PersonalVideo: High ID-Fidelity Video Customization without Dynamic and Semantic Degradation

Supplementary Material

8. Limitation

While our method demonstrates superior performance for identity-specific video generation, our approach still has some limitations. While it enables a plug-and-play injection into the pre-trained T2V model, the results are inherently constrained by the capabilities of the T2V model itself. For example, it fails to generate customized videos that contain multiple identities. One possible solution is to further decouple the attention map of each subject, which will be explored in our future work.

9. Implementation Details

Training Details. For AnimateDiff, we use Stable Diffusion 1.5 with Realistic Vision [3] during training and inference. During training, we learn the Isolated Identity Adapter for 800 iterations with a learning rate of 1e-4 with the batch size 1. We default to using AdamW optimizer with the default betas set to 0.9 and 0.999. The epsilon is set to the default 1e-8 and the weight decay is set to 1e-2. For Hunyuanvideo, we employ the AdamW optimizer configured with a learning rate of 2e-5 and a weight decay parameter of 1e-4 for 4000 training steps. During inference, we use 50 steps of DDIM sampler and classifier-free guidance with a scale of 7.5 for all baselines. We generate 16-frame videos with 512 × 512 spatial resolution for AnimateDiff and 61-frame videos with 720 × 1280 or 512 × 768 spatial resolution for HunyuanVideo. All experiments are conducted on a single NVIDIA A800 GPU.

Baseline Details. We compare our method with both optimization methods, such as MagicMe and Dreambooth-LoRA, and encoder-based methods such as IDAnimator and ConsisID. Specifically, Magic-Me is a recent T2V customization method that trains extended keywords on the Stable Diffusion and injects it into AnimateDiff. Besides, we compare with Dreambooth-LoRA, which uses traditional reconstructive loss during training. For a fair comparison, we train them for the same steps with PersonalVideo. For the encoder-based methods, we compare with ID-Animator and ConsisID, the recent encoder-based methods, which are both trained on the large video dataset.

10. More Comparison

We provide more comparison including more base models and different number of the references in Fig. 12, Fig. 13, and Fig. 14. As shown, both Dreambooth and MagicMe suffer from inferior ID fidelity. Besides, MagicMe has a severe misalignment of the prompt, *e.g.*, *tuning head*. In contrast, our PersonalVideo maintains higher ID fidelity without dynamic and semantic degradation, which is consistent with results on the DiT-based model.

	Face (↑)	**CLIP-T** (↑)	**Dynamic** (↑)
T2I w Aug	45.79	28.42	16.13
T2V w/o Aug	56.40	24.10	16.3
T2V w/ Aug	**61.05**	**28.59**	**17.85**

Table 2. Quantitative ablation of the non-reconstructive training.

	Face (↑)	**CLIP-T** (↑)	**Dynamic** (↑)
w/o SCR	61.08	26.38	13.22
w/ SCR	**61.05**	**28.59**	**17.85**

Table 3. Quantitative ablation of Semantic Consistency Reward.

11. More ablation study

Fig. 15 and Tab. 4b verify the improvement in semantic following of our Isolated Identity Adapter to inject the identity only on the spatial self-attention layer. As observed, injecting only on the cross-attention layer gets inferior ID fidelity with the reference images and disrupts the original capability of semantic following, such as the losing of *exquisite armor*. Although injecting on both self-attention and cross-attention slightly achieves better ID fidelity, it still damages to the semantic following.

12. More Results

As shown in Fig. 20, Fig. 21, Fig. 22, Fig. 23, and Fig. 24, we present more customization results of PersonalVideo, including few or just one reference image. They showcase it achieves high ID fidelity and preserves original motion dynamics and semantic following, which provides further evidence of its promising performance and robustness.

13. Reproducibility Statement

We make the following efforts to ensure the reproducibility of PersonalVideo: (1) Our training and inference codes together with the trained model weights will be publicly available. (2) We provide training details in the appendix (Sec. 9), which is easy to follow. (3) We provide the details of the human evaluation setups in the appendix (Sec. 5.4).

