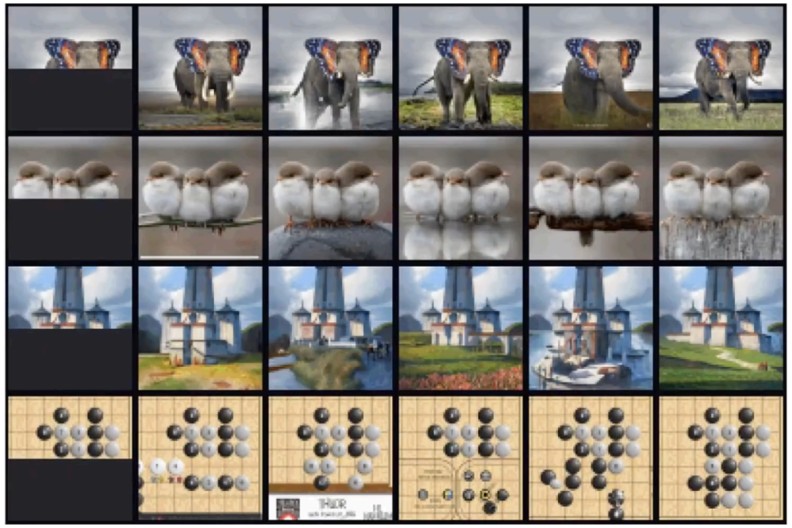
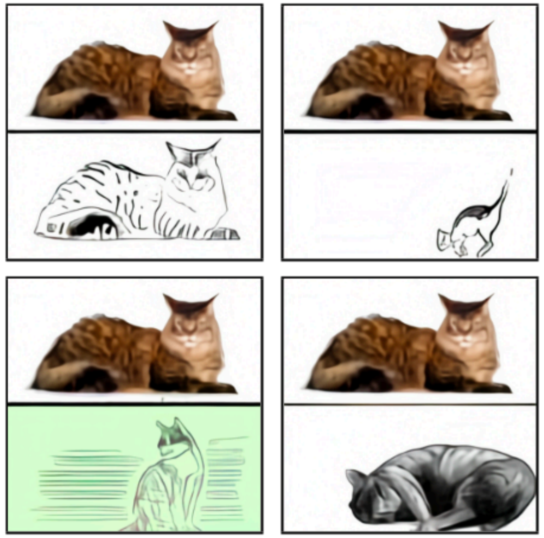
The Image Local Autoregressive Transformer

Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, Yanwei Fu, Xiangyang Xue
School of Data Science, Fudan University {20110980001,yanweifu}@fudan.edu.cn

Transformer-based image generation



(a) iGPT[1]



(d) the exact same cat on the top as a sketch on the bottom

(b) DALLE[2]

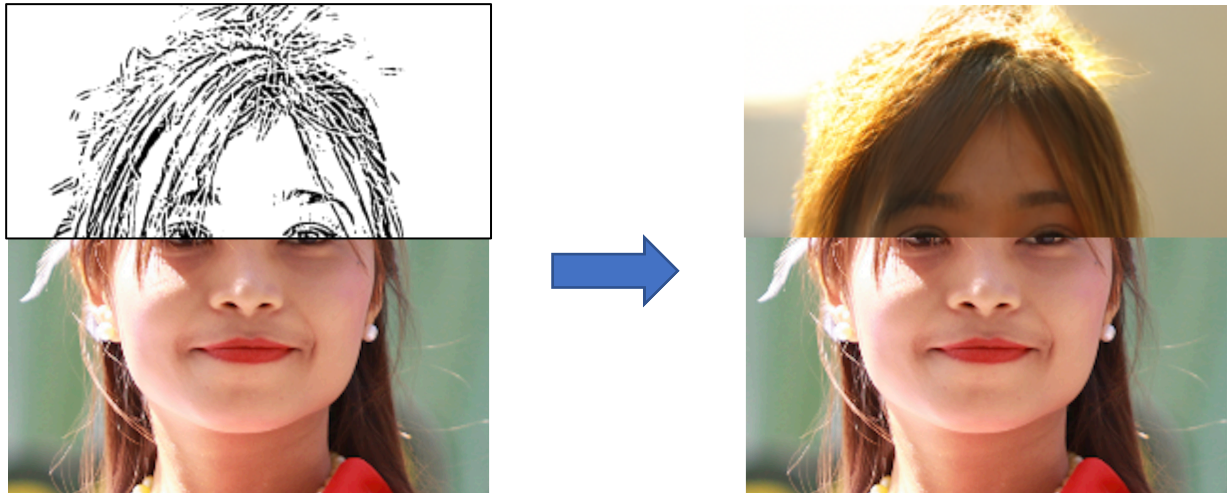
(c) Taming[3]

[1] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C] PMLR, 2020.

