FaceComposer: A Unified Model for Versatile Facial Content Creation

Jiayu Wang[*1], **Kang Zhao**[*1], **Yifeng Ma**[*2], **Shiwei Zhang**[1], **Yingya Zhang**[1], **Yujun Shen**[3], **Deli Zhao**[1], **Jingren Zhou**[1]

[1]Alibaba Group [2]Tsinghua University [3]Ant Group

{wangjiayu.wjy,zhaokang.zk,mayifeng.myf,zhangjin.zsw,yingya.zyy,jingren.zhou}@alibaba-inc.com

{shenyujun0302,zhaodeli}@gmail.com

FaceComposer

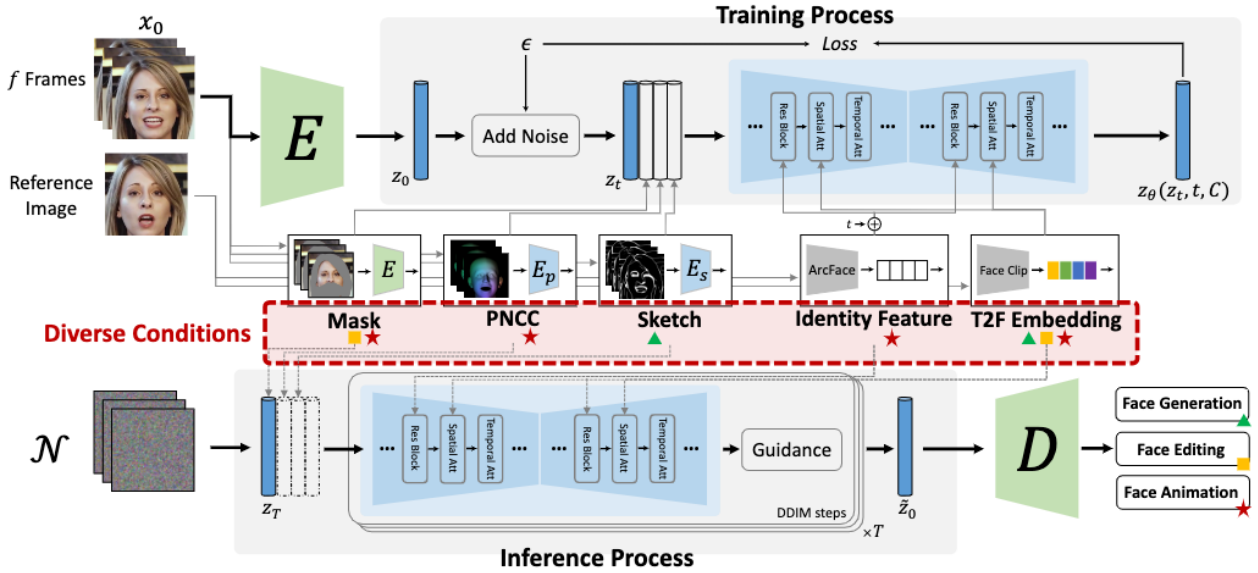


Figure 1: The framework of `FaceComposer`, which takes f frames and five face-related conditions as input, uses LDMs to predict the noise added in the latent space. We can combine diverse conditions to finish face generation/editing/animation or their combinations. For example, the green △ conditions are for face generation, yellow □ for face editing, and red ☆ for face animation.

- ✓ FaceComposer model various facial content creations as a multiple-condition-driven denoising process
- ✓ FaceComposer supports both static and dynamic content creations

Experiments Results

PNCCs
+
Anime
Girl

PNCCs
+
Cat
Woman

PNCCs
+
Harry
Potter

PNCCs
+
Elf

Motivations

- Existing face generative models are usually developed as highly customized systems, meaning that one model can only handle one task

- Limitations:

1. Hard to accomplish complex tasks, such as integrating face creating, editing and then animating the generated face in a single step
2. Redundant consumption of memory and computation. For example, one needs to train and save a number of models to build a multi-functional system, and perform complicated inference processes

