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/

Figure 1. **Concat-ID produces natural videos for identity-preserving video generation.** We select samples for (a) single-identity, (b) multi-identity, and (c) multi-subject scenarios, respectively.

Abstract

We present Concat-ID, a unified framework for identity-preserving video generation. Concat-ID employs variational autoencoders to extract image features, which are then concatenated with video latents along the sequence dimension. It relies exclusively on inherent 3D self-attention mechanisms to incorporate them, eliminating the need for additional parameters or modules. A novel cross-video pairing strategy and a multi-stage training regimen are introduced to balance identity consistency and facial editability while enhancing video naturalness. Extensive experiments demonstrate Concat-ID's superiority over existing methods in both single and multi-identity generation, as well as its seamless scalability to multi-subject scenarios, including virtual try-on and background-controllable generation. Concat-ID establishes a new benchmark for identity-preserving video synthesis, providing a versatile and scalable solution for a wide range of applications.

1. Introduction

Identity-preserving video generation, which seeks to create human-centric videos of a specific identity accurately matching a user-provided face image, has recently gained significant attention, as evidenced by the success of commercial tools such as Vidu [22] and Pika [18].

A primary challenge in this field is achieving a balance between maintaining identity consistency and enabling facial editability. Prior work [9, 13, 27, 29] fails to effectively preserve identity despite utilizing special face encoders and incorporating extra adapters to mitigate cross-modal disparities. To mitigate this limitation, some approaches [4, 28] substitute the spatially aligned reference image in pre-trained image-to-video models [2, 26] with facial images, leading to a significant improvement in identity consistency. However, they still face challenges in preventing the replication of facial expressions from the reference image. Moreover, the supplementary modules and parameters introduced by these methods contribute to increased complexity in both model training and inference.

In this work, we introduce Concat-ID, a concise, effective, and versatile framework for identity-preserving video generation. By unifying the model architecture, data processing, and training procedure, Concat-ID not only achieves single-identity video generation but also seamlessly integrates multiple identities and accommodates diverse subjects. Specifically, Concat-ID employs Variational Autoencoders (VAEs) to extract image features, which are then concatenated with video latents along the sequence dimension. This approach relies exclusively on 3D self-attention mechanisms, which are inherently present in state-of-the-art video generation models, to incorporate image features, thereby eliminating the need for extra modules or parameters. Furthermore, to effectively balance identity consistency and facial editability while enhancing video naturalness, we develop a novel cross-video pairing strategy and a multi-stage training regimen.

The quantitative and qualitative results, along with the user study (see Sec. 5.2), demonstrate that Concat-ID produces videos with the most consistent identity and superior facial editability across all baselines, for both single-identity and multi-identity video generation. Moreover, we illustrate that Concat-ID can seamlessly extend to multi-subject scenarios, including virtual try-on and background-controllable generation, while effectively preserving identity (see Sec. 5.3). These findings underscore Concat-ID's capability to scale effectively to diverse subjects, ensuring consistent high performance across various applications.

The principal contributions of this work are as follows:

- We propose Concat-ID, an effective framework for unified identity-preserving video generation across single-identity, multi-identity, and multi-subject scenarios.
- Concat-ID utilizes VAEs to extract image features and integrates them via inherent 3D self-attention mechanisms, without introducing additional parameters or modules.
- We develop a cross-video pairing strategy and a multi-stage training regimen to balance identity consistency and facial editability, while enhancing video naturalness.
- Concat-ID demonstrates superior identity consistency and facial editability in single and multi-identity scenarios, and seamlessly scales to multi-subject scenarios.

2. Related works

The rapid advancement of text-to-video and image-to-video diffusion models [8, 16, 19, 26, 30] has spurred significant interest in fine-tuning these models for downstream tasks, particularly identity-preserving video generation. Tuning-based methods [11, 17] adapt pre-trained video models for each new identity through test-time fine-tuning. Alternatively, tuning-free methods [9, 13, 27, 29] typically leverage face encoders [3, 20] to extract facial features and incorporate additional adapters to mitigate cross-modal discrepancies. Some approaches [4, 24, 28] further enhance identity consistency by integrating face features extracted from a Variational Autoencoder (VAE). For instance, ConsisID [28] and Ingredients [4] replace spatially aligned reference images in pre-trained image-to-video models for single-identity and multi-identity generation, respectively. Placing greater emphasis on enhancing video naturalness, Movie-Gen [19] refines the balance between identity consistency and facial editability for single-identity generation through cross-paired data construction. In this work, we

*Work done during the internship at Zhipu.

