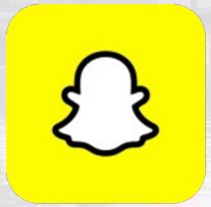
Video Alchemist

Multi-subject Open-set Personalization in Video Generation

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee,
Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, Sergey Tulyakov



CVPR Nashville JUNE 11-15, 2025

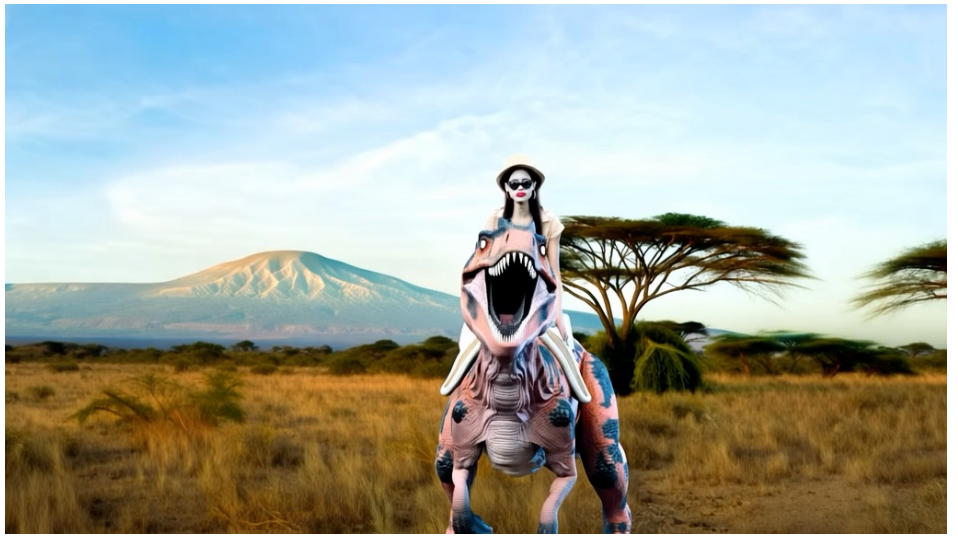
Overview

Text-to-Video Model

"A woman rides a dinosaur on a field."

Video Alchemist

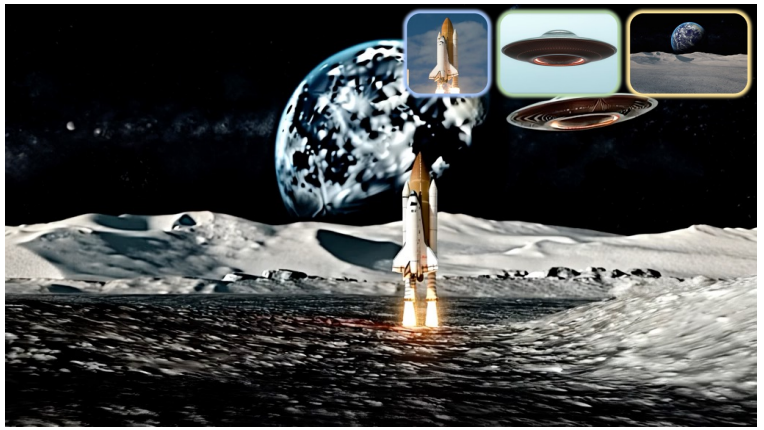
"A woman rides a dinosaur on a field."



Overview

"A man and a woman discuss something in a meeting room."

"A woman in a suit sits in a living room and drinks tea."



"A rocket launches from the Moon's surface with a UFO behind."

"A man pets a dog on a bridge."

