Concat-ID: Towards Universal Identity-Preserving Video Synthesis

Yong Zhong[1*] Zhuoyi Yang[2] Jiayan Teng[2] Xiaotao Gu[3] Chongxuan Li[1†]

[1] Gaoling School of AI, Renmin University of China, Beijing, China
[2] Tsinghua University [3] Zhipu AI
yongzhong@ruc.edu.cn, {yangzy22,tengjy24}@mails.tsinghua.edu.cn,
xiaotao.gu@zhipuai.cn, chongxuanli@ruc.edu.cn

Project page and code: https://ml-gsai.github.io/Concat-ID-demo/

Supplementary material

A. Experimental settings

A.1. Datasets

We remove reference images from the ConsistID-Benchmark that may appear in our training data using both manual and automated filtering methods. (1) Manual filtering: For each reference image in the ConsistID-Benchmark, we compute its cosine similarity with all training images and identify the most similar one. Human evaluators then determine whether the two images depict the same person. If so, all reference images of the corresponding identity are excluded. (2) Automated filtering: All reference images of an identity are discarded if any training image has a cosine similarity greater than 0.45 with one of its reference images.

A.2. Implementation details

In the first stage, we randomly select one reference image from a set of five for each video. Traditional data augmentation techniques, such as flipping, are not used for face images, as they can cause data augmentation leakage [4], leading the model to learn the augmented data distribution rather than the original distribution. For instance, horizontal flipping may result in incorrectly mirrored faces in generated videos.

A.3. Baselines

We try our best not to change original settings of baselines to maintain their original capabilities. IDAnimator [2] and ConsisID [5] can produce 16-frame and 49-frame videos at a resolution of 480×720, respectively. The multi-identity baseline Ingredients [1] generates 49-frame videos at a resolution of 480×720, integrating two distinct identities.

A.4. Training cost

The first, second, and third stages of training in the single-identity scenario required 3,260, 2,104, and 135 NVIDIA H800 GPU hours, respectively, with the cost of the third stage being negligible.

A.5. Limitations

Similar to common video generation models, our approach faces challenges in preserving the integrity of human body structures, such as the number of fingers, when handling particularly complex motions. In this paper, we focus on the single-identity scenario, and further improvement and evaluation of Concat-ID's performance in multiple-identity and multi-subject scenarios is left for future work.

*Work done during the internship at Zhipu.
†Correspondence to Chongxuan Li.

Method	Identity consistency		Text alignment	Facial editability
	ArcSim ↑	CurSim ↑	ViCLIP ↑	CLIPDist ↑
Concat-ID (Stage I)	**0.560**	**0.581**	0.237	0.274
Concat-ID (Stage II)	0.185	0.200	**0.248**	**0.434**
Concat-ID (Stage III)	<u>0.442</u>	<u>0.466</u>	<u>0.242</u>	<u>0.325</u>

Table 1. **Quantitative ablation.** Stage I, Stage II, and Stage III indicate the pre-training stage, cross-video stage, and trade-off stage of Concat-ID, respectively. The second-best result is underlined. Concat-ID in the third stage demonstrates the optimal balance.

B. Multiple identities and subjects

B.1. Multi-identity scenarios

Through the data construction process of pre-training pairs, we obtain approximately 300,000 videos featuring two identities. For each identity, we determine the sequence order by computing the mean horizontal position of face boxes across all reference images. We discard reference images where the face position does not align with the determined sequence order. Next, we construct cross-video pairs by independently processing each identity within a video. Finally, we collect around 8,000 videos, each of which contains identities that have corresponding cross-video reference images.

A similar strategy is used to construct three-identity training data, resulting in a final dataset of approximately 40,000 pre-training videos. For cross-video pairs, we retain videos in which at least two identities have corresponding cross-video reference images, resulting in about 2,000 videos.

For multiple identities, the pairing cosine similarity ranges between 0.87 and 0.97. We initialize the model using single-identity pre-training weights and train it only on the first two stages (i.e., the pre-training stage and cross-pair fine-tuning stage). Our findings indicate that single-identity pre-training facilitates multi-identity convergence and enhances identity consistency.

B.2. Multi-subject scenarios

We select a subset from the cross-video pairs of single-identity, comprising approximately 200,000 videos, where the pairing cosine similarity ranges between 0.87 and 0.97. To achieve virtual try-on, we use Grounded-SAM-2[1] to detect and segment the clothing of identities. For background-controllable generation, we extract the first frame and use Grounded-SAM-2 to obtain human masks. We then apply SDXL[2] to inpaint the masked areas to get bacground images, using a randomly selected classification label from YOLO[3] as input prompts.

We use weights from single-identity pre-training as initialization and apply only random horizontal flip augmentation to clothing images. Additionally, we introduce random noise to both the background and clothing images during training. In multi-subject scenarios, we only train models on the cross-pair fine-tuning stage.

In this paper, we focus on the single-identity scenario, and improving the performance of Concat-ID in multiple-identity and multi-subject settings is left for future work. To maximize model performance, we independently train different specialized models for specific tasks. The development of a comprehensive model capable of addressing multiple tasks simultaneously remains a direction for future research.

C. Ablation study

Tab. 1 presents the quantitative ablation study of Concat-ID. The pre-training stage achieves the best identity consistency (i.e., ArcSim and CurSim) but has the worst facial editability (i.e., CLIPDist). However, the cross-video stage significantly improves CLIPDist but degrades ArcSim and CurSim. In the third stage, Concat-ID obtains the second-best results across all metrics, demonstrating that it achieves an optimal balance. These results highlight the superiority of our multi-stage training strategy, which balances the knowledge learned in different stages to achieve optimal performance in the final stage.

Trade-off pairs can naturally enhance the identity consistency of Concat-ID, as they maintain better alignment between reference images and videos compared to cross-video pairs. An interleaved training strategy—alternating between Stage I for improving identity and Stage II for enhancing editability—can also achieve a favorable trade-off, a method similarly

[1] https://github.com/IDEA-Research/Grounded-SAM-2
[2] https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1
[3] https://github.com/ultralytics/ultralytics

adopted in Imagine-yourself [3]. However, our multi-stage training approach achieves an optimal balance just by adding a third stage where we carefully control identity consistency and sample quantity, showing that a simple design can be highly effective. Furthermore, by constraining the upper bound of identity consistency, we prevent the model from directly copying and pasting the reference image into the generated video.

References

[1] Zhengcong Fei, Debang Li, Di Qiu, Changqian Yu, and Mingyuan Fan. Ingredients: Blending custom photos with video diffusion transformers. *arXiv preprint arXiv:2501.01790*, 2025. 1

[2] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 1

[3] 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. 3

[4] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 1

[5] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 1