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

Feng Liang* [1], Haoyu Ma[2], Zecheng He[2], Tingbo Hou[2], Ji Hou[2], Kunpeng Li[2], Xiaoliang Dai[2], Felix Juefei-Xu[2], Samaneh Azadi[2], Animesh Sinha[2], Peizhao Zhang[2], Peter Vajda[2], Diana Marculescu[1]
[1]The University of Texas at Austin, [2]Meta GenAI
{jeffliang,dianam}@utexas.edu, {haoyuma,zechengh,stzpz}@meta.com
https://jeff-liangf.github.io/projects/movieweaver

Figure 1. We introduce **Movie Weaver**, a video diffusion model for personalized multi-concept video creation. Besides text prompts, our model allows users to input different combinations of reference images, *e.g.*, face, body, and animal images, to customize videos in a tuning-free manner. The left column displays different types of reference images, while the right column shows the generated videos, with *anchored prompt* listed beneath each video. We encourage readers to check our video results in the supplementary materials.

Abstract

Video personalization, which generates customized videos using reference images, has gained significant attention. However, prior methods typically focus on single-concept personalization, limiting broader applications that require multi-concept integration. Attempts to extend these models to multiple concepts often lead to identity blending, which results in composite characters with fused attributes from multiple sources. This challenge arises due to the lack of a mechanism to link each concept with its specific reference image. We address this with anchored prompts, which embed image anchors as unique tokens within text prompts, guiding accurate referencing during generation. Additionally, we introduce concept embeddings to encode the order of reference images. Our approach, Movie Weaver, seamlessly weaves multiple concepts—including face, body, and animal images—into one video, allowing flexible combinations in a single model. The evaluation shows that Movie Weaver outperforms existing methods for multi-concept video personalization in identity preservation and overall quality.

1. Introduction

Foundational text-to-video generation models [3–6, 9, 15, 17, 22, 27, 28, 38, 45, 46, 48, 51, 60, 62] have made substantial progress in the past few years. Leveraging these advancements, personalized video generation enables users to create customized videos with their images, offering huge potential for applications like consistent storytelling. However, prior efforts [19, 42, 56] primarily support single concept (face or object) personalization, limiting their use in complex, real-world scenarios. Practical applications often require multi-concept compositions, like interactions between two people or between a person and a pet. To meet this need, we introduce Movie Weaver, a video diffusion model that *weaves* diverse image combinations and text prompts to create personalized multi-concept videos, as illustrated in Figure 1.

Established video personalization methods [19, 24, 42, 56] typically extract vision tokens via an image encoder [43], then inject these tokens into diffusion models through cross attention. To support multiple face references, we extend this approach by directly feeding concatenated vision tokens from multiple references into the cross attention layer. While this direct extension works in some cases, it often encounters a severe identity blending issue [30, 57], generating composite characters that fuse attributes from both references. The issue is more severe when the character faces are close, *i.e.*, from the same gender or race. This is because such a direct extension lacks the ability to associate each reference image with their descriptions within the prompt. When vision tokens from different faces are similar, the model struggles to differentiate between them, resulting in identity blending.

To address this issue, we explicitly build the association between the concept description and the corresponding reference image. We introduce *anchored prompts* by injecting unique text tokens (`[R1]`) after each concept description, as shown in Figure 1. Upon encountering `[R1]`, the model links it with the matching reference image of `[R1]` and uses it as the visual supplement for the concept. This anchored prompt approach extends easily to multiple concepts by adding more anchors (*e.g.*, `[R2]`, `[R3]`) and corresponding images. Notably, anchored prompts only require input modifications without any architectural changes like identity-specific tuning [30, 31], predefined layout [16, 36] or masked attention [18, 25, 40, 57]. This preserves architectural simplicity and leverages scalability.

