeddings based on the spatial proportions of the facial region, ensuring robust semantic control. Furthermore, to circumvent the inherent background information loss caused by the variational autoencoder reconstruction process, we implement a pixel-level blending strategy that directly integrates the generated face with the original background in the pixel space. Experimental results demonstrate that our proposed model significantly outperforms baseline methods across key metrics, including FID, PSNR, LPIPS, and PickScore, successfully achieving both high-quality, prompt-driven face synthesis and perfect background preservation.

▸ **Key words**: Face swap, Latent diffusion model, Mask-weighted Loss, Face segmentation, Background preservation

• First Author: Moonsung Kang, Corresponding Author: Sangtae Ahn
*Moonsung Kang (rkdanstjd1018@gmail.com), School of Electronic and Electrical Engineering, Kyungpook National University
*Jihoon Lee (leejh98123@knu.ac.kr), School of Electronic and Electrical Engineering, Kyungpook National University
**Seungwon Jang (wkd8642@naver.com), School of Electronics Engineering, Kyungpook National University
*Suin Kim (tndls142@knu.ac.kr), School of Electronic and Electrical Engineering, Kyungpook National University
*Doheun Cha (chadoheun@knu.ac.kr), School of Electronic and Electrical Engineering, Kyungpook National University
***Sangtae Ahn (stahn@knu.ac.kr), School of Electronic and Electrical Engineering, Kyungpook National University
• Received: 2026. 03. 23, Revised: 2026. 05. 29, Accepted: 2026. 06. 01.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문은 기존 잠재 확산 모델(Latent Diffusion Model)이 고해상도 얼굴 변환 시 나타내는 비현실적인 회화적 질감과 텍스트 유도 성능의 불안정성을 해결하기 위해, Mask-weighted Loss와 Face-aware Text Adapter를 결합한 새로운 파이프라인을 제안한다. 먼저, 얼굴의 미세 디테일과 실사 질감을 복원하기 위해 Mask-weighted Loss를 활용하여 U-Net을 미세 조정하는 전략을 채택하였다. 하지만 이러한 최적화 과정은 시각적 충실도는 높이는 반면, 텍스트 의미 정보가 약화되거나 배경이 원치 않게 왜곡되는 부작용이 발생한다. 이를 해결하기 위해 본 연구는 Face-aware Text Adapter를 추가로 도입하여, 얼굴 영역의 공간적 비중에 따라 텍스트 임베딩의 강도를 동적으로 조절함으로써 강건한 의미론적 제어 능력을 유지하도록 설계하였다. 또한, Variational Autoencoder 재구성 과정에서 발생하는 고유한 배경 손실을 방지하기 위해, 최종 단계에서 생성된 얼굴과 원본 배경을 픽셀 공간에서 직접 합성하는 Pixel-level Blending 기술을 적용하였다. 실험 결과, 제안하는 모델은 FID, PSNR, LPIPS, PickScore 에서 기존 모델 대비 우수한 성능을 기록하였으며, 텍스트 조건에 부합하는 고품질 얼굴 생성과 배경 유지를 동시에 달성하였다.

▸**주제어**: 얼굴 변환, 잠재 확산 모델, 마스크 가중 손실, 얼굴 분할, 배경 보존

I. Introduction

텍스트 기반 얼굴 편집은 사용자가 제공한 텍스트 프롬프트에 따라 얼굴 이미지의 특정 속성을 자연스럽게 변환하는 기술로, 메타버스 기반의 가상 아바타 제작, 영상 편집, 맞춤형 광고 등에서 핵심 기반 기술로 주목받고 있다. 그러나 얼굴 이미지는 인간이 가장 예민하게 인지하는 시각 자극 중 하나로[1], 피부의 미세한 질감 왜곡, 조명 불일치, 기하학적 비대칭만으로 시각적 불쾌감을 유발한다. 따라서 원본 이미지의 조명이나 자세 등 시각적 맥락을 훼손하지 않으면서 타겟 얼굴 영역만을 자연스럽게 변환 및 편집하는 것은 컴퓨터 비전 분야의 핵심 과제로 자리 잡았다.

얼굴 이미지 편집 기술은 크게 두 가지 흐름으로 발전해 왔다. 첫째로, 참조 이미지 기반 방법[2,3]은 특정 참조 이미지를 활용해 타겟 얼굴의 신원이나 스타일을 전이하는 방식으로 Generative Adversarial Networks(GAN)[4] 기반 모델에서 시작하여 Latent Diffusion Model(LDM)[5] 기반 Image to Image 방식으로 발전하여 원본 구조 및 정체성을 정밀하게 전이하는 데 강력한 성능을 보인다. 그러나 이러한 방법은 반드시 참조 이미지가 필요하여 사용자의 창의적 편집 자유도가 제한된다. 둘째로, 텍스트 기반 Inpainting 방법들 [6,7,8,9]은 참조 이미지 없이 텍스트 프롬프트만으로 표정, 나이, 헤어스타일 등 특정 속성을 직관적으로 편집하는 방식이다. 최근 LDM의 발전에 힘입어 복잡한 전처리 없이 유연한 의미론적 제어가 가능해졌다.

그러나 범용 확산 모델을 이용한 텍스트 기반 방식을 고해상도 실사 얼굴 편집에 직접 적용할 경우 세 가지 고질적인 한계에 직면한다 [3,10,11,12,13]. 첫째, 고유의 실사 질감이 훼손되어 비현실적인 회화적 질감이 생성된다. 둘째, 텍스트 조건이 의도치 않게 배경까지 변형시키는 제어의 불안정성이 나타난다. 셋째, Variational Autoencoder(VAE)[14]를 통해 이미지 전체를 재구성하는 과정에서 고주파 정보가 손실되어, 텍스처와 색감이 왜곡되고 생성된 얼굴과 원본 배경 사이 경계에서 시각적 이질감이 나타나는 문제가 뒤따른다. 이러한 한계를 개선하고 텍스트 기반의 사실적인 얼굴 편집을 수행하기 위해 새로운 파이프라인 Fig. 1을 제안한다.

본 연구를 통해 고도화된 시각적 충실도가 요구되는 메타버스 가상 아바타, 고품질 영상 및 미디어 콘텐츠 편집, 맞춤형 광고 편집 등 다양한 산업 분야에서 핵심 기반 기술로 활용될 수 있을 것으로 기대된다.

II. Related Works

2.1 Face Swapping and Editing Methods

FaceShifter[2]나 SimSwap[15] 등 GAN[4] 기반 얼굴 변환 방법들은 얼굴 전체를 재합성하는 구조로 인해 배경 보존과 경계면 안정성에 한계를 보이며, 잠재 공간 조작의 한계로 인해 텍스트를 통한 세밀한 속성 제어가 어렵다. 실제로 CA-Edit[16]는 기존 확산 모델이 국소 얼굴 속성 편집에서 어려움을 겪는 핵심 원인으로, 학습 데이터에 국소

얼굴 속성을 정밀하게 기술하는 텍스트 캡션이 부족하다는 점을 실증적으로 지적한 바 있다. 최근 이러한 한계를 극복하기 위해 DiffFace[17]나 DiffSwap[18]과 같이 확산 모델을 적용한 얼굴 변환 및 편집 연구들이 다수 제안되어 시각적 품질이 크게 향상되었다. 나아가 Face-Adapter[19]는 사전학습된 확산 모델에 경량 어댑터를 결합하여 ID-속성-구조를 분리 제어함으로써 얼굴 재연과 편집을 하나의 모델에서 동시에 수행하였으며, Realistic and Efficient Face Swapping[20]은 얼굴 변환을 자기지도 인페인팅 문제로 재정의하고 학습 중 다단계 DDIM 샘플링을 도입하여 동일성 보존과 지각적 유사성을 강화하였다. 그러나 이러한 방법들 역시 주로 얼굴 영역의 생성 품질 향상에 초점을 맞추고 있으며, 원본 배경 보존과 경계 일관성 문제는 충분히 다루어지지 않았다. 특히 잠재 공간 기반 LDM의 Inpainting 방법론들 [21,22,23,24]에 의하면 후술할 VAE의 구조적 특성으로 인해 배경 텍스처 손실과 경계 불일치 문제가 반복적으로 발생하며, 이는 정교한 얼굴 편집을 방해하는 요소가 된다.

