



Video Alchemist — Multi-subject Open-set Personalization in Video Generation

Tsai-Shien Chen^{1,2}, Aliaksandr Siarohin¹, Willi Menapace¹, Yuwei Fang¹, Kwot Sin Lee¹,
Ivan Skorokhodov¹, Kfir Aberman¹, Jun-Yan Zhu³, Ming-Hsuan Yang², Sergey Tulyakov¹

¹Snap Inc., ²UC Merced, ³CMU



Multi-subject Open-set Video Personalization

Ref Images

Generated Video



"A woman rides a dinosaur on the field."



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



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



"A man pets a dog on bridge."



"A man pets a dog on desert."



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

Dataset Collection Pipeline

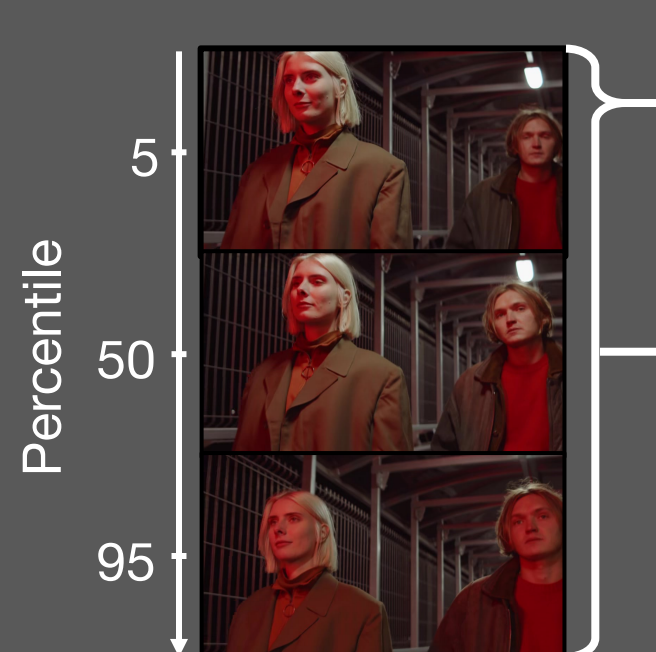
Original Data

"A woman with white hair in a dimly lit, gray room, wearing a brown jacket and orange sweater, walks with a man wearing a dark brown flight coat and red T-shirt."