†Correspondence to Chongxuan Li.

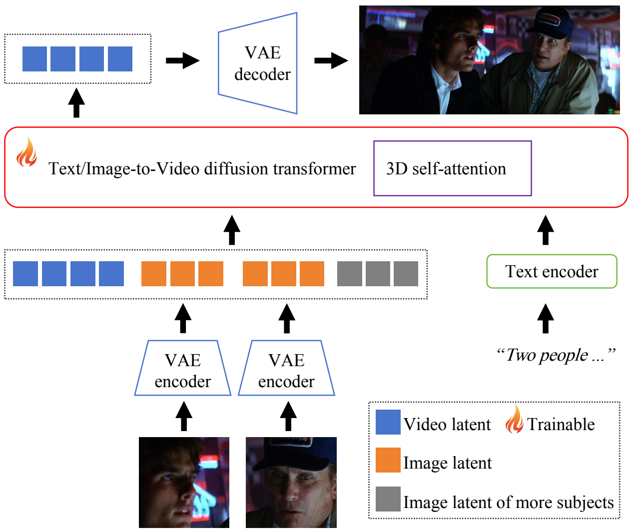


Figure 2. **The architecture of Concat-ID.** We utilize VAEs to extract image latents from reference images and concatenate them at the end of the video latents along the sequence dimension. Concat-ID relies solely on 3D self-attention mechanisms, which are commonly present in state-of-the-art video generation models, to integrate image features without adding extra modules or parameters.

explore a unified framework capable of handling single-identity, multi-identity, and multi-subject generation while maintaining a crucial balance between consistency and editability, without requiring test-time fine-tuning.

3. Preliminary

Existing state-of-the-art text-to-video and image-to-video models [8, 15, 16, 19, 26] generally consist of three main components: a 3D variational autoencoder (VAE) $\mathcal{E}$, text encoders $\mathcal{T}$, and a denoising transformer $\boldsymbol{\epsilon_\theta}$. Given a video $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ with N frames, $\mathcal{E}$ initially compresses the video into a latent representation $\mathbf{Z} \in \mathbb{R}^{T \times HW \times C}$ along the spatiotemporal dimensions, where HW denotes the spatial dimension, C represents the channel dimension, and T is the temporal dimension. To simplify, we refer to $T \times HW$ as the sequence dimension. The $\boldsymbol{\epsilon_\theta}$ then takes the noise-corrupted latent representation $\mathbf{Z}$ as its input, and applies a 3D (spatiotemporal) self-attention mechanism [8, 26] to model the distribution of video content. Additionally, a 3D relative positional encoding (i.e., 3D-ROPE) is incorporated within the 3D attention module to enhance the model's ability to capture both temporal and spatial dependencies in videos. Meanwhile, the text encoder $\mathcal{T}$ processes the text prompt and encodes it into a text representation c_{txt}. $\boldsymbol{\epsilon_\theta}$ typically integrates c_{txt} either through cross-attention layers [19] or by concatenating it with $\mathbf{Z}$ [26]. A mean squared error loss [5, 31] is commonly used to optimize $\boldsymbol{\epsilon_\theta}$.

4. Concat-ID

Given a reference image containing a human face, our goal is to generate identity-preserving videos based on user-provided text prompts, while also enabling the integration of additional identities or subjects. To address this challenge, we propose Concat-ID, a concise, effective, and versatile framework. As illustrated in Fig. 2, we introduce a unified architecture for extracting and injecting features from any number of identities and subjects without requiring extra modules or parameters (see Sec. 4.1). To balance identity consistency and facial editability while enhancing video naturalness, we further construct cross-video pairs as training data (see Sec. 4.2) and propose a novel multi-stage training strategy (see Sec. 4.3).