2.2 Diffusion Models & Latent Space Reconstruction

확산 모델(Diffusion Model)은 데이터에 Gaussian Noise를 점진적으로 추가하는 순방향 과정(Forward Process)과 이를 단계적으로 제거하여 원본 데이터를 복원하는 역과정(Reverse Process)을 학습한다. DDPM[25]은 노이즈 예측기 ϵ_θ를 최적화하기 위해 평균 제곱 오차(MSE) 기반 손실 함수인 식(1)을 사용하며, 이때 역과정의 핵심 백본으로 U-Net 구조가 널리 활용된다.

$$L = E_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad (1)$$

또한, DDIM[26]과 같은 결정론적 샘플링 기법은 Non-Markovian 과정을 도입하여 더 적은 추론 단계만으로도 고품질의 이미지 생성을 가능하게 하였다. Inpainting 분야에서는 RePaint[22] 와 같이 픽셀 공간에서 확산 과정을 제어하여 원본 배경을 유지하려는 시도가 있었으나, 이는 막대한 계산 비용을 수반한다는 한계가 있다.

이를 해결하기 위해 고해상도 픽셀 공간을 VAE[14]를 통해 압축된 저차원 잠재 공간으로 옮긴 잠재 확산 모델(LDM)[5]이 널리 사용되고 있다. 그러나 국소 편집(Local Editing) 시 마스크 영역만 수정하더라도 전체 이미지가 VAE 인코딩-디코딩 과정을 거쳐야 하므로 불가피한 재구성 오차가 발생한다. Blended Latent Diffusion(BLD)[3]은 잠재 공간 내 노이즈 혼합을 통해 경계를 완화하고자 하였으나, 최종 결과가 동일한 VAE 디코더를 통과하기 때문에 배경 열화 문제를 근본적으로 해결하지 못했다. 이러한 문제는 최근 연구에서도 지속적으로 보고되고 있다. ASUKA[27] 는 잠재 기반 인페인팅 모델에서 VAE의 불완전한 재구성과 생성 잠재 -실제 잠재 간 분포 격차로 인해 마스크 영역과 비마스크 영역 사이에 색상 불일치가 체계적으로 발생함을 실험적으로 분석하고, VAE 디코더를 지역 조화(local harmonization) 모듈로 재설계하여 색상 편차를 줄이는 후처리 기법을 제안하였다. 또한 Your Latent Mask is Wrong[28]은 잠재 공간에서의 마스크 기반 선형 보간이 픽셀 공간 합성과 본질적으로 비동치임을 실증적으로 분석하고, 경량 트랜스포머를 통해 마스크 경계 오차를 대폭 감소시켰다. 이는 잠재공간 조작만으로는 픽셀 수준의 배경 일관성을 보장하기 어렵다는 구조적 한계를 시사하며, 본 연구가 최종 합성을 픽셀 공간에서 수행하는 설계 근거를 뒷받침한다.

2.3 Mask-guided and Region-aware Learning for Image Editing

확산 모델 기반 Inpainting에서 배경 보존과 얼굴 세부 복원은 여전히 중요한 과제로 남는다. 식(1)에서 정의된 표준 확산 모델의 목적함수 L은 이미지 내 모든 픽셀의 오차를 동일한 가중치로 합산하므로, 얼굴과 같이 시각적으로 중요한 국소 영역이 충분히 강조되지 않는다.

이를 극복하기 위해 BLD[3] 이나 DiffEdit[29]은 추론 단계에서 공간적 마스크를 활용하여 생성 영역을 유도하거나 배경 노이즈를 치환하는 방식을 제안하였다. 또한 BiSeNet[30]과 같은 의미론적 분할 모델을 결합하여 영역 인식 능력을 향상시키려는 시도도 있었다. 최근에는 모델 구조 자체에 마스크 정보를 깊이 통합하는 방향으로 발전하고 있다. BrushNet[11]은 사전학습된 확산 모델에 plug-and-play 방식으로 결합 가능한 dual-branch 구조를 제안하여, 마스크 이미지 특징과 노이즈 잠재를 분리 처리하고 밀집 픽셀 단위 제어(dense per-pixel control)를 통해 학습 부담을 경감함으로써 이미지 품질, 마스크 영역 보존, 텍스트 일관성 등 다수의 핵심 지표에서 기존 모델을 능가하였다. 하지만 기존의 접근법들은 주로 추론 단계에서의 마스크 조작이나 범용 인페인팅을 위한 구조적 확장에 그쳐, 얼굴 중심 편집 환경에서 요구되는 정밀한 복원 성능을 보장하기에 부족하다.

이에 본 연구에서는 모델 학습 단계에서부터 얼굴 영역에 대한 학습 집중도를 동적으로 조절하는 마스크 가중 노이즈 예측 손실(Mask-weighted Loss)을 제안하여 국소 편집 품질을 향상시킨다.

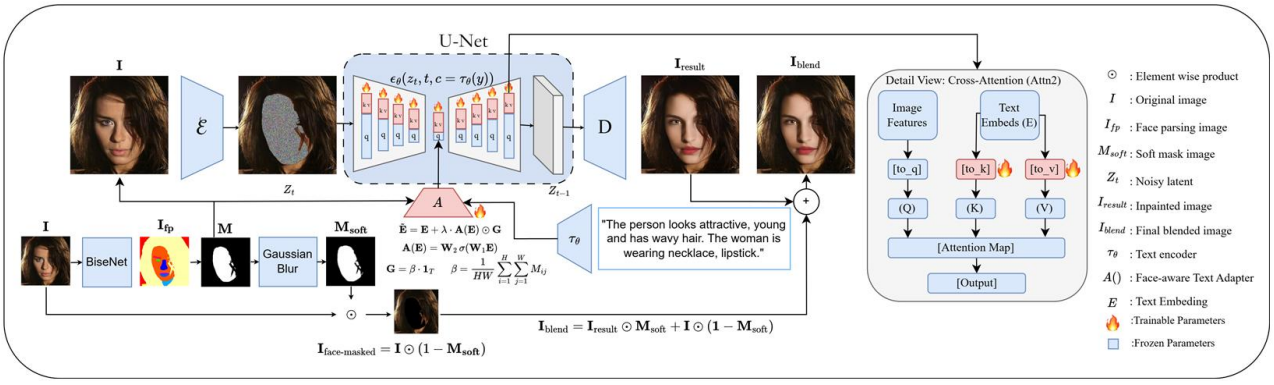


Fig. 1. Proposed Architecture. The pipeline is designed to transform only the facial region under text guidance while strictly preserving background integrity. To this end, it precisely isolates the facial area from the input image I, concentrates the learning signal on this region in the latent space, modulates text conditioning according to the facial-area proportion, and finally composes the output in pixel space. Specifically, the semantic face mask M extracted by BiSeNet is expanded into a soft mask M_{soft} via Gaussian blur, simultaneously serving as a weighting map for training and a blending matte for inference. During denoising of the latent representation Z_T, a mask-weighted loss prioritizes facial fidelity, while the Face-aware Text Adapter A rescales the text embedding E into Ê using a scale factor G derived from the facial-area ratio, balancing text guidance with spatial relevance. Training is restricted to this adapter and the cross-attention key/value projections, leaving the pre-trained generative prior intact. The denoised result I_{result} is then linearly blended with the original image I in pixel space using M_{soft} as an alpha matte, explicitly bypassing VAE decoding of the background and thereby producing the final output I_{blend} with consistent background texture and color.

2.4 Text Conditioning and Adaptation in Diffusion Models

텍스트 조건(Textual Condition)은 얼굴 편집 결과의 의미적 일관성을 좌우하는 핵심 요소이다. 기존 LDM은 고정된 CLIP[31] 텍스트 인코더를 사용하며, 텍스트 정보가 Cross-Attention[32] 메커니즘을 통해 이미지 전체 영역에 전역적으로 투영되는 특성을 갖는다[10]. 이로 인해 특정 도메인에 대한 미세 조정 시, 시각적 복원력과 의미적 표현 간의 불균형이 발생하거나 얼굴 속성 변경이 의도치 않게 배경까지 영향을 미치는 문제가 발생한다.