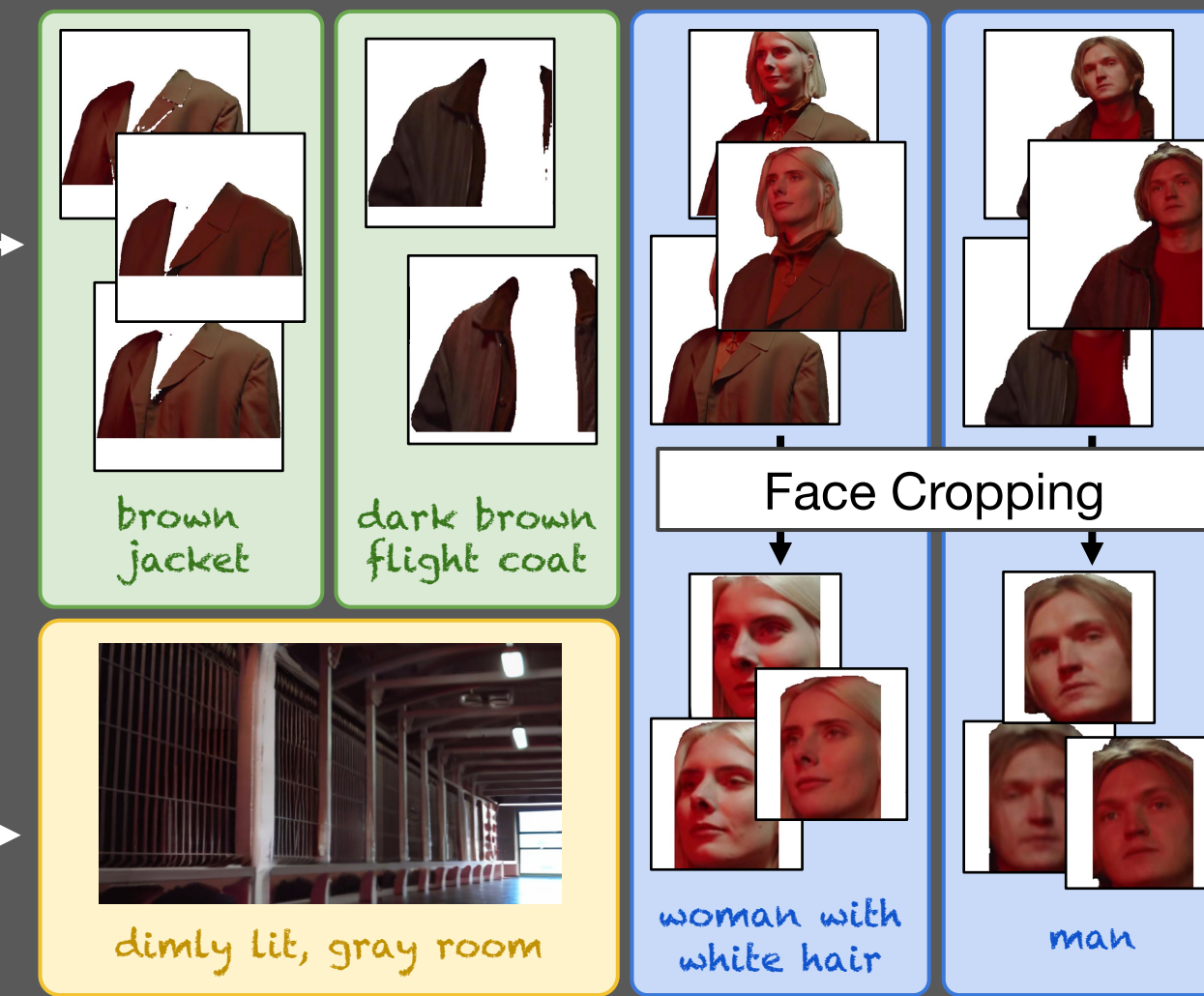


Data Processing

"A woman with white hair in a dimly lit, gray room, wearing a brown jacket and orange sweater, walks with a man wearing a dark brown flight coat and red T-shirt."



