This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Lynx: Towards High-Fidelity Personalized Video Generation

Shen Sang* Tiancheng Zhi* Tianpei Gu Jing Liu Linjie Luo

ByteDance Inc.

*Equal Contribution

{shen.sang, tiancheng.zhi, tianpei.gu, jing.liu, linjie.luo}@bytedance.com

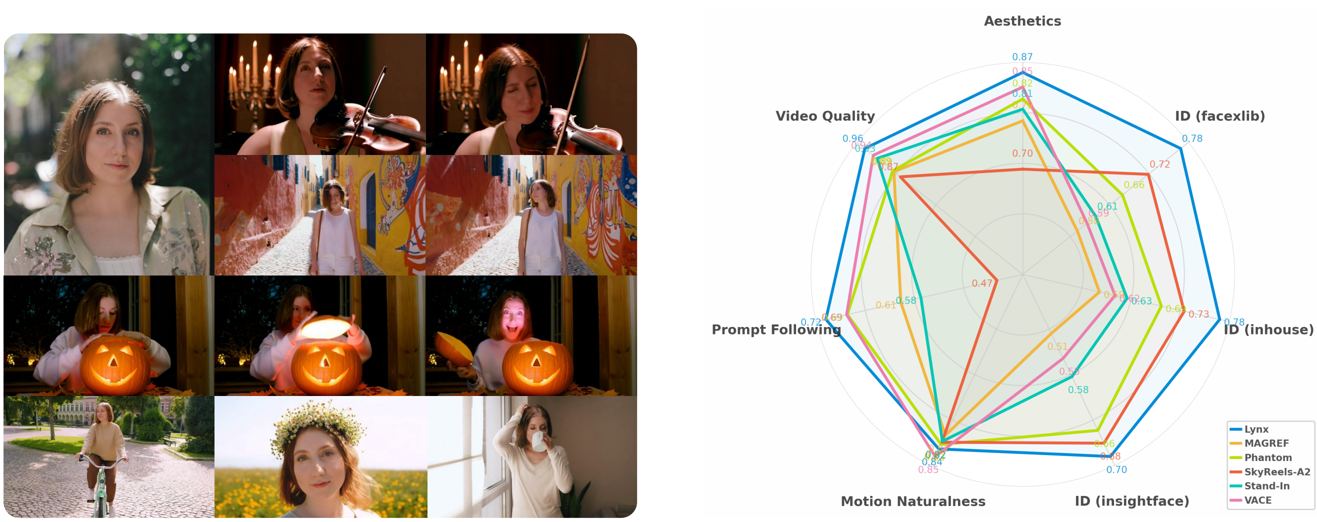


Figure 1. **Lynx teaser.** Left: Lynx consistently preserves facial identity with high fidelity, while producing natural motion, coherent lighting, and flexible scene adaptation (input shown at top-left). Right: Lynx demonstrates clear superiority in identity resemblance and perceptual quality, while remaining competitive in motion naturalness compared to other methods.

Abstract

*We present **Lynx**, a high-fidelity model for personalized video synthesis from a single input image. Built on an open-source Diffusion Transformer (DiT) foundation model, Lynx introduces two lightweight adapters to ensure identity fidelity. The ID-adapter employs a Perceiver Resampler to convert ArcFace-derived facial embeddings into compact identity tokens for conditioning, while the Ref-adapter integrates dense VAE features from a frozen reference pathway, injecting fine-grained details across all transformer layers through cross-attention. These modules collectively enable robust identity preservation while maintaining temporal coherence and visual realism. Through evaluation on a curated benchmark of 40 subjects and 20 unbiased prompts, which yielded 800 test cases, Lynx has demonstrated superior face resemblance, competitive prompt following, and strong video quality, thereby advancing the state of personalized video generation. Our project page:* https://byteaigc.github.io/Lynx/

1. Introduction

The field of visual content generation has witnessed rapid progress, largely propelled by the emergence of diffusion models [18, 37, 42], which offer a scalable and effective framework for high-fidelity synthesis across diverse modalities. Building upon early breakthroughs in text-to-image generation [3, 35, 40, 41, 44], the community has extended diffusion-based methods into the temporal domain, giving rise to text-to-video models [4, 16, 27, 28, 36, 45, 47, 51] capable of synthesizing dynamic visual content from natural language prompts. Recent advancements in backbone architectures—such as Diffusion Transformers (DiT) [37]—have further improved generation quality and scalability. Recent works have also explored improv-