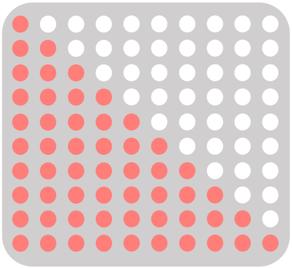
[2] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[J]. arXiv preprint arXiv:2102.12092, 2021.

[3] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C] CVPR, 2021.

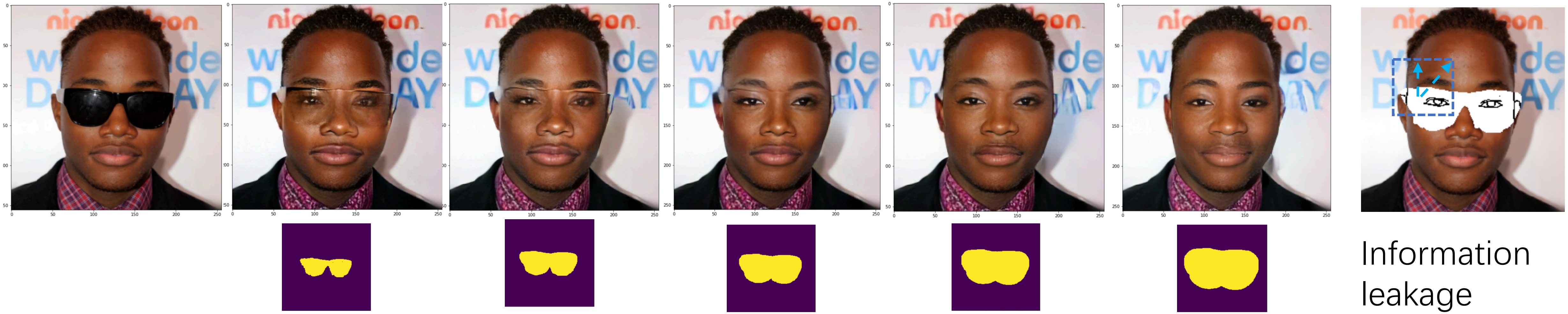
Problems



Causal Mask

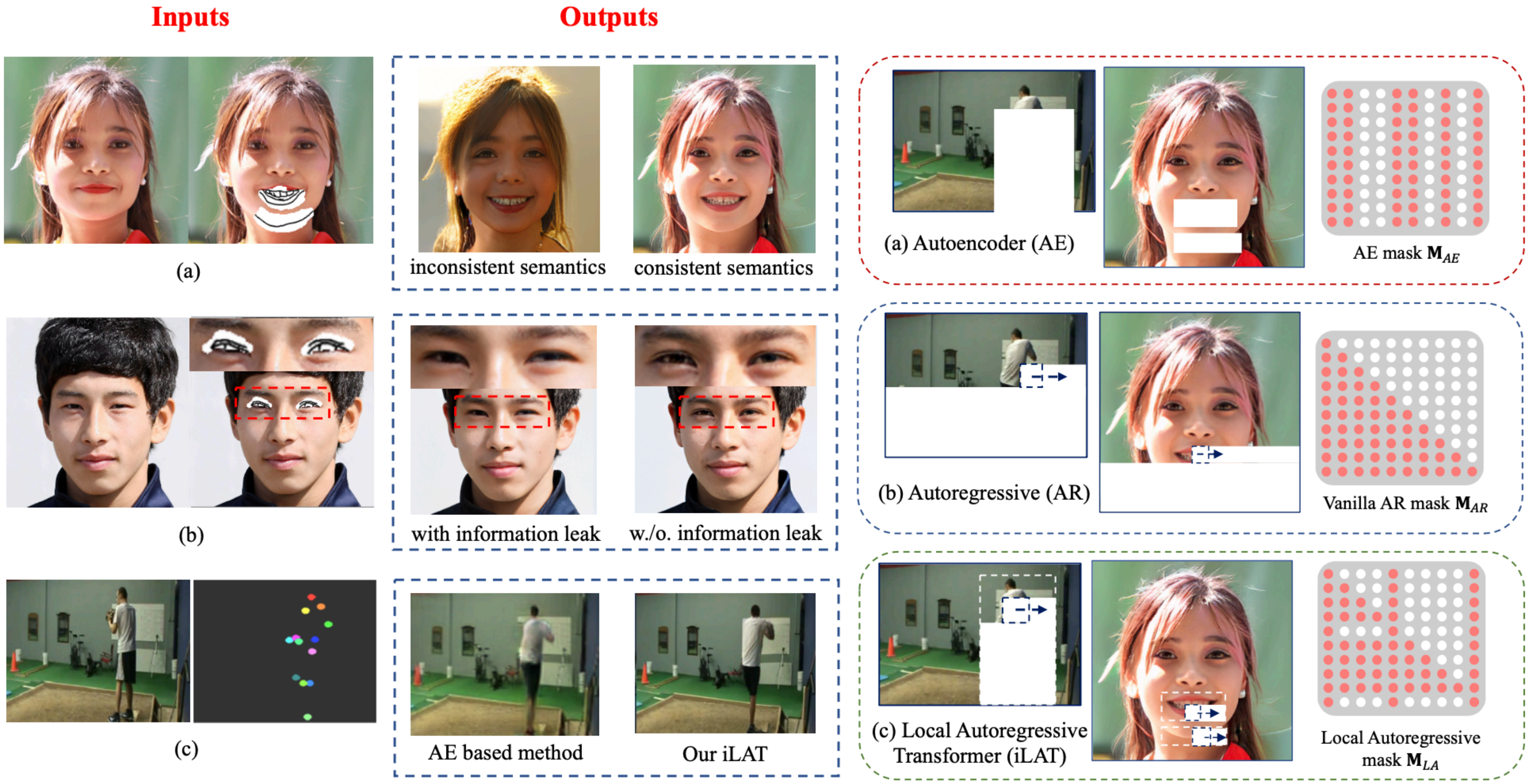


Inconsistent contexts



Information leakage

Introduction

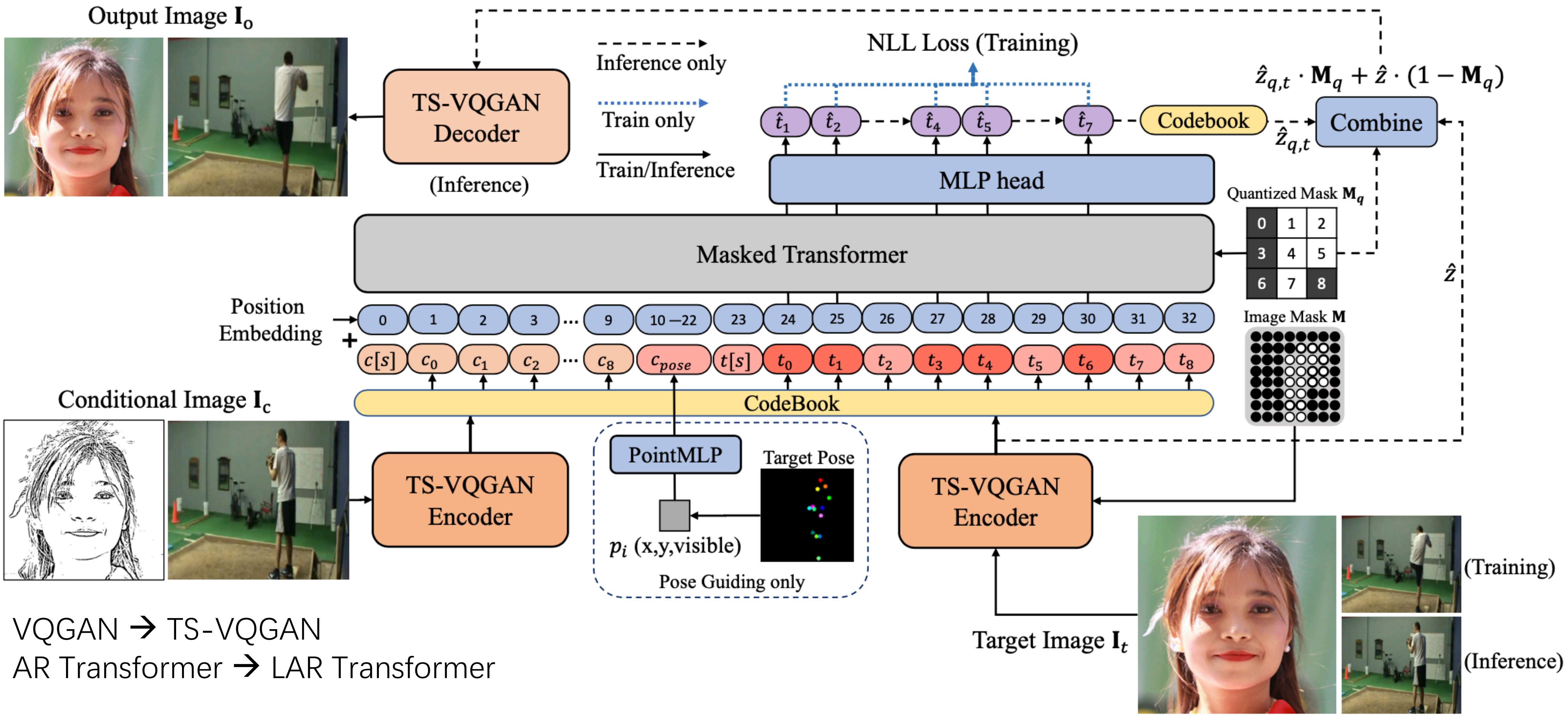


(A) Inputs and outputs of local generation compared with previous works (B) Comparison of different generative modes

