

MobileI2V: Fast and High-Resolution Image-to-Video on Mobile Devices

Shuai Zhang*, Bao Tang*, Siyuan Yu*, Yueling Zhu, Jingfeng Yao, Ya Zou,
 Shanglin Yuan, Li Yu, Wenyu Liu, Xinggang Wang†
 Huazhong University of Science and Technology

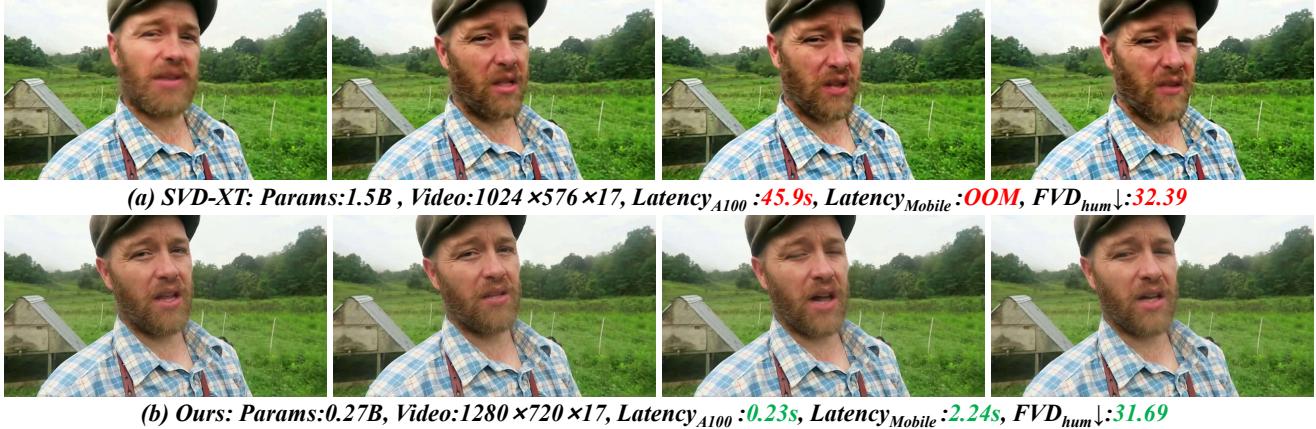


Figure 1. Compared with SVD-XT (1.5B), our $5.55 \times$ smaller MobileI2V (0.27B) achieves similar generation quality, using only 2.24 s on mobile and running 199 \times faster on an A100 GPU.

Abstract

Recently, video generation has witnessed rapid advancements, drawing increasing attention to image-to-video (I2V) synthesis on mobile devices. However, the substantial computational complexity and slow generation speed of diffusion models pose significant challenges for real-time, high-resolution video generation on resource-constrained mobile devices. In this work, we propose MobileI2V, a 270M lightweight diffusion model for real-time image-to-video generation on mobile devices. The core lies in: (1) We analyzed the performance of linear attention modules and softmax attention modules on mobile devices, and proposed a linear hybrid architecture denoiser that balances generation efficiency and quality. (2) We design a time-step distillation strategy that compresses the I2V sampling steps from more than 20 to only two without significant quality loss, resulting in a 10-fold increase in generation speed. (3) We apply mobile-specific attention optimizations that yield 2 \times speed-up for attention operations during on-device inference. MobileI2V enables, for the first time, fast 720p image-to-video generation on mobile de-

vices, with quality comparable to existing models. Under one-step conditions, the generation speed of each frame of 720p video is less than 100 ms. Our code is available at: <https://github.com/hustvl/MobileI2V>.

1. Introduction

With the development of video generation models [1, 2, 22, 55, 56], the demand for image-to-video (I2V) generation on mobile devices is constantly increasing. I2V on mobile devices can significantly enrich system applications, such as live wallpapers. Compared to cloud-based inference, on-device inference offers distinct advantages: it eliminates dependency on network connectivity, significantly cuts data transmission and latency, and better safeguards user privacy.

The core challenge of mobile I2V is that current diffusion models have high computational costs and slow inference speeds, which makes most models take a very long time to infer or cannot infer at all on mobile devices. The main reasons are: (1) Even after VAE compression, high-resolution videos still feed the denoiser a vast number of latent tokens. (2) The denoiser employs a DiT [30] built on softmax attention [37], whose $O(n^2)$ complexity causes computation to explode as the number of tokens grows. (3) Diffusion models require multi-step reasoning when generating videos. These reasons limit the application of video

* Equal contribution.

† Corresponding author (xgwang@hust.edu.cn).