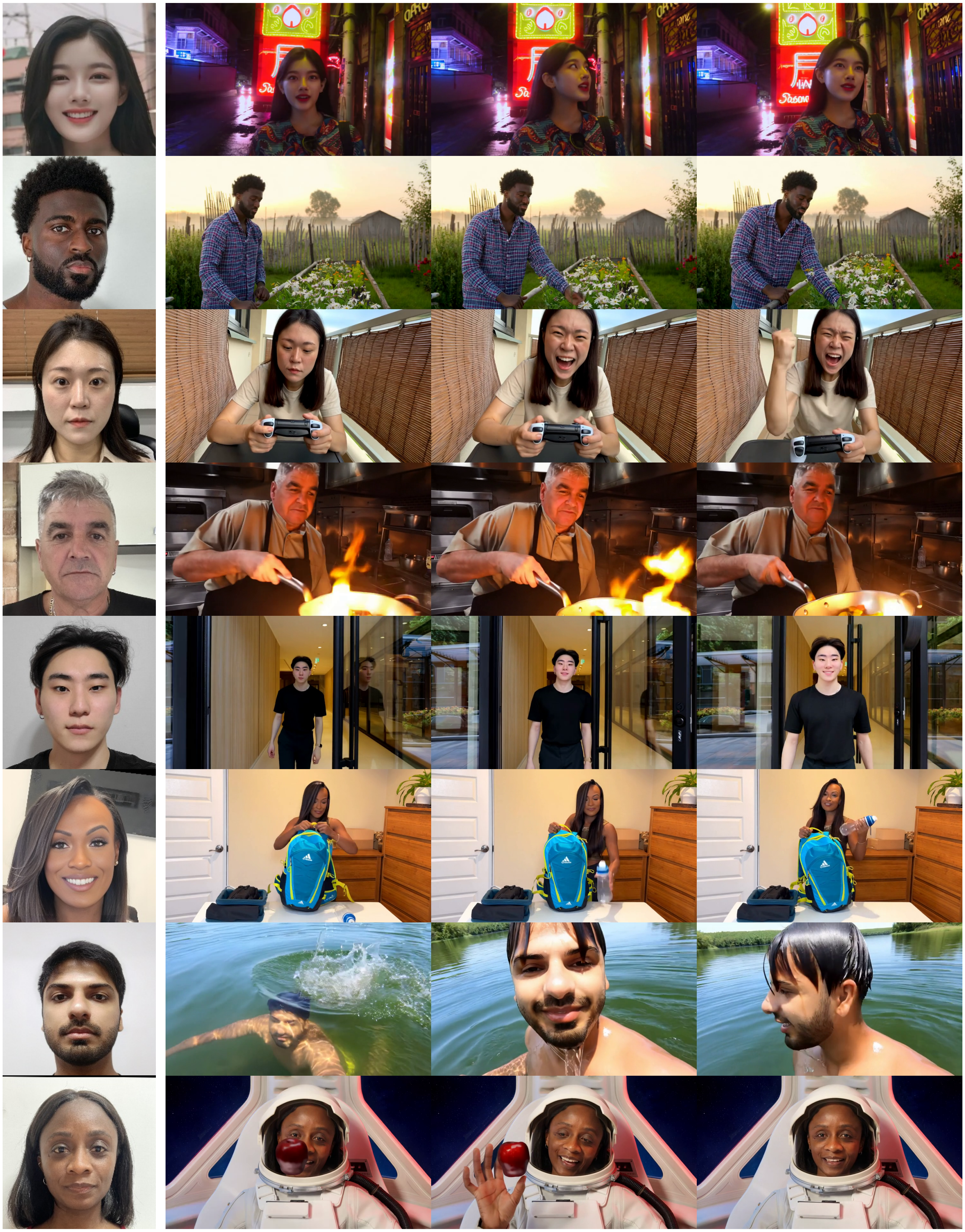


Figure 2. Videos generated from a single input image, showing strong identity preservation across expressive facial expressions (row 3), diverse lighting (rows 1, 4, 5), pose variations (rows 2, 6, 7), and object interactions (row 8).

ing video generation through reward-based optimization and alignment objectives for diffusion models [6, 38, 39]. Beyond foundational generation, there is growing interest in downstream tasks including video editing [29, 49, 57], multi-shot storytelling [25], and controllable motion synthesis [15, 24], reflecting the field's increasing demand for controllability, reusability, and efficiency. A key direction emerging from this trend is personalization, which aims to synthesize videos that faithfully preserve subject identity.

Personalized generation has been widely explored in the image domain, where pretrained diffusion models are adapted to user-provided reference images to achieve identity-consistent synthesis. Early approaches [13, 19, 43] relied on model fine-tuning or parameter-efficient adaptation techniques such as LoRA [19] to encode subject-specific information. While effective, these methods often require high computational cost and limit scalability to multiple identities. More recent works [48, 52, 56] adopt lightweight conditioning modules that inject identity features or embeddings directly into the diffusion process, avoiding full retraining and enabling more efficient personalization. These adapter-based methods have proven capable of maintaining strong identity fidelity while retaining generalization across diverse contexts.

Motivated by these advances, several studies have recently extended personalization to the video domain [10, 12, 23, 30, 33, 50, 53, 55]. Compared with image synthesis, personalized video generation introduces new challenges, including temporal consistency of identity features, motion-aware alignment across frames, and robust generalization under varying viewpoints and lighting conditions. Achieving temporally coherent identity preservation from limited references remains an open research problem.

