WithAnyone: Towards Controllable and ID Consistent Image Generation

Hengyuan Xu[1,2] Wei Cheng[2,†] Peng Xing[2] Yixiao Fang[2] Shuhan Wu[2]
Rui Wang[2] Xianfang Zeng[2] Daxin Jiang[2] Gang Yu[2,‡] Xingjun Ma[1,‡] Yu-Gang Jiang[1]

[1] Fudan University [2] StepFun

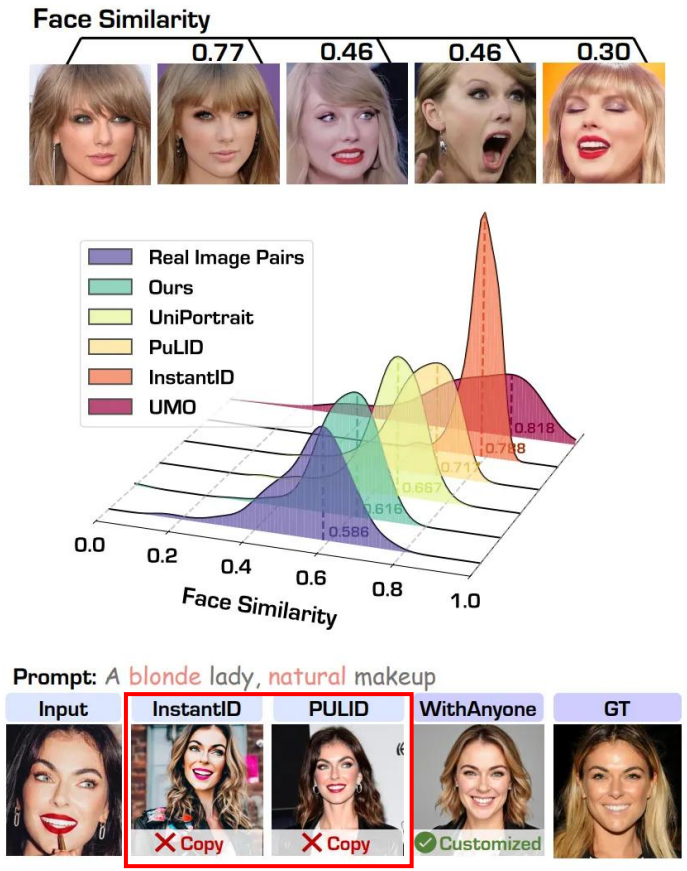
Abstract

- Single/Multi identity-consistent 생성 분야의 모델
- 기존 데이터 셋은 동일 인물이 포함된 다양한 이미지(조명, 자세, 표정 등)가 부족했었음
  - 따라서 이전 연구들은 reconstruction 기반 학습으로 진행함
- 이러한 이유로, **copy-paste** 문제가 나타남
  - **얼굴을 지나치게 복제하여, 다양한 조명, 자세, 표정등 표현력이 제한됨.**
- **WityAnyone**
  1. MultiID-2M 대규모 데이터 셋 구축
     - 동일 인물이 포함된 다양한 이미지 쌍을 구축함
  2. MultiID-Bench 평가지표 구축
     - Copy-paste 평가, fidelity 및 variation 의 trade-off 평가
  3. 새로운 학습 전략
     - Contrastive identity loss 제안 (fidelity 및 diversity 를 조절)
- 결과
  - Copy-paste 문제 완화 및 identity 유사성 보존 및 시각적 품질 향상
  - 다양성(조명, 자세, 표정) 생성 능력 향상

Figure 1. **Showcases of WithAnyone.** WithAnyone is capable of generating high-quality, controllable, and ID-consistent images by leveraging ID-contrastive training on the proposed **MultiID-2M** dataset.

