

ViPerID: Video Personalization via Disentangled Identity Learning

Supplementary Material

Appendix

The appendix contains the following sections:

- [More Data Details](#)
- [More Evaluation Cases](#)
- [More Abalation Cases](#)
- [Applications](#)

A. More Data Details

Category	Description
General	A {class_token} smiling warmly at the camera with a cityscape in the background.
Clothing	A {class_token} dressed in a stylish trench coat, smiling at the camera on a rainy day.
Accessory	A {class_token} wearing large sunglasses and a sunhat, smiling brightly.
Context	A {class_token} sitting at a picnic table with friends, surrounded by nature.
Style	A classical oil painting of a {class_token} standing with a gentle smile.
Action	A {class_token} reading a book , looking up with a smile.
Expression	A {class_token} laughing with joy , looking directly at the camera.
View	A {class_token} facing the camera with a beautiful landscape in the background.
Background	A {class_token} with a city skyline in the background, looking cheerful.

Table 1. Description templates for different categories.

As shown in Table 1, we provide a sample of the evaluation dataset, which includes nine categories of personalized attributes. The text prompts replace “class_token” based on the character’s gender. Each description is followed by detailed action instructions, as outlined in the main text. Each category contains 10 prompts, resulting in a total of 90 prompt sets per character for comprehensive evaluation.

B. More Evaluation Cases

We provide additional comparisons on the video reconstruction task. As shown in Figure 1, ViPerID demonstrates stronger adherence to the text prompt, and the personalized reconstructed videos remain closest to the original inputs.

Furthermore, as shown in Figure 2, we provide additional comparison results between ViPerID, ConsisID, and ID-Animator. In the first case, ConsisID generates relatively stiff movements with limited fluidity, while ID-Animator fails to preserve accurate identity information and produces restricted motion amplitude. In the second case, ConsisID exhibits reduced identity consistency, whereas ID-Animator does not successfully follow the text prompt “*gesturing with her hands as she talks.*” These results further highlight that ViPerID excels both in maintaining identity fidelity and in adhering to diverse textual prompts.

To further demonstrate ViPerID’s ability to handle complex textual prompts and generate natural, fluid results, we provide additional examples in Figure 3. These include both background modifications and designed action instructions. ViPerID consistently delivers outstanding performance in motion smoothness and textual adherence.

C. More Abalation Cases

In addition, we provide a visual comparison of component ablations in Fig. 4. Removing the Content-Preserving Loss introduces unnatural artifacts, while discarding the ID-Preserving Loss undermines identity consistency. On the other hand, eliminating VisualCA results in generations with larger motion amplitudes, whereas removing TextCA produces outputs more aligned with textual descriptions. By integrating both modules, our model achieves substantial improvements in both identity preservation and text fidelity.

D. Applications

Figure 5 showcases ViPerID’s stylization ability, effectively disentangling identity information and integrating it into diverse scenes (see appendix videos for better visualization). Figure 6 further illustrates ViPerID’s “multi-shot” generation, maintaining identity consistency across shots while accurately following action instructions.



The video features a young man with short, curly hair and a light beard, wearing a **plaid shirt** ... He is seated **in front of a dark background**, and appears to be engaged in a conversation or interview. ... His facial expressions remain **focused and serious**, with occasional **changes in his mouth and eye movements** to emphasize his points ...

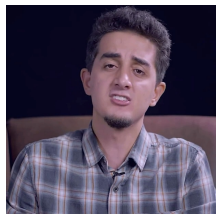
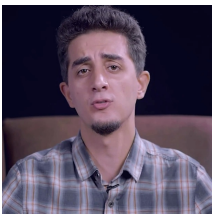
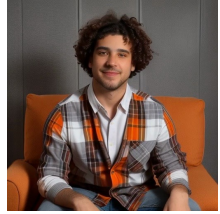
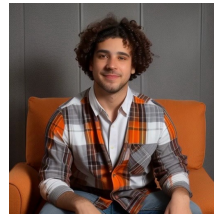
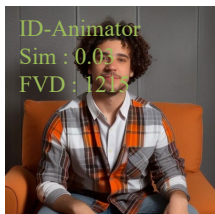
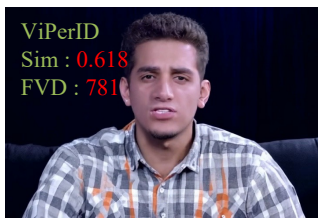


Figure 1. FVD evaluation of three models, indicating that ViPerID delivers the best overall performance.