Introduction

- **Motivation:**
- We propose an image **local autoregressive (LA)** transformer for local image synthesis, which enjoys both semantically consistent and realistic generative results.
- Two-stream convolutions and LA attention mask prevent both convolutions and transformer from **information leakage**, thus improving the quality of generated images.

Pipeline



VQGAN → TS-VQGAN

AR Transformer → LAR Transformer

Two-stream convolution based VQGAN

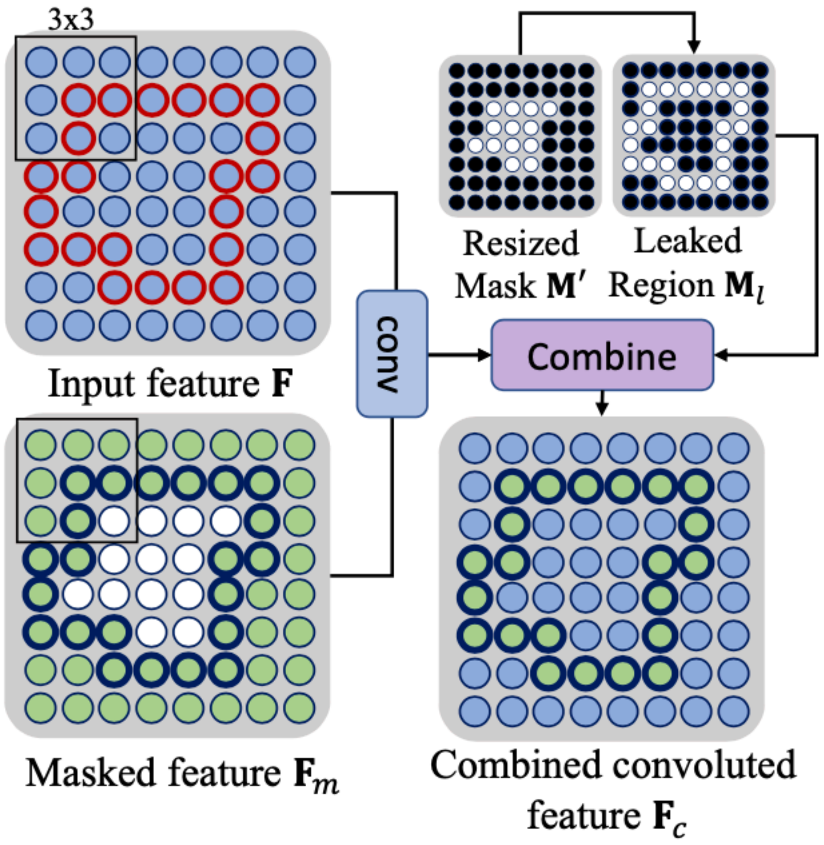
- For each convolution, replacing corrupted features with masked features.

$$\mathbf{M}_l = \text{clip}(\text{conv}_\mathbf{1}(\mathbf{M}'), 0, 1) - \mathbf{M}', \quad \mathbf{M}_l[\mathbf{M}_l > 0] = 1,$$

$$\mathbf{F}_c = \text{conv}(\mathbf{F}) \odot (1 - \mathbf{M}_l) + \text{conv}(\mathbf{F}_m) \odot \mathbf{M}_l.$$

- Unmasked features are directly encoded from the encoder, while masked features are replaced with the codebook vectors.

