

# [Study 5. 도진경] DreamID-V : Bridging the Image-to-Video Gap for High-Fidelity Face Swapping via Diffusion Transformer

- Title) **DreamID-V: Bridging the Image-to-Video Gap for High-Fidelity Face Swapping via Diffusion Transformer**
- Publication) 2026.01
- Reference)
  - [project page](#) | [github](#) | [paper](#)

## 1. Introduction

### 2. Related Works

### 3. Methodology

#### **3.1 SyncID-Pipe**

##### 3.1.1 Identity-Anchored Video Synthesizer (IVS)

##### 3.1.2 Bidirectional ID Quadruplets 구성

#### **3.2 DreamID-V Framework**

##### 3.2.1 Modality-Aware Conditioning (MC)

##### 3.3 DreamID-V Training Pipeline

### 4. Experiments

---

## 1. Introduction

- 주요 특징

- **고품질 얼굴 생성:** 이미지 기반 확산 모델 적용
- **일관성:** VFS에서의 일관성 유지
- **강화된 조건 입력/Transformer 구조:** 멀티모달 조건 처리 능력
- VFS(Video Face Swap)
  - : IFS(Image Face Swap) 대비 identity, 표정, 얼굴 움직임의 일관성 유지 어려움.
  - VividFace - 조건부 inpainting
  - DynamicFace - facial condition 조정
  - HiFiVFS - 세부 속성 추출 모듈 추가
  - CanonSwap - Canonical Space 기반 face swap

## 1. SyncID-Pipe

- IFS의 장점을 video domain에 적용
- IVS(Identity-Anchored Video Synthesizer) 을 pretrain한다.
  - : adaptive pose attention mechanism을 사용해서 First-Last-Frame video generation model 에 **포즈 정보를 주입**
  - : 처음부터 끝까지 consistent한 포즈를 표현하도록 만듦.
- Expression Adaptation Strategy 적용
  - : enhanced background recomposition mechanism 적용
  - : 생성 video와 target video 간 배경이 동일하게 구성됨.

## 2. DreamID-V

- DiT 모델 기반
- Modality-Aware Conditioning(MC) mechanism
  - : multi-modality 데이터 사용
  - : visual realism과 identity 일관성 유지
- **Identity Coherence Reinforcement Learning (IRL)** 🐧
  - : 다양한 환경/표정에서의 얼굴 표현 개선