4.1. A unified architecture

We focus on designing a unified model architecture capable of extracting and fusing the identity feature and readily extendable to multi-identity and multi-subject scenarios. Revisiting the role of VAEs, we recognize their ability to compress conditioning images into the same latent space as the video latent $\mathbf{Z}$. Consequently, our denoising transformer $\boldsymbol{\epsilon_\theta}$ can inherently interpret these features. Based on this insight, we adopt the VAE as our feature extractor.

Specifically, for M reference images $\{\mathbf{I}_i\}_{i=1}^M$, we encode each $\mathbf{I}_i$ to obtain the image feature $\mathbf{c}_i = \mathcal{E}(\mathbf{I}_i) \in \mathbb{R}^{1 \times HW \times C}$, and then concatenate these features with $\mathbf{Z}$ in sequence. Thus, the input to $\boldsymbol{\epsilon_\theta}$ is given by:

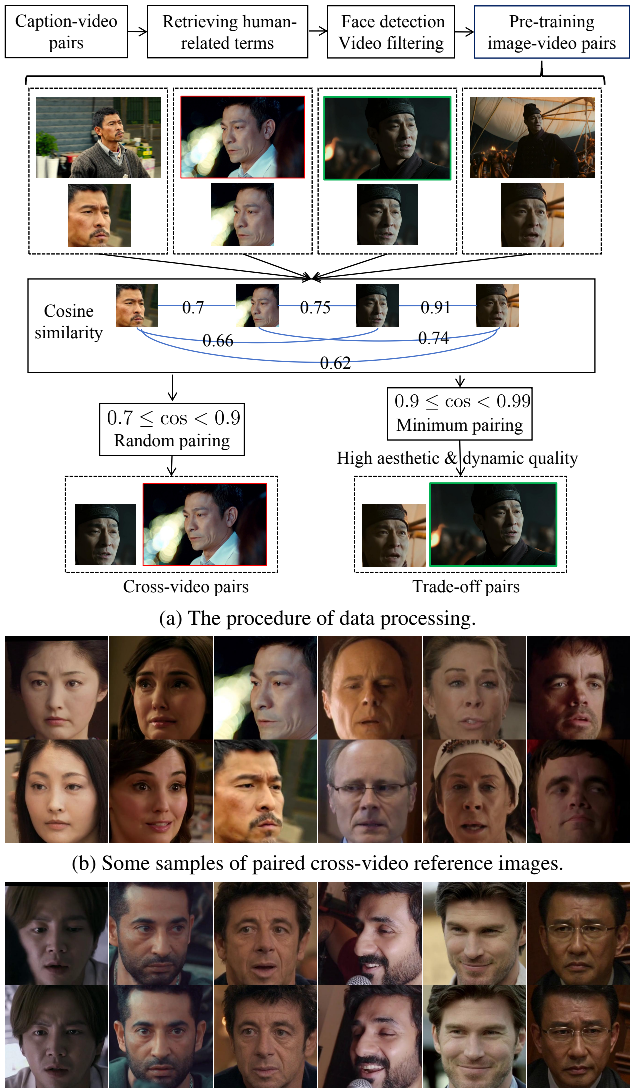
$$\mathbf{Z}' = \text{Concat}(\mathbf{Z}, \mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_M), \tag{1}$$

where $\text{Concat}(\cdot, \cdot, \cdots)$ denotes concatenation along the sequence dimension and $\mathbf{Z}' \in \mathbb{R}^{(N+M) \times HW \times C}$. As shown in Fig. 2, this feature injection through concatenation is compatible with any video generation model that utilizes 3D self-attention, which are generally present in state-of-the-art video generation models. Since $\mathbf{Z}$ and $\mathbf{c}_i$ are in the same latent space, $\boldsymbol{\epsilon_\theta}$ can seamlessly integrate identity-preserving features without the need for additional modules or parameters to address cross-modal disparities.