$$\mathbf{I}_o = D(z_q \odot \mathbf{M}_q + \hat{z} \odot (1 - \mathbf{M}_q)),$$



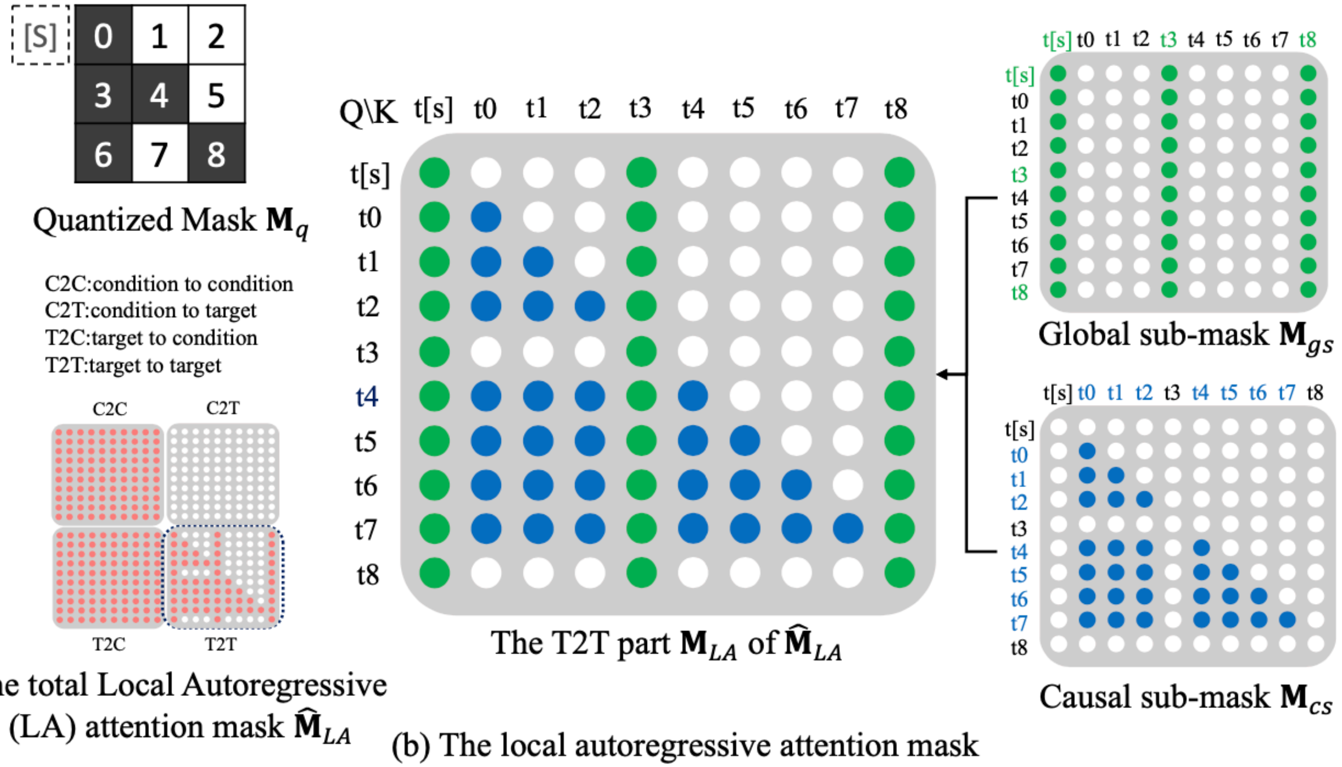
(a) The two-stream convolution

Local Autoregressive Mask

- Tokens are splited into global tokens and causal tokens.

$$p(t_m|c,t_u) = \prod_j p(t_{(m,j)}|c,t_u,t_{(m,<j)}).$$

$$\mathcal{L}_{NLL} = -\mathbb{E}_{t_m \sim p(t_m|c,t_u)} \log p(t_m|c,t_u).$$



Experiments

- Pose-guided generation of Penn Action (PA)
- Face-editing of Celeba-HQ and FFHQ
- Exploratory experiment: Synthetic DeepFashion (SDF) with complex backgrounds from Places2 for pose-guiding