이를 해결하기 위해 거대 모델의 파라미터를 효율적으로 튜닝하는 PEFT 기법들이 연구되었다. Houlsby[33]의 Adapter, ControlNet[21], T2I-Adapter[34] 등이 대표적으로, 각각 경량 어댑터 삽입, Zero-Convolution 기반 구조적 조건 주입 방식을 통해 조건부 제어 능력을 향상시켰다. 특히 IP-Adapter[35]은 이미지 특징과 텍스트 특징에 대한 Cross-Attention 레이어를 분리하는 설계를 통해 적은 파라미터만으로 사전학습된 확산 모델에 이미지 프롬프트 기능을 부여하면서도 기존 제어 도구와의 호환성을 유지하였다. 국소 영역 편집을 위해 Prompt-to-Prompt[10]는 Cross-Attention map을 직접 조작하는 방식을 제안했으나, 프롬프트 간의 엄밀한 단어 정렬이 요구된다는 제약이 있다. Textual Inversion[36]이나 DreamBooth[37] 역시 다수의 참조 이미지가 필요하다는 제약이 있다. 이후 Face2Diffusion[38]은 학습 과정에서 정체성 무관(identity-irrelevant) 정보를 제거하여 과적합을 방지하고, 인코딩된 얼굴의 편집 가능성을 향상시켰으나, 추론시 대상 인물의 참조 이미지를 반드시 필요로하며 텍스트 만으로 임의의 얼굴 속성을 유연하게 제어하는데에는 한계가 있다.

요약하면, 기존 연구들은 이미지 내에서 특정 타겟 영역의 공간적 비중에 따라 텍스트의 영향력을 동적으로 조절하는 메커니즘이 부족하다. 본 연구에서는 마스크 가중 학습과 결합하여 얼굴 영역 비율에 따라 텍스트 임베딩을 최적화하는 Face-aware Text Adapter를 도입한다.

III. The Proposed Scheme

3.1 Architecture overview

본 연구에서는 Latent Diffusion 기반 Inpainting 과정에서 발생하는 비현실적인 회화풍 텍스처와 배경 왜곡 문제를 줄이기 위해, 정밀한 얼굴 의미론적 분할, 마스크 가중 학습, 얼굴 인지 텍스트 조건화, 소프트 마스크 기반 픽셀 수준 합성을 통합한 얼굴 변환(Face Swap) 파이프라인(Fig. 1)을 제안한다. 본 파이프라인은 얼굴 고유의 구조적 디테일과 배경의 무결성을 동시에 보존하기 위해 다음과 같이 네 가지 핵심 단계로 구성된다.

첫째, 얼굴 마스크 추출 및 전처리: 기존의 랜드마크나 바운딩 박스 기반 방식은 비정형적인 얼굴 윤곽을 정교하게 분리하지 못해 합성 과정에서 잦은 경계 오류를 유발하는 한계가 있다. 이를 해결하기 위해 본 연구는 입력 이미지가 주어지면 가장 먼저 BiSeNet[30]을 활용하여 픽셀

단위의 정밀한 의미론적 분할(Semantic Segmentation)을 수행하고 이진 마스크 M을 추출한다. 나아가, 추후 합성 단계에서 발생할 수 있는 부자연스러운 경계 잡음을 방지하고자 마스크 외곽에 Gaussian Blur를 적용하여 부드러운 전환을 유도하는 소프트 마스크 M_{soft}를 생성한다. 이 과정은 픽셀 수준에서 타겟 영역(얼굴)과 보존 영역(배경)을 명확히 구분해주며, 생성된 마스크는 이후 파이프라인에서 학습 가중치 할당 및 추론 단계의 자연스러운 픽셀 블렌딩을 위한 핵심 기반으로 활용된다.

둘째, 잠재공간 변환 및 마스크 가중 노이즈 예측: 표준 잠재 확산 모델(LDM)은 손실 함수 계산 시 모든 픽셀에 동일한 중요도를 부여하므로, 모델의 한정된 표현력이 분산되어 얼굴의 핵심 디테일이 손실되는 현상이 발생한다. 이를 극복하기 위해 본 단계에서는 사전 학습된 VAE 인코더를 통해 입력 이미지를 잠재 표현 z_0로 변환하고 Gaussian 노이즈가 추가된 z_t를 생성한 뒤, U-Net을 거쳐 반복적인 디노이징을 수행한다. 특히 학습 단계에서는 앞서 추출한 얼굴 마스크 M을 잠재 공간 해상도의 마스크 M_Z으로 변환하여, 얼굴 영역의 노이즈 예측 오차에 더 큰 페널티를 부여하는 Mask-weighted Loss를 적용한다. 결과적으로 노이즈 예측 과정에서 얼굴 구조와 정체성 정보에 학습 신호가 강력하게 집중되며, 네트워크의 복원 능력을 타겟 영역에 우선 할당함으로써 고품질의 사실적인 얼굴 생성을 달성한다.

셋째, 얼굴 인지 텍스트 조건화: 앞선 마스크 가중 학습은 얼굴 영역의 픽셀 복원에 학습이 집중되므로, 상대적으로 텍스트 조건의 의미론적 반영이 약화되는 문제가 발생할 수 있다. 이를 해소하기 위해 본 연구에서는 사전 학습된 텍스트 인코더를 고정한 상태에서 경량 MLP 구조의 Face-aware Text Adapter를 도입한다. 해당 어댑터는 두 개의 선형 변환과 SiLU 활성화 함수로 구성되며, 원본 텍스트 임베딩을 얼굴 편집에 적합한 표현으로 변환한다. 또한 얼굴 마스크로부터 산출한 전역 스칼라 가중치를 모든 텍스트 토큰에 균등하게 적용하여, 얼굴 영역의 공간적 비중에 따라 텍스트 조건의 영향력을 동적으로 조절한다. 이를 통해 얼굴이 실제로 편집되는 경우에만 텍스트 의미가 효과적으로 강화되며, 얼굴 영역이 작거나 배경 비중이 큰 경우에는 과도한 텍스트 유도에 의한 불필요한 변형을 억제하여 전체 이미지의 시각적 일관성을 유지한다.

넷째, 픽셀 수준 블렌딩 및 최종 출력: 일반적인 Inpainting 파이프라인은 편집 대상이 아닌 배경 영역까지 VAE 디코더를 거치게 하여 필연적인 배경 텍스처 열화와 색감 변형을 수반한다. 본 연구는 이러한 정보 손실을 원천적으로 차단하기 위해 최종 합성을 잠재 공간이 아닌 픽셀 공간에서 수행한다. 디노이징이 완료된 잠재 표현을 VAE 디코더를 통해 복원한 뒤, 소프트 마스크 M_{soft}를 alpha matte로 삼아 생성된 얼굴 이미지와 원본 배경 이미지를 선형 보간 방식으로 결합한다. 이러한 과정으로 편집 대상이 아닌 배경 영역이 VAE 복원을 거치지 않도록 명시적으로 차단함으로써, 기존 잠재 확산 기반 Inpainting에서 반복적으로 발생하던 배경 텍스처 열화와 색감 변형 문제를 효과적으로 방지한다. 이와 같은 통합 파이프라인을 통해 본 연구는 확산 모델의 표현력을 유지하면서도, 얼굴 고유 특성 보존과 배경 일관성을 동시에 달성한다.

3.2 Reconstruction Ability

기존 Face Swap 파이프라인에서 활용되어 온 얼굴 랜드마크 검출이나 bounding box 기법은 얼굴의 외형을 단순한 다각형으로 근사한다. 이러한 방식은 계산 효율성은 높으나, 머리카락이나 귀 같은 비정형적인 경계면을 정교하게 분리해내지 못해 머리카락, 액세서리, 배경 일부가 얼굴로 포함되거나 반대로 실제 얼굴 구성 요소가 누락되는 문제가 발생한다. 이러한 오류를 줄이기 위해 BiSeNet[30]을 도입하여 얼굴 피부, 눈썹, 입술 등 얼굴 구성 요소를 독립적인 클래스로 Parsing하고, 이를 통해 이진 마스크 M을 생성한다. 이 과정은 실제로 편집해야 할 영역(얼굴)과 원본 상태로 보존해야 할 영역(배경)을 픽셀 단위에서 명확히 구분하는 역할을 한다. BiSeNet[30]으로 얼굴 영역을 추출한 결과와 실제 Ground Truth 이미지를 비교해 보면 실제 정답 이미지와 유사하게 이진 얼굴 마스크를 추출하는 것을 확인할 수 있다(Fig. 2).