In this work, we introduce **Lynx**, a high-fidelity personalized video generation framework that preserves subject identity from a single input image. Rather than fine-tuning the entire model, Lynx employs an adapter-based design with two specialized components: the *ID-adapter* and the *Ref-adapter*. The *ID-adapter* injects compact identity representations derived from facial embeddings into the generation process via cross-attention. Specifically, a face recognition network extracts embeddings that are transformed into a small set of identity tokens through a Perceiver Resampler, providing a rich and efficient representation of identity. The *Ref-adapter* complements this by incorporating detailed reference features from a pretrained VAE encoder of the base model. These features are propagated through a frozen copy of the diffusion backbone to extract intermediate activations from all DiT blocks, which are then fused into the main generation pathway to enhance detail fidelity.

To train Lynx effectively, we adopt a multi-stage progressive learning strategy combined with a spatio-temporal frame packing mechanism. This design enables efficient handling of mixed image and video data with diverse aspect ratios and temporal lengths, improving both spatial detail and temporal coherence.

For evaluation, we construct a benchmark comprising 40 subjects and 20 human-centric, unbiased text prompts, totaling 800 test cases. Identity preservation is measured using three state-of-the-art face recognition models, while prompt following and visual quality are assessed via an automated evaluation pipeline built on the Gemini-2.5-Pro API [1] . The model is instructed to score aesthetic quality, motion naturalness, prompt alignment, and overall perceptual quality. As shown in Table 1 and Table 2, Lynx achieves consistently superior performance compared to recent personalized video generation baselines, demonstrating strong identity fidelity, precise prompt adherence, and high perceptual quality.

2. Related Works

Video Foundation Models. Recent video foundation models are predominantly built on the diffusion framework, where variational autoencoders (VAEs) [26] compress raw videos into compact latent representations, enabling efficient training and generation. Early latent diffusion methods extended image foundation models with U-Net architectures to the video domain by incorporating temporal modules such as 3D convolutions and temporal attention [4, 17, 45]. As the demand for scalability and long-range temporal coherence increased, research shifted toward transformer-based architectures. Diffusion Transformers (DiT) [37] and their dual-stream variant MMDiT [11] demonstrated more expressive spatio-temporal modeling, leading to improved temporal consistency. These architectures now underpin state-of-the-art video foundation models, including CogVideoX [51], HunyuanVideo [28], Wan2.1 [47], Seedance [14], *etc.*, which achieve strong generalization through large-scale training data, substantial computational resources, and extended context length.

Identity-Preserving Content Creation. Identity-preserving generation is a central topic in content creation and has been extensively studied in the image domain. Early approaches [13, 19, 43] typically rely on model fine-tuning or optimization to obtain subject-specific models. However, such tuning-based methods are often impractical for real-world applications because of their computational cost and lack of scalability. For example, DreamBooth [43] and LoRA-based variants [19] require fine-tuning either the full base model or additional low-rank adapters. To overcome these limitations, tuning-free methods [48, 52] introduce lightweight ID-injection modules that avoid

[1] https://ai.google.dev/gemini-api/docs

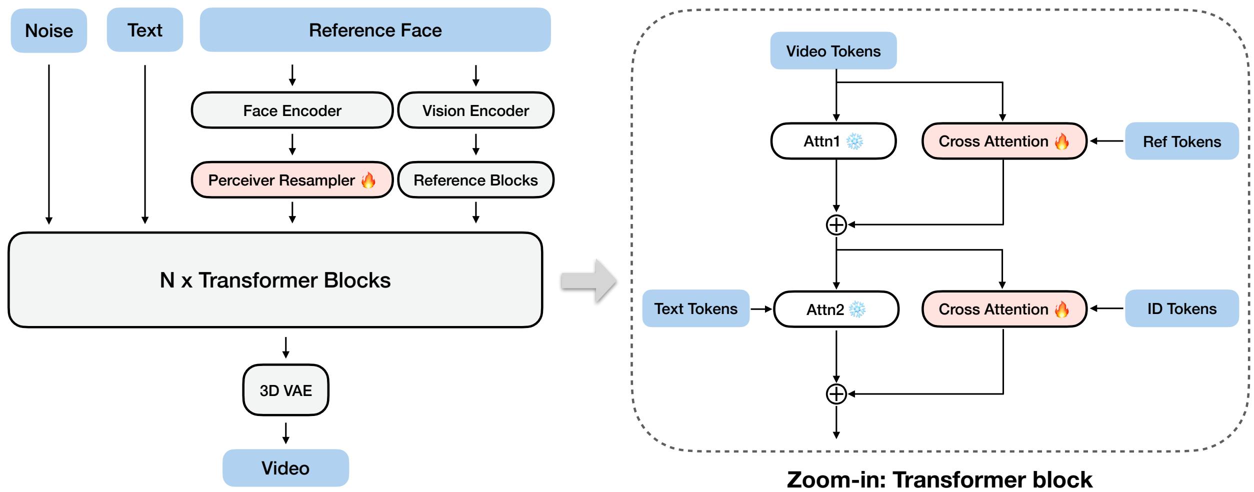


Figure 3. Architecture of Lynx. Built on a DiT-based video foundation model, Lynx introduces two adapter modules that inject identity features through cross-attention.

per-subject training. IP-Adapter [52] represents identity features with a face recognition encoder and injects them into the base model through adapters. Building on this idea, InstantID [48] incorporates a ControlNet [54] module for input decoupling and finer-grained control.