Concatenating $\mathbf{Z}$ and $\mathbf{c}_i$ along the channel dimension is another direct method for feature injection, as employed in ConsisID [28] and Ingredients [4]. However, this strategy introduces artifacts (see Fig. 4 and Fig. 5) due to spatial misalignment between face images and video latents. In contrast, by leveraging a 3D self-attention mechanism, our sequence concatenation promotes spatial interactions without compromising the quality of any generated frame. Furthermore, it scales efficiently to handle multi-identity and multi-subject scenarios (see Fig. 1).

4.2. Data construction

The task of identity-preserving video generation relies on image-video pairs as training data, where an image must de-



(a) The procedure of data processing.

(b) Some samples of paired cross-video reference images.

(c) Some samples of trade-off pairs.

Figure 3. **Constructing three types of image-video pairs for a single identity:** pre-training, cross-video and trade-off pairs.

pict a human face that matches the identity of corresponding videos. To progressively balance identity consistency and facial editability, as illustrated in Fig. 3, we construct three types of image-video pairs for a single identity: pre-training pairs $\mathcal{S}_{\text{pre}}$, cross-video pairs $\mathcal{S}_{\text{cross}}$, and trade-off pairs $\mathcal{S}_{\text{trade}}$.

Pre-training pairs. To ensure data quality, we filter out videos that are unrelated to humans, contain inconsistent numbers of individuals, or exhibit inconsistencies in identity. Specifically, to retrieve human-related videos from the caption-video pairs, we design a human term table that includes various categories such as basic human descriptors, gender, and occupation. We then exclude videos whose captions do not contain any human-related terms. Next, we uniformly sample two frames per second from each video, detect faces using SCRFD [6], and remove videos if more than 30% of the frames have inconsistent numbers of individuals.[1] Finally, for frames with the same face count, we compute the ArcFace cosine similarity [3] between consecutive frames and discard videos if more than 30% of the frames have a similarity score below 0.5.

The above processes yield 1.3 million videos featuring a single identity, and we uniformly select 5 face images per video, defining pre-training pairs $\mathcal{S}_{\text{pre}} = \{(\mathbf{I}_i^k, \mathbf{X}^k)\}_{i,k}$ where k denotes the video index and $\mathbf{I}_i^k$ represents the i-th reference image of $\mathbf{X}^k$. The self-supervised nature of this paired data, where images from the same video serve as labels, inherently limits facial editability. Specifically, models trained on such data may produce frames in which facial expressions unintentionally mirror those of the reference images (see Fig. 7), leading to unnatural content. This issue becomes particularly pronounced when a semantic gap exists between the reference images and the text prompts. To enhance facial editability and naturalness, we propose a cross-video image-video pairing strategy.

Cross-video pairs. The standard process for constructing video clips involves segmenting raw long videos into multiple shorter segments using various algorithms that detect scene transitions, such as motion variations and shot changes. Theoretically, many existing video clips in training sets feature varied facial expressions and head poses of the same person. To construct cross-video pairs where the reference image originates from a different video, we calculate the cosine similarity among images $\{\mathbf{I}_1^v\}_v$. For the k-th video, we randomly select an image $\mathbf{I}_1^j$ from $\{\mathbf{I}_1^v\}_v$ as its paired reference image, ensuring that $0.7 \leq \cos(\mathbf{I}_1^j, \mathbf{I}_1^k) < 0.9$, where the function $\cos(\cdot,\cdot)$ computes the cosine similarity. The final cross-video pairs $\mathcal{S}_{\text{cross}}$ include 0.8 million image-video pairs with 0.5 million reference images, indicating a reference image can correspond to multiple videos.

Personalized image generation can also synthesize reference images with the same identity as given videos but varied identity-irrelevant factors, as demonstrated in [10, 19]. However, this approach incurs high computational costs, particularly for large-scale image-video pairs. Additionally, existing personalized generation methods [7, 27, 32] often struggle to preserve detailed facial features, which limits their effectiveness. In contrast, as shown in Fig. 3b, our retrieval-based method efficiently gathers a large-scale set of real reference images that accurately match the identity of corresponding videos while exhibiting diversity across multiple dimensions, such as facial expressions, hairstyles, lighting conditions, and other identity-irrelevant factors.