Data Annotation

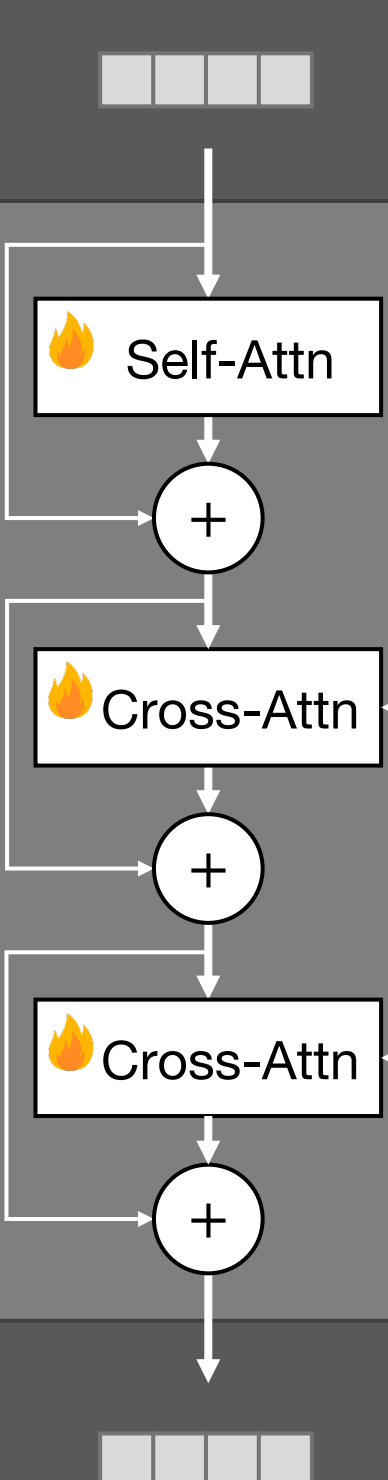


Entity Words and Images of ■ Subject ■ Object ■ Background

Model Architecture and Training

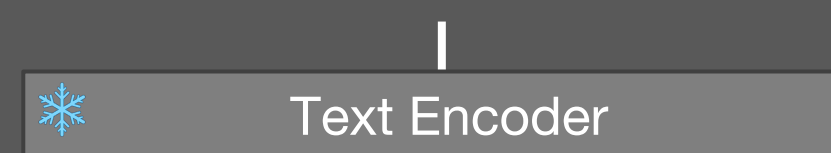


DiT Block

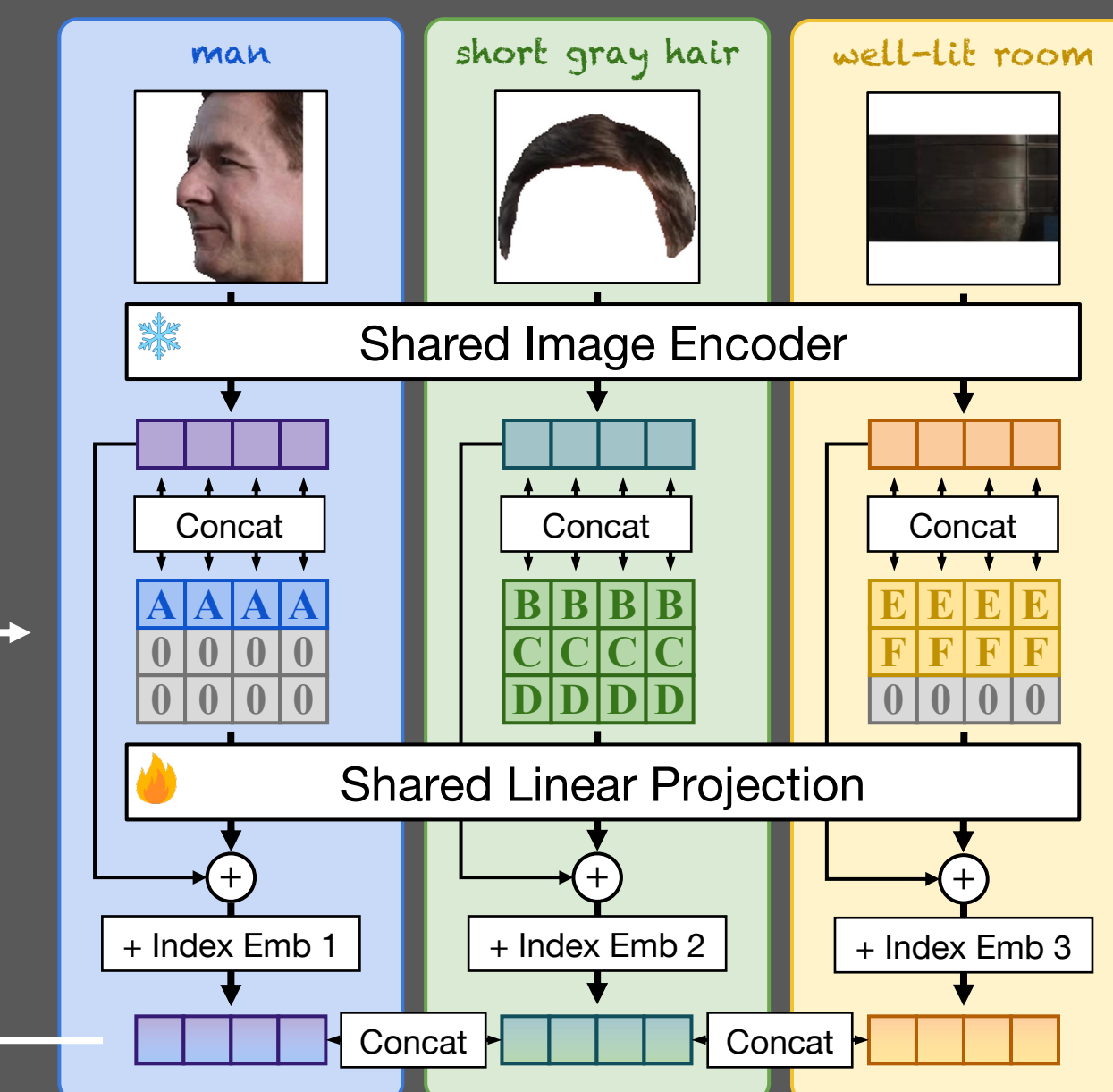


Text Conditioning

"A man in a white shirt with short gray hair stands still in a well-lit room, with black walls and shelves, crossing his arms ..."