With the advent of large video foundation models, research attention has shifted toward personalized video generation. For instance, ConsistID [53] enforces facial identity consistency via frequency decomposition. Concept-Master [22] employs a CLIP image encoder and a learnable Q-Former to fuse visual representations with corresponding text embeddings for each concept. HunyuanCustom [21] extends HunyuanVideo [28] with a multi-modal customization framework that integrates image, audio, video, and text conditions through modality-specific modules, achieving stronger identity consistency and controllable video generation. Another line of work (*e.g.*, SkyReels-A2 [12], VACE [23], Phantom [33]) concatenates reference conditions with noisy latents and processes the full sequence during denoising. However, balancing identity resemblance and editability has long been a persistent challenge. Our method significantly improves resemblance while maintaining strong prompt following and high video quality.

3. Architecture and Training Strategy

3.1. Model Architecture

We adopt Wan2.1-14B [47], one of the latest open-sourced video foundation models, as our base model. Wan is built upon the DiT architecture [37], combined with the Flow Matching [32] framework. Each DiT block first applies spatio-temporal self-attention over visual tokens, enabling joint modeling of spatial details and temporal dynamics, followed by cross-attention to incorporate text conditions.

Instead of restructuring and fine-tuning the full model, we introduce two adapter modules, *i.e.*, *ID-adapter* and *Ref-adapter*, to inject identity features and enable personalized video generation on top of the base model. The overall architecture and adapter design are illustrated in Figure 3.

ID-adapter. Prior works [48, 52] incorporated face recognition features [9] to achieve personalized generation in text-to-image models such as Stable Diffusion [41]. These methods typically attach additional adapter layers and introduce extra cross-attention modules to condition generation on identity features. Specifically, face image is passed through a face feature extractor to obtain a feature vector. To convert this vector into a sequence suitable for cross-attention, a Perceiver Resampler [1] (also known as the Q-Former [31]) is trained to map it into a fixed-length token embedding representation. We adopt the same paradigm. Given a face feature vector of dimension 512, the Resampler produces a sequence of 16 token embeddings of dimension 5120. The token embedding is concatenated with 16 additional register tokens [7] and cross-attended with the input visual tokens. The resulting representation is then added back to the main branch.

Ref-adapter. Several recent approaches [12, 33] use VAE features to enhance detail preservation during reference injection, taking advantage of the spatially dense representations produced by VAE encoders. Complementing the ID-adapter, our design also incorporates VAE dense features to enhance identity fidelity. Unlike prior approaches that directly place the feature map in front of noisy latents in an image-to-image-like generation fashion, we instead process the reference image through a frozen copy of the base model (with noise level as 0 and text prompt as ”image of a face”), similar to the design of ReferenceNet [20]. This allows spatial details from the reference image to be captured across all layers. As with ID-adapter, we apply separate cross-attention at each layer to integrate the corresponding reference tokens.

3.2. Training Strategy and Implementation Details

We describe here the strategies employed for large-scale training. Since training videos (and images) vary in both spatial resolution and temporal duration, we adopt the NaViT approach [8] to efficiently batch heterogeneous inputs. Multiple videos or images are packed into a single long sequence, with attention masks applied to separate samples. Training follows a progressive curriculum beginning with image pretraining, which leverages the abundance of large-scale image data, and is then extended to video training to restore temporal dynamics.

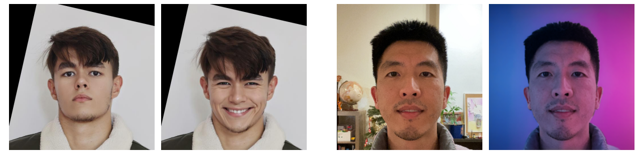
3.2.1. Spatio-Temporal Frame Pack

Traditional training in the image domain often relies on bucketing to handle multi-resolution inputs. Images are cropped and resized into a set of predefined aspect ratios and resolutions, and during training the data loader samples from a single bucket so that images within a batch share the same dimensions. While effective for images, this strategy does not generalize well to video, as the additional temporal dimension (frame length) introduces significant complexity. Bucketing by both resolution and duration reduces flexibility and limits the model’s ability to generalize to arbitrary aspect ratios and video lengths.

To overcome this limitation, inspired by Patch n’ Pack [8], we concatenate the patchified tokens of each video into a single long sequence, treating the collection a unified batch. An attention mask ensures that tokens only attend within their own video, preventing cross-sample interference. For positional encoding, we apply 3D Rotary Position Embeddings (3D-RoPE) [46] independently to each video. This design enables efficient batching of heterogeneous images and videos while preserving both spatial and temporal consistency.