Trade-off pairs. Similar to the construction of cross-video pairs, for the k-th video, we identify its reference image $\mathbf{I}_1^j$ with the smallest $\cos(\mathbf{I}_1^j, \mathbf{I}_1^k)$, ensuring that $0.9 \leq$

[1]The common person count across frames is considered the video's person count.

$\cos(\mathbf{I}_1^j, \mathbf{I}_1^k) < 0.99$. This forms our trade-off dataset $\mathcal{S}_{\text{trade}}$ with 160 thousand videos, improving consistency between reference images and videos compared to cross-video pairs. Additionally, we filter out videos where the facial region occupies less than 4% or more than 90% of the frame area and rank $\mathcal{S}_{\text{trade}}$ based on the weighted sum of aesthetics scores, optical flow scores, and motion scores [26]. Finally, we retain the top 50,000 videos for training.

In this section, we detail the data construction process for a single identity. However, this procedure can be seamlessly scaled to multi-identity by independently processing each identity within a video. Similarly, it can be extended to general subjects by replacing face detectors with open-set detectors, such as Grounding DINO [21], and substituting ArcFace cosine similarity with general feature similarity metrics, such as CLIP cosine similarity [20]. Please refer to Appendix B for further details on the training data construction for our multi-identity and multi-subject scenarios.

4.3. Training strategy

Building on our innovative data construction, we introduce a multi-stage training process: pre-training stage, cross-video fine-tuning, and trade-off fine-tuning. In the pre-training stage, we optimize a text-to-video model on $\mathcal{S}_{\text{pre}}$ to map facial details into generated videos. This self-supervised training method may constrain certain generated video frames to adhere strictly to the given condition images, potentially degrading the editability of facial expressions and the overall naturalness. The cross-video fine-tuning on $\mathcal{S}_{\text{cross}}$, using image-video pairs derived from different videos, can alleviate this issue. However, we observe that this fine-tuning enhances facial editability at the expense of identity fidelity (see Sec. 5.4).

A simple strategy to further balance fidelity and editability is to mix pre-trained pairs and cross-video pairs in a 1:1 ratio, a similar method adopted by Movie-Gen [19]. However, our initial experiments suggest that this approach results in unstable training due to varying identity consistency between pre-trained pairs and cross-video pairs. To address this issue while ensuring high-degree motion and high artistic quality, we ultimately fine-tune the model on $\mathcal{S}_{\text{trade}}$.

Throughout all training stages, we proportionally scale, pad, and center-crop images to match the video resolution. To ensure the model focuses on facial regions during training and prevents background leakage during inference, we segment and drop the background of reference images [25]. Additionally, to improve robustness and generalization, we introduce random noise to reference images during training, while omitting this noise during inference. To further differentiate the image latent $\mathbf{c}_i$ from the video latent $\mathbf{Z}$ and distinguish between different $\mathbf{c}_i$, we extend 3D-RoPE to incorporate multiple reference images along the sequence dimension. Specifically, we introduce a temporal bias N to define the 3D position of a token $\mathbf{t}_{h,w}$ in $\mathbf{c}_i$:

$$\text{3D-Pos}(\mathbf{t}_{h,w}) = (i + N, h, w), \tag{2}$$

where $\text{3D-Pos}(\cdot)$ denotes the 3D position and (h, w) are the spatial coordinates of the token.

Owing to the simplicity and efficiency of Concat-ID in both data construction and model architecture, our training strategy can seamlessly scale to multi-identity and multi-subject scenarios. Moreover, we establish that single-identity pre-training facilitates enhanced identity preservation in these downstream tasks (see Tab. 2).

5. Experiments

5.1. Experimental settings

Datasets. We evaluate all methods on the ConsistID-Benchmark [28], which consists of 172 reference images and 90 text prompts spanning nine categories. To ensure a fair comparison, we exclude reference images present in our training data using a combination of automated and manual filtering techniques. Consequently, our evaluation dataset comprises 873 prompt-image pairs, derived from 97 reference images, with one prompt randomly selected from each category for each image. For multi-identity evaluation, we additionally construct 14 distinct pairs of reference images and design 20 textual prompts using ChatGPT [1]. Please refer to Appendix A.1 for further details.