With the anchored prompts established, the next question is how to distinguish the reference images of `[R1]` and `[R2]`. Our baseline architecture [19, 24, 42] concatenates the vision tokens with text tokens and feeds them together to the cross attention layer of diffusion models. However, since cross attention is order-agnostic, we must inject some information about the order of reference images. Inspired by positional encoding [11, 49], we introduce *concept embedding*, adding a unique embedding to each set of vision tokens from the reference image. This is different from traditional positional encoding, where different embeddings are assigned to individual tokens. Our method applies a uniform concept embedding to all tokens from the same image.

We also propose an automatic data curation pipeline to get anchored prompts and ordered reference images. This pipeline supports diverse reference combinations (such as face, face-body, face-body-animal, two-face, and two-face-body) by leveraging a suite of foundations models [12, 26, 35, 43], yielding a dataset of 230K videos. Using the proposed anchored prompts and concept embeddings, we continue training Movie Weaver on a pre-trained single-face video personalization model [42]. As showcased in Figure 1, our model effectively generates high-quality videos with diverse reference combinations without additional tuning. Compared with proprietary Vidu 1.5 [50] and extended baseline, Movie Weaver exceeds in identity preservation and visual quality.

Our contributions are three-fold: (1) **Anchored Prompts:** We introduce anchored prompts to link specific reference images with concept description, resolving identity blending in multi-concept video personalization without architectural changes. (2) **Concept Embeddings:** We use unique embeddings for each reference image to maintain identity and order in multi-reference settings. (3) **Automatic Data Curation:** We implement a pipeline to curate

*Work partially done during an internship at Meta GenAI.

training data with diverse reference combinations, enabling high-quality, tuning-free multi-concept video generation.

2. Related Work

Personalized image generation Personalized generation begins with identity-specific tuning methods that further finetune a text-to-image model on a set of reference images. For instance, Textual Inversion [14] finetunes special text tokens for the target identity. DreamBooth [44] further conducts end-to-end model finetuning besides tuning the special text tokens. Custom Diffusion [30] extends a parameter-efficient finetuning technique to incorporate multiple concepts. However, these tuning-based methods require separate optimizations for every concept, which does not scale well in real applications. Recent tuning-free methods train one base personalization model and then use it for arbitrary reference images in inference. For example, ELITE [54], PhotoMaker [32], PhotoVerse [8], IP-Adapter [61] , InstantID [52], and Imagine Yourself [10, 20] all leverage a vision encoder to extract visual tokens from the reference image and inject them to the diffusion process. Our method Movie Weaver falls in the second tuning-free method which targets multi-concept video personalization.

Personalized video generation While personalized image generation has shown promising results, personalized video generation remains an unsolved problem. Compared to static images, personalized videos require more diverse and complex modifications on the reference image, *e.g.*, turning the head, changing poses, and camera motion movements. Pioneering work VideoBooth [24], DisenStudio [7], Magic-Me [39], DreamVideo [55], CustomVideo [53] and CustomCrafter [56] use identity-specific finetuning to inject identity into a video generation model. However, these methods require separate fine-tuning for every identity, which limits their applications. Another line of work includes tuning-free methods. ID-Animator [19] and MovieGen [42] train a vision encoder [43] to inject the reference image features into the diffusion models. However, all the aforementioned methods don't tackle the multi-concept video personalization in the tuning-free setting.

Multi-concept personalization Multi-concept personalization aims to generate harmonized content with multiple reference concepts. Identity-specific tuning methods, *e.g.*, Custom Diffusion [30], Break-A-Scene [2], Concept Weaver [31] and MuDI [23], achieve multi-concept personalization by finetuning the multiple text embeddings and model weights. Tuning-free methods, *e.g.*, FastComposer [57], MoA [40], InstantFamily [25] and UniPortrait [18], train on large-scale text-image datasets and inject multiple image embeddings directly into diffusion process during inference. Mix-of-show [16] and OMG [29] merge single-concept models but are limited by the availability of community models, primarily covering well-known intellectual property (IP) or celebrities. Unlike these methods that separate concepts using predefined layout [16, 36] or masked attention [25, 25, 40, 57], Our Movie Weaver proposes anchored prompts and concept embeddings to link concept with matched reference image without architectural change. Concurrent work ViMi [13] is the closest to our setting but requires complex data retrieval to curate data and an additional multimodal LLM [34] for the image-text process. In contrast, our data curation pipeline operates without supplementary retrieval data, while maintaining simplicity in architecture.