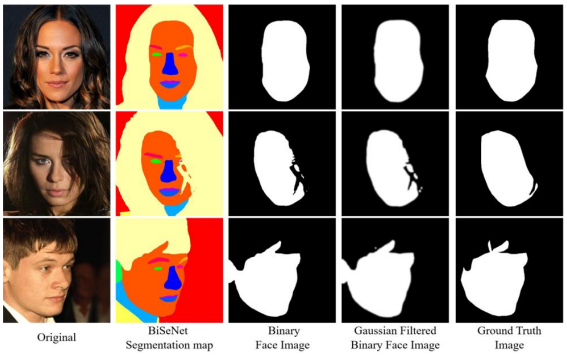


Fig. 2. Binary Face Image using BiSeNet[30]

이진 마스크는 경계에서 0과 1이 급격히 전환되는 hard edge를 가지므로, 그대로 사용할 경우 합성 단계에서 경계 잡음과 톤 불연속성이 발생하기 쉽다. 이를 완화하기 위해 마스크 경계에 Gaussian smoothing을 적용하여 소프트 마스크(soft mask) M_{soft}을 얻는다.

$$M_{soft} = G_\sigma * M \qquad (2)$$

여기서 G_σ는 표준편차 σ를 갖는 가우시안 커널이며, $*$는 합성곱 연산을 의미한다. 이 연산을 통해 마스크 경계는 [0,1]범위의 연속적인 값을 가지게 되고, 이후 inference 블렌딩 단계에서 알파 매트(alpha matte)로 활용되며, 생성 얼굴과 원본 피부 톤이 점진적으로 섞이는 자연스러운 합성을 유도한다.

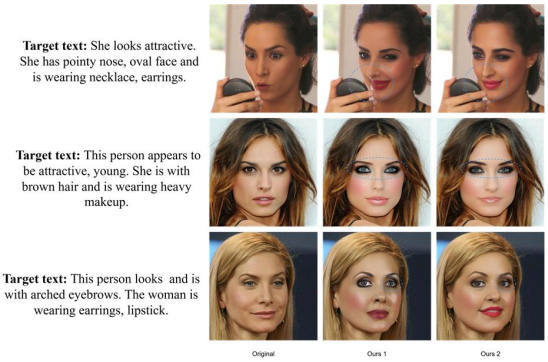


Fig. 3. Face Editing Results with Mask-weighted Loss

3.3 Face-focused Fine-tuning Strategy

기존 Stable Diffusion 기반 Inpainting 모델은 얼굴 편집 결과가 실제 사진과 같은 사실적인 품질을 완전히 재현하지 못하고, 회화적 질감이 나타나는 한계가 존재한다. 이러한 문제를 완화하기 위해 고해상도 실사 얼굴 데이터셋인 CelebA-HQ를 학습에 사용한다. 구체적으로, BiSeNet[30]을 이용한 의미론적 분할을 통해 얼굴 마스크를 정교하게 추출함과 동시에, 각 이미지에 대해 제공되는 어노테이션 정보를 기반으로 약 10개의 텍스트 프롬프트 후보를 구성하고, 학습시에는 이들 후보 중 하나를 샘플링하여 조건부 노이즈 예측 학습을 수행한다. 이와 같은 고품질 사진 데이터와 다변화된 텍스트 조건을 결합한 학습 방식은, Inpainting 과정에서 생성되는 그림 같은 질감을 개선하고, 실제 데이터 distribution과 유사한 Face Swap을 가능하게 한다.

또한, 얼굴 편집 작업에서 핵심 과제는 얼굴 영역의 디테일과 고유한 특징이다. 하지만 표준 LDM 학습 과정에서는 모든 픽셀이 동일한 중요도로 취급되어, 모델의 한정된 표현력이 얼굴과 배경에 균등하게 분산되어 얼굴 영역의 복원에 충분한 학습 신호가 집중되지 못하는 문제가 발생한다. 기본 노이즈 예측 MSE 손실은 다음과 같다.

$$L_{MSE} = E_{x_0,t,\epsilon}[\frac{1}{N}\sum_{i=1}^{N}(\epsilon_\theta(z_t,t,c)_i - \epsilon_i)^2] \qquad (3)$$

여기서 $\epsilon_\theta\ (z_t, t, c)$는 U-Net이 예측한 노이즈 값, ϵ는 타겟 노이즈를 나타낸다. 이를 해결하기 위해 BiSeNet[30]으로 얼굴 마스크 M을 추출한 뒤, latent level인 M_{latent}로 변환하여 얼굴 영역에 더 많은 학습 신호를 집중시키는 마스크 가중 노이즈 예측 손실(Mask-Weighted Loss)을 적용한다. 또한 위치별 가중치 w를 도입한 가중 MSE 손실 L_W을 정의한다.

$$w = M_{Latent} + \alpha \qquad (4)$$

$$L_{MSE} = E_{x_0,t,\epsilon}[\frac{1}{N}\sum_{i=1}^{N} w_i \odot (\epsilon_\theta(z_t,t,c)_i - \epsilon_i)^2] \qquad (5)$$

여기서 $\odot$는 요소별 곱(element-wise product)을 의미한다. 식(4)의 Latent 마스크 M_{latent}는 얼굴 영역에서 1, 배경 영역에서 0의 값을 가진다. 본 실험에서 기본 가중치 α를 0.05로 설정했다. 따라서 최종 가중치 w는 얼굴 영역에서 $1.05(1+\alpha)$, 배경 영역에서 $0.05(\alpha)$가 된다. 배경 영역을 $w \approx 0$으로 설정하여 배경에 대한 과도한 최적화를 피하면서 완전히 무시되지 않도록 학습을 유지한다. 이는 배경 영역보다 얼굴 영역의 노이즈 예측 오차에 상대적으로 훨씬 더 큰 페널티를 부여하여 U-Net이 눈, 코, 입, 윤곽과 같은 중요한 디테일을 더 높은 품질로 생성하도록 유도한다.

배경에 대한 과도한 최적화를 피하면서 완전히 무시되지 않도록 학습을 조율하며 기본 MSE Loss와 Mask-weighted Loss를 비교하는 실험을 수행했다. 그 결과, 제안한 Mask-weighted Loss를 적용했을 때 얼굴 디테일과 사실성이 뚜렷하게 향상되었으며, 특히 실험 결과 Fig. 3에서 파란색으로 표시된 영역을 통해 얼굴 가장자리, 눈 등과 같은 얼굴의 주요 특징에서 개선 효과가 확인된다. 결과적으로, 마스크 가중 손실은 한정된 모델 표현력을 얼굴 영역에 우선 할당함으로써 고품질 얼굴 이미지를 생성하며, 이후 픽셀 블렌딩 단계에서 얼굴과 얼굴 이외의 영역을 자연스럽게 연결해주는 효과를 제공한다.

3.4 Face-aware Text conditioning Strategy

기존 Stable Diffusion Inpainting 모델에서 텍스트 조건은 텍스트 인코더의 출력 임베딩을 cross-attention을 통해 U-Net 전반에 전달하는 방식으로 반영된다. 이때 텍스트 임베딩은 이미지 내 모든 공간 위치에 전역적으로 투

영되므로[10], 특정 국소 영역에 대한 편집 집중도와 텍스트 조건의 영향력 간의 공간적 불균형이 구조적으로 내재한다. 그러나 본 연구와 같이 얼굴 영역에 초점을 둔 마스크 가중 학습(Mask-weighted fine-tuning)을 수행할 경우, 이는 마스크 가중 손실로 인해 얼굴 영역의 텍스처 및 디테일 시각적 복원이 우선적으로 최적화되면서, 텍스트 조건을 통해 전달되는 의미 정보가 충분히 반영되지 않는 문제가 발생한다.

이를 해결하기 위해 본 연구에서는 Face-aware Text Conditioning 전략을 제안한다. 구체적으로, 사전 학습된 텍스트 인코더는 고정(freeze)한 상태로 유지하면서, 텍스트 임베딩을 얼굴 편집 작업에 적합한 방향으로 재조정하는 경량의 텍스트 적응 모듈(Face Text Adapter)을 추가한다. 먼저, 텍스트 인코더로부터 추출된 토큰 임베딩을 다음과 같이 정의한다. $E \in R^{T \times d}$ 여기서 T는 텍스트 토큰 길이이며, d는 텍스트 임베딩 차원을 의미한다. 본 연구에서는 Stable Diffusion의 기본 설정을 따라 T=77, d=768을 사용한다.

