FLAG Control: 얼굴 랜드마크 기반 구조 가이던스를 활용한 Training-Free Control Diffusion 모델 성능 향상

서유정[1], 김광수[2]
[1]성균관대학교 소프트웨어학과 학부생
[2]성균관대학교 소프트웨어융합대학 교수

syj000229@skku.edu
kim.kwangsu@skku.edu

FLAG Control: Enhancing Training-Free Control Diffusion with Face Landmark-Aware Spatial Guidance

Yujeong Seo[1], Kwangsu Kim[2]
[1]Dept. of Computer science and engineering, Sungkyunkwan University
[2]College of Computing and Informatics, Sungkyunkwan University

요 약

FreeControl은 재학습 없이 다양한 조건에 따라 텍스트-이미지 생성을 제어할 수 있는 diffusion 기반 모델로, 포즈나 형태 등은 구조 가이던스, 스타일 · 색감 · 이미지 구도 등은 외향 가이던스를 통해 제어된다. 그러나 인물 생성 시 얼굴의 정체성(눈 · 코 · 입의 위치, 비율, 구도 등)을 유지하는 데 한계가 있으며, 배경이나 손 등 다른 시각적 요소와 혼재되어 얼굴 표현이 왜곡되는 문제가 발생한다. 이는 기존 가이던스만으로는 얼굴과 같은 정밀한 영역을 충분히 제어하지 못하고, 전경(foreground)에 대한 집중도가 낮기 때문이다. 본 논문에서는 얼굴 랜드마크 기반의 새로운 가이던스인 FLAG(Face Landmark-Aware Guidance)를 제안한다. FLAG는 참조 이미지에서 추출한 얼굴 랜드마크를 기반으로 피처의 일관성을 유지하도록 유도하며, Soft Attention Mask를 통해 얼굴 영역에 집중된 손실을 적용함으로써 자연스럽고 안정적인 인물 표현을 가능하게 한다.

1. 서론

최근 텍스트-투-이미지 모델의 급속한 발전으로, 텍스트 입력만으로도 고품질의 이미지를 생성할 수 있는 다양한 확산 기반 프레임워크가 등장하였다. 이 과정에서 정체성 보존을 위한 접근은 사전 학습된 모델에 추가 학습 여부에 따라 학습 기반(training) 방식과 비학습 기반(training-free) 방식으로 나눌 수 있다. 전자는 DSG(Disentangled Semantic Guidance) [1], BLIP(Bootstrapping Language-Image Pre-training) [2] 등과 같이 추가적인 파인 튜닝을 활용하여 구조 정보를 보존한다. 그러나 이 방식은 적용에 앞서 별도의 학습 과정이 필요하며, 조건이 바뀔 때마다 유연한 대응이 어렵다는 한계를 가진다. 반면, FreeControl [3]은 사전 학습된 확산 모델을 재학습 없이 사용하는 비학습 기반 방식으로, 다양한 조건에 따라 유연하게 이미지를 제어할 수 있는 프레임워크이다. 구조 정보는 구조 가이던스로, 스타일이나 색감은 외향 가이던스로 보존하면서 효과적인 공간 제어를 가능하게 한다. 그러나 얼굴처럼 세밀한 구조 보존이 중요한 경우에는 한계가 존재한다. 구조가 복잡하거나 조건 정보가 부족한 경우, 가이던스가 전경에 집중되지 않아 얼굴 특징이 배경이나 손, 팔 등의 다른 신체 부위와 혼재되어 형태가 뭉개지거나 부자연스럽게 연결되는 문제가 발생한다. 이러한 현상은 복잡한 배경에서 더욱 두드러진다.

이에 본 연구에서는 별도 학습이나 모델 수정 없이 얼굴 일관성을 향상시킬 수 있는 FLAG(Face Landmark-Aware Guidance)를 제안한다. FLAG는 참조 이미지로부터 추출한 얼굴 랜드마크 좌표를 기반으로 초기 잠재 피처와의 일관성을 유지하도록 랜드마크 일관성 손실을 도입하고, 랜드마크 주변에만 국소적으로 가중치를 부여하는 가우시안 기반 Soft Attention Mask를 적용하여 구조 보존과 자연스러운 이미지 생성을 동시에 달성했다. 제안된 FLAG는 기존 가이던스 방식에 보완적으로 작용하며, 추가 학습 없이 기존 프레임워크에 쉽게 통합 가능하다는 실용적 강점을 갖는다.

2. FLAG (Face Landmark-Aware Guidance)

입력 이미지 I_g가 주어졌을때, 얼굴 영역에서 총 N 개의 랜드마크 좌표 $\{(x_i, y_i)\}_{\{i=1\}^N}$를 추출하였다. 이는 coords_tensor $\in R^{N\times 2}$ 로 구성되며, 피처 맵의 해상도에 맞춰 스케일 조정된 후 해당 위치의 피처를 인덱싱 하는데 사용되었다. 랜드마크의 경우 얼굴의 눈 · 코 · 입 등 주요 부위에서 추출했으며, 이 위치에서의 피처 일관성을 유지하는 것이 핵심 목적이다.