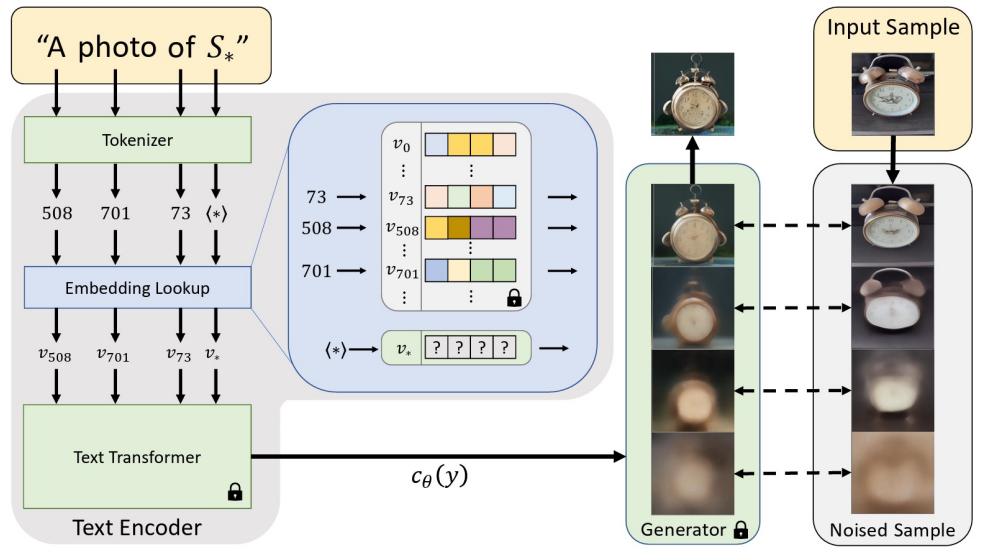
"A man pets a dog on desert."

"A man pets a dog on the Moon's surface."

Previous Methods

[Optimization-based Methods] Optimize a text token representing a new concept

Textual Inversion[1]



Other optimization-based methods: DreamBooth[2], Custom Diffusion[3]

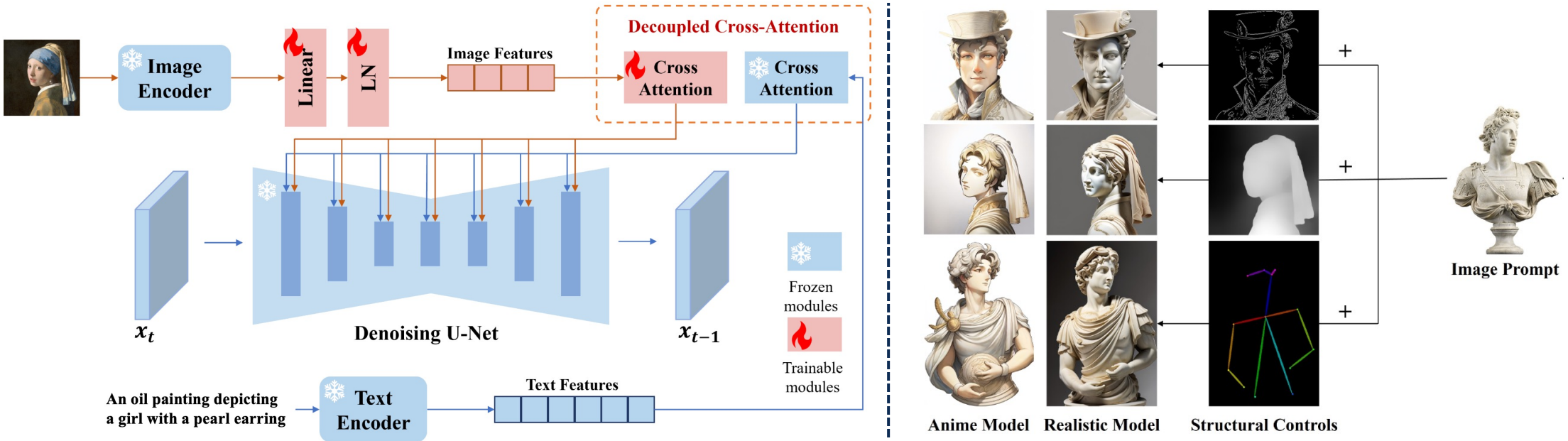
[Drawback] Require time-consuming optimization to adapt a new concept.

[1] Rinon Gal, et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*, 2023.
[2] Nataniel Ruiz, et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 2023.
[3] Nupur Kumari, et al. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*, 2023.