3.4.1 Face-aware Text Adapter

텍스트 조건을 얼굴 편집 작업에 보다 효과적으로 반영하기 위해, 경량 MLP 구조의 Face-aware Adapter A를 설계하였다. 해당 어댑터는 두 개의 선형 변환과 SiLU 활성화 함수로 구성되며, 원본 텍스트 임베딩을 얼굴 중심 표현으로 변환한다.

$$A(E) = W_2 \sigma(W_1 E) \quad (6)$$

여기서 $W_1, W_2 \in R^{d \times d}$는 학습 가능한 파라미터이며, $\sigma(\cdot) = SiLU$는 활성화 함수를 나타낸다. 이 어댑터는 텍스트 조건의 전체 의미 구조를 유지하면서도, 얼굴 생성에 중요한 표현 성분을 강조하도록 학습된다.

3.4.2 Mask-Guided Token Weighting

얼굴 영역의 공간적 중요도를 텍스트 조건에 반영하기 위해, 얼굴 마스크 $M \in (0,1)^{H \times W}$를 이용하여 전역 스칼라 가중치 β를 계산한다.

$$\beta = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{ij} \quad (7)$$

해당 값은 이미지 전체 대비 얼굴 영역의 비율을 나타내며, 이를 모든 텍스트 토큰에 동일하게 적용하기 위해 길이 T의 토큰 가중치 벡터 G를 다음과 같이 정의한다.

$$G = \beta \cdot 1_T \quad (8)$$

여기서 $1_T \in R^{T \times 1}$는 모든 원소가 1인 벡터이다. 이 설계는 얼굴의 시각적 중요도를 텍스트 조건 전반에 균등하게 전달함으로써, 특정 토큰에 대한 수동적 가중 조정 없이도 안정적인 조건 제어를 가능하게 한다.

본 연구에서는 토큰별 중요도를 별도로 예측하는 복잡한 메커니즘 대신, 얼굴 영역의 전역적 시각적 중요도를 하나의 스칼라값으로 요약하여 모든 토큰에 동일하게 적용하였다. 이는 과도한 학습 파라미터 증가를 억제하고, 제한된 데이터 환경에서도 안정적인 텍스트 조건 제어를 가능하게 한다.

3.4.3 Final Textual Conditioning Injection

최종적으로, 얼굴 마스크 기반 가중치를 적용한 텍스트 임베딩은 다음과 같이 구성된다.

$$\acute{E} = E + \lambda \cdot A(E) \odot G \quad (9)$$

여기서 λ는 텍스트 조건의 영향력을 조절하는 하이퍼파라미터이며, $\odot$는 요소별 곱을 의미한다. 본 연구에서는 실험적으로 λ=0.3으로 설정하였다. 제안하는 구조는 얼굴 영역의 공간적 비중에 따라 텍스트 조건의 영향을 동적으로 조절함으로써, 실제로 얼굴이 편집되는 경우에만 텍스트 의미가 효과적으로 강화되도록 설계되었다. 특히 얼굴 영역이 작거나 배경 비중이 큰 이미지에서 과도한 텍스트 유도에 의한 불필요한 변형을 억제하여, 전체 이미지의 생성 안정성과 시각적 일관성을 유지한다.

또한 U-Net 전체를 미세 조정하는 대신, 텍스트 조건과 직접적으로 상호작용하는 cross-attention 모듈의 key 및 value projection 층(attn2의 to_k, to_v)만을 선택적으로 fine-tuning함으로써, 사전 학습된 Stable Diffusion 모델의 표현 능력을 보존하면서도 얼굴 중심 텍스트 조건 반영 능력을 효율적으로 향상시킨다.

나아가, 제안한 Face-aware Text Adapter는 학습 단계와 추론 단계에서 동일한 방식으로 적용되며, 얼굴 마스크 기반 가중치를 통해 텍스트 조건의 강도를 일관되게 조절한다. 이를 통해 학습-추론 간 조건 불일치를 방지하고, 다양한 얼굴 크기와 구도에서도 안정적인 텍스트 기반 얼굴 편집 성능을 달성한다.

3.5 Pixel-level Blending Strategy

LDM 기반 inpainting의 핵심적인 한계는, 얼굴만 수정하더라도 전체 이미지가 다시 VAE 디코더 D를 통과하면서 배경 디테일이 손실된다는 점이다. 이 문제를 근본적으로 방지하기 위해, 최종 합성을 잠재 공간(latent space)

이 아닌 픽셀 공간(pixel space)에서 수행하는 픽셀 수준 블렌딩 전략(pixel-level blending)을 채택하였다. 생성된 이미지 I_{result}과 원본 이미지 I 그리고 소프트 마스크 M_{soft}에 대해 최종 결과 I_{blend}은 다음과 같이 정의된다.

$$I_{blend} = I_{result} \odot M_{soft} + I \odot (1 - M_{soft}) \quad (10)$$

여기서 $\odot$는 요소별 곱(Element-wise product)을 의미한다. 제안하는 픽셀 수준 블렌딩 전략(Pixel-level Blending Strategy)이 얼굴 합성 품질 및 배경 보존에 미치는 영향을 확인하기 위해, 정성적 시각화 분석과 정량적 지표 분포 분석을 수행하였다. 그 결과는 다음과 같다.

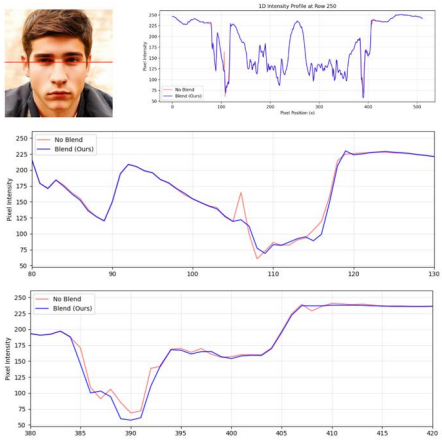


Fig. 4. Visual Continuity of the Boundary Region. The 1D intensity profile demonstrates that the proposed blending strategy ensures a smooth transition at the boundaries, effectively eliminating the spikes and artifacts observed in the non-blended version.

경계 영역의 시각적 연속성 (Visual Continuity of the Boundary Region) Fig. 4는 소프트 마스크를 활용한 경계 완화 기법은 합성 경계면에서 발생하는 아티팩트를 효과적으로 억제하는 것을 보여 준다. 1차원 픽셀 강도(1D Intensity Profile) 분석 결과에 따르면, 블렌딩을 적용하지 않은 경우(No Blend) 경계 부분에서 급격한 강도 변화와 스파이크가 발생하여 부자연스러운 외곽선이 나타난다. 반면, 제안하는 블렌딩 기법(Blend)을 적용했을 때 픽셀 강도가 훨씬 매끄럽게 연결되어, 시각적으로 자연스러운 전이를 보여준다.

배경 보존 및 구조적 무결성(Background Preservation and Structural Integrity) Fig. 5 실험을 통해 일반적인 U-Net 기반 생성 모델은 지정된 마스크 영역뿐만 아니라 수정이 불필요한 배경 영역의 픽셀값까지 함께 변형시켜 미세하지만 노이즈와 구조적 왜곡을 유발한다. 이는 디코더가 이미지의 전체 공간 범위를 재구성하면서 생기는 문제이다. 'U-Net Result - Blend'에서 확인할 수 있듯이 모델 출력물에는 미세한 노이즈와 변형이 포함되어 있다. 이에 반해, 제안하는 소프트 마스크 블렌딩 전략은 Fig. 5의 'No Blend - Blend'에서 나타나듯이, 타겟 영역의 생성 결과는 유지하면서도, 마스크 외부의 배경을 무손실로 보존함으로써 구조적 무결성을 확보한다.

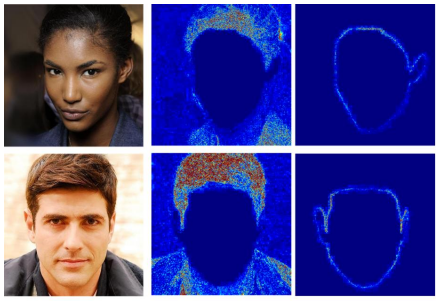


Fig. 5. Background Preservation and Structural Integrity. The proposed blending strategy effectively suppresses artifacts outside the target mask, maintaining the structural integrity of the original background.