Introduction

- 동일 인물에 대해 자세, 표정, 메이크업, 조명등의 변화로 인해 **Identity 유사도는 상당히 달라진다.**
  - 동일 인물 얼굴 이미지 쌍 간의 유사도 분포를 측정했을 때, Real Image Pairs(보라색) 과 가장 유사했음.
- 이전 모델들은 copy-paste 문제가 나타남
  - ✓ 이러한 문제를 WithAnyone 에서 공식화하고,
  - ✓ 정량화하기 위한 평가 메트릭을 개발하고,
  - ✓ 문제를 완화하기 위한 학습 전략을 제시함.



Related Work

- Single-ID preservation
  - Unet/Stable Diffusion 에서는 **학습된 임베딩(CLIP, ArcFace)**를 **cross-attention, adapter** 방식으로 주입했었음
- Multi-ID preservation
  - Xverse, UMO 에서는 **VAE 에서 얻어진 얼굴 임베딩을 모델 입력과 연결하여** 사용했음
  - **copy-paste 문제 유발 및 control도 저하시킴**
- ID-Centric Datasets and Benchmarks
  - 평가 메트릭이 미흡한 상태임
  - 다른 연구에서는 CelebA에서 인물을 샘플링하여 테스트 셋을 구성하는데, 이는 재현성을 떨어뜨림.

MultiID-2M: Paired Multi-Person Dataset Construction

1. **Single-ID**
   - 데이터 수집, ArcFace 임베딩 및 클러스터링하여 clean 데이터 구축(3천 명 인물, 1백만 장의 레퍼런스)
2. **Multi-ID**
   - 다중 이름 및 장면 인식 쿼리를 통해 그룹 사진을 검색하고 얼굴을 찾음.(2천만 장)
3. **ID Image Pairing**
   - Cosine similarity 를 사용하여 ArcFace 임베딩을 single-ID cluster 와 매칭함으로써 identity 를 할당함
4. **Post-preprocessing**
   - Recognize Anythin, Aesthetic scoring, OCR 기반 워터마크 제거, LLM 기반 캡션 생성

- 최종 구성
  - 레퍼런스 데이터와 매칭된 multi-id 이미지 약 50만장
  - reconstruction 학습을 위한 추가 unidentified 된 150만 장으로 구성
  - 다양한 국적과 인종을 가진 약 25,000명 의 identity 로 구성

MultiID-Bench: Comprehensive ID Customization Evaluation

- Multi-ID 생성을 위한 통합 벤치마크
  - identity fidelity 와 생성 품질을 평가 / 테스트 셋 435쌍 (레퍼런스 – 프롬프트 – GT)
- SIM_{GT}: 생성 이미지 – GT / SIM_{Ref}: 생성 이미지 – 레퍼런스 유사도 평가
  - 기존 연구에서는 SIM_{Ref} 만 사용했음 → 이것이 copy-paste 유발
  - MultiID-Bench 는 프롬프트가 지정한 실제 identity 와의 유사성(SIM_{GT})을 주요 지표로 사용함
  - Copy-paste 메트릭 ⬇️: 자연스러운 변형(자세, 표정 등)이 예상될 때, 과도한 복제를 높은 점수를 프롬프트에 맞게 생성한 경우 보상을 제공함(낮은 점수).
    - $\boldsymbol{r}$: 레퍼런스 얼굴 임베딩
    - $\boldsymbol{t}$: 타겟(GT) 얼굴 임베딩
    - $\boldsymbol{g}$: 생성된 얼굴 임베딩

$$\mathrm{M_{CP}}(\mathbf{g} \mid \mathbf{t}, \mathbf{r}) = \frac{\theta_{gt} - \theta_{gr}}{\max(\theta_{tr}, \varepsilon)} \in [-1, 1], \qquad \theta_{ab} = \arccos(\mathrm{Sim}(a, b))$$

$$\mathrm{Sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|},$$

  - 1점 : 생성 이미지 – 레퍼런스 와 완전히 일치 (copy-paste)
  - -1점 : 생성 이미지 – GT 와 완전히 일치

WithAnyone: Controllable and ID- Consistent Generation

- MultiID-2M 대규모 및 paired-reference supervision 으로 reconstruction 보다 우수하며, 강력한 identity-conditioned 합성을 가능하게 하는 3가지 학습 loss 구축