Personalization Conditioning



"Copy-and-Paste" Effect

Propose personalization-specific augmentation to guide the model to focus on the subject's identity.

New Benchmark MSRVTT-Personalization

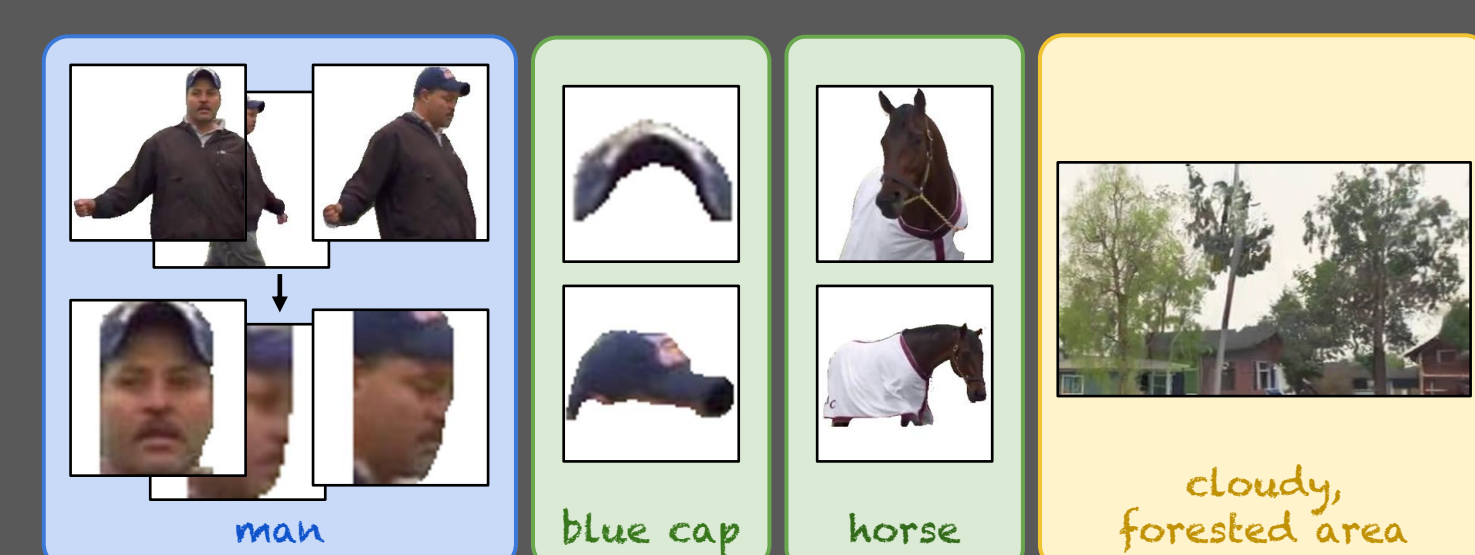
Ground Truth



Metrics

- Text similarity
- Video similarity
- Subject similarity
- Face similarity
- Dynamic degree

Personalization Annotations



Learn More?