## 3. IDBench-V

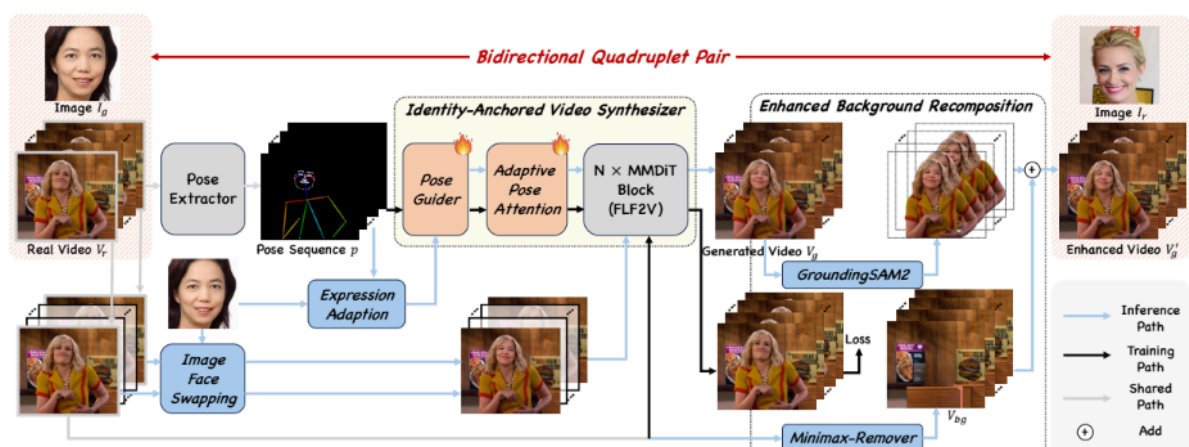
- 얼굴 표정, 포즈, 조명에 따른 benchmark 제안

## 2. Related Works

- IFS의 장점을 Reinforcement Learning에 적용
- identity similarity와 attribute preservation 문제를 근본적으로 개선
- 기존 VFS 모델과의 차별점
  - 모델 구조
    - 기존 모델: U-Net + 3D Conv/Temporal Attention을 사용한 시간 처리
    - DiT 기반 모델
      - : Transformer를 사용한 품질/일관성 개선
      - : 한계점 - I2V, T2V, Keyframe Interpolation 기반 방법으로 디테일 표현에 한계
  - Face Swap
    - IFS는 GAN → Diffusion으로 발전, 정체성 유사도와 시각적 품질 향상
    - VFS는 temporal consistency 유지를 위한 모델 다수
      - Identity/Attribute consistency 유지 ↓

## 3. Methodology

### 3.1 SyncID-Pipe



**Figure 2 Overview of SyncID-Pipe.** We pre-train the Identity-Anchored Video Synthesizer and combine it with the Image Face Swapping model to construct Bidirectional Quadruplet Pair data.

#### 3.1.1 Identity-Anchored Video Synthesizer (IVS)

- **First-Last-Frame 기반 비디오 생성 (FLF2V)**

$$\hat{V}_r = \text{IVS}(v_1, v_T, p)$$

- ID는 keyframe에 고정, motion은 pose로 정의 → ID quadruplets 구성으로 사용

- **Adaptive Pose-Attention**

: 포즈 기반 motion 정보 주입

① Pose feature 정렬

$$P = \text{PoseGuider}(p)$$

- 간단한 CNN구조
- P(포즈의 latent vector)를 Z(video의 noisy latent feature)와 동일 차원으로 정렬하도록 학습

② DiT block ❄️

$$Q = ZW_q, \quad K = ZW_k, \quad V = ZW_v$$

- Self-Attention
  - Q (Query): 비디오 생성 공식 - "프레임 내 얼굴 표현을 위해 어떤 정보가 필요한가?"
  - K (Key): 생성 기준 - pose
  - V (Value): pose feature

③ Pose-Attention

$$K' = PW'_k, \quad V' = PW'_v$$

- P를 latent attention 공간으로 projection ( W 학습)

④ 최종 Attention 출력

$$Z_{\text{new}} = \text{Attn}(Q, K, V) + \lambda \text{Attn}(Q, K', V')$$

$$\mathbf{Z}_{\text{new}} = \text{Softmax} \left( \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V} + \lambda \cdot \text{Softmax} \left( \frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}} \right) \mathbf{V}'$$

- 사전학습된 비디오 생성 기능 ❄️
- pose 기반 모션 제어
- $\lambda$  : pose 조건 강도 조절
- 학습 방식
  - 대용량 portrait dataset + pose 추출 시퀀스
  - Flow Matching 기반 optimization

### 3.1.2 Bidirectional ID Quadruplets 구성

- pre-trained **IVS**를 **DreamID** (IFS)와 결합, **Bidirectional Quadruplet Pair** 데이터 구축
  - IFS - ID 정답 생성
  - IVS - motion 보존된 video 생성
  - bidirectional quadruplet -  $\text{img} \leftrightarrow \text{vid} \mid \text{real} \leftrightarrow \text{gen}$  간 ID 공간 일치

[개념]

- **bidirectional**
  - $\text{IFS} \rightarrow \text{VFS} \{I_r, V_g, V_r\}$ 
    - IFS 결과가 VFS 결과와 ID feature 공간에서 가까워지도록 학습 (identity 유지)
  - $\text{VFS} \rightarrow \text{IFS} \{I_r, V_r, V_g\}$ 
    - VFS의 시간·표정 다양성을 IFS에 반영
    - 비디오 embedding을  $V_r$  기준으로 ID 표현 공간 alignment
- **ID quadruplets**

: 다음 4가지 값을 정의 (  $I_r/V_r/I_g$ 를 사용해서 생성된  $V_g$  )

  - **$I_r$**  = Source ID (image) - 바꿔 넣을 얼굴
  - **$V_r$**  = Target frame (video) - 바뀌기 전 비디오 프레임

- **Ig** = result (image) - IFS 결과

$$I_{\text{refl}} = \text{IFS}(v_1^{(r)}, I_g), \quad I_{\text{ref2}} = \text{IFS}(v_T^{(r)}, I_g)$$