Previous Methods

[Encoder-based Methods] Generate image/video using the embeddings of reference images as conditions

IP-Adapter[1]

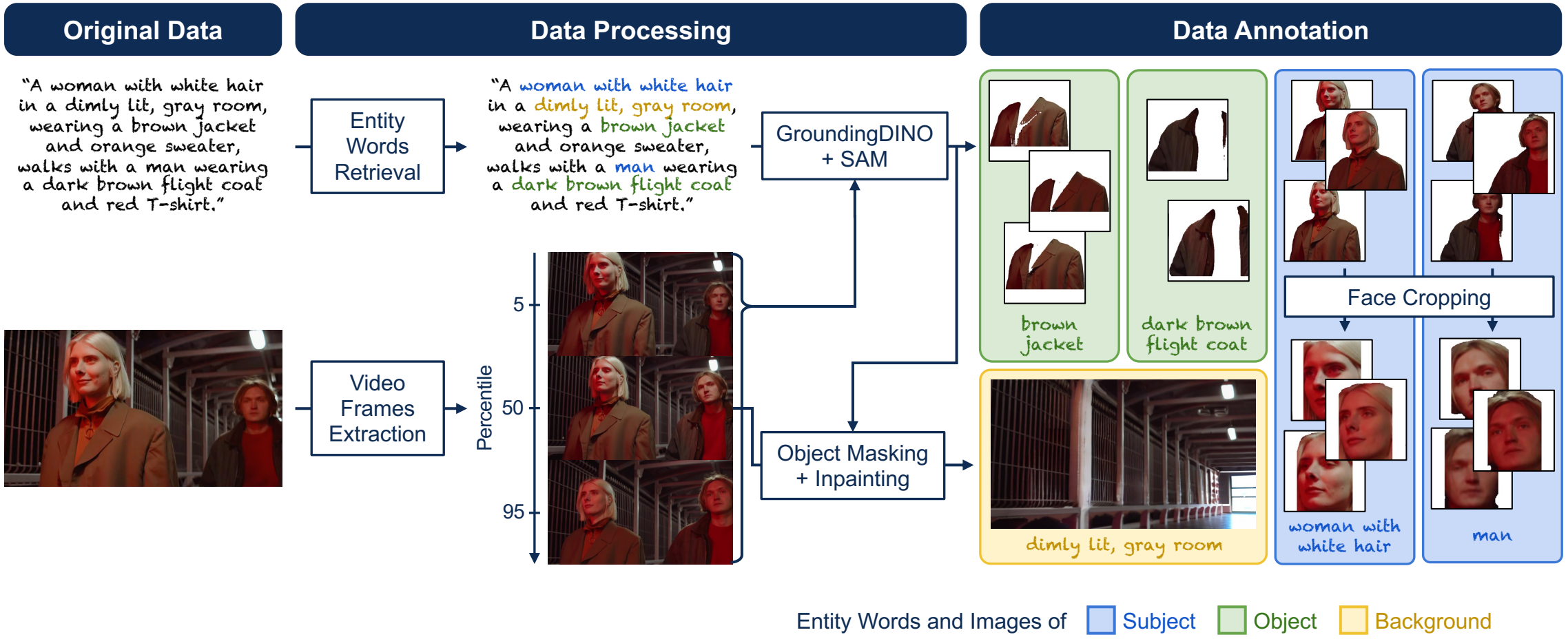


Other encoder-based methods: InstanceBooth[2], VideoBooth[3]