1. **Diffusion loss (flow matching loss)**

$$\mathcal{L}_{\text{diff}} = \left\| v_\theta(x_t^{(i)}, t^{(i)}, c^{(i)}) - (x_1^{(i)} - x_0^{(i)}) \right\|_2^2,$$

2. **Ground-truth-Aligned ID Loss**
   - ArcFace 임베딩은 **landmark detection, alignment** 를 필요로 함
   - 생성 이미지에서 랜드마크 추출하는 것은 위험함 (노이즈가 있기 때문)
   - → 저자들은 **GT 랜드마크를 사용하여** 생성된 이미지를 alignment 진행함.
     - 생성 이미지 및 GT 의 얼굴 임베딩을 측정 후 cosine distance 를 측정하여 최소화함.

$$\mathcal{L}_{\text{ID}} = 1 - \cos(\mathbf{g}, \mathbf{t}) \qquad \mathcal{L}_{\text{id}} = 1 - \cos(f(g(\mathbf{T}), \mathbf{G}), f(g(\mathbf{T}), \mathbf{T})). \qquad (6)$$

   - $\boldsymbol{t} = f(g(\boldsymbol{T}), \boldsymbol{T}),\ \boldsymbol{g} = f(g(\boldsymbol{G}), \boldsymbol{G})$
     - $g(\cdot)$: $detection\ model\ (RetinaFace), f(\cdot)$: $recognition\ model\ (ArcFace)$
     - $g(\boldsymbol{G})$ 대신 $g(\boldsymbol{T})$ 를 사용함 → $\boldsymbol{g} = f(g(\boldsymbol{T}), \boldsymbol{G})$
     - $g(\boldsymbol{G})$ 는 생성할 때의 노이즈가 포함되어 있으므로, 추정하는데 영향을 끼침

WithAnyone: Controllable and ID- Consistent Generation

3. ID Contrastive Loss With Extended Negatives.

- **identity 유지를 강인하게 하기 위해** ID Contrastive loss 도입
- 생성된 이미지를 얼굴 임베딩 space 에서 **레퍼런스 이미지는 가깝게 하고**, 다른 identity 는 멀어지게 함.

$$\mathcal{L}_{\mathrm{CL}} = -\log \frac{\exp(\cos(\mathbf{g}, \mathbf{r})/\tau)}{\sum_{j=1}^{M} \exp(\cos(\mathbf{g}, \mathbf{n}_j))/\tau)}, \qquad (7)$$

- $\boldsymbol{r}$: 생성된 이미지와 동일 ID를 가진 레퍼런스 이미지의 얼굴 임베딩
- $\boldsymbol{n}$: M개의 다른 ID를 가진 negative sample 얼굴 임베딩

- 전체 loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{ID}}\mathcal{L}_{\mathrm{ID}} + \lambda_{\mathrm{CL}}\mathcal{L}_{\mathrm{CL}}, \qquad (8)$$

Training pipeline (4 phase)