Metrics. We evaluate all methods on identity consistency, text alignment, and facial editability. (1) Identity consistency: Following [28], we use FaceSim-Arc (ArcSim) and FaceSim-Cur (CurSim) to assess the average cosine similarity between reference images and generated videos based on ArcFace [3] and CurricularFace [12], respectively. These face recognition models are specifically designed to disentangle identity-related features from identity-unrelated ones. (2) Text alignment: We adopt ViCLIP [23] to compute the similarity between text prompts and generated videos, following [14, 19]. (3) Facial editability: We calculate the cosine distance of CLIP image embeddings [20] (CLIPDist) between reference images and video frames. CLIP effectively captures comprehensive facial features, and thus a larger CLIPDist indicates improved facial editability.

Implementation details. We use the text-to-video model CogVideoX-5B [26] as our base model. The learning rates are set to 1.0×10^{-5}, 5.0×10^{-6}, and 5.0×10^{-6} for the first, second, and third training stages, respectively. We fine-tune all model parameters with a linear learning rate decay across all stages. The training data resolution is maintained at 480×720 pixels with 49 frames per video. Text and image prompts are independently dropped with a probability of 0.1. Further details are provided in Appendix A.2.

Baselines. For a comprehensive comparison, we use three representative open-source approaches as baselines. (1)

Method	Identity consistency		Text alignment	Facial editability
	ArcSim ↑	CurSim ↑	ViCLIP ↑	CLIPDist ↑
Single identity				
ID-Animator [9] ‡	0.289	0.304	0.204	0.297
ConsisID [28]†‡	0.432	0.451	0.237	0.303
Concat-ID (Ours) †	**0.442**	**0.466**	**0.242**	**0.325**
Multiple identities				
Ingredients [4] †‡	0.293	0.316	**0.199**	0.407
Concat-ID (Ours) †	**0.492**	**0.514**	0.190	**0.410**

Table 1. **Quantitative results for single-identity and multi-identity generation.** † denotes that these methods share the same video model. ‡ indicates corresponding methods introduce additional adapters and auxiliary loss. Concat-ID achieves superior identity consistency and facial editability while maintaining better or comparable text alignment relative to the baselines.

Single-identity personalization methods: ID-Animator [9] and ConsisID [28]. (2) Multi-identity personalization methods: Ingredients [4]. ID-Animator, ConsisID, and Ingredients all incorporate additional adapters and auxiliary loss functions to enhance identity consistency. Notably, Concat-ID, ConsisID, and Ingredients are all built upon the same video model, CogVideoX-5B.

5.2. Main results

We demonstrate the effectiveness of Concat-ID through quantitative metrics, qualitative assessments, and the user study for single-identity and multi-identity generation.

Quantitative comparisons. Table 1 presents the quantitative results for single-identity and multi-identity generation. For single-identity generation, ID-Animator performs the worst, exhibiting the lowest ArcSim, CurSim, and CLIPDist scores. This suggests that it achieves the least effective balance between identity preservation and facial editability. Moreover, ID-Animator, ConsisID, and Ingredients incorporate additional adapters and auxiliary loss functions to enhance identity consistency, increasing the complexity of both training and generation processes.

In contrast, for both single-identity and multi-identity generation, Concat-ID achieves superior identity consistency simply by concatenating image latents after video latents, highlighting the effectiveness of our architecture. Furthermore, by constructing cross-video pairs, Concat-ID attains a higher CLIPDist score than ID-Animator, ConsisID, and Ingredients, demonstrating an optimal balance between identity preservation and facial editability.

Qualitative comparisons. Fig. 4 presents qualitative comparisons for single-identity generation. ID-Animator fails to maintain facial characteristics. ConsisID achieves better identity consistency, but some frames replicate facial expressions of reference images. In contrast, Concat-ID mitigates this issue while preserving identity by leveraging advantages of cross-video pairs. For multi-identity generation, as shown in Fig. 5, Concat-ID produces videos that more accurately match identities in given images compared to Ingredients, demonstrating its effectiveness and scalability.