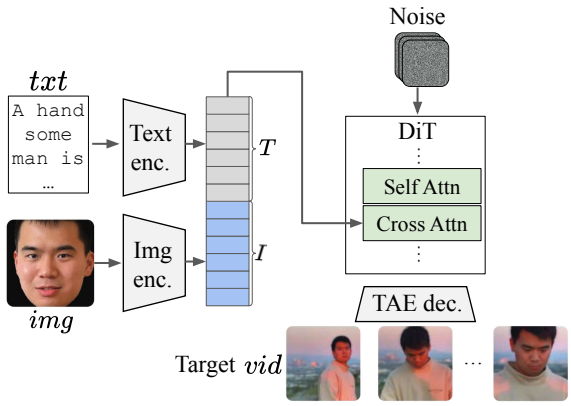


Figure 2. **Single-concept personalization architecture.** Building on a pre-trained text-to-video model, this approach adds an image encoder to process reference images. Image and text tokens are concatenated and fed into cross attention layer.

3. Challenges in Extending Single-Concept to Multi-Concept Personalization

3.1. Singe-concept video personalization

Based on a pre-trained text-to-video diffusion model, prior single-concept video personalization methods [19, 24, 42, 56] typically introduce an additional CLIP image encoder [43, 58] to process reference images. As shown in Figure 2, given a dataset triplet $\{txt, img, vid\}$—representing text prompt, reference image, and target video—these methods extract text tokens T and image tokens I via text and image encoders. The tokens I and T are concatenated and then fed into the cross attention layer of the diffusion transformer [41].

$$\text{CrossAttn} = \text{softmax}\left(\frac{Q_Z\left[K_T, K_I\right]^{Tr}}{\sqrt{d}}\right)\left[V_T, V_I\right] \quad (1)$$

Here, the query is derived from the video latent Z, and the key/value is derived from the concatenation of I and T.

Reference faces Frame from generated video

Figure 3. **Identity blending** generates composite faces with characteristics from both references. Text prompt: "`A woman in wheelchair discussing with a woman nurse.`"

$[\cdot]$ and ^{Tr} denotes the concatenation and transpose operation, respectively. Key and value are derived through linear projection, so $[V_T, V_I] = V_{[T,I]}$. For simplicity, we maintain formal notation without specifying this further.

3.2. Naive extension to multi-concept

To support multi-concept personalization, a straightforward approach is preparing image tokens from multiple reference images. We began by experimenting with two-face personalization. According to Equation 1, this results in $K_I = [K_{I_1}, K_{I_2}]$ and $V_I = [V_{I_1}, V_{I_2}]$, where I_1 and I_2 represents tokens of two different faces. This naive approach, however, led to severe identity blending [30, 57], where the characteristics of two faces would fuse together, resulting in a composite one as seen in Figure 3. The issue arises because the model cannot effectively link each concept description to its corresponding image. In cross attention, a query latent in Q_Z can attend to the entire prompt, including two text tokens of "`woman`", and cannot distinguish between vision tokens I_1 and I_2 due to order-agnostic processing. For accurate video generation, the model must link "`A woman in wheelchair`" to the first image and "`a woman nurse`" to the second image.

4. Movie Weaver

Unlike traditional approaches that require concept-specific tuning or architectural modifications, Movie Weaver aims to enable multi-concept video personalization with architectural simplicity and flexibility. Movie Weaver introduces two novel components, namely anchored prompts and concept embeddings, alongside an automatic data curation pipeline, all of which allow accurate, tuning-free multi-concept video generation.