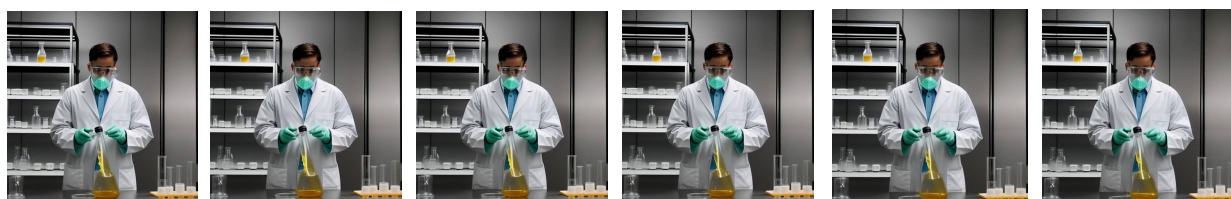
A man in a lab coat is standing at a lab bench ... **adds reagents to a beaker**,... The camera captures his **focused expression and steady hands**, as the liquid reacts with a burst of color. The lighting highlights his concentration and the smooth, confident **flow of his movements**, showcasing his expertise and dedication to the experiment.



ViPerID



ConsistID



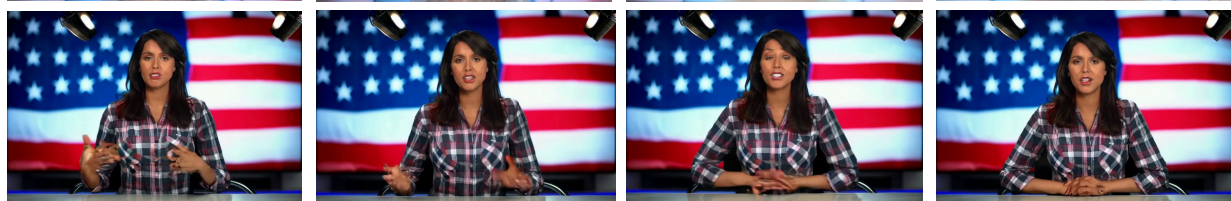
ID-Animator



A woman sitting at a desk in front of a large screen displaying an **American flag**, wearing a **plaid shirt** and appears to be delivering a news report. She speaks with confidence, **gesturing with her hands as she talks**, set in a newsroom or studio environment ...



ViPerID

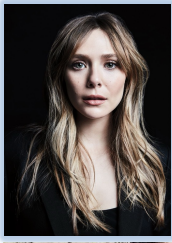


ConsistID



ID-Animator

Figure 2. Further comparison with state-of-the-art models. The results demonstrate that ViPerID achieves superior performance in detailed text adherence and identity preservation.



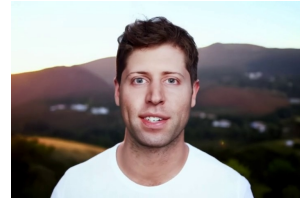
A woman adorned with a flower crown, standing amidst **a field of gently swaying wildflowers** ... Her eyes sparkle with a serene gaze, and a faint smile graces her lips, suggesting a moment of peaceful contentment. The shot is framed from the waist up, **highlighting the gentle breeze lightly tousling her hair**. The background reveals an expansive meadow under a bright blue sky ...



ViPerID



A man facing the camera with a **beautiful landscape in the background** ... with rolling hills bathed in the warm glow of the golden hour creating an idyllic backdrop. Bathed in soft light and the play of shadows, the scene captures a sense of serenity and momentary bliss. **A gentle breeze ruffles** through the man's hair ...



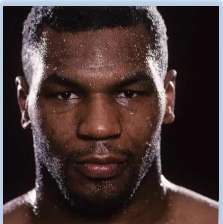
ViPerID



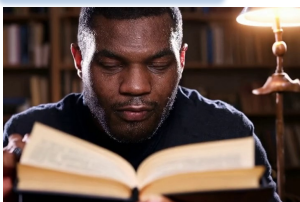
A man holding flowers, with a park behind them ... his figure **bathed in the warm glow of the afternoon sun**. Behind him, **a lively park unfolds with children playing**, joggers passing by ... The dappled sunlight creates intricate patterns of light and shadow on the grass, enhancing the serene atmosphere. Soft laughter mingles with the distant hum of bees ...



ViPerID



A man holding a book and wearing reading glasses, looking scholarly ... carefully **adjusting his reading glasses** as he examine the pages under the warm glow of a rustic library lamp. **The ambient light** casts gentle shadows across his focused expression ... Occasionally, he pause to jot down notes in the margins ...