	Face (↑)	Dynamic (↑)	CLIP-T(↑)
All steps	62.37	13.93	26.95
1/2 steps	60.36	16.22	25.63
1/4 steps	**63.90**	**18.00**	**27.47**

(a) Different steps to inject the identity.

	Face (↑)	CLIP-T (↑)	Dynamic (↑)
Cross	42.68	26.20	17.70
Self + Cross	**62.99**	23.35	17.33
Self	62.61	**27.87**	**17.80**

(b) Different layers to inject the identity in the Unet-based model.

Table 4. Quantitative ablation studies of Isolated Identity Adapter.

Figure 12. **Qualitative comparison on Animatediff.** As observed, both Dreambooth and MagicMe suffer from inferior ID fidelity. Besides, MagicMe has a semantic following degradation, *e.g.*, *tuning head*. In contrast, our PersonalVideo maintains high ID fidelity and preserve the original motion dynamics and semantic following, significantly surpassing others.

14. Impact Statement

Our main objective in this work is to empower novice users to generate visual content creatively and flexibly. However, we acknowledge the potential for misuse in creating fake or harmful content with our method. Thus, we believe it's essential to develop and implement tools to detect biases and malicious use cases to promote safe and equitable usage.

Figure 13. **More comparisons on HunyuanVideo.** As observed, Dreambooth suffers from inferior ID fidelity, while our PersonalVideo maintains higher ID fidelity without dynamic and semantic degradation, which is consistent with Fig. 12.

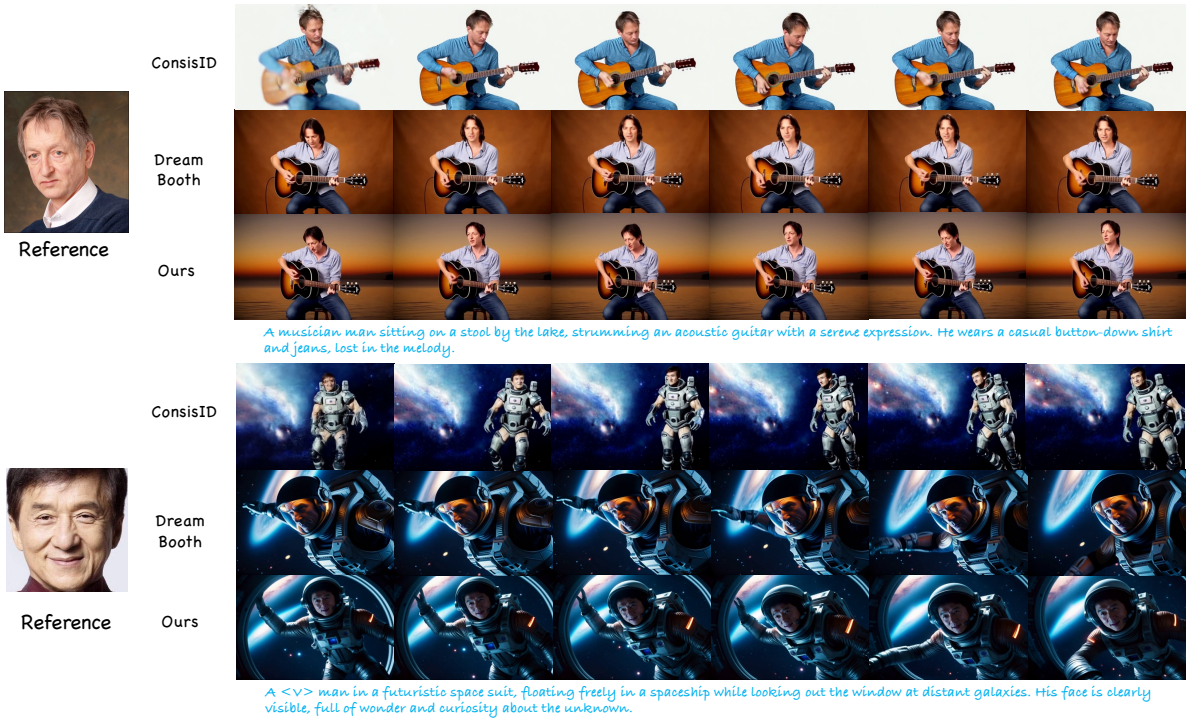


Figure 14. **More comparison for a single reference.** While ConsisID and Dreambooth suffer from the inferior ID fidelity, as well as severe degradation of motion dynamics and semantic following, *e.g. the stool by the lake*, our PersonalVideo achieves robust customization with high ID fidelity and preserved motion dynamics and semantic following.