- **Vg** = VFS result (video) - VFS 결과

$$V_g = \text{IVS}(I_{\text{ref1}}, I_{\text{ref2}}, \tilde{p}_r)$$

- **bridge the gap between VFS and IFS**

- IMG ↔ VID | SWAP 유 ↔ 무 | 정답 ↔ 모델출력 차이 최소화
- VFS가 IFS 수준의 정체성 일관성을 갖도록 만드는 학습
  - IFS: 정적인 이미지 내 얼굴 교체 - identity 유지 쉬움, 상대적으로 높은 품질
  - VFS: identity/attribute의 시간적 일관성, 표정/조명 변화 표현 어려움.

[주요 특징]

① Source Data Curation

- talking-head 데이터셋 사용 - 다양한 환경/얼굴 condition 적용

② Expression Adaptation

$$\mathcal{F}_t = \text{Reconstruct}(\alpha_{id}^g, \alpha_{exp}^r(t), \alpha_{pose}^r(t))$$

- ID와 표정, pose 요소를 분리 (3D Face Disentanglement)
  - ID from Ig / expression, pose from Vr

③ Enhanced Background Recomposition

1. foreground mask:

$$M_r, M_g = \text{SAM2}(V_r, V_g)$$

2. background 추출:

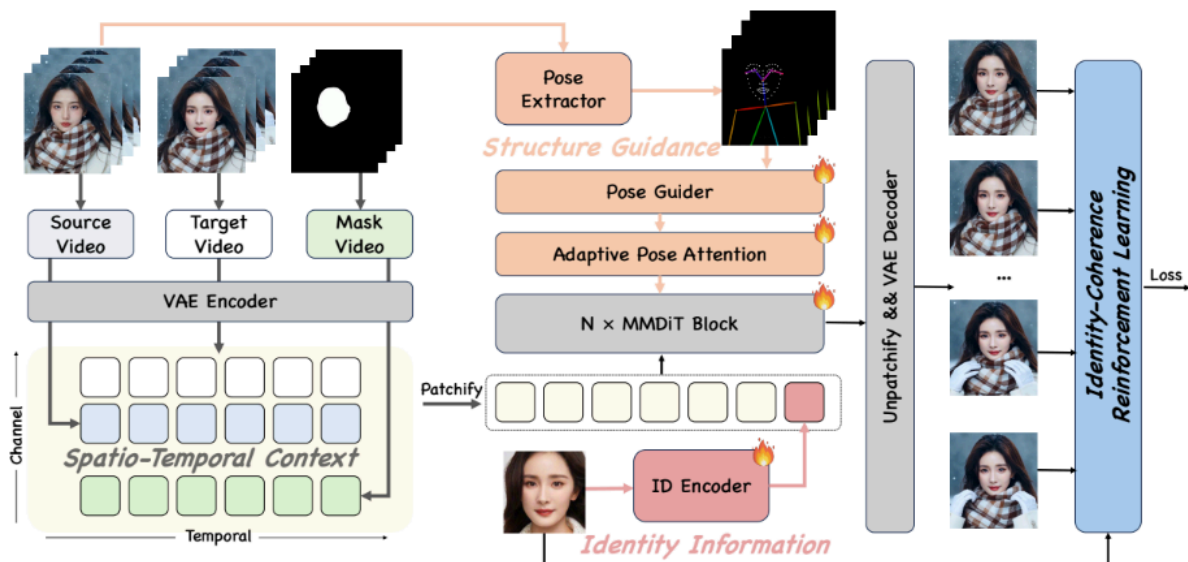
$$V_{bg} = \text{RemoveFG}(V_r)$$

3. 합성:

$$V'_g = M_g \odot V_g + (1 - M_g) \odot V_{bg}$$

- SAM2
- Foreground / Background 분리

## 3.2 DreamID-V Framework



- 서로 성질이 완전히 다른 정보(ID/Attribute)를 diffusion 모델 안에 넣되, 섞이지 않게(disentangle) 하면서도 필요할 때는 충분히 상호작용(fusion)
  - **Identity Transfer**  
: source 얼굴의 ID를 target video에 transfer