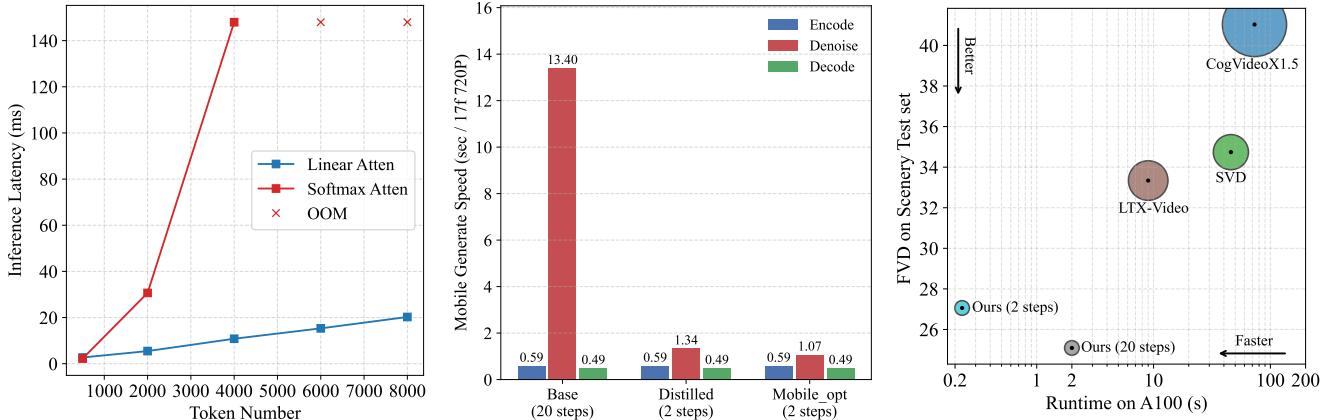


Figure 2. Softmax Attention vs Linear Attention speed in mobile devices (Left). The I2V generation speed of the proposed model (Middle). Comparison results of existing I2V models (Right).

diffusion models, especially in resource-limited mobile scenarios. Therefore, this paper attempts to solve these problems and proposes a video generation model that quickly generates high-resolution videos on mobile devices.

VAE with a high compression rate significantly reduces the number of tokens. LDM [31] proposes for the first time to compress images using VAE and perform the diffusion process in the latent space. DCAE [4] proposes a VAE with 32×32 spatial down-sampling. With the high sampling rates of VAE, SANA [41] achieves fast generation of high-resolution images. LTX-Video [11] further extended DCAE to the video generation domain, proposing a $32 \times 32 \times 8$ spatial-temporal down-sampling rate VAE, which significantly improved the inference speed of T2V/I2V tasks. Therefore, we adopted this high compression ratio VAE and replaced the decoder with a lighter model [60].

Lightweight denoisers with linear complexity can result in faster inference speeds. As shown in Fig. 2, we found that the linear attention [18] has a significant advantage over softmax attention in terms of inference speed on mobile devices. As the length of the sequence increases, the performance gap between linear attention and softmax attention becomes increasingly apparent, especially on mobile devices. Adopting a fully linear attention architecture can significantly enhance the speed of the model, but it may also reduce its accuracy. Therefore, we designed a hybrid architecture that balances the speed and accuracy of the model.

Time-step distillation significantly reduces the number of inference steps. Currently, the inference of common video generation models requires around 20 to 50 sampling steps. After time-step distillation, the inference steps can be reduced to only 1–4 steps. Existing approaches such as LADD [33], LCM [25], DMD [51] and SF-V [53] have been applied to text-to-image and text-to-video generation. We further employed time-step distillation to reduce the I2V model’s inference steps from 20 to 2.

To this end, we propose MobileI2V, an I2V model that

can perform fast inference on mobile devices. We use linear hybrid attention to construct a denoiser with only 270M parameters. As shown in Fig. 2, after optimization, the proposed model can generate 17 frames of 720p video within 2 seconds on the iPhone 16 PM. Our contributions are as follows:

1. We propose a hybrid architecture diffusion model with only 270M parameters for 720P mobile I2V tasks. We combined the advantages of linear attention and softmax attention to balance the speed and quality of device-side generation, and for the first time explored the efficiency performance of linear attention in device-side video generation.

2. We propose a composite timestep distillation scheme for lightweight I2V models, which reduces the number of inference steps from 20 to 1–2 while maintaining comparable generative quality. This enables over 10 \times acceleration on mobile devices.

3. On mobile devices, we introduce three inference optimizations for linear attention: 4D channels-first layout and operator lowering, head tiling, and reduced data movement. This improves the inference speed of linear attention and softmax attention mechanisms by 2 \times on mobile devices.

4. To the best of our knowledge, we have achieved high-resolution (720P) I2V on a mobile device for the first time. Under one-step inference conditions, the generation time for a single frame at 720P is less than 100 ms.

2. Related Work

Video Diffusion Model. The development of video generation models is advancing rapidly [19, 38, 47]. DiT [30] first introduced the Vision Transformer into diffusion models, demonstrating the scalability of Transformers for image/video generation tasks. Latte [28] experimented with various Transformer variants for video generation, and subsequent experiments validated the optimal Transformer variant. Opensora [55] and Opensora-Plan [22], as outstanding open-source models within the community, have

garnered widespread attention. CogVideo [15] fine-tunes pre-trained text-to-image generation models, avoiding expensive pre-training. There are currently some acceleration methods [48, 49] for DiT, but generating videos on mobile devices still requires some powerful lightweight designs.