4.1. Anchored prompts

The success of multi-concept personalization lies in accurately associating each concept with its corresponding image. Previous methods often rely on predefined layouts [16, 36] or complex masked attention modules [18, 25, 40, 57] to establish these associations, which increases model complexity and limits flexibility. Our Movie Weaver introduces a streamlined solution with anchored prompts. For prompt in Figure 3, we use Llama-3 [12] to identify the concept descriptions "`A woman in wheelchair`" and "`a woman nurse`", we then append unique tokens (*e.g.*, `[R1]`, `[R2]`) after each description, creating the prompt "`A woman in wheelchair [R1] discussing with a woman nurse [R2].`" Ordered reference images are then linked to these unique tokens, allowing the model to associate each description precisely with its reference image. This approach offers two key advantages: (1) Flexibility: anchored prompts can easily extend to different descriptions, such as body or animal descriptions. Moreover, it allows to append multiple references, such as face and body images, on the same person.; and (2) Simplicity: this approach requires only input modifications, allowing Movie Weaver to retain an architecture similar to single-concept models.

4.2. Concept embeddings

While anchored prompts establish explicit associations, we also need to encode the order information of reference images. The cross attention mechanism, as outlined in Equation 1, is inherently order-agnostic, *i.e.*, swapping the order of two reference images yields identical results. To address this, Movie Weaver introduces concept embeddings, a novel adaptation of positional encoding [49] tailored to multi-concept personalization. Specifically, given I_1 and I_2 as image tokens from two different reference images, we add the same concept embedding Pos_1 and Pos_2 to each set of tokens, respectively:

$$I_1' = I_1 + Pos_1, I_2' = I_2 + Pos_2, ... \quad (2)$$

Here, I_1 and I_2 have dimensions $[N, C]$, while Pos_1 and Pos_2 are of shape $[1, C]$, where N is the number of tokens and C is the feature dimension. Using the broadcasting of Pytorch, the same concept embedding is added to the entire set. This is different from traditional positional encoding where distinct embeddings are added to individual tokens. We also experimented with per-token positional encoding, but it produced less effective results compared to using a concept embedding for each set of vision tokens.

4.3. Data curation pipeline

Starting with text-video pairs, we curate data by leveraging a set of foundation models. We take the preparation of two-face configuration as an example in Figure 4 (a). We first use in-context learning of Llama-3 [12] to extract concept descriptions from the original prompt and get a rewritten anchored prompt. For the reference frame, typically the first frame of the video clip, we use a detection model Detic [63] and a segmentation model SAM [26] to extract the subject masks. Using the detection results, we can also tell the number of people in the video and other objects in the video, which helps us filter the data. Then, a pre-trained

