Lynx: Towards High-Fidelity Personalized Video Generation

Supplementary Material

This supplementary consolidates implementation details and additional experiments and evaluations that complement the main paper.

1. Implementation Details

1.1. Training Details

Lynx is built on top of the Wan2.1-14B [47] video diffusion transformer. The backbone remains frozen during training and only the two adapter modules (ID-adapter and Ref-adapter) are optimized.

Training follows a two-stage progressive strategy. We first perform image pretraining by treating each image as a single-frame video, which stabilizes identity learning. The model is then further trained on large-scale video data to learn temporal dynamics and motion patterns.

We use the AdamW [34] optimizer with learning rate 1×10^{-5} and weight decay 0.01. Training is conducted on 128 GPUs with 80GB memory. Following the native resolution style packing strategy described in the main paper, training scale is measured in tokens instead of batch size, with approximately 33,600 tokens processed per GPU per iteration.

1.2. Identity Conditioning

The ID-adapter extracts a 512-dimensional ArcFace [9] embedding from the reference image and converts it into 16 identity tokens using a Perceiver Resampler [1]. These tokens are injected into each transformer block via cross-attention.

The Ref-adapter extracts dense features from the reference image using the pretrained VAE encoder and a frozen copy of the backbone network. The resulting reference tokens are also injected through cross-attention to enhance fine-grained identity details.

1.3. Trainable Parameters and Inference Efficiency

The only additional trainable parameters are the two adapters. They introduce 4.20B parameters, corresponding to a 29.35% increase over the base model.

Method	Latency (s)	Peak per-GPU VRAM (GB)	#H100 GPUs	#Params (B)
SkyReels-A2	410	52.81	1	14.29
Phantom	147	64.19	8	14.29
Lynx (ours)	563	50.73	1	18.49

Table 1. Inference-time statistics.

2. Additional Experimental Results

2.1. Expanded Benchmark to 100 Subjects

To evaluate robustness with respect to benchmark size, we expand the evaluation from 40 to 100 subjects and re-run Lynx. Due to resource constraints, we do not re-run other methods at the expanded scale. As shown in the table below, the resulting metrics remain highly consistent (gap $< 5\%$), indicating that the quantitative results are not sensitive to the number of subjects.

#Sub	facexlib	insightface	in-house	PF	AQ	MN	VQ
40	0.779	0.699	0.781	0.722	0.871	0.837	0.956
100	0.781	0.702	0.781	0.691	0.866	0.825	0.943

Table 2. Benchmark size. PF: prompt following. AQ: aesthetic quality. MN: motion naturalness. VQ: overall video quality.

2.2. Human Evaluation

We conduct a user study with 18 participants. Each participant rates generated videos on a 1 to 5 scale across multiple dimensions. As shown in table, Lynx achieves the highest scores across all criteria.

Methods	FR	PF	AQ	MN	VQ
SkyReels-A2	3.54	2.87	3.12	3.11	3.19
VACE	3.12	3.33	3.29	3.40	3.26
Phantom	3.35	3.41	3.29	3.34	3.31
MAGREF	3.33	3.26	3.17	3.01	3.19
Stand-In	3.47	3.21	3.28	3.18	3.23
Lynx (ours)	**3.83**	**3.57**	**3.54**	**3.41**	**3.55**

Table 3. Human evaluation. FR: face resemblance. PF: prompt following. AQ: aesthetic quality. MN: motion naturalness. VQ: overall video quality.

3. Additional Resources

To facilitate reproducibility and further analysis, we release the following resources:

- **Code.** The inference code is publicly available at: https://github.com/bytedance/lynx.
- **Models.** We release our models on Hugging Face: https://huggingface.co/ByteDance/lynx.
- **Project Page.** Our project page provides additional qualitative results and video comparisons with baseline methods: https://byteaigc.github.io/Lynx.