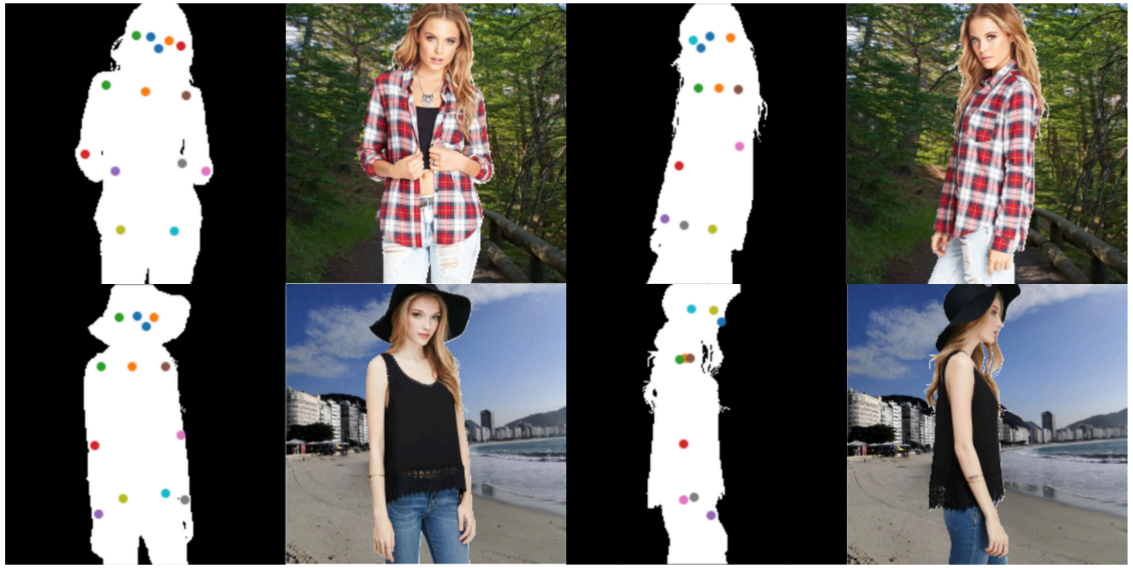


Figure 2: The illustration of the SDF dataset. Columns 1 and 3 are masks and pose landmarks (18 landmarks with -1 indicating invisible points), while columns 2 and 4 are related synthetic pictures.

Quantitative results

Table 1: Quantitative results in PA (left) and SDF (right). ↑ means larger is better while ↓ means lower is better. iLAT* indicates that iLAT trained without two-stream convolutions.

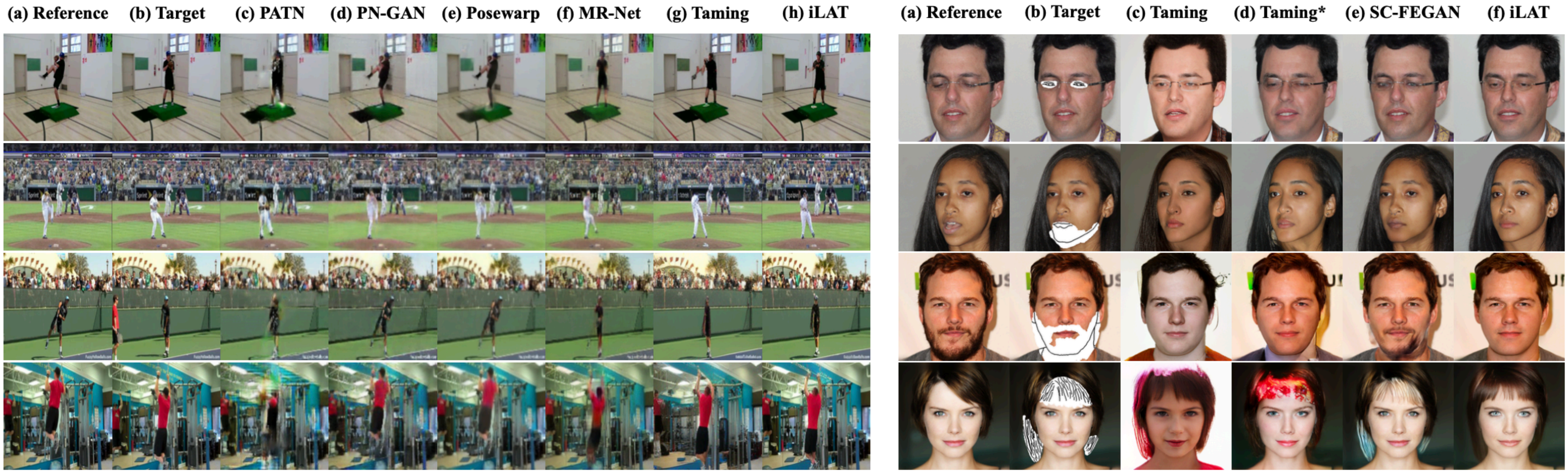
	PATN	PN-GAN	Posewarp	MR-Net	Taming	iLAT*	iLAT
PSNR↑	20.83	21.36	21.76	21.79	21.43	21.68	**22.94**
SSIM↑	0.744	0.761	0.794	0.792	0.746	0.748	**0.800**
MAE↓	0.062	0.062	0.053	0.066	0.057	0.056	**0.046**
FID↓	82.79	64.43	93.61	79.50	33.53	31.83	**27.36**