Fig. 7 및 Fig. 8의 분포 분석과 Table. 1의 정량적 수치는 제안하는 픽셀 수준 블렌딩 전략의 유효성을 일관되게 뒷받침한다. PSNR 분포 Fig. 7에서 블렌딩 미적용(No Blend)에 대비 블렌딩 적용 시(Blend) 데이터 분포가 더 높은 수치(오른쪽)로 뚜렷하게 이동한 것을 확인할 수 있다. 이는 배경 픽셀이 원본 이미지와 비교하여 정보 손실 없이 그대로 보존됨을 정량적으로 증명한다. LPIPS 분포 Fig. 8역시 Blend 적용시 No Blend 대비 낮은 LPIPS 값(왼쪽)에 밀집된 분포를 형성하며, Table. 1의 정량적 수치에 의하면 배경의 질감과 세부 디테일을 매우 높은 지각적 충실도로 보존하고 있음을 시사한다.

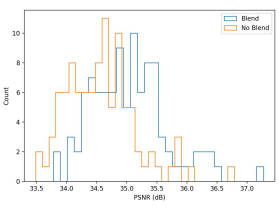


Fig. 7. Quantitative Comparison of Background PSNR Distribution.

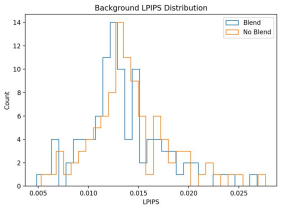


Fig. 8. Quantitative Comparison of Background LPIPS Distribution.

본 연구에서 전경 영역만을 정밀하게 분리하는 과정은 전통적인 영상 합성의 Rotoscoping과 유사한 역할을 수행한다. 이는 복잡한 배경에 분산될 수 있는 U-Net의 생성 역량을 오직 Target Text가 지시하는 얼굴영역에만 집

중시키기 위한 핵심적인 과정이다.

다만, 본 픽셀 수준 블렌딩 전략은 원본의 조명 환경과 피부 톤이 어느 정도 유지되는 텍스트 조건에서 최적의 성능을 발휘하며, 원본과 형태 및 광학적 특성이 극단적으로 달라지는 변환에서는 향후 조명 일치화(Lighting Harmonization)와 같은 추가적인 톤 매핑 기법이 요구될 수 있다.

Table 1. Background Preservation Results

	PSNR	LPIPS
No Blend	34.59 ± 0.58	0.0142 ± 0.004
Blend	**35.04 ± 0.63**	**0.0134 ± 0.004**

IV. Experiments

4.1 Implementation Details

본 실험에서는 텍스트 기반 얼굴 편집 성능을 평가하기 위해 고해상도 얼굴 이미지 데이터셋인 CelebA-HQ[39]와 CelebAMask-HQ[40]에서 제공하는 40개의 이진 속성 어노테이션을 텍스트로 변환하여 학습 및 평가에 활용하였다. 실험의 재현성을 보장하기 위해 난수 생성 시드를 42로 고정하였으며, 전체 데이터 30,000장 중 20,000장을 학습 데이터셋으로, 나머지 10,000장은 평가 데이터셋으로 분할하였다. 이는 모델 학습에 충분한 데이터를 확보하는 동시에, 다양한 얼굴 속성을 포함하는 대규모 평가 풀을 구성하기 위함이다. 최종 정량 평가는 평가 테스트셋에서 동일 시드로 무작위 선정한 텍스트-이미지 100쌍을 대상으로 수행하였다. CelebA 기반 선행 연구에서도 제한된 수의 테스트 이미지를 활용한 정량 평가가 수행된 바 있다[41].

본 연구의 모든 모델 학습 및 평가 실험은 NVIDIA RTX A6000 GPU (49GB VRAM) 환경에서 수행되었으며, 학습 과정에서 약 31.1GB의 VRAM이 사용되었다. 제안 모델은 CUDA 11.8 환경에서 PyTorch 2.5.1 및 Torchvision 0.20.1을 기반으로 구현되었으며, 잠재 확산 모델 과 어텐션 메커니즘은 각각 diffusers 0.31.0과 transformers 4.47.0을 활용하여 구성하였다.

전처리 과정에서는 모든 이미지와 마스크를 동일한 해상도 512×512크기로 조정한 후. 확산 모델의 학습 안정성을 위해 RGB 이미지의 픽셀 값을 평균 0.5, 표준편차 0.5를 기준으로 정규화하여 [-1, 1] 범위로 변환하였다. 모델의 일반화 성능을 향상시키기 위해 학습 단계에서 이미지-마스크 쌍에 대해 50%의 확률로 무작위 수평 뒤집기를 적용하였다. 편집 영역을 지정하는 마스크 데이터의 경우 픽셀값 0.5를 임계값으로 하여 이진화하여 편집 대상 영역을 명확하게 정의하였다.

학습 단계에서는 안정적인 수렴을 위해 U-Net 의 교차 어텐션(Cross-attention) 레이어와 Face-aware Text Adapter의 가중치만을 미세 조정하였다. 옵티마이저는 AdamW를 적용하였으며, 학습률은 U-Net: 5×10^{-5}, Face-aware Text Adapter: 1×10^{-4}로 설정하였다. 배치 크기는 16으로 총 10 에폭 학습을 수행하였으며, 전체 학습 소요 시간은 약 12시간이다.

4.2 Quantitative Comparisons

성능 평가는 생성 품질과 배경 보존을 함께 고려하기 위해 FID(Fréchet Inception Distance)[42], PSNR(Peak Signal-to-Noise Ratio)[43], LPIPS(Learned Perceptual Image Patch Similarity)[44], PickScore[45] 네 가지 평가 지표로 진행하였다. FID는 생성된 얼굴의 사실성 및 분포 유사도를 평가하는 지표로 값이 낮을수록 성능이 우수하며, PickScore[28]는 입력 프롬프트와 생성 이미지 간 의미론적 정합성을 측정하는 지표로 값이 높을수록 좋다. 배경 보존능력은 PSNR[43]과 LPIPS[44]를 통해 평가하며, PSNR[43]은 픽셀 단위 신호 대 잡음비를 나타내며 높을수록 배경 정보가 잘 보존되었음을 의미한다. LPIPS[27]는 인간의 시각적 인지에 기반한 지각적 유사도 지표로 값이 낮을수록 원본과의 유사성이 높음을 나타낸다. 본 연구에서는 이미지 편집 태스크의 특성을 고려하여 VGG 백본 기반을 사용하였으며 값이 낮을수록 원본과의 유사성이 높음을 나타낸다.

정량 평가를 위한 비교 모델로는 Stable Diffusion Inpainting 모델 v1.5, Blended Latent Diffusion(BLD)[3], BrushNet[11], Diptych Prompting[46]을 선정하였다. 모든 비교 모델은 공개된 코드 기반으로 동일한 추론 환경에서 실행되었다. 평가는 CelebA-HQ[24] 데이터셋 10,000장 중 무작위로(seed=42) 추출한 100장의 이미지를 대상으로 수행하였으며, 각 이미지에 대응하는 CelebAMask-HQ[40] 속성 기반 텍스트 프롬프트를 편집 조건으로 활용하여 각 지표를 산출하였다. Table. 2는 그 결과를 나타낸다.

Table 2. Quantitative Comparisons results

CelebA	FID ↓	PSNR[dB] ↑	LPIPS ↓	Pick Score ↑
SD v1.5 inpaint	91.785	30.779	0.266	15.226
BLD	94.442	30.551	0.322	9.177
BrushNet	91.895	33.216	0.168	17.449
Diptych Prompting	75.690	28.054	0.347	**33.842**
Ours	**63.981**	**33.538**	**0.143**	24.306

Table 2.의 정량적 평가 결과에 따르면, 제안하는 방법이 전반적으로 가장 우수한 성능을 달성하였다. 비교 모델 중 Diptych Prompting이 가장 높은 PickScore를 기록하며 텍스트 프롬프트의 의미를 정확하고 강력하게 반영하는 우수한 성능을 보였다. 다만, 해당 모델은 PSNR과 LPIPS 지표에서는 상대적으로 낮은 수치를 기록하였는데, 이는 텍스트 반영도를 극대화하는 과정에서 원본 이미지의 조명, 피부 톤 및 배경 문맥이 일정 부분 변형되는 트레이드오프를 나타낸다.