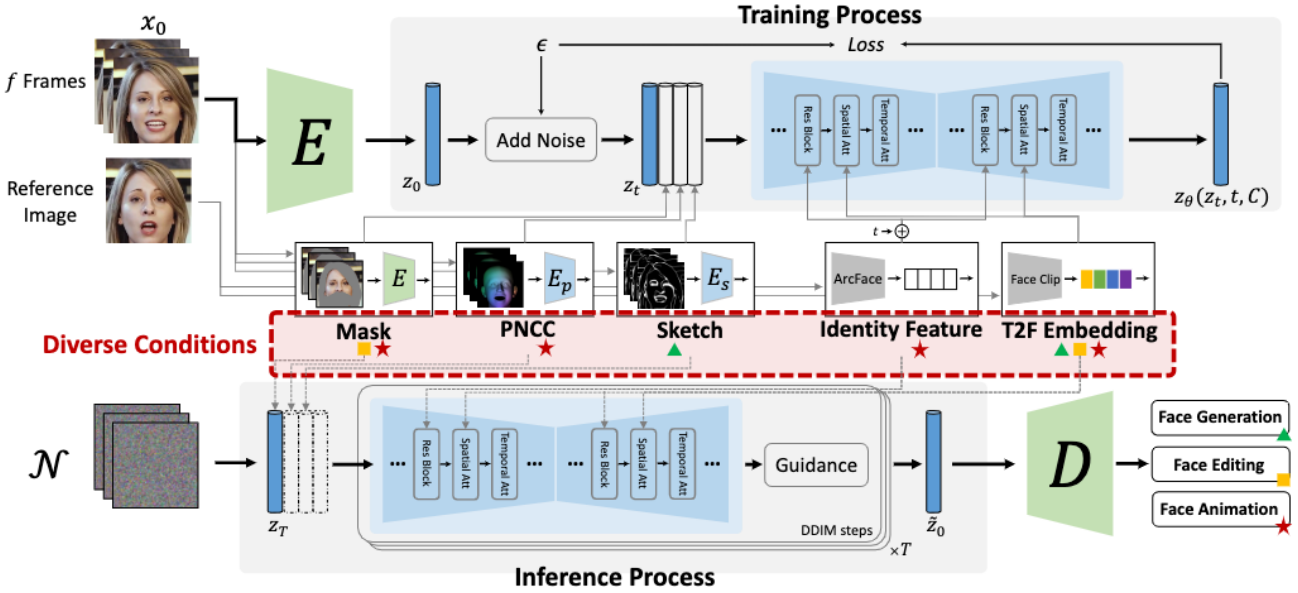
Versatile Creations

✓ **We propose compositional FaceComposer, a unified model that is capable of simultaneously tackling versatile facial tasks**

Table 1: Versatile creations based on condition compositions. *M*, *S*, *PNCCs*, *ID* and *T2F* are short for Mask, Sketch, PNCC sequence, Identity Feature, and T2F Embedding, respectively.

Single Creation		Versatile Creations	
Task	Conditions	Task	Conditions
face generation	① *T2F* ② *S* ③ ...	face generation+animation	① *PNCCs+T2F* ② *PNCCs+ID* ③ ...
face editing	① *M+T2F* ② *M+S* ③ ...	face generation+editing	① *ID+M* ② *ID+T2F* ③ ...
face animation	① *M+T2F+PNCCs* ② ...	face generation+editing+animation	① *ID+T2F+PNCCs* ② ...

FaceComposer



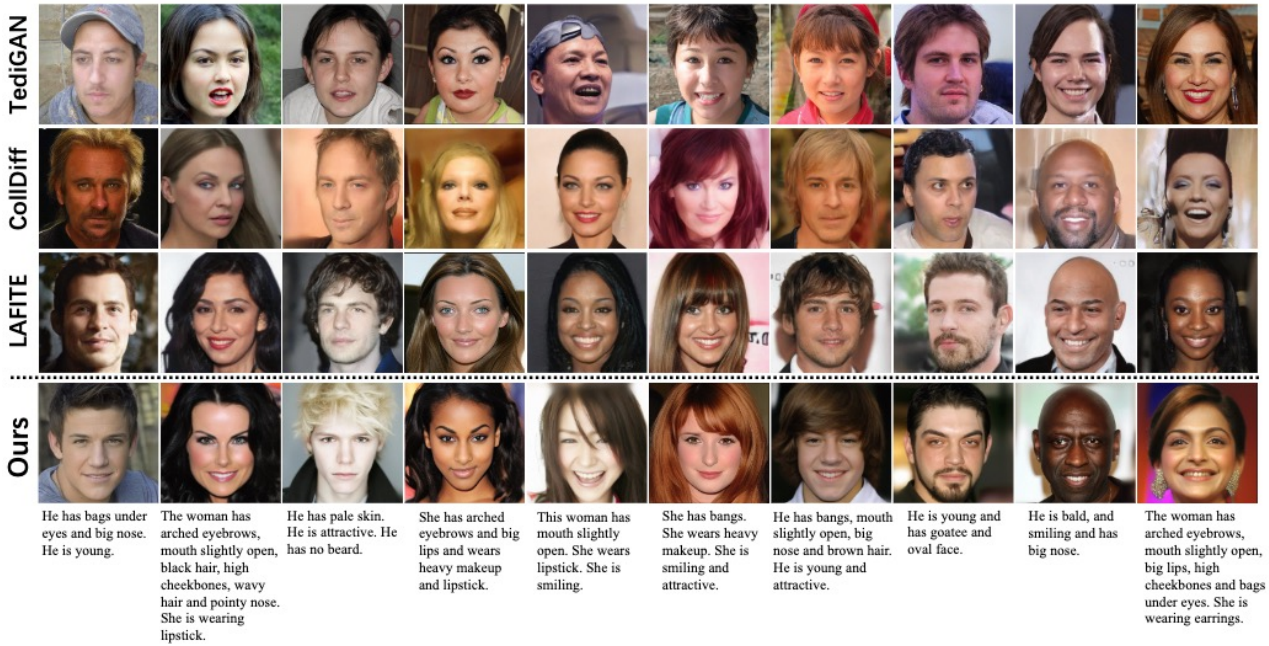
✓ FaceComposer takes f frames and five face-related conditions as input, uses LDMs to predict the noise added in the latent space

