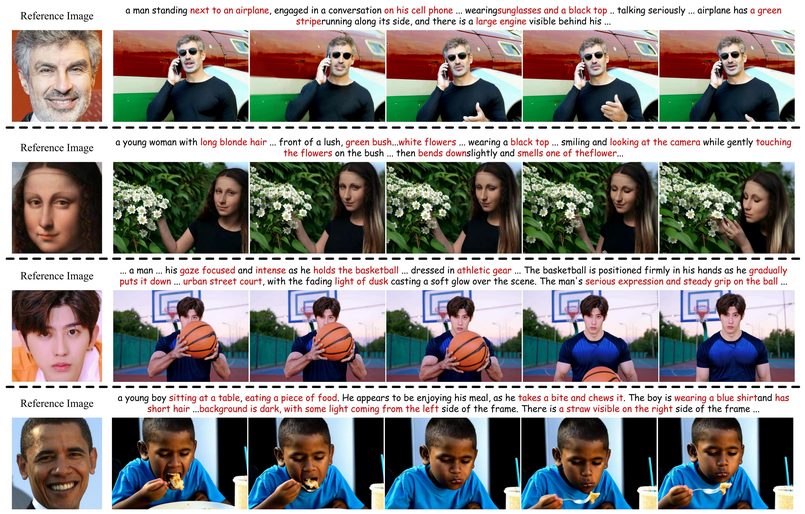
Identity-Preserving Text-to-Video Generation by Frequency Decomposition

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, Li Yuan

CVPR Nashville JUNE 11-15, 2025

Contributions

- A tuning-free pipeline without tedious case-by-case finetuning.
- A frequency-aware heuristic identity-preserving Diffusion Transformer (DiT) -based control scheme.
- A novel data pipeline to process and construct a high quality identity preservation video dataset



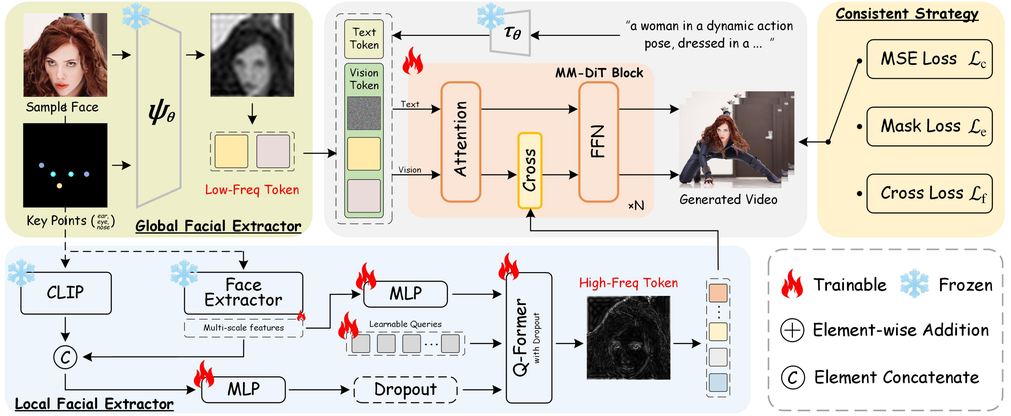
Key Findings

These findings are not merely synthesized from and validated by existing diffusion and ViT literature, but with additional support by our experiments.

- **Finding 1.** Low-frequency features are essential for pixel-level prediction tasks in diffusion models, as they ease model training. *This can solve the problem of DiT being difficult to train.*
- **Finding 2.** Transformers have limited perception of high-frequency information, which is important for preserving facial features. *Inspires us to design local extractor to enhance performance.*