Figure 15. **Ablation for different layers to inject the identity.** As observed, injecting it on the cross-attention layer disrupts the ability of semantic following, *e.g.*, the losing of *exquisite armor*.

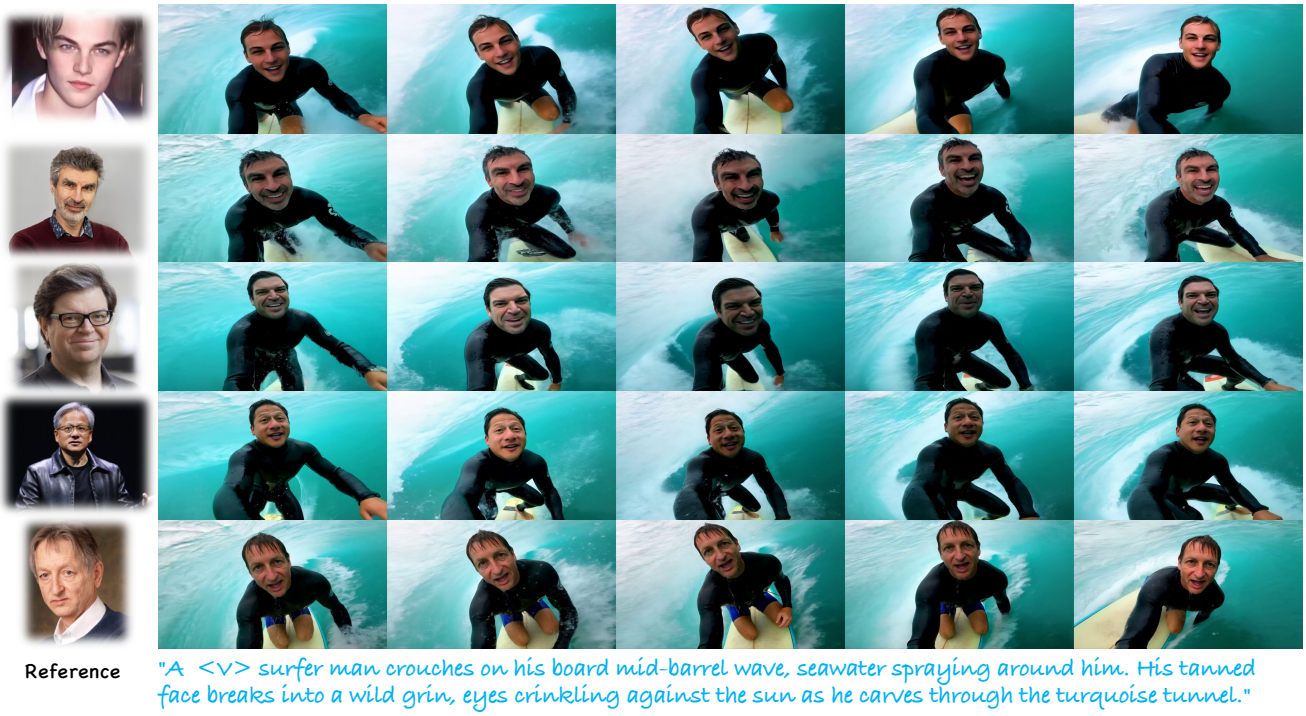


Figure 16. **More results of PersonalVideo**.

Reference

"A <V> musician man sitting on a stool, strumming an acoustic guitar with a serene expression. He wears a casual button-down shirt and jeans, lost in the melody."