반면, 제안한 방법은 PSNR, LPIPS 지표에서 최고 성능을 달성하며, 원본 이미지의 공간적 구조와 조명 정보를 매우 높은 수준으로 보존함을 입증하였다. PickScore에서는 Diptych Prompting에 미치지 못하나, PSNR 및 LPIPS에서의 우수한 성능을 함께 고려하면 제안 모델이 원본의 시각적 일관성과 텍스트 반영 사이에서 가장 안정적인 균형을 달성하였음을 확인할 수 있다.

4.3 Computational Efficiency

생성품질과 더불어 실제 서비스 배포 가능성을 평가하기 위해 단일 이미지 추론 시 요구되는 VRAM 및 추론 시간을 비교 분석하였다 Table. 3

Table 3. Comparison of Computational Efficiency

	VRAM [MiB]	Inference Time [sec]
SD v1.5 inpaint	7,532	1.37
BLD	**3,394**	**1.6**
BrushNet	4,484	3.5
Diptych Prompting	46,114	180
Ours	7,794	2.5

앞선 평가 지표 중 PickScore 에서 가장 높은 수치를 보인 Diptych Prompting의 경우, 텍스트 정합성이 높은 고품질의 이미지를 생성하지만, 복잡한 파이프라인 구조로 인해 단일 이미지 추론에 약 180초의 시간과 46,114 MiB의 VRAM을 요구하여 실용적인 배포 환경에서의 활용에 제약이 따른다.

반면, 제안하는 방법은 추론 시간 2.5초, VRAM 사용량 7,794 MiB로, Diptych Prompting 대비 추론 시간은 약 72배 단축, VRAM 사용량은 약 5.9배 감소하였다. 이는 SD v1.5 Inpaint(1.37초, 7,532 MiB), BLD(1.6초, 3,394 MiB), BrushNet(3.5초, 4,484 MiB)와 비교하여 기존 경량 베이스라인 모델들의 연산 효율성에 준하는 결과이다. 즉, 제안 모델은 일반 소비자용 하드웨어 환경에서도 구동 가능하며, 앞서 입증된 고품질의 생성 품질 및 배경 보존 능력을 실용적인 연산 비용 내에서 제공할 수 있음을 보여준다.

4.4 Qualitative Comparison

정량 평가 결과를 시각적으로 검증하기 위해 Fig 9.와 Fig 10.에서는 대표적인 정성 비교 결과를 제시한다.

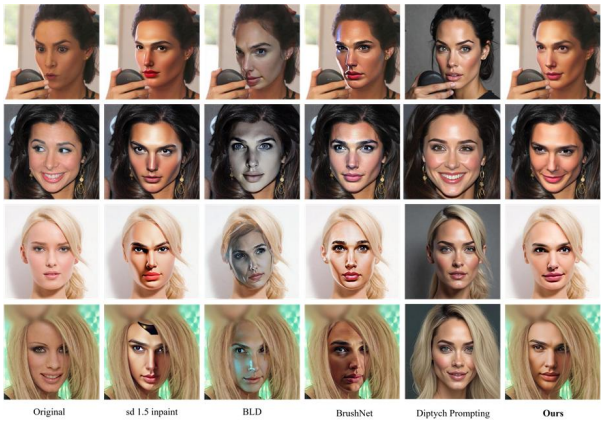


Fig. 9. Visual comparison with prompt “a photorealistic face of Gal Gadot, strong facial features, natural makeup, realistic studio lighting”

Fig. 9에서 CelebA-HQ 이미지와 프롬프트 “a photorealistic face of Gal Gadot, strong facial features, natural makeup, realistic studio lighting”을 조건으로 하여 Stable Diffusion Inpainting v1.5, BLD[3], BrushNet[11], Diptych Prompting[46] 그리고 제안한 방법의 결과를 비교한다. SD v1.5 Inpaint와 BrushNet는 얼굴 정체성이 불안정해지거나 텍스처가 회화적으로 변형되는 경향을 보이며, BLD는 얼굴 톤이 비자연스럽게 회색화되거나 세부 디테일이 손실되는 현상이 관찰된다.

최신 방법론인 Diptych Prompting의 경우, 텍스트에 명시된 속성을 가장 강력하게 반영하여 시각적으로 뚜렷하고 사실적인 변화를 만들어낸다. 다만 프롬프트의 시각적 요소를 극대화 하는 과정에서 원본 이미지 고유의 조명 톤이나, 헤어스타일, 배경 등의 시각적 문맥이 전반적으로 변형되는 트레이드오프가 확인되었다. 반면 제안한 방법은 얼굴 영역에서 텍스트에 명시된 속성을 충실히 반영하면서도, 피부 톤과 질감이 자연스럽게 유지되어 원본 이미지와의 시각적 일관성을 높게 보존하였다.

Fig. 10은 “round face”, “high nose bridge”, “full lips”와 같이 얼굴 형태와 직접적으로 관련된 국소적 속성을 텍스트 조건으로 추가한 결과를 비교한다. 이 설정에서 BLD와 BrushNet은 얼굴 영역이 과도하게 변형되거나 비

정상적인 색상 분포를 보이며, Diptych Prompting은 원본의 배경 및 조명 정보가 크게 변화하는 경향을 보였으며, 전반적인 분위기 생성 능력에 비해 세밀한 국소 형태 제어에서 한계를 나타냈다. 반면, 제안한 방법은 원본 이미지와의 시각적 일관성을 유지함과 동시에, 프롬프트로 지정된 코의 높이나 입술의 형태적 속성을 주변 구조와 이질감 없이 현실적인 형태로 반영됨을 확인할 수 있다.

Fig. 10. Visual Comparison with face detail prompt.

V. Conclusions

본 논문에서는 잠재 확산 모델 기반 얼굴 편집 시 발생하는 비현실적인 텍스처와 배경 왜곡 문제를 해결하기 위해, 정밀한 영역 분할과 공간 인지적 조건화를 결합한 새로운 얼굴 변환 파이프라인을 제안하였다.

제안하는 방법은 학습 단계와 추론 단계의 핵심 전략으로 구성된다. 학습 단계에서는 마스크 가중 손실(Mask-weighted Loss)을 적용하여 모델의 학습 역량을 얼굴 영역에 집중시키고, 텍스트 조건이 배경까지 전역적으로 투영되는 문제를 해결하기 위해 Face-aware Text Adapter를 도입하였다. 추론 단계에서는 BiSeNet[30] 기반 정밀한 의미론적 분할 마스크와 Gaussian Blur를 결합한 픽셀 수준 블렌딩(Pixel-level Blending)을 수행하여 VAE[디코딩 과정의 재구성 오차로 인한 배경 손실을 효과적으로 억제하였다. CelebA-HQ 데이터셋 기반 실험에서 제안 모델은 이미지 생성 품질 (FID), 배경 보존 (PSNR, LPIPS) 측면에서 비교모델 대비 최고 성능을 달성하였으며, 텍스트 정합성(PickScore) 측면에서도 경쟁력 있는 수준을 유지하면서 시각적 일관성과 텍스트 반영 간의 안정적인 균형을 달성하였음을 입증하였다.

다만, 본 연구는 Stable Diffusion의 사전 학습된 내부 지식에 의존하는 구조적 특성상, 학습 데이터에 충분히 포함되지 않은 인물에 대해서는 정체성 보존 및 생성 성능이 저하될 수 있다는 한계를 지닌다. 또한, 의미론적 분할 네트워크의 초기 마스크 정확도에 따라 블렌딩 경계의 품질이 영향을 받을 수 있다는 점도 한계로 남는다.

이상의 결과는 확산 모델의 의미론적 제어 능력을 활용하면서도 구조적 왜곡을 최소화하는 안정적인 파이프라인의 실용적 가능성을 보여주며, 향후 고해상도 정밀 이미지 편집 기술의 발전에 기여할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

This research, undertaken at Kyungpook National University, was supported by the Regional Innovation System & Education (RISE) program through the Daegu RISE Center, funded by the Ministry of Education (MOE) and the Daegu Metropolitan City, Republic of Korea (2025-RISE-03-001). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2025-02214941).