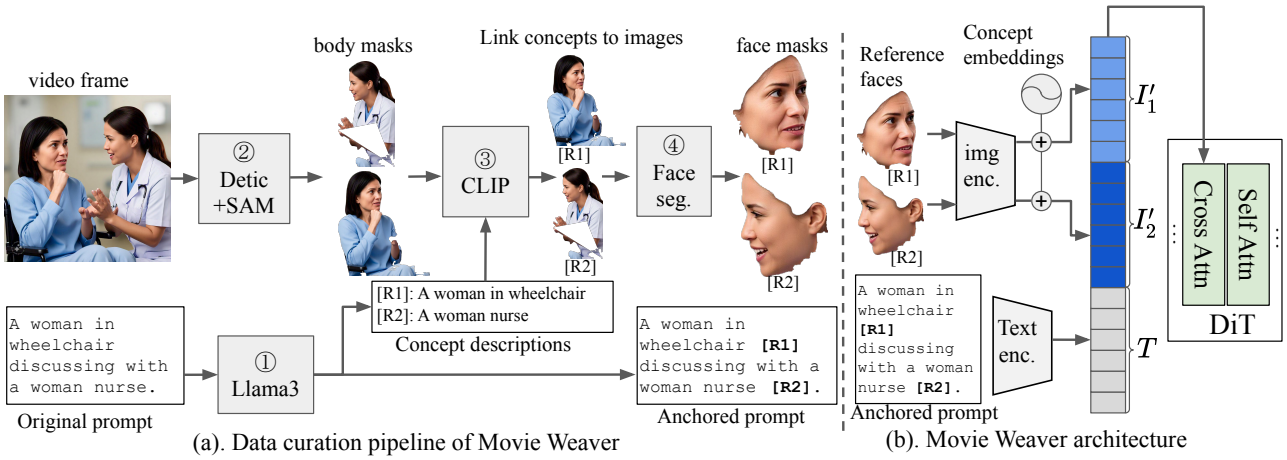


Figure 4. (a) **Data curation.** For a video-text pair, ① concept descriptions and anchored prompts are generated via in-context learning with Llama-3. After ② extracting body masks, ③ CLIP links each concept to its corresponding image. ④ Finally, face images are obtained using a face segmentation model. (b) **Movie Weaver architecture.** Compared to the single-concept baseline, reference images are arranged in a specific order for concept embedding, and anchored prompts are utilized. Shared components are omitted for simplicity.

CLIP [43] assigns the concept descriptions to the body images, establishing the link between `[R]` and reference images. Lastly, we use a face segmentation model to extract face masks from body images. While this example only illustrates the two-face scenarios, the approach can be naturally extended to other combinations with more `[R]`s and reference images. More examples can be found in the supplementary materials.

With ordered reference images and rewritten anchored prompts prepared, our Movie Weaver architecture is illustrated in Figure 4(b). Compared with single-concept architecture in Figure 2, our model requires reference images in a specific order to apply concept embeddings effectively. We also need to have corresponding anchored prompts to strengthen the results.

5. Experiments

5.1. Implementation details

Pretraining data Our Movie Weaver supports 5 configurations: face, face-body, face-body-animal, two-face and two-face-body, as showcase in Figure 1. We collect all our videos from Shutterstock Video [1]. For the face and face-body configurations, we curated 100K videos featuring only a single person performing various activities. In the two-face and two-face-body configurations, videos were selected based on the presence of exactly two individuals, verified via a detection model, resulting in a dataset comprising 118K videos. For the face-body-animal setup, we found it more challenging to source videos featuring both a person and an animal, ultimately assembling a collection of 10K videos. We conduct mix-pretraining, where we equally sample examples for each configuration. More information can be found in Section 5.3.2.

Finetuning data Existing research [10, 21, 42] suggests that further finetuning on a small-scale, very high quality data significantly enhances visual quality and subject motion. We follow this principle by manually selecting videos with large human motion, rich human iterations and high visual aesthetic. Same as pretraining data, we source different videos for our different configurations. In summary, our finetuning set has 291 videos for face and face-body, 175 videos for two-face and two-face-dody, 185 videos for face-body-animal.

Model details We adopt the Movie Gen architecture [42], with two versions: a 4B parameter model for the ablation study and a 30B parameter model for the final results. Unless specified otherwise, Movie Weaver refers to the 30B model. Movie Weaver uses MetaCLIP [58] as image encoder and uses three text encoders: MetaCLIP [58], ByT5 [59] and UL2 [47]. The diffusion model contains 48 layers of diffusion transformers [41]. The temporal VAE has a compression rate of $8\times8\times8$, represents $8\times$ dimension reduce for spatial height/width and temporal frame. Using an additional $2\times2\times1$ patchification, we compressed each 128-frame landscape video with 544×960 resolution into a token sequence of length 32K.

Training hyperparameters We initialize Movie Weaver from a pre-trained single-face personalization Movie Gen checkpoint. The model is trained with a learning rate of 1e-5 using the AdamW [37] optimizer for 20K iterations with a batch size of 32. The training objective is flow matching [33] with optimal transport path. It took around 5 days to do pretraining on a cluster of 256 H100 GPUs. Following this, we performed supervised finetuning with a reduced learning rate of 2e-6 for an additional 2K iterations.