Figure 17. **More results of PersonalVideo**.

Reference

A <V> woman in a tactical medic uniform stabilizes a patient amidst battlefield chaos, holographic triage interfaces orbiting her head. Her steady hands contrast with the desperation in her eyes as plasma fire lights the smoke around them.

Figure 18. **More results of PersonalVideo**.

Reference

"A <V> serene woman in a cozy sweater and jeans kneels on a lush green meadow, gently petting a friendly golden retriever. The dog's tail wags happily as its fur gleams in the soft sunlight. She smiles warmly, her hand moving tenderly over the dog's head and back. Rolling hills, vibrant blooming wildflowers, and a clear blue sky create a tranquil and picturesque backdrop, while the retriever leans into her touch, full of affection and trust."

Reference

"A <V> man dressed in medieval knight's armor, holding a shield and sword while riding a horse through a dense forest. His face is visible, showing a focused expression as he prepares for battle."

Reference

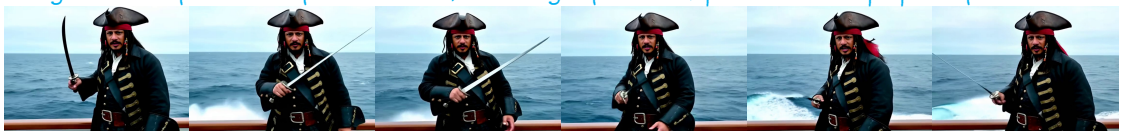
"A <V> woman with curly auburn hair takes center stage in a luxurious opera house. Dressed in a flowing emerald green gown with intricate gold embroidery, she sings passionately, surrounded by velvet red curtains and lavish gold designs."

Figure 19. **More results of PersonalVideo**.

A <v> man street at night, hands in his pockets, gazing up at the stars with a thoughtful expression. He standing on a quiet e wears a leather jacket and jeans.

A <v> man dressed in a pilot's outfit, sitting confidently in the cockpit of a jet as he prepares for takeoff. His face is visible, showing excitement and focus as he looks ahead at the sky.

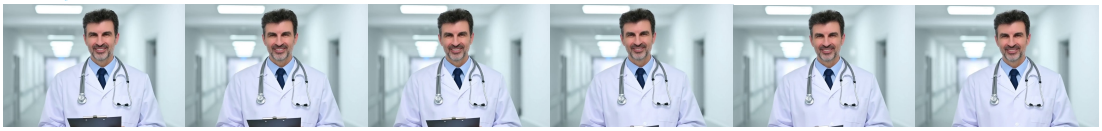
A <v> man dressed in a medieval knight's armor, holding a shield and sword while riding a horse through a dense forest. His face is visible, showing a focused expression as he prepares for battle.



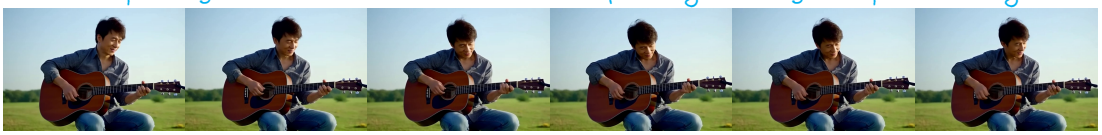
A <v> man dressed as a pirate, holding a sword and standing on the deck of a ship as waves crash around him. His determined face is visible as he looks out over the vast ocean, ready for adventure.

Figure 20. **More results of PersonalVideo**.

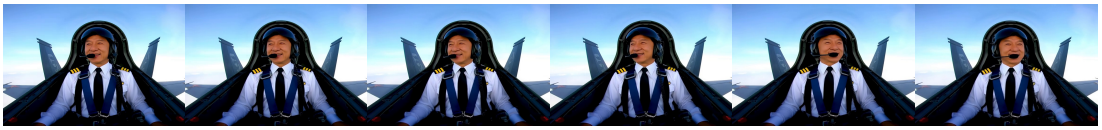
A <v> business man in a sleek gray suit, adjusting his wristwatch while standing in front of a skyscraper. His confident smirk reflects ambition and success.