REFERENCES

[1] Mori, Masahiro, Karl F. MacDorman, and Norri Kageki. "The uncanny valley [from the field]." IEEE Robotics & automation magazine 19.2 (2012): 98-100. DOI: https://doi.org/10.1109/MRA.2012.2192811

[2] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping", arXiv preprint arXiv:1912.13457, December 2020. DOI: https://doi.org/10.48550/arXiv.1912.13457

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski, "Blended Latent Diffusion", ACM Transactions on Graphics (TOG), Vol. 42, No. 4, pp. 1-11, August 2023. DOI: https://doi.org/10.1145/3592450

[4] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020). DOI: 139-144.https://doi.org/10.1145/3422622

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer, "High-resolution image synthesis with latent diffusion models", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695, June 2022. DOI: https://doi.org/10.48550/arXiv.2112.10752

[6] Patashnik, Or, et al. "Styleclip: Text-driven manipulation of stylegan imagery." Proceedings of the IEEE/CVF international

conference on computer vision. 2021. DOI: https://doi.org/10.48550/arXiv.2103.17249

[7] Xia, Weihao, et al. "Tedigan: Text-guided diverse face image generation and manipulation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. DOI: https://doi.org/10.48550/arXiv.2012.03308

[8] Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. "Diffusionclip: Text-guided diffusion models for robust image manipulation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. DOI: https://doi.org/10.48550/arXiv.2110.02711

[9] Yue, Dongxu, et al. "Chatface: Chat-guided real face editing via diffusion latent space manipulation." arXiv preprint arXiv:2305.4742 (2023). DOI: https://doi.org/10.48550/arXiv.2305.14742

[10] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022). DOI: https://doi.org/10.48550/arXiv.2208.01626

[11] Ju, Xuan, et al. "Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024. DOI: https://doi.org/10.1007/978-3-031-72661-3_9

[12] Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. DOI: https://doi.org/10.48550/arXiv.2211.09800

[13] Kawar, Bahjat, et al. "Imagic: Text-based real image editing with diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. DOI: https://doi.org/10.48550/arXiv.2210.09276

[14] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes", Proceedings of the International Conference on Learning Representations (ICLR), 2014. DOI: https://doi.org/10.48550/arXiv.1312.6114

[15] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, "SimSwap: An Efficient Framework For High Fidelity Face Swapping", Proceedings of the 28th ACM International Conference on Multimedia (MM '20), pp. 2003–2011, October 2020. DOI: https://doi.org/10.1145/3394171.3413630

[16] Xian, Xiaole, et al. "CA-edit: Causality-aware condition adapter for high-fidelity local facial attribute editing." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 8. 2025. DOI: https://doi.org/10.1609/aaai.v39i8.32928

[17] Kim, Kihong, et al. "Diffface: Diffusion-based face swapping with facial guidance." Pattern Recognition 163 (2025): 111451. DOI: https://doi.org/10.1016/j.patcog.2025.111451

[18] Zhao, Wenliang, et al. "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. DOI: https://doi.org/10.1109/CVPR52729.2023.00828

[19] Han, Yue, et al. "Face-adapter for pre-trained diffusion models with fine-grained id and attribute control." European conference on computer vision. Cham: Springer Nature Switzerland, 2024. DOI: https://doi.org/10.48550/arXiv.2405.12970

[20] Baliah, Sanoojan, et al. "Realistic and efficient face swapping: A unified approach with diffusion models." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025. DOI: https://doi.org/10.1109/WACV61041.2025.00112

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847, October 2023. DOI: https://doi.org/10.48550/arXiv.2302.05543

[22] Lugmayr, Andreas, et al. "Repaint: Inpainting using denoising diffusion probabilistic models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. DOI: https://doi.org/10.48550/arXiv.2201.09865

[23] Wang, Su, et al. "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. DOI: https://doi.org/10.48550/arXiv.2212.06909

[24] Xie, Shaoan, et al. "Smartbrush: Text and shape guided object inpainting with diffusion model." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. DOI: https://doi.org/10.48550/arXiv.2212.05034

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 6840–6851, December 2020. DOI: https://doi.org/10.48550/arXiv.2006.11239

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models", International Conference on Learning Representations (ICLR), May 2021. DOI: https://doi.org/10.48550/arXiv.2010.02502

[27] Wang, Yikai, et al. "Towards enhanced image inpainting: Mitigating unwanted object insertion and preserving color consistency." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. DOI: https://doi.org/10.48550/arXiv.2312.04831

[28] Bradbury, Rowan, and Dazhi Zhong. "Your Latent Mask is Wrong: Pixel-Equivalent Latent Compositing for Diffusion Models." arXiv preprint arXiv:2512.05198 (2025). DOI: https://doi.org/10.48550/arXiv.2512.05198

[29] Couairon, Guillaume, et al. "Diffedit: Diffusion-based semantic image editing with mask guidance." arXiv preprint arXiv:2210.11427 (2022). DOI: https://doi.org/10.48550/arXiv.2210.11427

[30] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation", Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341, September 2018. DOI: https://doi.org/10.48550/arXiv.1808.00897

[31] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021. DOI: https://doi.org/10.48550/arXiv.2103.00020

[32] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017). DOI: https://doi.org/10.48550/arXiv.1706.03762

[33] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-Efficient Transfer Learning for NLP", Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 2790–2799, June 2019. DOI: https://doi.org/10.48550/arXiv.1902.00751

[34] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie, "T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4296–4305, June 2023. DOI: https://doi.org/10.48550/arXiv.2302.08453

[35] Ye, Hu, et al. "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models." arXiv preprint arXiv:2308.06721 (2023). DOI: https://doi.org/10.48550/arXiv.2308.06721

[36] 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", Proceedings of the International Conference on Learning Representations (ICLR), May 2023. DOI: https://doi.org/10.48550/arXiv.2208.01618

[37] 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", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510, June 2023. DOI: https://doi.org/10.48550/arXiv.2208.12242

[38] Shiohara, Kaede, and Toshihiko Yamasaki. "Face2diffusion for fast and editable face personalization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. DOI: https://doi.org/10.48550/arXiv.2403.05094

[39] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017). DOI: https://doi.org/10.48550/arXiv.1710.10196

[40] Lee, Cheng-Han, et al. "Maskgan: Towards diverse and interactive facial image manipulation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. DOI: https://doi.org/10.48550/arXiv.1907.11922

[41] Martin, Ségolene, et al. "Pnp-flow: Plug-and-play image restoration with flow matching." International Conference on Learning Representations. Vol. 2025. 2025. DOI: https://doi.org/10.48550/arXiv.2410.02423

[42] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017). DOI: https://doi.org/10.48550/arXiv.1706.08500

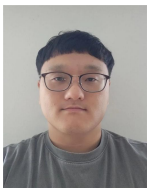
[43] Wikipedia contributors: Peak signal-to-noise ratio — Wikipedia, the free encyclopedia (2024). DOI: *https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio*

[44] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. DOI: https://doi.org/10.48550/arXiv.1801.03924

[45] Kirstain, Yuval, et al. "Pick-a-pic: An open dataset of user preferences for text-to-image generation." Advances in neural information processing systems 36 (2023): 36652-36663. DOI: https://doi.org/10.48550/arXiv.2305.01569

[46] Shin, Chaehun, et al. "Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. DOI: https://doi.org/10.48550/arXiv.2411.15466

Authors



Moonsung Kang received the B.S. degree in Electronics Engineering from Kyungpook National University, Daegu, South Korea, in 2025, where he is currently pursuing the M.S. degree. His current research interests include AI and computer vision.

Jihoon Lee received the B.S. and M.S. degrees in Electronics Engineering from Kyungpook National University, Daegu, South Korea. He is currently pursuing the Ph.D. degree in the School of Electronic and Electrical Engineering at Kyungpook National University. His current research interests include artificial intelligence and computer vision.

Seungwon Jang is currently pursuing the B.S. degree in the School of Electronics Engineering at Kyungpook National University, Daegu, South Korea. His current research interests include AI and computer vision.

Suin Kim received the B.S. degree in Electronics Engineering from Kyungpook National University, Daegu, South Korea, in 2025, where he is currently pursuing the M.S. degree. His current research interests include AI and computer vision.

Doheun Cha received the B.S. and M.S. degrees in electronics engineering from Kyungpook National University, Daegu, South Korea. He is currently pursuing the Ph.D. degree in the Department of Electronic and Electrcial Engineering at Kyungpook National University. His current research interests include AI and spiking neural networks.

Sangtae Ahn is an Associate Professor at the School of Electronic and Electrical Engineering, Kyungpook National University, South Korea. He received his Ph.D. degree at Gwangju Institute of Science and Technology, South Korea, in 2016.
His research interests include brain-inspired AI, generative AI and physical AI.