To maximize the potential of image-to-video models, ConsisID and Ingredients concatenate the reference image with the first latent frame along the channel dimension. However, this feature injection approach can introduce artifacts in the first generated frame due to spatial misalignment between faces images and generated videos, as evident in the initial frames of all videos. As a comparison, Concat-ID excels in identity preservation without compromising the quality of any generated frames, highlighting the validity of our concatenation along the sequence dimension.

User study. According to both quantitative and qualitative results, we compare Concat-ID with the strongest baseline, ConsisID, through human evaluation. Specifically, we generate 100 videos using 10 reference images and 10 prompts designed by ChatGPT [1] to focus on expression and head pose variation. For each video group, voters answer three questions, selecting the video that: (1) best matches the reference image in facial similarity (identity consistency), (2) best aligns with the facial expressions and head poses described in the prompt (facial motion alignment), and (3) exhibits the most natural and smooth facial motion (facial motion naturalness). With 100 video groups, three types of questions, and three voters participating, we collect a total of 900 video comparison results. As shown in Fig. 6, Concat-ID surpasses ConsisID by a significant margin in identity consistency and motion alignment and naturalness, demonstrating the effectiveness of our architecture and the advantages of cross-video pair construction.

5.3. Multiple identities and subjects

We demonstrate that the architecture, data construction, and training strategy of Concat-ID make it seamlessly extendable to multi-identity and multi-subject scenarios.

Multi-identity scenarios. As illustrated in Fig. 1b, when provided with face images of different individuals, Concat-ID can generate multi-person videos while preserving their identities, without requiring any additional parameters or modules compared to single-identity generation. Notably, despite being trained on only 40,000 videos, Concat-ID can generate three-identity videos while maintaining distinct identities, leveraging the prior knowledge from two-identity pre-training and a powerful 3D self-attention mechanism that effectively captures both temporal and spatial dependencies. Moreover, Concat-ID determines the spatial position of each identity in the generated videos based on the concatenation sequence of the reference images.

Multi-subject scenarios. As illustrated in Fig. 1c, by sequentially concatenating a face image with a clothing image, Concat-ID enables virtual try-on while preserving both the given identity and intricate clothing details, such as lo-

Figure 4. **Qualitative comparisons for single-identity generation.** ID-Animator fails to preserve facial details, while ConsisID replicates the expressions of the reference images, particularly in the third case, where the semantic gap between texts and reference is significant. Concat-ID effectively preserves identity, while simultaneously preventing the direct replication of facial expressions from reference images.

gos and textures. This capability also highlights Concat-ID's potential in simulating interactions between people and objects. Furthermore, the background-controllable identity-preserving generation achieved by Concat-ID demonstrates its ability to manipulate spatial layouts in generated videos by integrating spatially aligned conditions.

In this section, we introduce two-identity and three-identity generation, along with two additional subjects (*i.e.* clothing and background). Further details on training and data are provided in Appendix B. We posit that Concat-ID's architecture, characterized by its simplicity and effectiveness, coupled with the generalizability of its data construction and training strategy, enables effective scalability to more identities and diverse subjects, ensuring consistent high performance across a wider range of applications.

5.4. Ablation study

Fig. 7 present the qualitative ablation of Concat-ID. The pre-training stage achieves the best identity consistency but results in low facial editability. For example, facial expressions of some frames in the pre-training stage closely resemble those in reference images. However, the cross-video stage enhances editability at the expense of identity consistency, aligning with the findings in [19]. In the third stage, Concat-ID further refines the matching threshold of cross-video pairs to better balance identity preservation and facial editability. Leveraging prior knowledge from both pre-training and cross-video fine-tuning, the trade-off stage achieves an optimal balance using only 50,000 videos. These results underscore the effectiveness of each stage in

Figure 5. **Qualitative comparisons for multi-identity generation.** Concat-ID better maintains different identities.