✓ FaceComposer can combine diverse conditions to finish face generation / editing / animation or their combinations

Face Generation

Table 2: Results of face generation.

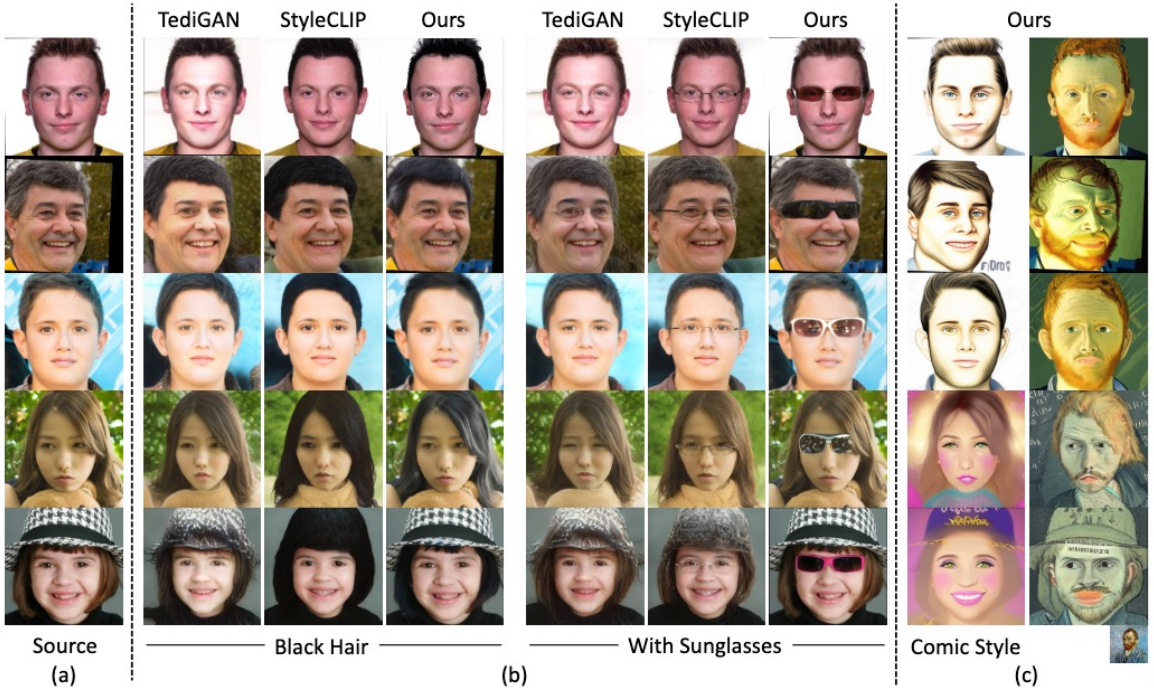
Method	FID↓	R-precision(%)↑	Accuracy↑	Realism↑
TediGAN [51]	107.14	44.96	2.80	3.55
CollDiff [13]	98.76	67.41	3.43	2.96
LAFITE [63]	12.54	81.94	4.07	3.88
Ours	**11.34**	**86.63**	**4.86**	**4.67**



Face Editing

Table 3: Results of face editing.

