# 안무 동영상의 생성 품질 향상을 위한 Attention 구조에 관한 연구

박노갑
SK Telecom

tony.nokap.park@sk.com

# A Study on the Attention Mechanism for Enhancement of Dance Video Generation Quality

No Kap Park
SK Telecom

요 약

본 논문은 기존 안무 생성 모델에서 Unet 기반 확산모델이 CLIP 이미지 임베딩만 사용하는 한계를 보완하고자, 얼굴(identity) 특성을 고도화 하여 융합하는 attention 구조를 제안한다. 제안된 구조는 CLIP 인코더에서 추출한 임베딩 이외에 별도로 InsightFace 인코더에서 추출한 임베딩을 CLIP 임베딩 공간으로 매핑하여 변환한 후 두가지의 임베딩을 함께 고려한다. 두 임베딩은 각각 분포가 다르기 때문에 평균 및 표준편차 단위의 정규화를 통해 분포를 맞춘 뒤 융합한다. 이를 통해 얼굴 특성 정보가 확산 모델의 생성 과정 내에서 주요 컨텍스트와 일관성 있게 유지되도록 하여, 얼굴 identity 보존 성능을 향상시킨다.

## Ⅰ. 서 론

안무 생성은 음악의 시간적 흐름에 따라 인물의 동작 시퀀스와 공간적 구성을 설계하는 창의적이며 예술적인 과정이다. 이를 자동화하는 안무 생성 모델은 음악과 안무 간의 복잡한 상관관계를 학습하여, 주어진 음악에 자연스럽고 표현력 있게 어울리는 동작 시퀀스를 생성하는 것을 목표로 한다.

최근 연구에서는 음악 입력을 안무 시퀀스로 변환하는 Music-to-Pose (M2P) 인코더를 도입하여 모듈화된 구조로 음악과 안무 시퀀스 간의 복잡한 관계를 효과적으로 학습하였다. 그러나 Diffusion 기반 U-Net 모델이 CLIP 이미지 임베딩에만 의존할 경우, 개별 인물의 identity 을 안정적으로 유지하는 데 한계가 존재한다.

본 논문은 이러한 한계를 극복하고자 기존 U-Net 기반 확산 모델 구조를 확장한다. 특히, 참조(reference) 인물의 identity 보존을 강화하기 위해 얼굴에서 identity 특성을 추출하여 Unet 에 융합할 수 있는 새로운 attention 메커니즘을 제안한다.

## Ⅱ. 모델 구조

### 1. Face Embedding 과 CLIP Embedding 의 병합

본 연구에서는 얼굴에서 추출한 identity feature 보존을 위해 [2] 에서 제안한 identity 임베딩 추출 파이프라인을 참고하였다. 이 파이프라인은 얼굴 탐지, 정렬, 임베딩 생성의 세 단계로 구성된다.

입력 이미지는 InsightFace 를 활용하여 얼굴을 탐지하며, 복수의 얼굴이 검출될 경우 면적이 가장 큰 얼굴을 선택한다. 얼굴 탐지에 성공하면 해당 위치에서 임베딩 벡터가 추출된다.

정렬된 얼굴 이미지 혹은 탐지된 얼굴 영역은 cross-attention 기반의 얼굴 인코더에 입력되어 512 차원 크기의 identity 임베딩이 산출된다. 이 임베딩은 얼굴의 고유 특징을 정량적으로 표현하여 다양한 얼굴 각도, 표정 변화, 조명 조건에서도 견고한 식별력을 제공한다.

생성된 identity 임베딩은 CLIP 기반 이미지 임베딩 공간으로 매핑되며, 이 과정에서 identity 임베딩과 CLIP 이미지 임베딩 간 분포 차이를 줄이고 상호 표현이 효과적으로 정렬된다.

### 2. Face Identity Attention Branch 도입

앞서 생성된 N 개의 identity 임베딩은 CLIP 임베딩 토큰과 Concatenation 되어 U-Net 에 입력된다. 제안하는 U-Net 구조에서는 identity 임베딩을 나타내는 마지막 N 개의 토큰을 분리하여 별도 identity attention branch 에서 처리한다(그림 1). 연구[3] 에서는 cross-attention 레이어를 수정하여 입력 특성과 독립적으로 ID 토큰/임베딩을 다양한 스케일에서 결합함으로써, 이미지 생성 과정에서 얼굴 특성의 누락 또는 왜곡 문제를 해결할 수 있음을 보였는데 이를 참조하였다.