	Taming	iLAT
PSNR↑	16.25	**16.71**
SSIM↑	0.539	**0.599**
MAE↓	0.107	**0.096**
FID↓	72.77	**70.58**

Table 2: Average inference time (sec/image) in PA, SDF, and FFHQ of the vanilla AR transformer based generation (Taming) and iLAT. We also show the average masked rate of three datasets.

	masked rate	Taming	iLAT
PA	31.97%	8.551	**3.426**
SDF	28.09%	8.372	**3.898**
FFHQ	6.64%	8.183	**1.180**

Qualitative results

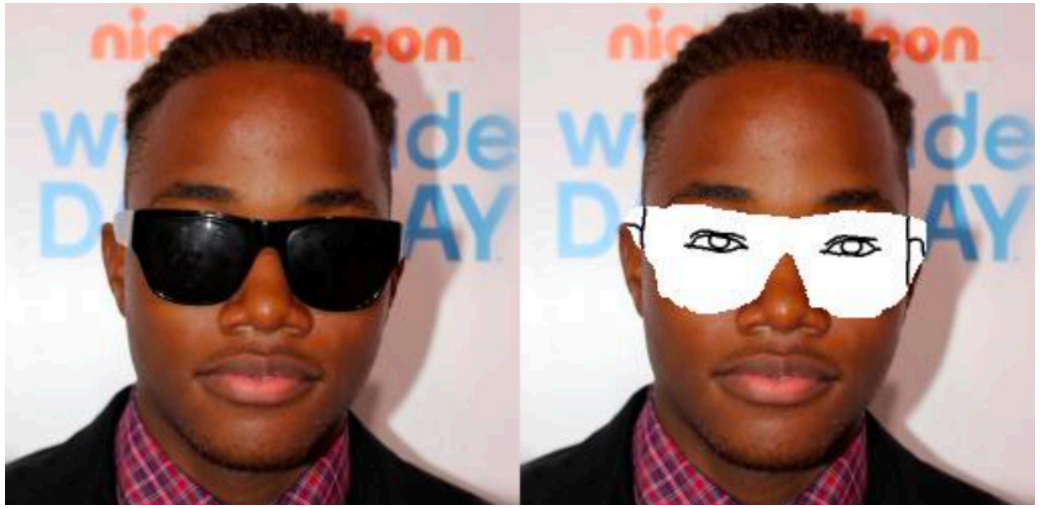


(A) Pose-Guided Generation in PA.

(B) FFHQ (row 1, 2) and CelebA (row 3, 4).

Figure 4: Qualitative results. Targets in (B) are combined with masks and XDoG sketches. Taming* means that the Taming transformer tested with our LA attention mask. Please zoom-in for details.