- **Phase 1: Reconstruction pre-training with fixed prompt**
  - 백본을 initialize 하기 위해 reconstruction 학습 진행
  - 완전한 identity-condition 생성보다 간단하고, 대규모 레이블이 없는 데이터를 활용할 수 있음.
  - 처음 몇 천 step 동안 캡션을 일정한 dummy prompt(ex: “two people”) 로 고정되어 모델이 **텍스트 condition styling 보다는 identity-conditioning 경로를 학습하는데 우선순위들 둠.**
- **Phase 2: Reconstruction pre-training with full captions**
  - identity 학습을 **텍스트 condition 생성을 포함하여 학습함.**
- **Phase 3: Paired tunning**
  - copy-paste 를 억제하기 위해 학습 샘플의 50% 를 **MultiID-2M**의 500k 레이블이 지정된 이미지에서 가져온 쌍으로 된 데이터 대체
  - 각 쌍은 input 및 target(GT) 이 동일한 ID 이고 reference는 해당 ID의 레퍼런스 셋에서 랜덤하게 선택함.
  - 이것이 low-level copying(픽셀, 구체적인 특징) 보다 **high-level copying(코의 형태, 골격 구조, 눈과 입의 비율 등) 의 identity 임베딩에 의존하도록 강요함.**
- **Phase 4: Quality tunning**
  - 고품질 이미지 셋 활용함. (+ augment: 모델이 생성한 스타일 변형)
  - 시각적 품질을 향상시키고 스타일 견고성 및 전이성(transferability) 를 향상시킴
  - 텍스처, lighting, stylistic adaptability 향상

Experiments

- 평가 대상 모델
  - 일반적인 커스터마이징 모델: OmniGen/OmniGen2, Qwen-Image-Edit, FLUX.1 Kontext, UNO, USO, UMO, GPT-4o
  - 얼굴 커스터마이징 모델: UniPortrait, ID-Patch, PuLID, InstantID
- copy-paste 현상이 많이 발생하면, 얼굴 유사도 점수가 높은 현상이 기존 방법들에서 많이 나타남.

Table 1. **Quantitative comparison on the single-person subset of MultiID-Bench and OmniContext.** , , and indicate the first-, second-, and third-best performance, respectively. For Copy-Paste ranking, only cases with $\mathrm{Sim(GT)} > 0.40$ are considered.

a **MultiID-Bench**

Method	Identity Metrics			Generation Quality		
	Sim(GT) ↑	Sim(Ref) ↑	CP ↓	CLIP-I ↑	CLIP-T ↑	Aes ↑
DreamO	0.454	0.694	0.303	0.793	0.322	4.877
OmniGen	0.398	0.602	0.248	0.780	0.317	5.069
OmniGen2	0.365	0.475	0.142	0.787	0.331	4.991
FLUX.1 Kontext	0.324	0.408	0.099	0.755	0.327	5.319
Qwen-Image-Edit	0.324	0.409	0.093	0.776	0.316	5.056
GPT-4o Native	0.425	0.579	0.178	0.794	0.311	5.344
UNO	0.304	0.428	0.141	0.765	0.314	4.923
USO	0.401	0.635	0.286	0.790	0.329	5.077
UMO	0.458	0.732	0.359	0.783	0.305	4.850
UniPortrait	0.447	0.677	0.265	0.793	0.319	5.018
ID-Patch	0.426	0.633	0.231	0.792	0.312	4.900
InfU	0.439	0.630	0.233	0.772	0.328	5.359
PuLID	0.452	0.705	0.315	0.779	0.305	4.839
InstantID	0.464	0.734	0.337	0.764	0.295	5.255
Ours	0.460	0.578	0.144	0.798	0.313	4.783
GT	1.000	0.521	-0.999	N/A	N/A	N/A
Ref	0.521	1.000	0.999	N/A	N/A	N/A

b **OmniContext Single Character Subset**

Method	Quality Metrics		Overall
	PF ↑	SC ↑	Overall ↑
DreamO	8.13	7.09	7.02
OmniGen	7.50	5.52	5.47
OmniGen2	8.64	8.50	8.34
FLUX.1 Kontext	7.72	8.60	7.94
Qwen-Image-Edit	7.66	8.16	7.51
GPT-4o Native	7.98	9.06	8.12
UNO	7.22	7.72	7.04
USO	6.96	7.88	6.70
UMO	6.56	7.92	6.79
UniPortrait	6.62	6.00	5.55
ID-Patch	N/A	N/A	N/A
InfU	7.69	4.62	4.70
PuLID	6.62	6.83	5.78
InstantID	4.89	5.49	4.35
Ours	7.43	7.04	6.52