3.2.2. Progressive Training

Image Pretraining. We begin with image pretraining given the large amount of available image data. To ensure consistency across training stages, each image is treated as a single-frame video and the same frame pack strategy described above is applied. In our experiments, training the Perceiver Resampler from scratch yielded unsatisfactory results: no facial resemblance was observed even after substantial training, suggesting that the model either fails to converge or requires prohibitively longer training. Instead, we found that initializing the Resampler from an image-domain pretrained checkpoint (*e.g.*, InstantID [48]) leads to



(a) Expression Augmentation (b) Portrait Relighting

Figure 4. Examples of our augmentation strategies: (a) expression augmentation via X-Nemo [58], and (b) portrait relighting via LBM [5].

much faster convergence. With this initialization, recognizable facial resemblance emerges after only 10k iterations, while the complete first stage runs for 40k iterations.

Video Training. Image pretraining alone tends to produce videos that are largely static, as the model primarily learns to preserve appearance rather than capture motion. To restore temporal dynamics, a second stage that exposes the model to large-scale video data is necessary. This stage enables the network to learn motion patterns, scene transitions, and temporal consistency while retaining and enhancing the strong identity conditioning established during image pretraining. Training proceeds for 60k iterations.

Hyperparameters. We use the AdamW [34] optimizer, with learning rate 1e-5 and weight decay 0.01. We use 128 80G GPUs for training. Since we pack all tokens as mentioned in Sec. 3.2.1, we use number of tokens instead of batch size to measure the amount of data. The number of tokens processed in each iteration is 33,600 per GPU.

4. Data Pipeline

The goal of our data pipeline is to construct high-quality person–text–video triplets. While text prompts can be readily obtained through captioning models (e.g., Qwen 2.5-VL [2]), the main challenge lies in establishing reliable person–video pairs, i.e., pairing an image of a person as the ID condition with a target video of the same individual.

Our raw data consist of images and videos collected from both publicly available datasets and in-house sources. We categorize the data into four types: (1) single images; (2) single videos; (3) multi-scene image collections of the same person; and (4) multi-scene video collections of the same person. To construct image–image and image–video pairs, where one image serves as the ID condition and the other image or video serves as the generation target, a straightforward approach is to crop faces directly from images or videos. However, this often leads to overfitting of expression and lighting. Meanwhile, multi-scene data, which are essential for robust training, are inherently scarce.

To address these limitations, we adopt two augmentation strategies, illustrated in Figure 4:

- **Expression Augmentation.** We employ X-Nemo [58] to edit a source face so that it matches the target expression, thereby enriching expression diversity (Figure 4a).

- **Portrait Relighting.** We apply LBM [5] to relight faces and replace backgrounds under varying illumination conditions, enhancing robustness to lighting variation (Figure 4b).

After augmentation, we perform identity verification using a face recognition model and discard pairs with low resemblance to ensure high-quality ID consistency. Resemblance filter is also applied to raw multi-scene data without augmentation.

Finally, our pipeline constructs a total of 50.2M pairs, consisting of 21.5M single-scene pairs, 7.7M multi-scene pairs, and 21.0M augmented single-scene pairs. For single-scene pairs where the condition image is directly cropped from the target, we also apply background augmentation by segmenting the subject and replacing the background. During training, these different types of pairs are retrieved through weighted sampling to balance data diversity.

5. Experiments

5.1. Benchmark and Metrics

We construct an evaluation benchmark comprising 40 subjects and 20 unbiased text prompts, resulting in a total of 800 test videos. The subject set consists of: 10 celebrity photos, 10 AI-synthesized portraits, and 20 in-house licensed photos spanning diverse demographic groups to capture racial and ethnic diversity. The text prompts are generated using ChatGPT-4o, guided by carefully designed in-context examples, and explicitly crafted to avoid bias with respect to race, age, gender, motion, and other attributes.

We evaluate Lynx along three key dimensions: *face resemblance*, *prompt following*, and *video quality*.

Face resemblance. To measure identity fidelity, we compute cosine similarity using three independent feature extractors. These include two publicly available ArcFace implementations, facexlib[2] and insightface[3], together with our in-house face recognition model. Employing multiple extractors reduces reliance on a single feature space and yields a more reliable assessment of identity preservation.

Prompt following and video quality. To assess semantic alignment and perceptual quality, we construct an automated evaluation pipeline based on the Gemini-2.5-Pro API. In this pipeline, Gemini is instructed with task-specific prompts to assign scores across four dimensions: (1) *prompt alignment*, which evaluates consistency between the generated video and the input text description, (2) *aesthetic quality*, which measures visual appeal and composition, (3) *motion naturalness*, which captures the smoothness and realism of temporal dynamics, and (4) *general video quality*, which provides an overall judgment that integrates multiple aspects of perceptual fidelity. This evaluation framework allows scalable and multi-faceted assessment of generated videos beyond traditional expert-model-based metrics.