Ablations

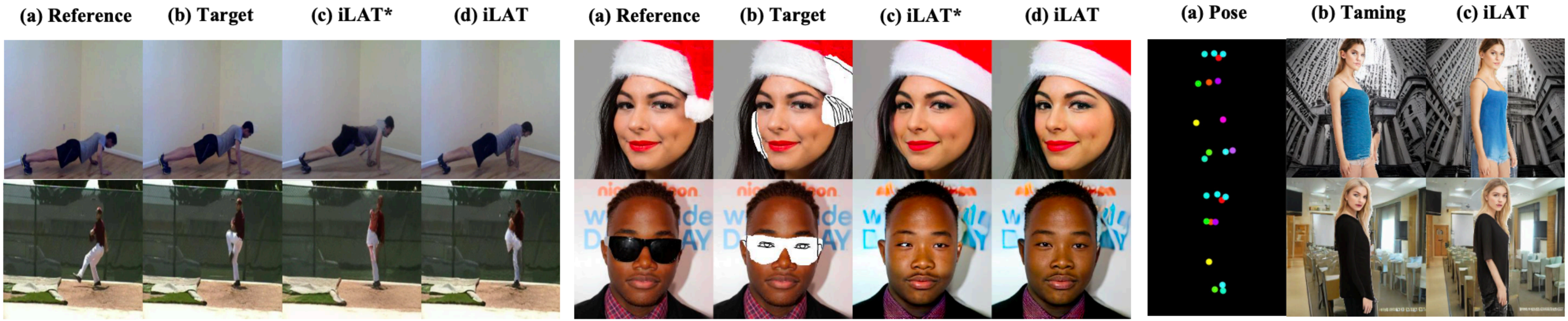


Trained in CelebaHQ | Trained in FFHQ

with TS
w/o mask dilation

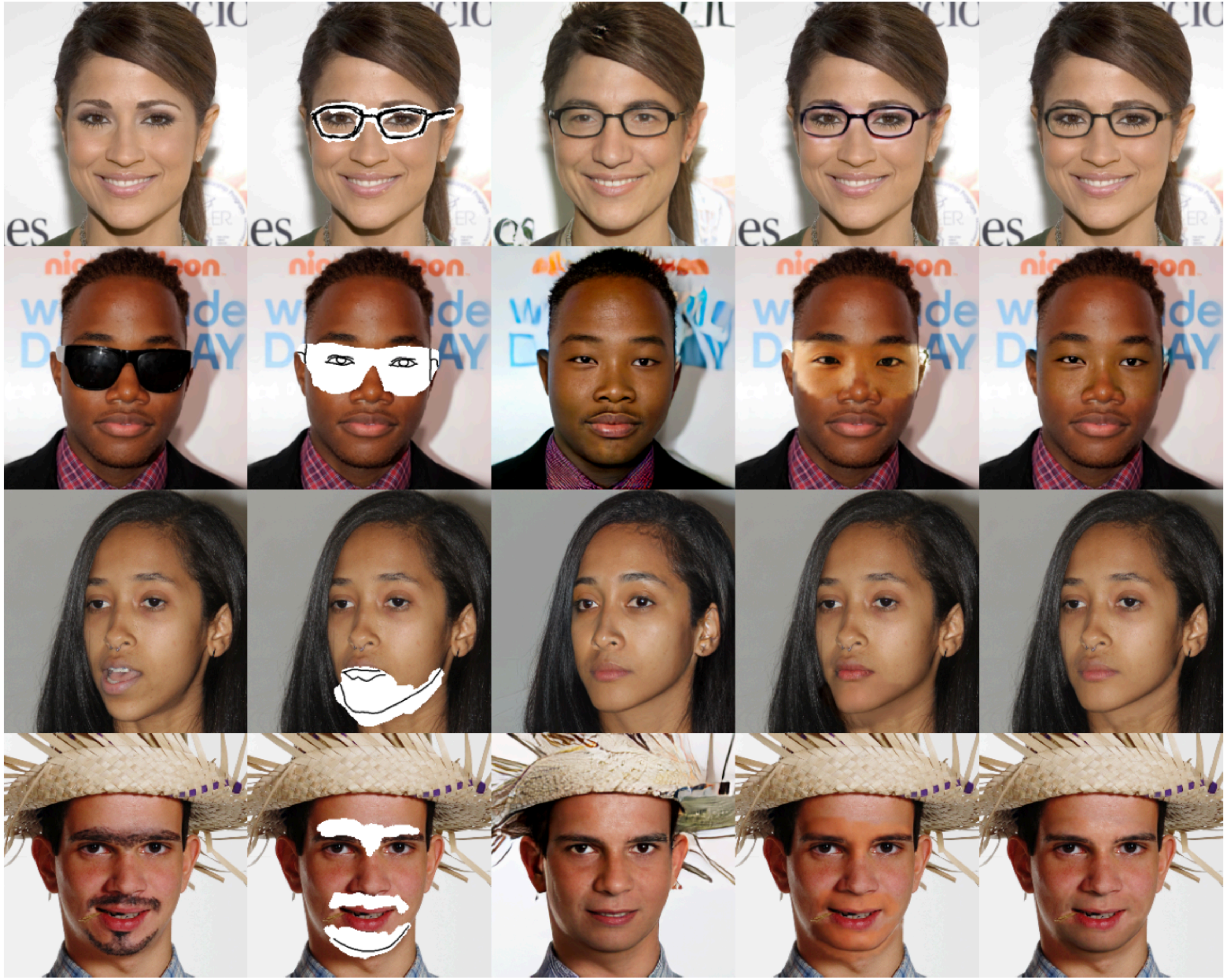
w/o TS
with mask dilation

Ablations



(A) Ablation in pose guiding (B) Ablation in face editing (C) Qualitative results in SDF

Figure 5: Ablation study for two-stream convolutions (A, B) and qualitative results in SDF (C). iLAT* means iLAT without two-stream convolutions. Please zoom-in for details.



(a) Reference (b) Target (c) iLAT (d) Taming+Reference (e) iLAT+Reference

Conclusions

- This method leverages a novel LA attention mask to enlarge the receptive fields of AR, which achieves not only semantically consistent but also realistic generative results.
- A two-stream convolution is proposed to learn a discrete representation learning without information leakages.

- Codes: https://github.com/ewrfcas/iLAT