A <v> doctor man in a white lab coat, smiling warmly while holding a clipboard. His stethoscope hangs around his neck as he stands confidently in a bright hospital hallway.

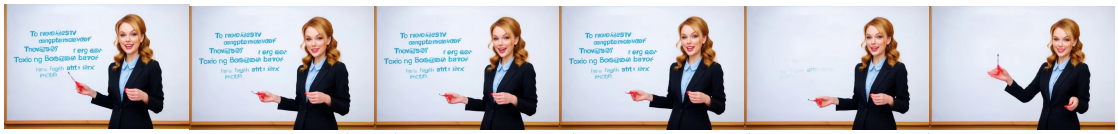


A <v> musician man sitting on a stool, strumming an acoustic guitar with a serene expression. He wears a casual button-down shirt and jeans, lost in the melody.



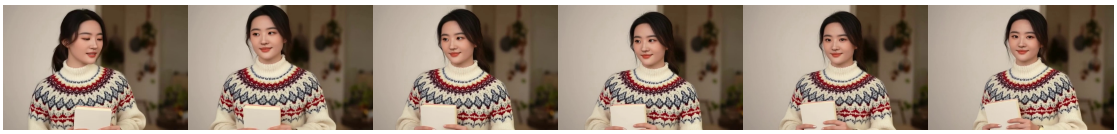
A <v> man dressed in a pilot's outfit, sitting confidently in the cockpit of a jet as he prepares for takeoff. His face is visible, showing excitement and focus as he looks ahead at the sky.

Figure 21. **More results of PersonalVideo**.

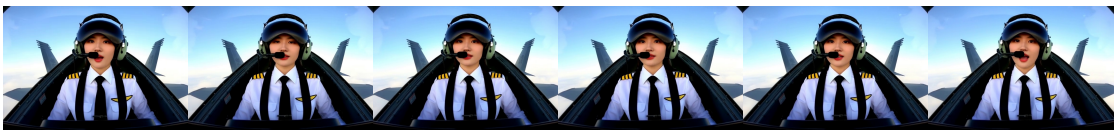


A <v> teacher woman standing in front of a whiteboard, gesturing as she explains a concept. She wears professional attire and has a warm, encouraging smile.

A <v> woman in a vibrant superhero outfit, flying through the clouds with her fists clenched. Her face is visible as she zooms above a city, her cape trailing behind her.



A <v> woman in a cozy sweater, holds a book while smiles softly

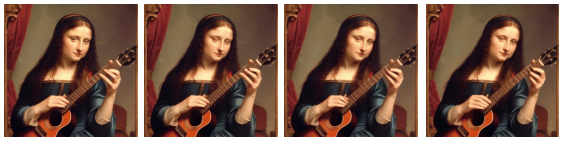


A <v> woman dressed in a pilot's outfit, sitting confidently in the cockpit of a jet as she prepares takeoff. Her face is visible, showing excitement and focus as she looks ahead at the sky.

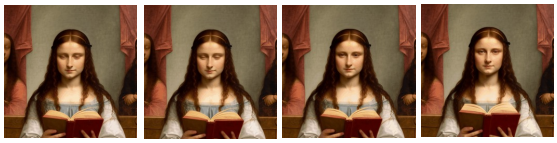
Figure 22. **More results of PersonalVideo**.

Single Reference

A <v> woman expertly pours a glass of wine, savoring the aroma



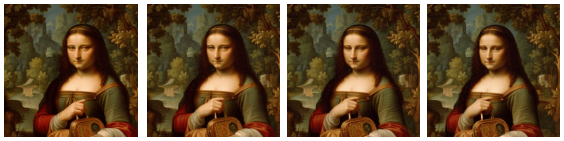
A <v> woman playing the guitar



A <v> woman reading a book in the classroom

A <v> woman, cherry blossoms sway in the breeze

A <v> woman playing basketball



A <v> woman walking in the forest

Figure 23. **More results of PersonalVideo** with only just one image.