- **Attribute Preservation**

: Source 비디오의 pose/expression/background/light/motion 유지

---

### 3.2.1 Modality-Aware Conditioning (MC)

- VFS 조건을 3가지 modality로 분리

$$\mathcal{C} = \{C_{\text{ctx}}, C_{\text{str}}, C_{\text{id}}\}$$

- Spatio-Temporal Context (Cctx)
  - 픽셀 단위로 정확히 정렬되어야 하는 값 - 배경, 조명 등
  - diffusion latent와 공간/시간 일대일대응
- Structural Guidance (Cstr)
  - 구조·기하적 정보 - pose, expression, motion 등
  - temporal alignment 필요
  - IVS의 Pose-Attention 그대로 사용
- Identity Information (Cid)

$$\{z_1, \dots, z_N\} \cup \{e_{\text{id}}\} \in \mathbb{R}^{(N+1) \times d}$$

- token-level global attention
  - 프레임의 모든 시간 x 모든 공간에 영향을 주는 token 단위 ID 유지  
[Study 5] DreamID - Face Encoder
- 

### 3.3 DreamID-V Training Pipeline

- 단계적 학습 + 난이도 기반 재가중 학습
  - Identity similarity (ID가 얼마나 정확히 유지되는가)
  - Attribute preservation (pose, expression, background)
  - Visual realism
  - Temporal identity consistency (시간에 따라 ID가 흔들리지 않는가)

## ① Synthetic Training

- 안정적이고 빠른 ID Alignment
- forward-generated paired data  $\{I_r, V_g, V_r\}$  사용  
: video foundation model이 이해하기 쉬운  $V_g$ 를 사용
- Flow Matching Loss 사용

$$\mathcal{L}_{\text{syn}} = \mathbb{E}_{t, \epsilon} \left[ \|(z - \epsilon) - v_{\theta}((1 - t)z + t\epsilon, t, y)\|^2 \right]$$

- $V_g$ 를 supervision으로 사용

## ② Real Augmentation Training

$$\mathcal{L}_{\text{real}} = \mathbb{E}_{t, \epsilon} \left[ \|(z - \epsilon) - v_{\theta}(\cdot)\|^2 \right]$$

- 다양한 배경/환경 품질 개선
- backward-real paired data  $\{I_r, V_r, V_g'\}$  사용
  - background preservation 유지
  - $V_r$ 를 supervision으로 사용

## ③ Identity-Coherence Reinforcement Learning (IRL)

- temporal ID flickering 방지
  - 정면 프레임 → ID similarity 높음
  - 측면, 큰 motion → ID similarity 낮음
- 강화학습 관점의 component
  - policy:  $\pi_{\theta}$  = diffusion generator
  - action:  $x^*_0$  = 생성된 프레임
  - state:  $y$  = 조건
- Q-value 정의; id 표현 난이도

$$Q(y, \hat{x}_0) = \frac{1}{\cos(E(\hat{x}_0), E(I_t)) + \delta}$$

- ID가 다를수록 높은 Q-value  
: 어려운 프레임 = 높은 보상
- IRL loss

$$\mathcal{L}_{\text{IRL}}(\theta) = \sum_{c=1}^C \mathbb{E}_{t, \epsilon} \left[ Q_c \cdot \|(\mathbf{z}_c - \epsilon) - \mathbf{v}_\theta((1-t)\mathbf{z}_c + t\epsilon, t, y)\|^2 \right]$$