2.1. Soft Attention Mask

단순히 랜드마크 위치에서 피처를 비교하는 것만으로는 주변 정보의 간섭을 충분히 억제하기 어렵다. 이에 따라, 랜드마크 위치 중심으로 국소적인 중요도를 부여하여 비교 효과를 높이기 위해 가우시안 기반의 Soft Attention Mask를 도입하였다. 이 마스크는 다음과 같이 정의된다:

$$M(x, y) = \sum_{i=1}^{N} \exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}\right) \quad (1)$$

여기서 (x_i, y_i)는 얼굴 랜드마크 좌표이며, σ는 마스크의 확산 범위를 조절하는 하이퍼파라미터이다. 소프트 마스크는 참조 피처 F_{ref}^{g}와 생성 중 피처 F_t 양쪽에 적용되어 공간적으로 가중치가 부여된 비교를 가능하게 했다. 이는 얼굴 영역이 배경이나 다른 요소와 혼재되는 현상을 줄여, 보다 자연스러운 이미지 생성을 달성하였다. 특히, 실험을 통해 낮은 해상도의 피처 layer보다는 초기 layer에서 마스크를 적용할 때 더욱 정밀한 구조 보존 성능을 확인하였다.

2.2. 피처 일관성 손실 (Feature Consistency Loss)

참조 피처 F_{ref}^{g}와 생성 중 피처 F_t 간의 유사도를 다음과 같은 손실 함수로 계산하였다:

$$g_l = \text{MSE}(F_t \odot M,\ F_{\text{ref}}^{g} \odot M) \quad (2)$$

여기서 랜드마크를 적용한 피처와 소프트 마스크 M와의 요소별 곱(Element-wise product)을 통해 랜드마크 영역에 더 큰 가중치를 두고 비교가 수행된다. 이를 통해 전역적인 구조 보존과 더불어 지역적인 얼굴 특징도 세밀하게 유지될 수 있도록 하였다.

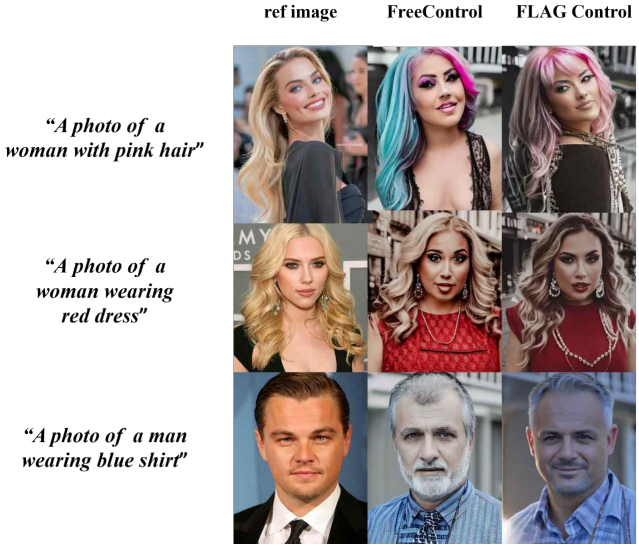
2.3. 생성 과정에서의 가이던스 통합

기존 FreeControl의 구조 가이던스 g_s와 외형 가이던스 g_a에 더해, 얼굴 일관성 유지를 위한 FLAG(g_l)를 새롭게 도입하였다. 최종적으로 사용된 노이즈 예측 식은 다음과 같다:

$$\hat{\epsilon}_t = (1+s)\epsilon_\theta(x_t; t, c) - s\epsilon_\theta(x_t; t, \emptyset) + \lambda_s g_s + \lambda_a g_a + \lambda_l g_l \quad (3)$$

3. 실험 결과

(그림 1)에서 FreeControl은 얼굴이 뭉개지거나 프롬프트 반영이 부족한 모습을 보인 반면, FLAG Control은 구조적 일관성을 유지하며 보다 자연스러운 결과를 생성한다.



(그림 1) 정성적 결과

4. 결론

본 연구에서는 얼굴 일관성 유지를 위해 FLAG(Face Landmark-Aware Guidance)를 제안하였다. 이를 통해 얼굴의 피처 일관성을 유도함으로써, 더 자연스럽고 안정적인 얼굴 이미지를 생성할 수 있었다. 본 방법은 추가적인 학습 없이도 적용 가능하며, 기존 FreeControl 구조에 쉽게 통합될 수 있는 방식으로 동작한다.

사사문구

이 논문은 2025년도 정부(개인정보보호위원회)의 재원으로 한국인터넷진흥원의 지원을 받아 수행된 연구임(No.RS-2023-00231200, 자율주행 환경에서 AI 학습 가능한 개인영상정보 프라이버시 보존 기술개발)

참고문헌

[1] Epstein, D., Yin, X., Zhang, L., & Ma, T., "Diffusion Self-Guidance for Controllable Image Generation," Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA, 2023, Vol. 36, pp. 16222–16239.

[2] Li, D., Li, J., & Hoi, S., "BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing," Advances in Neural Information Processing Systems (NeurIPS), New Orleans, USA, 2023, Vol. 36, pp. 30146–30166.

[3] MoMo, S., Ma, Q., Yang, Y., Chen, X., Zhou, P., & Li, Z., "FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024, pp. 4539–4550.