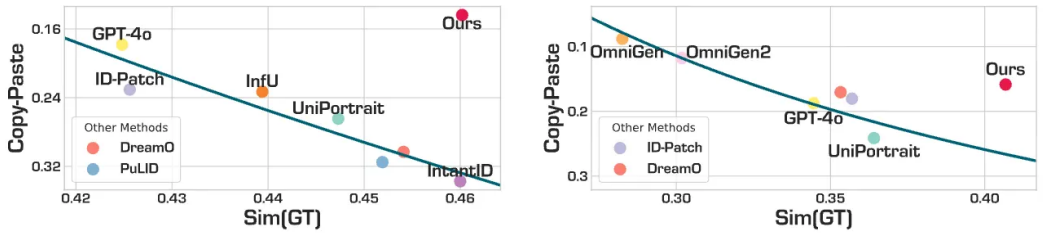
Table 2. **Quantitative comparison on the multi-person subset of MultiID-Bench.** , , and indicate the first-, second-, and third-best performance, respectively. For Copy-Paste ranking, only cases with $\mathrm{Sim(GT)} > 0.35$ are considered. GPT exhibits prior knowledge of identities from TV series in subsets with more than two IDs, leading to abnormally high similarity scores.

a **2-people Subset**

Method	Identity Metrics				Generation Quality		
	Sim(GT) ↑	Sim(Ref) ↑	CP ↓	Bld ↓	CLIP-I ↑	CLIP-T ↑	Aes ↑
DreamO	0.359	0.514	0.179	0.105	0.763	0.319	4.764
OmniGen	0.345	0.529	0.209	0.110	0.750	0.326	5.152
OmniGen2	0.283	0.353	0.081	0.112	0.763	0.334	4.547
GPT	0.332	0.400	0.061	0.092	0.774	0.328	5.676
UNO	0.223	0.274	0.043	0.082	0.735	0.325	4.805
UMO	0.328	0.491	0.176	0.111	0.743	0.316	4.772
UniPortrait	0.367	0.601	0.254	0.075	0.750	0.323	5.187
ID-Patch	0.350	0.517	0.183	0.085	0.767	0.326	4.671
Ours	0.405	0.551	0.161	0.079	0.770	0.321	4.883

b **3-and-4-people Subset**

Method	Identity Metrics				Generation Quality		
	Sim(GT) ↑	Sim(Ref) ↑	CP ↓	Bld ↓	CLIP-I ↑	CLIP-T ↑	Aes ↑
DreamO	0.311	0.427	0.116	0.081	0.709	0.317	4.695
OmniGen	0.345	0.529	0.209	0.110	0.750	0.326	5.152
OmniGen2	0.288	0.374	0.099	0.071	0.734	0.329	4.664
GPT	*0.445*	*0.484*	*0.048*	*0.044*	0.815	0.320	5.647
UNO	0.228	0.276	0.046	0.065	0.717	0.319	4.880
UMO	0.318	0.465	0.180	0.070	0.717	0.309	4.946
UniPortrait	0.343	0.517	0.178	0.048	0.708	0.323	5.090
ID-Patch	0.379	0.543	0.195	0.059	0.781	0.329	4.547
Ours	0.414	0.561	0.171	0.045	0.771	0.325	4.955



a Single-ID subset b Multi-ID subset

Figure 5. **Trade-off between Face Similarity and Copy-paste.** Except for WithAnyone, the other models fall roughly on a fitted curve, illustrating a clear trade-off between face similarity and copy-paste. Upper-right corner is desired.

Experiments

- Single-ID

Prompt: "a woman wearing a white hooded jacket with a black inner garment. Her hair is styled loosely, and she has minimal makeup. The woman is posing with her head slightly tilted, showcasing a calm and composed demeanor. Her expression is neutral.