[2] https://github.com/xinntao/facexlib

[3] https://github.com/deepinsight/insightface

Table 1. Quantitative comparison of Lynx with recent personalized video generation models on *face resemblance*. Scores are computed with three independent evaluators: facexlib, insightface, and our in-house face recognition model. Lynx achieves the best overall identity consistency across all evaluators, while SkyReels-A2 ranks second but shows weak prompt following due to reliance on copy–paste mechanisms, as shown in Table 2.

Model	Face Resemblance		
	facexlib	insightface	in-house
SkyReels-A2 [12]	<u>0.715</u>	<u>0.678</u>	<u>0.725</u>
VACE [23]	0.594	0.548	0.615
Phantom [33]	0.664	0.659	0.689
MAGREF [10]	0.575	0.510	0.591
Stand-In [50]	0.611	0.576	0.634
Lynx (ours)	**0.779**	**0.699**	**0.781**

5.2. Qualitative Results

Figure 5 presents qualitative comparisons between Lynx and state-of-the-art baselines. As shown, existing methods frequently struggle with identity preservation, producing faces that drift away from the reference subject or lose fine-grained details (row 1 example 1, row 3 example 2). Moreover, they often generate unrealistic behaviors (row 1 example 2), copy-pasting effects of background (row 4 example 2) or lighting (row 5 example 2). In contrast, Lynx successfully maintains strong identity consistency across diverse prompts, while achieving natural motion, coherent visual details, and high-quality scene integration. These results demonstrate that our model effectively balances identity preservation, prompt alignment, and video realism, outperforming existing approaches both in terms of fidelity and controllability.

5.3. Quantitative Results

Table 1 reports quantitative comparisons across *face resemblance*, *prompt following*, and *video quality*. On identity preservation, Lynx consistently outperforms all baselines, achieving the highest resemblance scores.SkyReels-A2 ranks second on identity resemblance, but its reliance on copy–paste generation introduces visual artifacts and leads to weak semantic alignment, as reflected in its poor prompt following performance as shown in Table 2. Phantom demonstrates strong prompt alignment at the expense of identity fidelity, suggesting a trade-off between semantic consistency and subject preservation. In contrast, Lynx achieves the best balance, combining superior identity fidelity with competitive prompt alignment, highlighting the advantage of our model design.