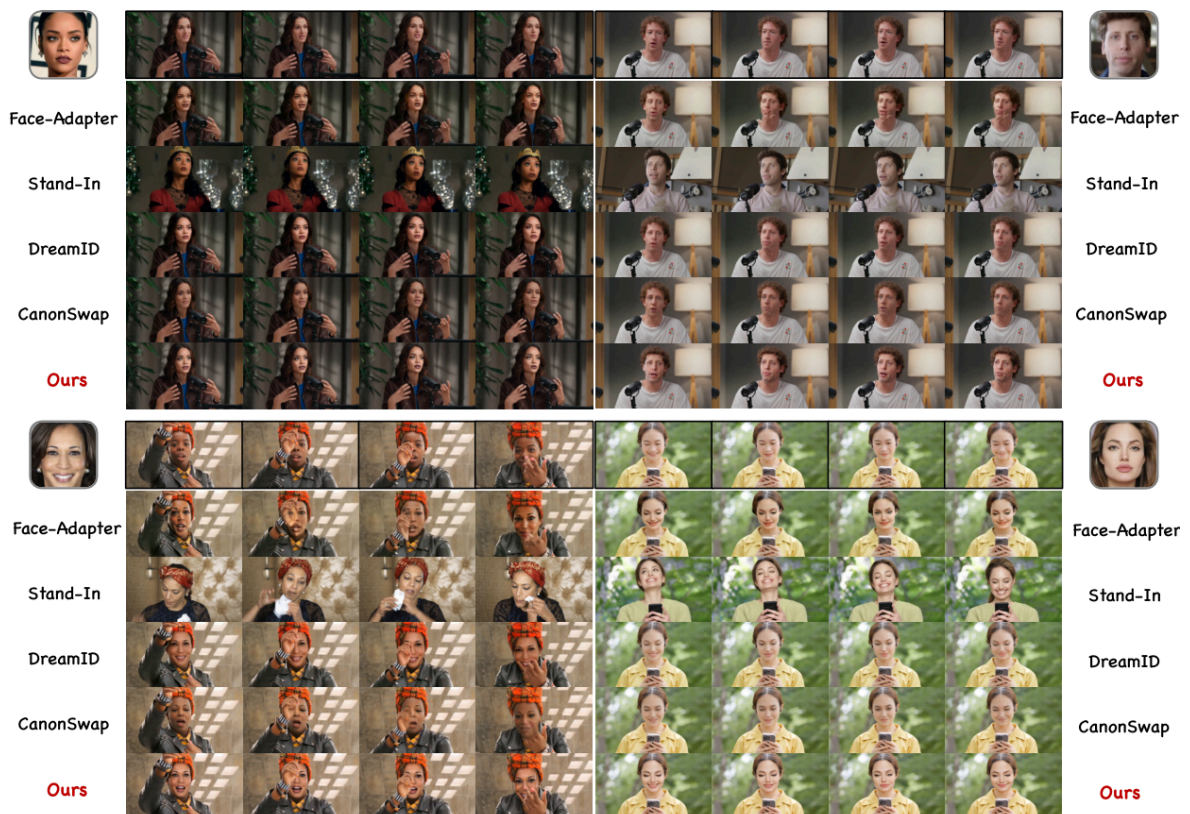
- Q value를 프레임 VAE 기반 chunk 단위로 평균, 가중치로 사용
- ID가 약한 chunk에 gradient ↑
- 사전학습된 VAE = WAN
- Sync-to-Real Curriculum: ①+②+③

▶ 이런 paired data를 사용하면 다양한 swapping task로 확장 가능

## 4. Experiments

- **IDBench-V Benchmark**  
: 실사용 환경 중 VFS가 실패하기 쉬운 경우만 모음.
  - 200개 (source video, target image) 쌍
  - 포함 난이도:
    - 작은 얼굴

- extreme head pose
  - occlusion
  - 복잡한 표정
  - 여러 인물·복잡한 배경
- Training Dataset
  - **OpenHumanVid**
  - 동일 ID 비디오 쌍만 남기도록 ID similarity 기반 필터링
- 타 VFS와 비교
  - IFS - FSGAN, REFace, Face-Adapter, DreamID
  - VFS - Stand-In, CanonSwap / VividFace, DynamicFace



- 평가 방법
- 1. Identity Consistency

- a. ID 유사도
    - : ArcFace, InsightFace, CurricularFace
  - b. Temporal stability
    - : 시간흐름에 따른 Face Encoder 결과 분포
- 2. Attribute Preservation
  - a. pose / expression 차이 비교
- 3. Video Quality
  - a. VBench metrics
    - i. Background consistency ↑
    - ii. Subject consistency ↑
    - iii. Motion smoothness ↑
- 요소별 비교
  - Identity Consistency
    - DreamID-V가 가장 좋음
  - Attribute Preservation
    - CanonSwap 이 가장 좋음
  - Video Quality
    - VFS ( DreamID-V / CanonSwap / Stand-In) 대체로 좋음.

### [Study 5] DreamID - Face Encoder