Prompt: "a woman with long, dark hair flowing dynamically. She wears a white and blue geometric patterned top with a shawl-like drape. Her posture is poised, showcasing elegant jewelry and a subtle smile. The background features a blurred circular pattern in shades of gray.

Prompt: "a woman in a black leather jacket holding a red microphone. She is smiling and appears to be performing or speaking, with her head slightly tilted and her mouth open as if she is in the middle of talking. Her long brown hair is styled straight..

Experiments

- Multi-ID

Prompt: "a man in a dark suit holding a coffee mug and a woman in a light blue sweater resting her head on her hand. They appear to be in a kitchen, looking concerned or surprised. The man is standing, while the woman is seated at a counter.

Prompt: "A couple posing together. The woman wears a blue, sleeveless, V-neck dress, while the man dons a light blue, semi-buttoned shirt. Both are smiling and standing close, with the man's arm around the woman, indicating a friendly or intimate relationship.

Prompt: "three people, two women and one man, posing closely together. The woman on the left wears a white blazer, while the younger woman in front has a strapless top. The man has a white shirt. All are smiling warmly at the camera.

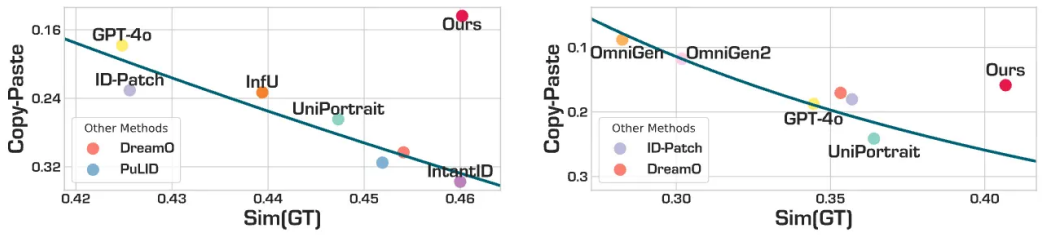
Prompt: "four people dressed in white shirts posing together. The group includes three males and two females, with one male and one female in the center. They are smiling and standing closely, suggesting a family or close-knit group. The attire is casual and coordinated.

Conclusion

- 기존 ID 커스터마이징 모델에서는 copy-paste 문제가 나타났었음.
- 이것은 표정, 자세, 조명등 자연스러운 변화를 반영하지 못했음.
  - 특히 loss 가 face-similarity 때문에 더욱 악화시켰음.
- 해결 방법
  - MultiID-2M 데이터 셋 구축
  - 새로운 학습 전략 및 손실 함수 도입
    - contranstive identity loss, paired-training 전략
- 결과
  - copy-paste 문제 크게 억제, identity similarity 능력 유지, fidelity 및 copy-paste 의 trade-off 문제 해결함.

WithAnyone: Towards Controllable and ID Consistent Image Generation

Figure 1. **Showcases of WithAnyone.** WithAnyone is capable of generating high-quality, controllable, and ID-consistent images by leveraging ID-contrastive training on the proposed **MultiID-2M** dataset.



a Single-ID subset b Multi-ID subset

Figure 5. **Trade-off between Face Similarity and Copy-paste.** Except for WithAnyone, the other models fall roughly on a fitted curve, illustrating a clear trade-off between face similarity and copy-paste. Upper-right corner is desired.

- Multi identity-consistent generation model
- Problem: 얼굴을 그대로 복제하는 "Copy-paste" artifact
- WithAnyone
  - MultiID-2M 대규모 데이터 셋 구축
  - 메트릭 구축: copy-paste 평가
  - phase-4 학습 전략
  - GT-aligned ID loss 및 ID Contrastive loss 구성
- 결과
  - 이전 연구들의 copy-paste 현상을 억제하며, 다양한 표정, 자세등 identity consistent하게 생성할 수 있음
  - 레퍼런스 얼굴의 표정, 조명등 컨트롤에 자유롭게 프롬프트로 생성 가능함