Method	IDS↑	Accuracy↑	Realism↑
TediGAN [51]	0.62	2.92	2.09
StyleClip [29]	0.75	4.27	2.87
Ours	**0.94**	**4.58**	**4.47**



Face Animation

Table 4: Results of face animation on HDTF and MEAD-Neutral.

Methods	HDTF / MEAD-Neutral						
	SSIM↑	CPBD↑	F-LMD↓	M-LMD↓	Sync_{conf}↑	SSIM-M↑	CPBD-M↑
MakeItTalk [62]	0.63/0.74	0.19/0.10	4.10/3.88	4.22/5.41	3.07/2.02	0.62/0.69	0.15/0.06
Wav2Lip [30]	0.75/0.79	0.18/0.12	2.01/2.38	2.54/2.95	**5.27/3.99**	0.68/0.78	0.08/0.03
PC-AVS [61]	0.51/0.51	0.23/0.07	3.64/4.76	3.52/3.91	4.16/3.09	0.60/0.67	0.10/0.05
AVCT [46]	0.73/0.77	0.17/0.10	2.85/2.68	3.53/4.46	3.81/2.56	0.70/0.73	0.16/0.08
EAMM [14]	0.59/0.41	0.08/0.08	4.16/7.39	4.19/5.03	2.30/1.40	0.60/0.71	0.13/0.05
StyleTalk [26]	**0.78**/0.79	0.23/0.12	2.10/2.35	2.40/2.80	4.17/3.05	0.76/0.80	0.16/0.09
SadTalker [58]	0.61/0.73	0.21/0.12	3.98/3.67	3.46/4.09	4.05/2.62	0.61/0.69	0.15/**0.10**
StyleSync [7]	0.77/0.80	0.21/0.12	1.93/2.22	2.36/2.76	4.21/3.10	0.76/0.80	0.17/**0.10**
Ground Truth	1/1	0.23/0.20	0/0	0/0	4.52/3.57	1/1	0.21/0.12
FaceComposer	**0.78/0.84**	**0.27/0.14**	**1.84/2.16**	**2.25/2.70**	4.27/3.12	**0.78/0.83**	**0.18/0.10**

Table 5: User study results of different methods on HDTF and MEAD-Neutral for the face animation. *LS*, *VQ*, *OR* stand for user study metrics *LipSync*, *VideoQuality* and *OverallRealness*, respectively.

Methods	HDTF / MEAD-Neutral		
	LS↑	VQ↑	OR↑
MakeItTalk	1.71/2.20	1.87/2.38	1.44/1.74
Wav2Lip	2.93/3.33	1.02/1.10	1.10/1.12
PC-AVS	2.88/3.20	1.97/2.46	1.73/1.98
AVCT	2.04/2.76	2.60/2.66	2.46/2.62
EAMM	1.90/2.58	1.30/1.78	1.62/1.94
StyleTalk	3.10/3.60	3.08/3.00	2.44/2.82
SadTalker	3.22/3.68	2.82/2.92	1.97/2.42
Ground Truth	4.34/4.56	4.06/4.26	4.22/4.40
FaceComposer	**3.53/3.96**	**3.38/3.50**	**2.93/3.73**

Face Animation

Figure 4: Qualitative results of face animation. It can be seen that `FaceComposer` not only achieves accurate lip-sync but also produces high-fidelity results in the mouth area.

More Results

Figure R2: Results are conditioned on Identity Feature and Mask.