Figure 5. **Qualitative results of Movie Weaver.** Movie Weaver supports different combinations of reference images and can generate high-quality videos with high identity preservation. We encourage readers to check our video results in the supplementary materials.

5.2. Qualitative results

5.2.1 Performance highlight

We demonstrate three configurations of Movie Weaver: face-body-animal, two-face, and two-face-body, in Figure 5. For the same text prompt, we generate two different videos with two different sets of reference images. Notably, no facial, clothing, or dog descriptions are included in the prompts; all identity information is derived solely from the reference images. We highlight key features of our Movie Weaver: (1) Identity Preservation: Face, body, and animal details are accurately maintained in the generated videos. Even small clothing details, such as the logo on the gray T-shirt in the first video and the tear on the white T-shirt in the sixth, are faithfully preserved. In challenging two-face scenarios with same-gender, same-race pairs, Movie Weaver effectively retains each identity. (2) Flexibility with References: Movie Weaver can adapt reference images to match the prompt without having to strictly follow. For instance, in the second video, the reference body image shows a man standing, yet the prompt requires him to sit. Our Movie Weaver selectively uses the upper body to align with the prompt. Also for the fifth video, the standing woman body reference image is adapted to appear seated on a beach. (3) Rich iteration between subjects: Beyond preserving identity, the generated videos capture dynamic interactions between subjects. In the first and second videos, the person interacts well with the dog, while in third through sixth videos, the two people display rich engagement.

5.2.2 Comparison with existing methods

We compare Movie Weaver with the proprietary Vidu 1.5 [50], the only existing video model, to the best of

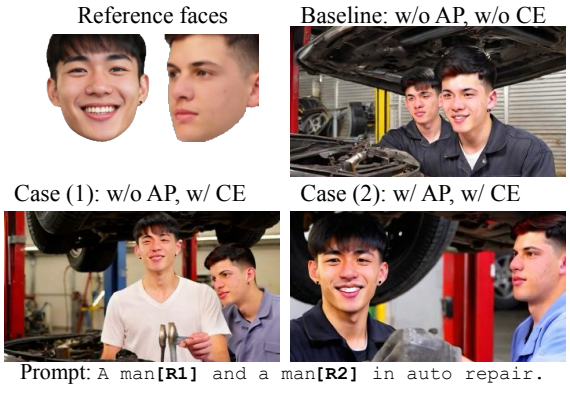
Figure 6. **Comparison with state-of-the-art.** Compared to proprietary Vidu 1.5, Movie Weaver demonstrates superior identity preservation for both human and animal reference images.

our knowledge, that supports multi-concept personalization. Following Vidu's guidance, we prepared its reference images without masking. In the first face-body-animal example of Figure 6, Movie Weaver shows better identity preservation for both the African American woman and the dog. While Vidu 1.5 correctly identifies the dog as a Corgi, it fails to reconstruct the distinctive white spots on the dog's face, which are crucial to the dog's identity. In the second two-face-body example, because Vidu 1.5 supports a maximum of 3 reference images, we only input the face of the first character. Vidu 1.5 suffers from severe identity blending issues with the two generated characters looking and wearing similarly to each other, whereas Movie Weaver maintains clear distinctions between the two individuals.

5.3. Ablation study

5.3.1 Anchored prompts and concept embeddings

In this section, we analyze the effects of the proposed Anchored Prompts (AP) and Concept Embeddings (CE). We conduct a human study on a two-face personalization evaluation dataset with 300 image pairs. As in the top part of Figure 7, the baseline suffers from identity blending issue, reflected by a low sep_yes score, which indicates percentage of cases where the two generated faces are separable. Introducing concept embeddings, as seen in case (1), raises



Case	Modules		Human study metrics		
	AP	CE	sep_yes↑	face1_sim↑	face2_sim↑
Baseline			42.9	3.4	3.0
(1)		✓	98.2	58.8	41.9
(2)	✓	✓	99.3	66.8	66.1

Figure 7. **Ablation study of Anchored Prompts (AP) and Concept Embeddings (CE).** The top part shows the effect of AP and CE, while the bottom presents results from a human study. Metric sep_yes indicates the percentage of cases where the two generated faces are distinguishable (*i.e.*, no identity blending), face1_sim and face2_sim represent where a similar face to the left or right reference face, respectively, is found in the generated video.