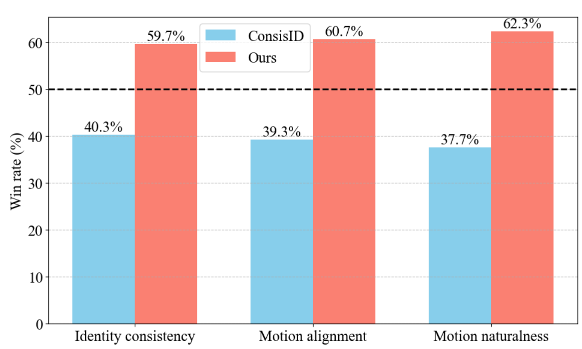


Figure 6. **Human evaluation.** Concat-ID produces more precise and natural videos while effectively preserving identity.

our training strategy. Moreover, the quantitative analysis in Appendix C consistently supports our findings.

Figure 7. **Qualitative ablation.** Stage I, Stage II, and Stage III indicate the pre-training stage, cross-video stage, and trade-off stage.

Method	Identity-1		Identity-2	
	ArcSim ↑	CurSim ↑	ArcSim ↑	CurSim ↑
No single-identity pre-training	0.514	0.535	0.526	0.550
Concat-ID (Pre-training)	**0.629**	**0.650**	**0.651**	**0.674**

Table 2. **The effect of single-identity pre-training on multi-identity pre-training.** The single-identity pre-training enhances identity consistency in downstream tasks.

We also investigate the influence of single-identity pre-training on multi-identity and multi-subject pre-training. Specifically, we conduct a comparative analysis of Concat-ID with and without single-identity pre-training. Although the two-identity generation is pre-trained on approximately 0.3 million videos, as presented in Tab. 2, single-identity pre-training still results in improved ArcSim and CurSim scores across all identities. This enhancement indicates that single-identity pre-training effectively strengthens identity preservation in downstream tasks. These findings provide empirical support for the scalability of our architecture, data construction methodology, and training strategy.

6. Conclusions

In this paper, we introduce Concat-ID, a unified framework for identity-preserving video generation. Concat-ID relies solely on 3D self-attention mechanisms, which are commonly used in state-of-the-art video generation models, without introducing additional modules or parameters. We also present a novel cross-video pairing strategy and a multi-stage training regimen to balance identity consistency and facial editability while enhancing video naturalness. Thanks to its architecture, data construction, and training strategy, Concat-ID can scale seamlessly to multi-identity and multi-subject scenarios.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 6

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 4, 5

[4] 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. 2, 3, 6

[5] Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. 3

[6] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021. 4

[7] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 4

[8] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 3

[9] 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. 2, 6

[10] 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. 4

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[12] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 5

[13] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025. 2

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[15] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 3

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[17] 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. 2

[18] Pika. Pikascenes. `https://pika.art/ingredients`, 2024. 2

[19] 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, 4, 5, 7

[20] 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, 5

[21] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 5

[22] Vidu. Reference to video. `https://www.vidu.com/create/character2video`, 2024. 2

[23] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 5

[24] Tao Wu, Yong Zhang, Xiaodong Cun, Zhongang Qi, Junfu Pu, Huanzhang Dou, Guangcong Zheng, Ying Shan, and Xi Li. Videomaker: Zero-shot customized video generation with the inherent force of video diffusion models. *arXiv preprint arXiv:2412.19645*, 2024. 2

[25] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and

efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 5

[26] 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, 3, 5

[27] 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. 2, 4

[28] 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. 2, 3, 5, 6

[29] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic mirror: Id-preserved video generation in video diffusion transformers. *arXiv preprint arXiv:2501.03931*, 2025. 2

[30] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025. 2

[31] Yong Zhong, Min Zhao, Zebin You, Xiaofeng Yu, Changwang Zhang, and Chongxuan Li. Posecrafter: One-shot personalized video synthesis following flexible pose control. *arXiv preprint arXiv:2405.14582*, 2024. 3

[32] Yufan Zhou, Ruiyi Zhang, Kaizhi Zheng, Nanxuan Zhao, Jiuxiang Gu, Zichao Wang, Xin Eric Wang, and Tong Sun. Toffee: Efficient million-scale dataset construction for subject-driven text-to-image generation. *arXiv preprint arXiv:2406.09305*, 2024. 4