ViPerID

Figure 3. Additional results of videos generated by ViPerID. ViPerID demonstrates satisfactory performance in motion smoothness and identity consistency.



A man enjoying a **cup of coffee**, sitting by a window in a cozy cafe. A man with a thoughtful expression sips a steaming cup of coffee while **gazing out the window** of a cozy cafe, morning **light cascading gently** through the glass to illuminate the pages of an open book resting on the table, as subtle shadows dance across his face and the warm aroma of freshly brewed coffee fills the air, creating a serene and inviting atmosphere.



VideopID



w/o VisionCA

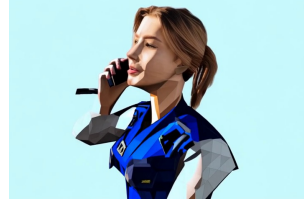


w/o TextCA

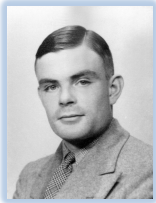
Figure 4. More visual examples from the ablation study. The absence of text cross-attention affects semantic alignment between visual and text, while the absence of visual cross-attention reduces facial similarity.



A woman police officer talking on the radio in low poly 3D, geometric reduction, **faceted style** ... Her uniform, carefully faceted, **captures the essential details in a minimalist** yet recognizable form ... angular highlights emphasize her composed expression



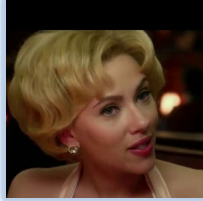
ViPerID



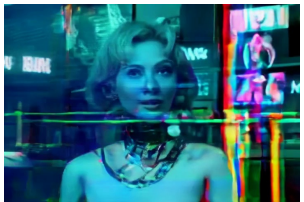
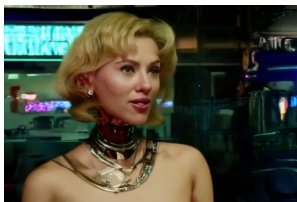
A man washing the **dishes in low poly 3D**, geometric reduction, faceted style ... His movements are expressed through bold geometric reductions, each gesture forming a sequence of crisp, faceted shapes. The dishes he cleans gleam with smooth, simplified reflections, while **water splashes in carefully structured** polygonal arcs. The background, composed of minimalist kitchen element ...



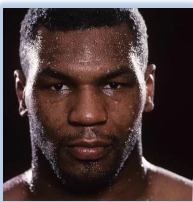
ViPerID



A woman looking around in electronic glitch art, neon, **cyberpunk aesthetic style** ... her figure **briefly fragmenting** into digital glitches as **she looks around. Holographic advertisements flicker above**, casting an eerie glow on her face, her ... Each time she **shifts her gaze**, the scene distorts momentarily, as if reality itself is a **fragile digital construct** ...



ViPerID



... A watercolor portrait of a man captured mid-laughter as he stand surrounded by a **lush garden**, the sunlight filtering through the leaves creating patterns of **light and shadow on his face**, with a gentle breeze rustling the nearby foliage ...



ViPerID

Figure 5. Stylization Capability Results. ViPerID effectively disentangles identity information from textual prompts.



A serene woman with delicate features wearing a flowing white blouse:
 (1) intensely focused on her task at a wooden table...
 (2) stands at her wooden workbench...
 (3) her hands skillfully chiseling details into the wood...
 (4) measuring and cutting wood with precise movements...
 (5) is seen in a spacious workshop, surrounded by wooden furniture...



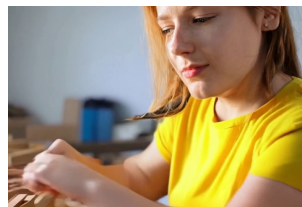
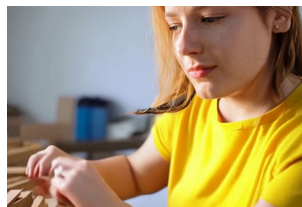
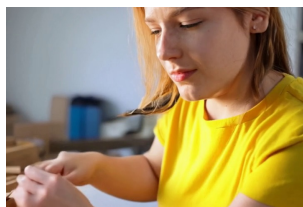
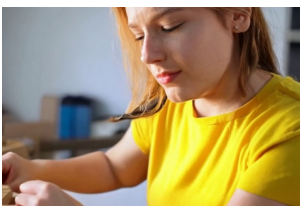
Lens One



Lens Two



Lens Three



Lens Four



Lens Five

Figure 6. ViPerID’s multi-shot generation capability. The model maintains strong identity consistency and accurate text adherence across different shots.