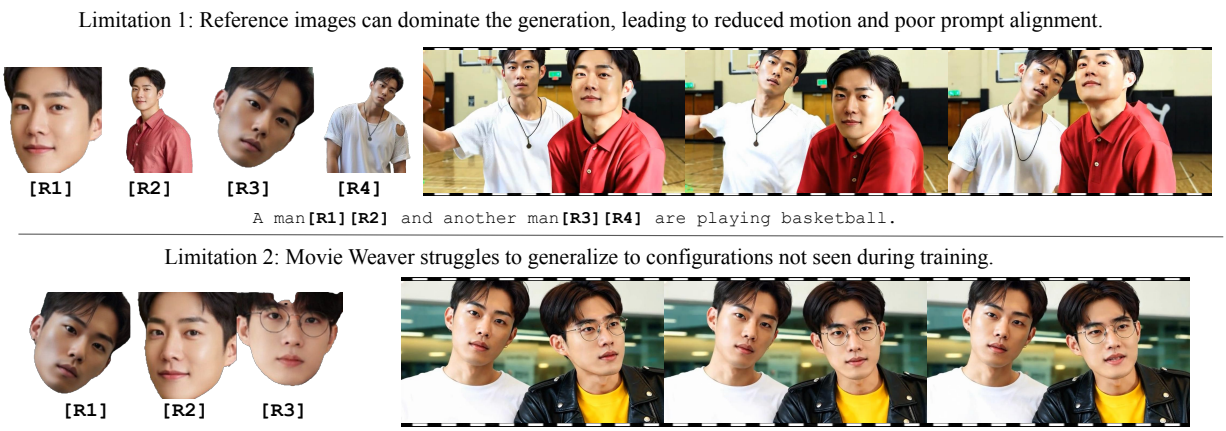


Figure 8. **Limitations of Movie Weaver.** Reference images can dominate generation, resulting in 'big-face' videos. Our model also struggles to generalize to configurations not seen during training.

Table 1. **Ablation study of mixed training with multiple reference configurations.** 1F, 2F and F-B-A represents one-face, two-face and face-body-animal data, respectively. The metric face_sim and face_cons represents the similarity of character to the reference face and face consistency throughout the video.

Case	Pretraining data			Human study metrics	
	1F	2F	F-B-A	face_sim↑	face_cons↑
1	✓			13.1	70
2	✓	✓		29.5	81.3
3	✓	✓	✓	46	91

the sep_yes score from 42.9% to 98.2%. In case (2), anchored prompts further enhance identity preservation, increasing face1_sim and face2_sim scores. These scores represent the percentage of cases where a similar face to the left (face1) and right (face2) reference face is identifiable in the video. More details about human study can be found in the supplementary materials.

5.3.2 Mixed training

Our Movie Weaver is trained on various combinations of reference images, and this section examines the impact of this mixed training approach. For a fair comparison, instead of initializing from a single-face personalization checkpoint, we start with a text-to-video base model. As shown in Table 1, we try different data configurations (1F: one-face, 2F: two-face and F-B-A: face-body-animal) and evaluate with a single-face personalization human study, which comprises 300 datapoints. Compared to case (1), which uses only one-face data, mixed training with additional configurations improves both face_sim (similarity of the generated character to the reference face) and face_cons (consistency of the face throughout the video). We attribute this improvement to the increased data diversity provided by mixed training.