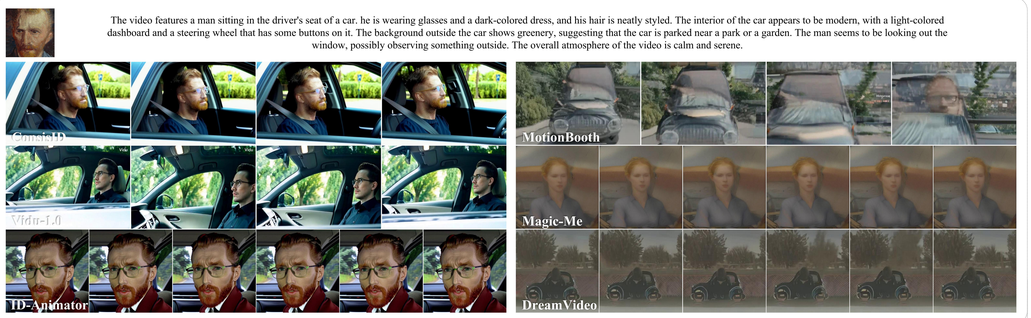
Framework

- *ConsisID is designed based on both high-frequency and low-frequency perspectives and is complemented by the Consistency Training Strategy.*



Comparison

- *We surpasses existing tuning-free/based, open/closed-source and I2V methods.*

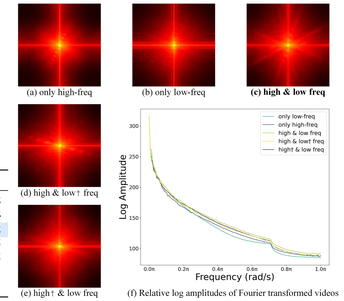


	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
Vidu 1.5 [6]	0.36	0.39	32.87	215.42
ConsisID	**0.52**	**0.54**	**33.08**	**163.68**
DreamVideo [72]	0.03	0.03	26.03	237.91
MotionBooth [73]	0.05	0.06	24.42	287.90
Magic-Me [46]	0.09	0.10	23.14	237.35
ConsisID	**0.46**	**0.47**	**27.45**	**181.97**

	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
CogVideoX-5B-I2V [78]	0.37	0.38	**28.53**	201.69
EasyAnimate v4 [76]	0.15	0.15	27.95	235.68
OpenSora-Plan v1.3 [37]	0.31	0.32	27.25	224.99
DynamiCrafter512 [74]	0.25	0.26	29.76	212.13
ConsisID	**0.46**	**0.47**	27.45	**181.97**

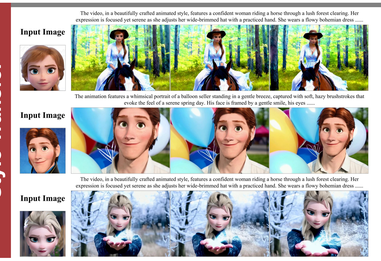
Validate Key Findings

- Training instability when low-frequency features are omitted.
- Injecting high-frequency feature can lead to improvements.



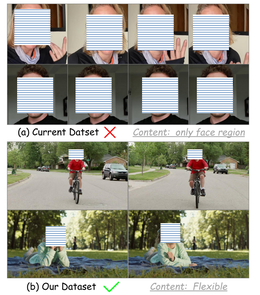
Plan	FaceSim-Arc ↑	FaceSim-Cur ↑	CLIPScore ↑	FID ↓
a	0.05	0.05	34.86	269.88
b	0.66	0.68	34.48	**104.34**
c	**0.73**	**0.75**	**36.77**	127.42
d	0.64	0.68	30.69	177.65
e	0.62	0.66	33.61	164.15
f	*unstable training process*			
g	*unstable training process*			

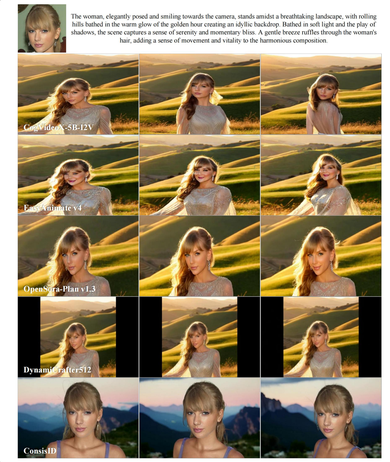
Style Transfer



- *ConsisID can be zero-shot transferred to animation generation.*

Dataset Quality





Video Page

Paper Code