Project Page



Github

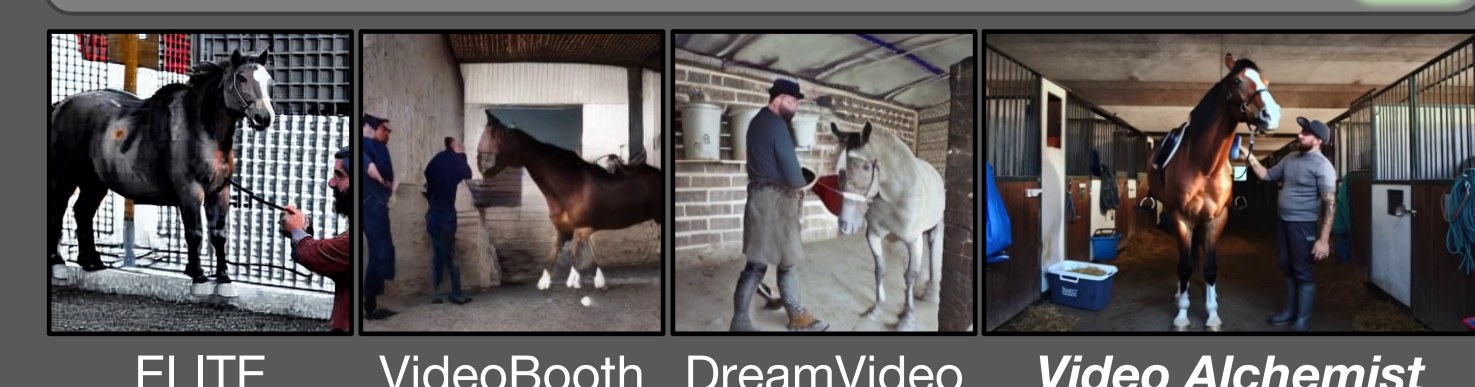


Comparisons with the SOTAs

Open-set Subject Mode

Method	Text-S $\uparrow$	Vid-S $\uparrow$	Subj-S $\uparrow$	Dync $\uparrow$
ELITE	0.245	0.620	0.359	-
VideoBooth	0.222	0.612	0.395	0.448
DreamVideo	0.261	0.611	0.310	0.311
Video Alchemist	0.269	0.732	0.617	0.466

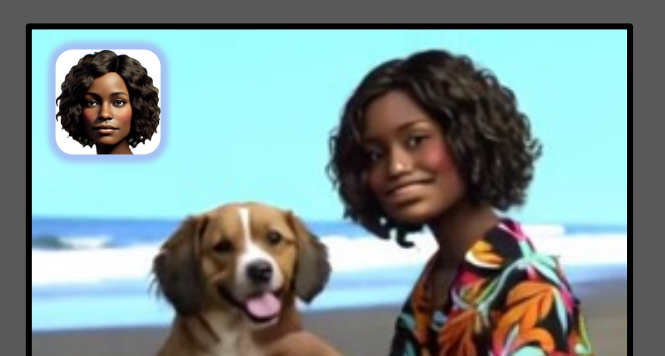
"A bearded man in gray clothes brushes a brown [...]."



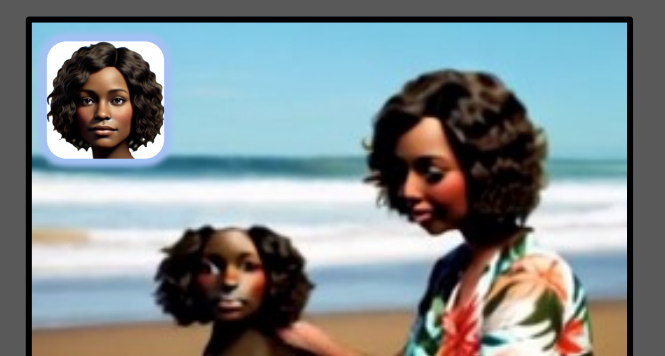
Face Mode

Method	Text-S $\uparrow$	Vid-S $\uparrow$	Face-S $\uparrow$	Dync $\uparrow$
IP-Adapter	0.251	0.648	0.269	-
PhotoMaker	0.278	0.569	0.189	-
Magic-Me	0.251	0.602	0.135	0.418
Video Alchemist	0.273	0.687	0.382	0.424

Ablation Study



Use CLIP



No word token



No augmentation



Video Alchemist