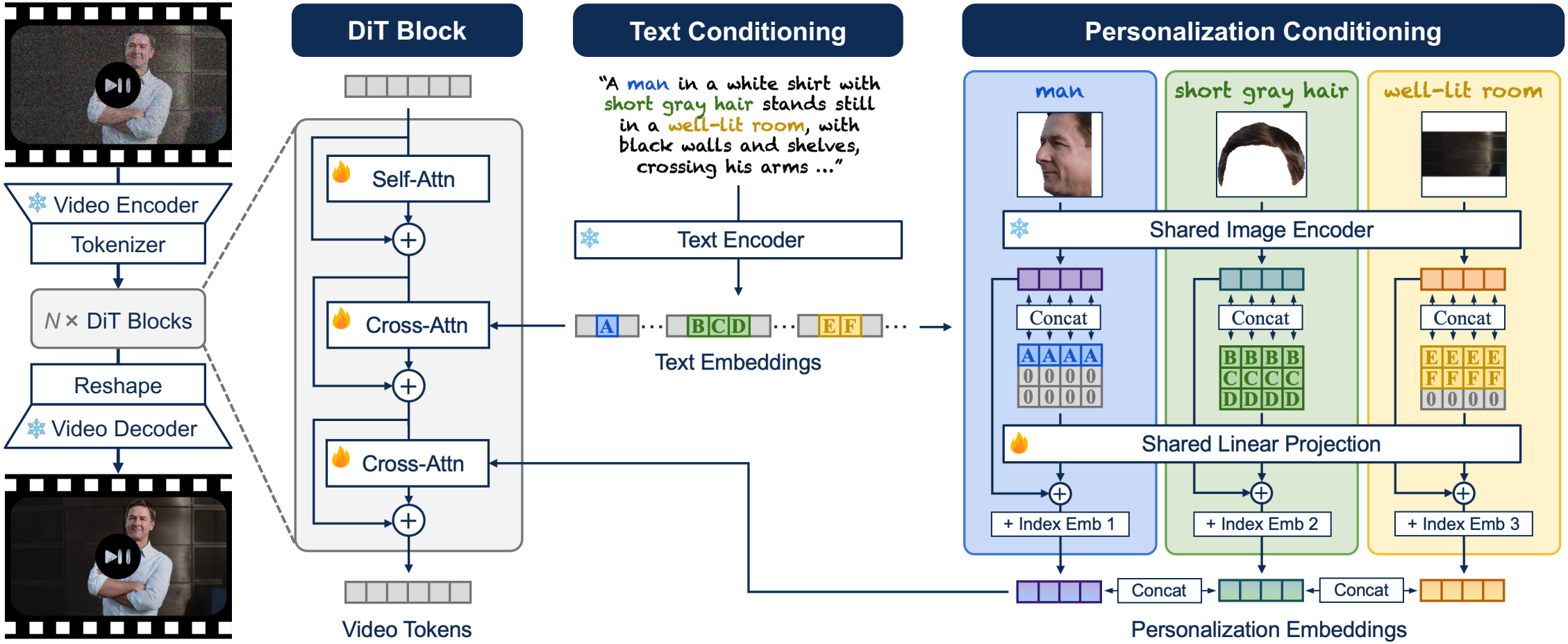
We also adapt encoder-based method to eliminate expensive test-time optimization and further extend it to more generic personalization.

[1] Hu Ye, et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *Arxiv preprint*, 2023.
[2] Jing Shi, et al. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. In *CVPR*, 2024.
[3] Yuming Jiang, et al. VideoBooth: Diffusion-based Video Generation with Image Prompts. In *CVPR*, 2024.

Methodology: Dataset



Methodology: Model



Methodology: Reducing Model Overfitting

Copy-and-Paste Issue

Reference Image

IP-Adapter[1]

Model Overfitting		Proposed Augmentations
If the reference image is high-resolution, the model generates a large subject close to the camera.	→	Downscaling + Gaussian blurring
If the reference image is cropped, the model places the subject at the edge, causing it to be cropped by the video boundary.	→	Random crop
The model often replicates the subject's pose and lighting conditions from the reference image.	→	Color jittering + Brightness adjustment
If multiple reference images represent the same subject with similar poses, the model produces a subject with minimal motion.	→	Random horizontal flip + Image shearing & rotation

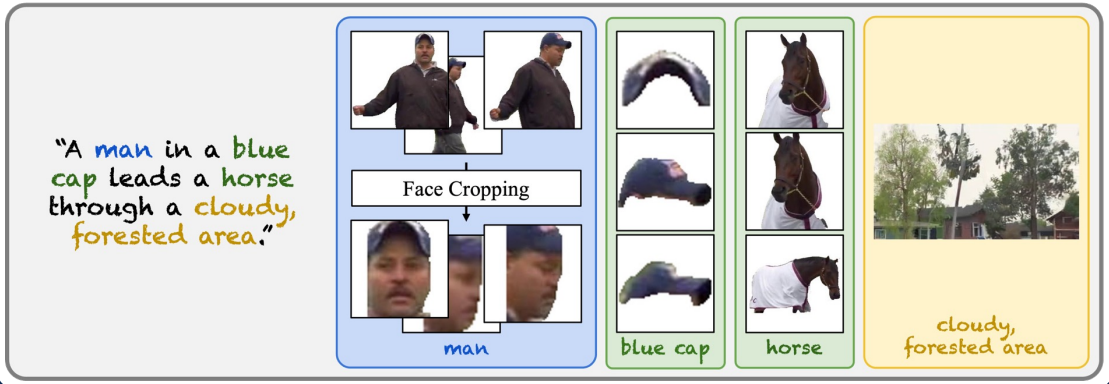
[1] Hu Ye, et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *Arxiv preprint*, 2023.