5.4. Limitations

While Movie Weaver demonstrates strong multi-concept personalization capabilities, it has certain limitations. Firstly, the personalized videos often have limited overall motion compared to the results of base text-to-video model, and we sometimes see 'big-face' videos where faces occupy a large portion of the video frame. We believe this occurs because the reference images can dominate the generation, leading to reduced motion and poor prompt alignment. For instance, in the first example in Figure 8, we provide two half-body reference images with a prompt involving playing basketball. The generated video reproduces the half-bodies but fails to align well with the action described in the prompt. The underlying reason is Movie Weaver struggles to balance influence of reference images and text prompts when they are mismatched. During training, videos of sports activities like basketball typically include full-body references, whereas during inference, users may provide less aligned inputs. Addressing the balance between reference images and prompts remains an area for future improvement. Secondly, Movie Weaver struggles to generalize to configurations not seen during training. In the second example in Figure 8, the goal is to generate a video with three people talking, but the result only features two people. This is because we don't have any training videos that contain more than two people. Addressing this issue will require incorporating additional data configurations during pretraining, which we identify as a direction for future work.

6. Conclusion

We present Movie Weaver to support tuning-free multi-concept video personalization. We alleviate the identity blending issue by explicitly associating the concept descriptions with reference images. With the pursuit of architectural simplicity and flexibility, we propose anchored

prompts to inject unique tokens within text prompts and concept embeddings to encode the order of reference images. Our results show that Movie Weaver can generate high quality personalized videos with diverse reference types, including face, body and animal images.

Acknowledgments

This research was supported in part by ONR Minerva program, iMAGiNE - the Intelligent Machine Engineering Consortium at UT Austin, and a UT Cockrell School of Engineering Doctoral Fellowship.

References

[1] Stock footage video, royalty-free hd, 4k video clips, 2023. 5

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024.

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[7] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. *arXiv preprint arXiv:2405.12796*, 2024. 3

[8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. PhotoVerse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 3

[9] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6441–6451, 2024. 2

[10] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 3, 5

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] 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. 2, 4, 11

[13] Yuwei Fang, Willi Menapace, Aliaksandr Siarohin, Tsai-Shien Chen, Kuan-Chien Wang, Ivan Skorokhodov, Graham Neubig, and Sergey Tulyakov. Vimi: Grounding video generation through multi-modal instruction. *arXiv preprint arXiv:2407.06304*, 2024. 3

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[15] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2

[16] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[17] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[18] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024. 2, 3, 4

[19] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2, 3

[20] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, Li Chen, Ankit Jain, Ning Zhang, Peizhao Zhang, Roshan Sumbaly, Peter Vajda, and Animesh Sinha. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 3

[21] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine

yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 5

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[23] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024. 3

[24] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 2, 3

[25] Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024. 2, 3, 4

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4

[27] KlingAI. Kling AI, 2024. 2

[28] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 2

[29] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *arXiv preprint arXiv:2403.10983*, 2024. 3

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 3, 4

[31] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8880–8889, 2024. 2, 3

[32] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. *arXiv preprint arXiv:2312.04461*, 2023. 3

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[36] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 57500–57519, 2023. 2, 3, 4

[37] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[38] LumaLabs. Dream Machine, 2024. 2

[39] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 3

[40] Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. MoA: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024. 2, 3, 4

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 5

[42] 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. 2, 3, 5, 11, 12

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[45] RunwayML. Gen-3 Alpha, 2024. 2

[46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[47] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022. 5

[48] Genmo Team. Mochi 1: A new sota in open-source video generation. https://github.com/genmoai/models, 2024. 2

[49] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 4

[50] VIDU Studio. Character2Video, 2024. Accessed: 2024-11-14. 2, 6

[51] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 2

[52] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[53] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 3

[54] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 3

[55] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024. 3

[56] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 2, 3

[57] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 2, 3, 4

[58] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 3, 5

[59] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. 5

[60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[62] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[63] 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. 4, 11

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 [63] 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 [42], we first employ the LLaMa3-Video [12] 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 [42]. 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. More results analysis

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 9.

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