CrossID

Cross-Supervised Spatio-Temporal Gated Fusion for Personalized Portrait Generation

Dongxu Yue, Qixin Yan, Shiao Yang, Xiaoqiang Zhou, Hao Liu, Zhihai He, Chun Yuan

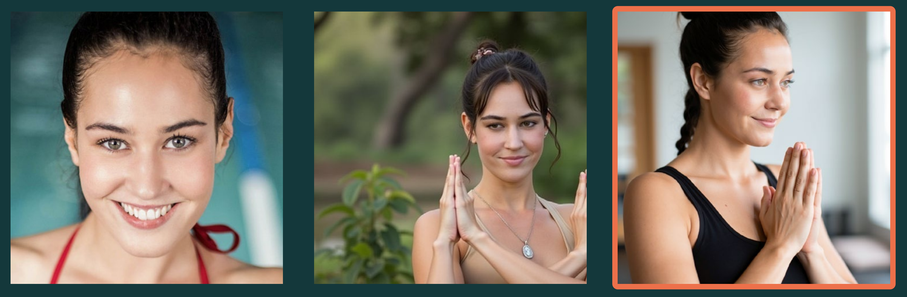
NeurIPS 2026
MAIN CONFERENCE

The same person. A new possibility.

Keep identity and facial detail while following changes in pose, expression and gaze.

ONE REFERENCE. NO SUBJECT TUNING.

01 / CHANGE THE VIEW



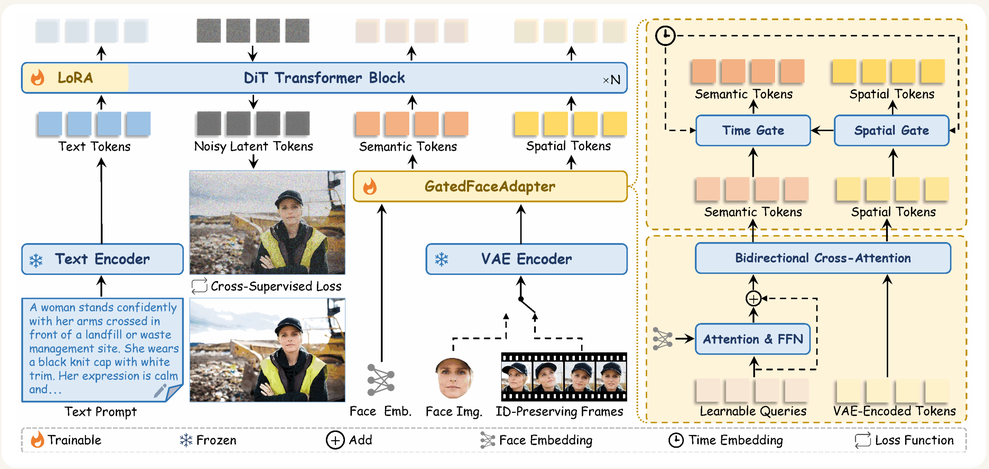
Yoga pose, hands in prayer, **3/4 profile face**.

02 / KEEP THE DETAILS



Taking a selfie, **looking at the phone screen**.

01 Learning identity through change.



GatedFaceAdapter + cross-supervised training

A Fuse identity + detail

ArcFace identity embeddings and VAE spatial tokens interact through bidirectional attention.

B Gate when + where

Time gates balance both branches. Spatial gates suppress identity-irrelevant content.

C Change the reference

After reconstruction pretraining: same identity, different frame. Freeze the adapter; train LoRA.

02 Evidence across benchmarks.

Best FaceSim, CLIP-T, SC and IQ

On CrossID-Bench and Unsplash-50; compared methods.

Method	FaceSim	CLIP-T	NI	SC	IQ
PuLID-FLUX	.6274	.2422	4.287	4.094	4.001
InfiniteYou	.5769	.2496	4.609	4.060	3.587
DreamO	.6253	.2381	3.542	3.865	3.450
WithAnyone	.5646	.2309	4.827	4.305	4.104
CrossID	**.6291**	**.2508**	**4.733**	**4.362**	**4.316**

CrossID-Bench, selected baselines (Table 1). Higher is better. NI: Natural Integration; SC: Semantic Correctness; IQ: Image Quality (NI / SC / IQ: 1-5).

Cross-supervision improves NI



Table 2

11M portrait images

0.9M ID-preserving video clips

100 x 20 identities x prompts

ydx25@mails.tsinghua.edu.cn

CrossID-11M / CrossID-Bench

PAPER