Benchmark: *MSRVTT-Personalization*

Video Sample

[Source] MSR-VTT dataset[1]

Personalization Annotations



Evaluation Metrics

- Text similarity[2]
- Video similarity[3]
- Subject similarity
- Face similarity
- Dynamic degree[4]

Check the benchmark!

[1] Jun Xu, et al. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 2016.
[2] Chenfei Wu, et al. GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. *Arxiv preprint*, 2021.
[3] Rinon Gal, et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*, 2023.
[4] Ziqi Huang, et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *CVPR*, 2024.

Experiments: Comparisons with SOTA

Comparisons for the Subject Mode of *MSRVTT-Personalization*

Method	Reference Images		Text-S↑	Vid-S↑	Sub-S↑	Dync-D↑
	Subject	Background				
ELITE*[1]	single	✗	0.245	0.620	0.359	-
VideoBooth[2]	single	✗	0.222	0.612	0.395	0.448
DreamVideo[3]	single	✗	0.261	0.611	0.310	0.311
Video Alchemist	single	✗	0.269	0.732	0.617	0.466
DreamVideo[3]	multiple	✗	0.253	0.604	0.256	0.303
Video Alchemist	multiple	✗	0.268	0.743	0.626	0.473
Video Alchemist	multiple	✓	0.254	0.780	0.570	0.506

*For text-to-image models, outputs are treated as single-frame videos without evaluating temporal quality.

[1] Yuxiang Wei, et al. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *ICCV*, 2023.
[2] Yuming Jiang, et al. VideoBooth: Diffusion-based Video Generation with Image Prompts. In *CVPR*, 2024.
[3] Yujie Wei, et al. DreamVideo: Composing Your Dream Videos with Customized Subject and Motion. In *CVPR*, 2024.

Experiments: Comparisons with SOTA

Comparison for the Face Mode of *MSRVTT-Personalization*

Method	Reference Images: Face Crop	Text-S↑	Vid-S↑	Face-S↑	Dync-D↑
IP-Adapter*[1]	single	0.251	0.648	0.269	-
PhotoMaker*[2]	single	0.278	0.569	0.189	-
Magic-Me[3]	single	0.251	0.602	0.135	0.418
Video Alchemist	single	0.273	0.687	0.382	0.424
PhotoMaker*[2]	multiple	0.275	0.582	0.216	-
Magic-Me[3]	multiple	0.248	0.618	0.153	0.385
Video Alchemist	multiple	0.272	0.694	0.411	0.402

*For text-to-image models, outputs are treated as single-frame videos without evaluating temporal quality.

[1] Hu Ye, et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *Arxiv preprint*, 2023.
[2] Zhen Li, et al. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *CVPR*, 2024.
[3] Ze Ma, et al. Magic-Me: Identity-Specific Video Customized Diffusion. *Arxiv preprint*, 2024.

Experiments: Comparisons with SOTA

User Preference Study for the Subject / Face mode of *MSRVTT-Personalization*

Method	Preference Ratio↑	
	Quality	Fidelity
ELITE	2.7%	0.6%
VideoBooth	0.3%	0.8%
DreamVideo	0.5%	0.5%
Video Alchemist	96.5%	98.1%

Method	Preference Ratio↑	
	Quality	Fidelity
IP-Adapter	10.4%	20.2%
PhotoMaker	37.5%	7.4%
Magic-Me	4.4%	4.0%
Video Alchemist	47.6%	68.4%

Experiments: Comparisons with Pika[1]

[1] https://pika.art/

Experiments: Ablation Study

Method	Image Encoder	Use Word Token	Image Augmentations	Text-S↑	Vid-S↑	Sub-S↑	Dync-D↑
Use CLIP	CLIP[1]	✗	✗	0.269	0.768	0.569	0.552
No word token	DINOv2[2]	✗	✗	0.256	0.790	0.566	0.569
No augmentation	DINOv2[2]	✗	✗	0.251	0.781	0.609	0.506
Video Alchemist	DINOv2[2]	✓	✓	0.257	0.790	0.600	0.570

"A woman smiles and looks at a dog on a beach with waves lapping."

Use CLIP

No word token

No augmentation

Video Alchemist

[1] Alec Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.
[2] Maxime Oquab, et al. DINOv2: Learning Robust Visual Features without Supervision. In *TMLR*, 2024.

Video Alchemist

Multi-subject Open-set Personalization in Video Generation

Project Page

Github