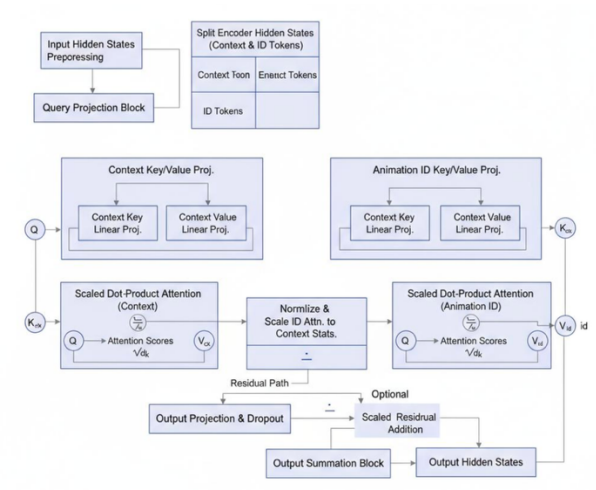


그림 1. 제안한 Attention 구조

각 attention 단계에서는 main attention branch 와 identity attention branch 가 별도로 연산되며, latent feature 의 통계치(평균, 표준편차)에 기반해 정규화 (normalization) 및 스케일링 후 토큰 단위로 결합된다.

이는 identity 보존에 기여하는 정보가 기존 CLIP attention 만 지날 때보다, 확산과정에 더욱 직접적으로 주입됨을 뜻한다.

3. 분포 정규화 기반 ID 융합

각 attention 브랜치의 결과(즉, 필터링된 main latent feature $h$ 및 identity latent feature $I$ )는 평균과 표준편차가 다르므로 바로 더하면 밸런스를 해칠 수 있다. 따라서 identity latent feature 의 분포를 main latent feature 분포에 맞추어 정규화 해준다.

$$\mathbf{i}' = (\mathbf{i} - \mu_i)\frac{\sigma_h}{\sigma_i + \epsilon} + \mu_h$$

여기서 $\epsilon$ 은 안정화를 위한 작은 상수이다. 이후 정규화된 $\mathbf{i}'$ 에 스케일 파라미터 $\alpha$ 를 곱해 최종 융합 특성 $\mathbf{h} + \alpha\mathbf{i}'$ 를 얻어 최종적으로 identity 정보를 유지하면서 생성 품질 저하 및 feature 왜곡을 방지한다.

4. Face Identity 적용 영상

그림 2 는 본 논문에서 제안한 Attention 구조를 적용한 결과를 보여준다. 제안한 모델은 Reference 와 더 흡사한 영상을 생성하며, 얼굴 특성의 추가가 얼굴이외에 다른 부위의 품질도 높이는 결과를 관찰하였다. 정량적 평가는 향후 진행 예정이다.

## Ⅲ. 결론

본 논문에서는 음악에 맞춘 안무 생성 과정에서 참조 인물의 identity 를 안정적으로 보존하기 위한 확장된 U-Net 기반 확산 모델 구조를 제안하였다. Identity 임베딩과 CLIP 이미지 임베딩의 효과적인 융합을 위해 새로운 attention 메커니즘인 identity attention branch 를 도입하였으며, 분포 정규화 기반 ID 융합 기법을 적용하여 identity 정보가 확산 과정에서 균형 있게 반영되도록 하였다. 실험 결과, 제안한 모델은 참조 인물의 얼굴 특성을 보다 정확히 유지하면서 안무 영상의 전체적인 품질 향상에도 기여하는 것처럼 보인다. 향후에는 정량적 평가를 통해 모델의 성능을 체계적으로 검증하고, 다양한 음악 장르 및 인물 유형에 대한 확장 가능성을 모색할 예정이다.


그림 2. 제안한 Attention 사용 Inference 결과


## ACKNOWLEDGMENT

본 연구는 문화체육관광부 및 한국 콘텐츠 진흥원의 2024 년도 문화체육관광연구 개발사업으로 수행되었음 (과제명: 전통예술 가무악의 융복합 공연제작 활성화를 위한 융복합 공연 기획/제작 플랫폼 기술 개발, 과제번호: RS-2024-00398536)


## 참 고 문 헌


[1] Nokap Tony Park (2025), “M2PE-Diff: Music-to-Pose Encoder for Dance Video Generation Leveraging Latent Diffusion Framework”, ACM MM 2025

[2] Tu, S., Xing, Z., Han, X., Cheng, Z.-Q., Dai, Q., Luo, C., & Wu, Z. (2024). StableAnimator: High-quality identity-preserving human image animation. In Proceedings of the IEEE/CVPR 2025.

[3 Mohamed, S., Han, D., & Li, Y. (2024). Fusion is all you need: Face fusion for customized identity-preserving image synthesis. arXiv. https://doi.org/10.48550/arXiv.2409.19111