Movie Weaver: Tuning-Free Multi-Concept Video Personalization with Anchored Prompts

Supplementary Material

A. Appendix

A.1. Data curation

Video processing and filtering. For a given video, we uniformly sample five frames and apply a large-vocabulary object detector [7] to each frame. The intersection of all detected objects across these frames is used to determine the objects present throughout the video. Using these detection results, we filter videos based on specific criteria. For example, to select videos featuring two people, we require two 'person' bounding boxes in the detection results. Similarly, for videos with one person and an animal, we ensure there is exactly one 'person' bounding box along with a 'dog' or 'cat' bounding box.

Two-face data curation. After obtaining the two-person video data, we utilize a suite of foundational models to generate anchored prompts and ordered reference images, as described in Section 4.3 of the main paper. Building on the approach of Movie Gen [6], we first employ the LLaMa3-Video [2] model to produce detailed text prompts for the video clips. These prompts follow a structured format, enabling the use of in-context learning to extract concept descriptions. For example, given the input prompt: `Dentist Appointment. Senior woman smiling listening to her dentist during consultation.`, the outputs are two concept phrases: `[Senior woman smiling, her dentist]` and the anchored prompt: `Dentist Appointment. Senior woman smiling <ID1> listening to her dentist <ID2> during consultation.` Additional examples can be found in `in_context_twoface.txt`. Here, `<ID1>` and `<ID2>` represent `[R1]` and `[R2]`, respectively. We further refine the output by ensuring that the concept phrases contain exactly two items and that both `<ID1>` and `<ID2>` appear in the anchored prompt.

Two-facebody data curation. After generating the two-face anchored prompt, creating the two-facebody prompt is straightforward. This involves replacing the original `<ID2>` with `<ID3> <ID4>` and `<ID1>` with `<ID2> <ID2>`. Additionally, we prepare the ordered two-facebody reference images to align with the updated prompt structure.

Face-body-animal data curation. We filter videos that feature one person with a pet (dog or cat). We use in-context examples to add three anchors to the original prompt. Examples can be found in `in_context_facebodyanimal.txt`

A.2. Human evaluation

A.2.1 Two-face human evaluation

We conduct a human evaluation with 300 evaluation samples to ablate the effectiveness of the proposed anchored prompts and concept embeddings in Section 5.3.1. We provide the evaluation guidance as below. Besides the text guideline, we also include some visual examples to better help the annotators to judge.

Guidance. This document describes how to do Movie Weaver two-face character consistency evaluation on generated video and their reference faces. The focus is on personalized video generation, where two reference faces are used to create a video, and the evaluation assesses how well the two generated characters maintain a consistent visual appearance compared to the two reference faces. We will be primarily focused on human characters (realistic or stylized).

Task description. Annotators will be shown a set of two-faces and a generated video. They are then asked to rate the character consistency level on the set of generated frames based on a few different questions related to the visual appearance of the person(s) in the reference image(s).

Questions

- In the worst frame (they are not separable), are the two faces separable in the generated video (no fusion within two faces):
  1 - Totally separable
  2 - Somewhat separable
  3 - Not separable
  4 - Only one face or no face or more than two faces generated or visible
  Note: In the specific example in Figure 3, annotators are expected to give the answer "not separable"
- For the LEFT face in the reference, how well does the best aligned generated character's face capture the person likeness? (Please first try the best to locate the best aligned character for the left reference face):

1 - Really similar
2 - Somewhat similar
3 - Not similar
4 - Only one face or no face or more than two faces generated or visible

Note: In this specific example in Figure 3, annotators are expected to give the answer "Not similar"

- For the RIGHT face in the reference, how well does the best aligned generated character's face capture the person likeness? (Please first try the best to locate the best aligned character for the right reference face):
  1 - Really similar
  2 - Somewhat similar
  3 - Not similar
  4 - Only one face or no face or more than two faces generated or visible

  Note: In this specific example in Figure 3, annotators are expected to give the answer "Not similar"

A.2.2 One-face human evaluation

We perform a human evaluation with 300 samples to assess the effectiveness of mixed training, as discussed in Section 5.3.2. The evaluation protocol closely follows that of single-face personalized Movie Gen [6]. Specifically, annotators are provided with a reference image and a generated video clip and asked to rate two aspects: Face similarity (face_sim): How well the generated character's face matches the reference person in the best frame. Face Consistency Score (face_cons): How visually consistent the faces are across all frames containing the reference person. Ratings are given on an absolute scale: "really similar," "somewhat similar," and "not similar" for identity, and "really consistent," "somewhat consistent," and "not consistent" for face consistency. Annotators are trained to adhere to specific labeling guidelines and are continuously audited to ensure quality and reliability.

A.3. Additional results

A.3.1 Comparison with multi-concept image personalization.

We also compare with representative multi-concept image personalization methods in Figure 1. For Tweediemix [4], we first fine-tune the base SDXL [5] model for each reference concept using LORA [3], then conduct multi-concpet sampling using Tweedie's formula. Because Tweediemix requires background reference, we select one of its pretrained garden LORA weights. Freecustom [1] is a tuning-free method, so we follow its practice by preparing two reference faces. We select the first frame of our Movie Weaver to compare with these image methods. As shown in Figure 1, our Movie Weaver preserves a much better identity and has higher visual quality when compared with TweedieMix and FreeCustom.

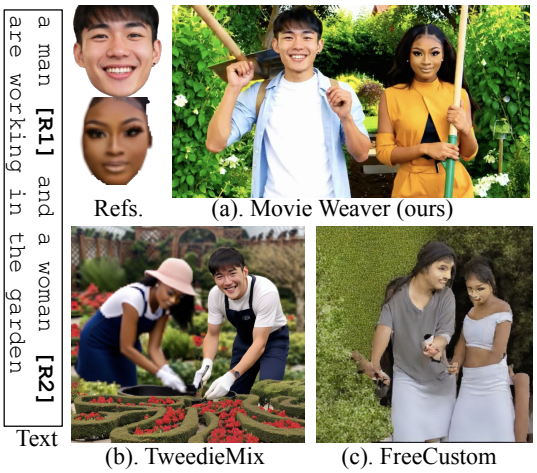

Figure 1. **Comparison with multi-concept image methods.** Movie Weaver has a better identity preserving and visual quality.

Table 1. **Ablation study of Anchored Prompts (AP) and Concept Embeddings (CE) on "two-face-body" config.**

Case	Modules		Human study metrics		
	AP	CE	sep_yes↑	human1_sim↑	human2_sim↑
Baseline			54.8	12.3	16.5
(1)		✓	98.8	66.7	69.4
(2)	✓	✓	98.0	72.3	71.1

A.3.2 Ablation on two-face-body configuration

As shown in Table 1, ablation with "two-face-body" showed similar trends to "two-face" configurations. However, clothing details, like small logos in Figure 1 in the main paper, are harder to retain, likely due to the 256px reference resolution. Higher-resolution references may enhance clothing detail preservation.

A.3.3 Order of reference images

In this section, we examine how the order of reference images influences the final output. Since the order information is incorporated through concept embeddings, altering the sequence of reference images results in different videos, even with the same prompt. This effect is illustrated in Figure 2.

References

[1] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In

Anchored Prompt: `A man wearing a black leather jacket` **`[R1]`** `is sitting on a motorcycle next to a man in a yellow T-shirt` **`[R2].`**

Figure 2. By changing the order of reference images, we can assign certain face to certain attributes.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9089–9098, 2024. 2

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[4] Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. *arXiv preprint arXiv:2410.05591*, 2024. 2

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[6] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1, 2

[7] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 1