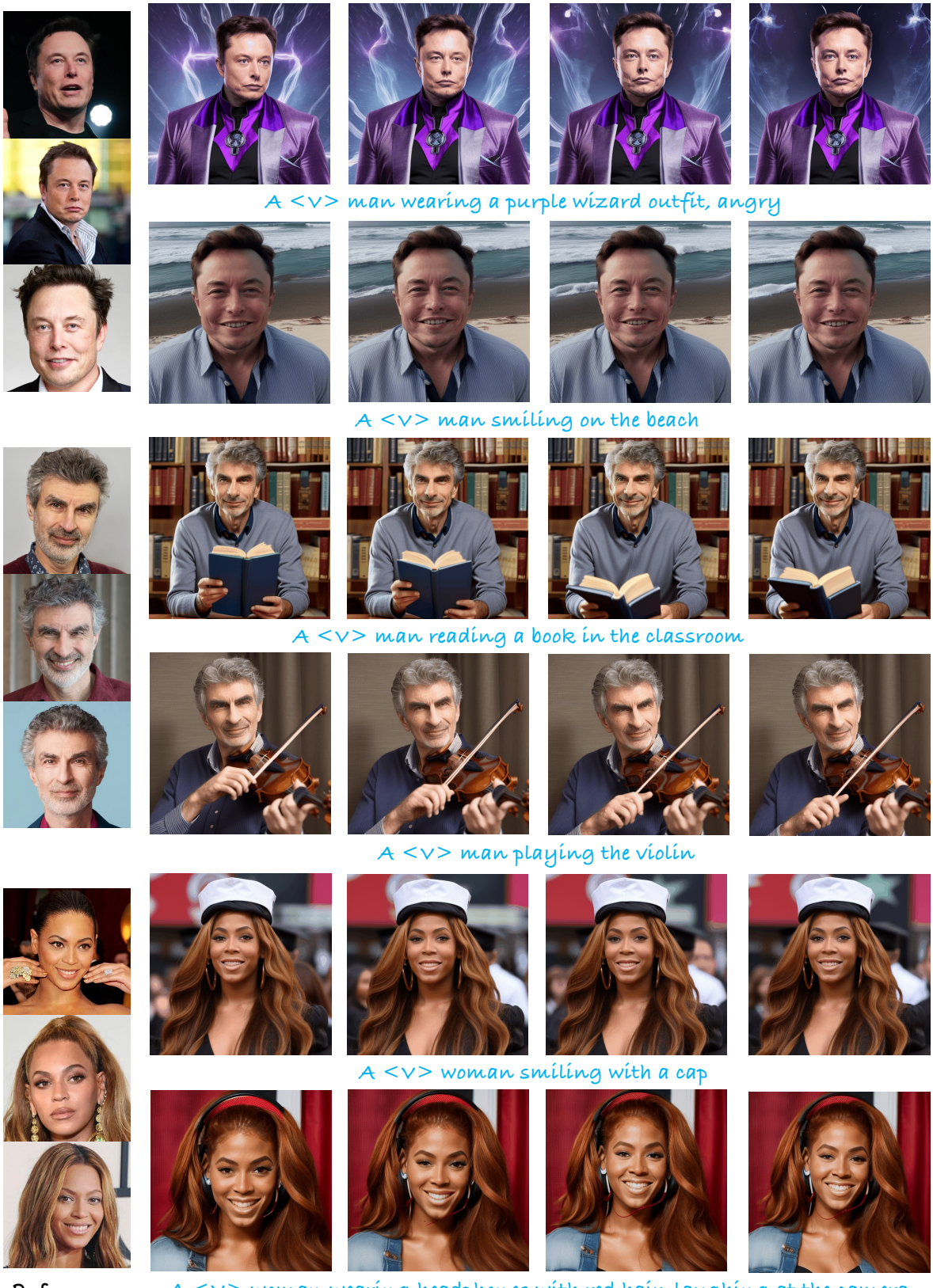


Figure 24. **More results of PersonalVideo** with few images.