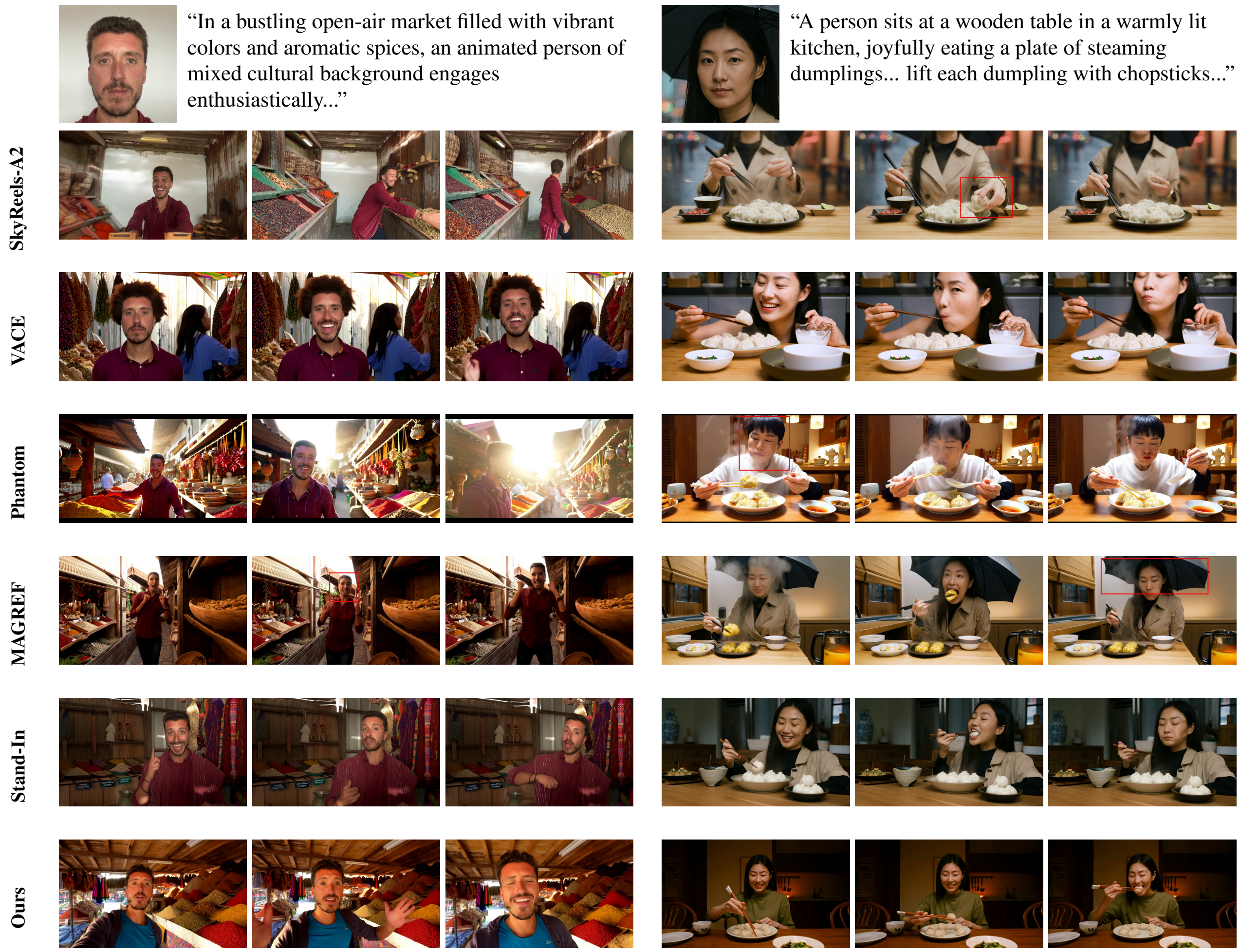


Figure 5. Qualitative comparison with baseline methods. As annotated with red boxes, competing methods often exhibit issues such as unrealistic actions (row 1 example 2), copy-pasting effects of background (row 4 example 2) or lighting (row 5 example 2), or poor identity resemblance (row 3 example 2, row 4 example 1). In contrast, Lynx consistently preserves facial identity with high fidelity, while producing natural motion, coherent lighting, and flexible scene adaptation.

Table 2 further evaluates *prompt following*, *aesthetic quality*, *motion naturalness*, and *overall video quality* using the Gemini-2.5-Pro evaluation pipeline. Lynx delivers the best performance in four out of five metrics, including prompt alignment, aesthetics, and overall video quality, which demonstrates the perceptual quality of our outputs. VACE attains the highest score in motion naturalness, reflecting its strong temporal modeling capability, while Phantom and Stand-In perform competitively across most dimensions but lag behind in overall video quality. These results show that Lynx not only preserves identity more effectively but also produces videos that are semantically accurate, visually appealing, and of high perceptual quality.

Figure 1 (right) provides a visual summary of these comparisons, where Lynx demonstrates consistent superiority across identity resemblance and perceptual quality dimensions. The combined evidence from multiple evaluators underscores the robustness of our approach and establishes Lynx as a new state of the art. The user study in supplementary materials also supports this claim.

5.4. Ablation Study

To evaluate the contribution of each component, we perform ablation studies by removing either the ID-adapter or the Ref-adapter from the full Lynx model. As shown in Table 3, excluding the Ref-adapter (*Lynx-id-only*) slightly affects identity resemblance but causes a noticeable drop in prompt following. Conversely, removing the ID-adapter (*Lynx-ref-only*) slightly improves prompt following (0.722 $\rightarrow$ 0.738)

Table 2. Quantitative comparison of Lynx with competing methods on *prompt following*, *aesthetic quality*, *motion naturalness*, and *overall video quality*, evaluated using the Gemini-2.5-Pro pipeline. Lynx achieves the highest performance in three out of four metrics, with particularly strong results in prompt alignment and overall quality.

Model	Prompt Following	Aesthetic Quality	Motion Naturalness	Video Quality
SkyReels-A2 [12]	0.471	0.704	0.824	0.870
VACE [23]	0.691	0.846	**0.851**	0.935
Phantom [33]	0.690	0.825	0.828	0.888
MAGREF [10]	0.612	0.787	0.812	0.886
Stand-In [50]	0.582	0.807	0.823	0.926
Lynx (ours)	**0.722**	**0.871**	0.837	**0.956**

Table 3. Ablation study on module design. We compare the full Lynx model with two variants: *Lynx-id-only* and *Lynx-ref-only*. Results are reported across three metrics: identity resemblance, prompt following, and overall video quality.

Model	ID-insightface	Prompt Following	Video Quality
Lynx-id-only	0.655	0.624	0.925
Lynx-ref-only	0.523	**0.738**	0.921
Lynx (full)	**0.699**	0.722	**0.956**

but significantly degrades identity consistency (0.699 → 0.523), indicating that while the Ref-adapter enhances semantic controllability, it alone cannot preserve subject appearance. The full Lynx model achieves the best trade-off across all metrics, combining strong identity preservation, faithful prompt alignment, and high perceptual quality. The visual results in Figure 6 demonstrates the same conclusion. These results confirm that the ID- and Ref-adapters offer complementary benefits, jointly enabling robust and semantically accurate personalized video generation.

6. Conclusion

In this work, we introduced **Lynx**, a high-fidelity framework for personalized video generation that preserves subject identity from a single reference image. Lynx employs two lightweight adapters: the *ID-adapter*, which encodes ArcFace-derived identity tokens to represent distinctive facial traits, and the *Ref-adapter*, which integrates VAE-based dense features through a frozen diffusion pathway to enhance visual detail. Together, these modules enable strong identity fidelity while maintaining motion realism and temporal coherence.