More Results

Wav2Lip

MakeItTalk

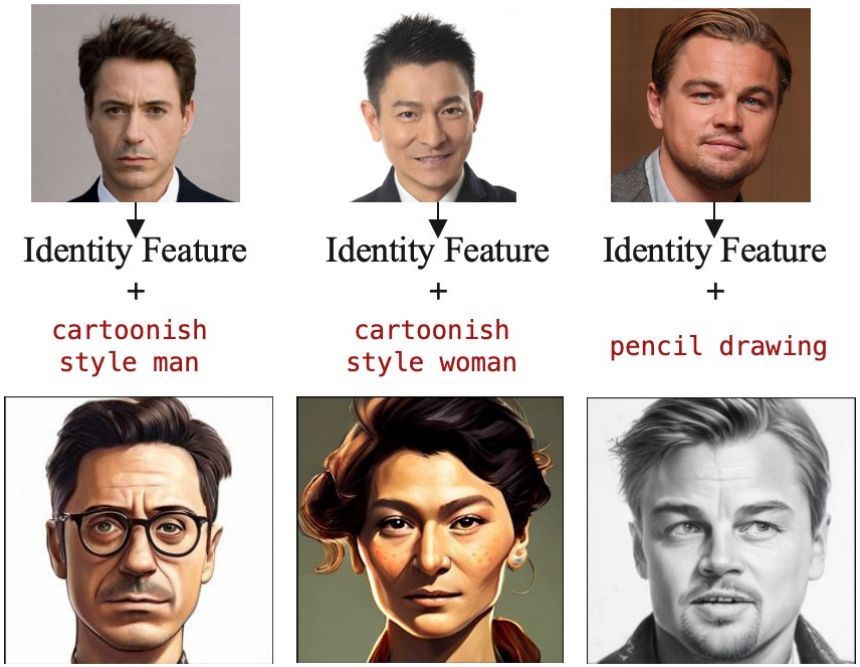
PC-AVS

AVCT

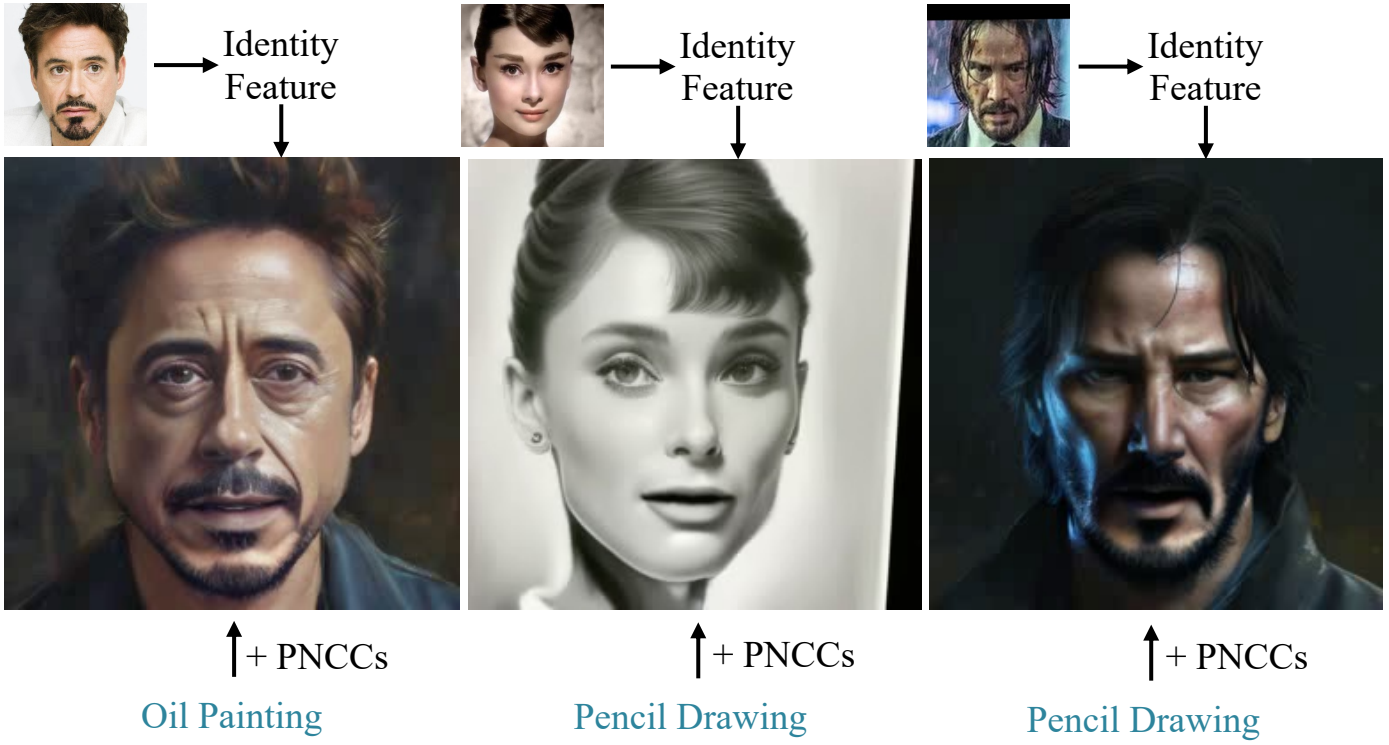
EAMM

FaceComposer

More Results



More Results



Q&A