Comprehensive evaluation shows that Lynx achieves state-of-the-art identity resemblance and competitive video quality. Future works can include multi-subject and cross-modal personalization, enabling coherent interactions and generalization across visual and auditory modalities. Further integration of controllable motion editing may improve realism and user controllability. Overall, Lynx provides a scalable adapter-based solution that advances personalized video generation toward practical, flexible, and identity-consistent synthesis.

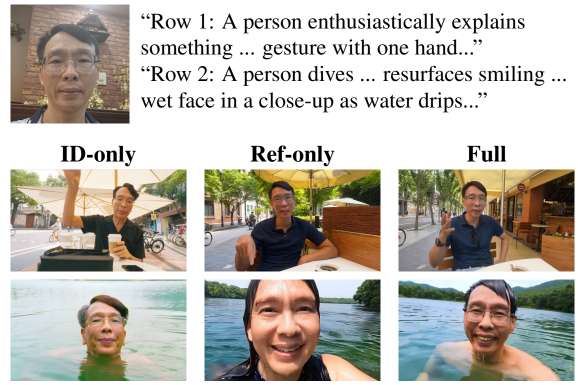


Figure 6. Qualitative ablation study. As shown in Column 1, when using only the ID-adapter, the hairstyle is exaggerated, and the hand gesture described in the prompt is not as clear as in the full method. As shown in Column 2, when using only the Ref-adapter, the identity resemblance (row 2) is low compared to the full method.

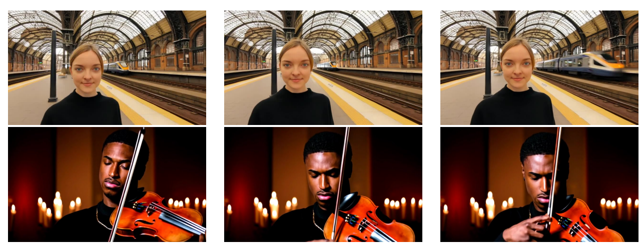


Figure 7. Limitations. Top: the train dynamics is unreasonable, as it moves backward prior to moving forward. Bottom: the person is playing the violin, but the bow does not touch the strings.

Limitations We have observed limitations where injecting ID into could lead to unreasonable dynamics or interactions. Figure 7 illustrates two examples. In the first case, a train moves backward before moving forward. In the second example, the person appears to be playing the violin, but the bow does not touch the strings. The unreasonable dynamics stem from both the base Wan2.1 model and identity conditioning. Identity injection may amplify such effects in certain cases. These issues could potentially be mitigated by adding regularization during training.

Societal Impact Our work advances the technical development of personalized video generation to improve fidelity and controllability. Misuse of such technology could lead to ethical or privacy concerns, and it should not be applied to inappropriate or non-consensual scenarios such as identity manipulation or deceptive content creation.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4, 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[3] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1

[4] 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. 1, 3

[5] Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. *arXiv preprint arXiv:2503.07535*, 2025. 5, 6

[6] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2024. 3

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4

[8] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 5

[9] 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. 4, 1

[10] Yufan Deng, Xun Guo, Yuanyang Yin, Jacob Zhiyuan Fang, Yiding Yang, Yizhi Wang, Shenghai Yuan, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation. *arXiv preprint arXiv:2505.23742*, 2025. 3, 6, 8

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[12] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 3, 4, 6, 8

[13] 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

[14] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 3

[15] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 3

[16] Google DeepMind. Veo: Advanced text-to-video generation. https://deepmind.google/technologies/veo/, 2025. Accessed: 2025-08-30. 1

[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. 3

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[19] 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. 3

[20] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 5

[21] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 4

[22] 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. 4

[23] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3, 4, 6, 8

[24] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025. 3

[25] Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] 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. 1

[28] 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. 1, 3, 4

[29] Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv: 2506.05046*, 2025. 3

[30] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12737–12746, 2025. 3

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[33] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 3, 4, 6, 8

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5, 1

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[36] OpenAI. Sora: A text-to-video diffusion model. `https://openai.com/sora`, 2024. Accessed: 2025-08-30. 1

[37] 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. 1, 3, 4

[38] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023. 3

[39] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients, 2024. 3

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 4

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[43] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[45] 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. 1, 3

[46] J Su, Y Lu, S Pan, A Murtadha, B Wen, and YL Roformer. Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2023. 5

[47] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3, 4

[48] 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, 4, 5

[49] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2589–2598, 2025. 3

[50] Bowen Xue, Qixin Yan, Wenjing Wang, Hao Liu, and Chen Li. Stand-in: A lightweight and plug-and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025. 3, 6, 8

[51] 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. 1, 3

[52] 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, 4

[53] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025. 3, 4

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 4

[55] Yunpeng Zhang, Qiang Wang, Fan Jiang, Yaqi Fan, Mu Xu, and Yonggang Qi. Fantasyid: Face knowledge enhanced id-preserving video generation, 2025. 3

[56] Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. Id-patch: Robust id association for group photo personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[57] Zhenghao Zhang, Zuozhuo Dai, Long Qin, and Weizhi Wang. Effived: Efficient video editing via text-instruction diffusion models. *arXiv preprint arXiv:2403.11568*, 2024. 3

[58] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025. 5