Linear Generative Models. The linear attention mechanism [13, 14] has been widely attempted to be applied to generative tasks, such as image super-resolution [21], image generation [41, 59], and video generation [6, 9, 16], due to its $O(n)$ computational complexity. Arflow [17] applies a hybrid linear model to an autoregressive flow model to improve image generation efficiency. SANA-Video [6] significantly improves video generation efficiency by utilizing Block Linear Diffusion. In this paper, we will explore the enormous potential of linear attention mechanisms for mobile generation.

Mobile-Side Generative Models. MobileDiffusion [54] can generate high-quality 512×512 images within 0.2 seconds on iPhone. MobileVD [44] generating latents for a $14 \times 512 \times 256$ px clip in 1.7 seconds on a Xiaomi-14 Pro. SnapGen [5] demonstrates the generation of 1024^2 px images on a mobile device around 1.4 seconds. SnapGen-V [40] has only 0.6M parameters and can generate a 5-second video on an iPhone 16 PM within 5 seconds. These methods all employ the U-Net architecture, and the resolution of generative models on mobile devices is relatively low. In this paper, we have achieved for the first time the implementation of a 720P I2V task on a mobile device.

Time-step Reduction. Time-step distillation compresses a teacher model’s multi-step trajectory into a student model that needs only 1–4 steps. Representative techniques include: Simple direct distillation [7, 32] that regress the T-step chain into a one-step mapping; Consistency-based methods (e.g., CM [35], LCM [25], sCM [24]) that enforce self-consistency along the probability-flow ODE; Adversarial distillation (e.g., ADD [34], LADD [33]) that uses a discriminator to align the one-step output with the teacher distribution in pixel or latent space; VSD-based variants (e.g., DMD [51], DMD2 [50], SiD [57], SIM [26]) that match distributions via variational score distillation while suppressing mode collapse. Most of these methods are used for text-to-image tasks, and recent work such as SF-V [53] is transferring them to video generation by adding spatio-temporal discriminators to maintain frame-wise consistency.

3. Preliminaries

Flow-based Diffusion Models. The input video data is $V \in \mathbb{R}^{H \times W \times 3 \times T}$, where H, W, and T represent the length, width, and number of frames of the video, respectively. Compress to latent $L \in \mathbb{R}^{\hat{H} \times \hat{W} \times 3 \times \hat{T}}$ via VAE, where $\hat{H} = H/32, \hat{W} = W/32$ and $\hat{T} = T/8$. The noise

$N \in \mathbb{R}^{\hat{H} \times \hat{W} \times 3 \times \hat{T}}$ is denoised by a denoiser and then decoded into video by a decoder. We use the SD3 [8] optimizer to train and infer our model. We define a forward process, corresponding to a probability path between p_0 and $p_1 = \mathcal{N}(0, 1)$, as

$$z_t = a_t x_0 + b_t \epsilon, \quad \text{where } \epsilon \in \mathcal{N}(0, I). \quad (1)$$

The objective function of the model is:

$$L(x_0) = \mathbb{E}_{t, \epsilon} [w_t \lambda'_t ||\epsilon(z_t, t) - \epsilon||^2], \quad (2)$$

where $w_t = -\frac{1}{2} \lambda'_t b_t^2$, $\lambda_t = \log \frac{a_t^2}{b_t^2}$ and $\lambda'_t = 2(\frac{a'_t}{a_t} - \frac{b'_t}{b_t})$.

Rectified Flows [23] defines the forward process as a straight line path between the data distribution and the standard normal distribution. So for it, $a_t = 1 - t$, $b_t = t$.

Softmax Attention and Linear Attention. Given tokens of length N , the self attention calculation method is as follows [12]:

$$O_i = \sum_{j=1}^N \frac{\text{Sim}(Q_i, K_j)}{\sum_{j=1}^N \text{Sim}(Q_i, K_j)} V_j, \quad (3)$$

where $Q = xW_Q$, $K = xW_K$, $V = xW_V$. $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$ are projection matrices and $\text{Sim}(\cdot)$ denotes the similarity function. For Softmax attention, $\text{Sim}(Q, K) = \exp(QK^T / \sqrt{d})$. For Linear Attention, we refer to the approach used in SANA [41], $\text{Sim}(Q, K) = \text{ReLU}(Q_i)\text{ReLU}(K_j)^T$.

4. Method

4.1. Model Architecture

Our framework diagram is shown in Fig. 3. We design a hybrid linear DiT as a noise reducer and use a $32 \times 32 \times 8$ high compression ratio VAE to reduce the number of tokens.

Hybrid Linaer DiT. We follow SANA [41]’s architecture. We used 16 layers of DiT, with only 0.27M parameters. Due to the $O(n^2)$ complexity of the ordinary Softmax Attention, the complexity increases quadratically when processing high resolution. As shown in Fig. 2, the performance of linear attention is significantly faster than the traditional vanilla attention on mobile devices. Therefore, in this paper, we replaced part of the softmax attention with linear attention. We drew inspiration from the hybrid architecture in MiniMax-01 [20], where a softmax attention follows every seven linear attentions. The experiment found that introducing two layers of softmax attention can significantly compensate for the accuracy loss caused by linear attention. Therefore, we adopted a hybrid architecture to balance the model’s performance and accuracy. We added positional encoding to the model based on Allegro [58].

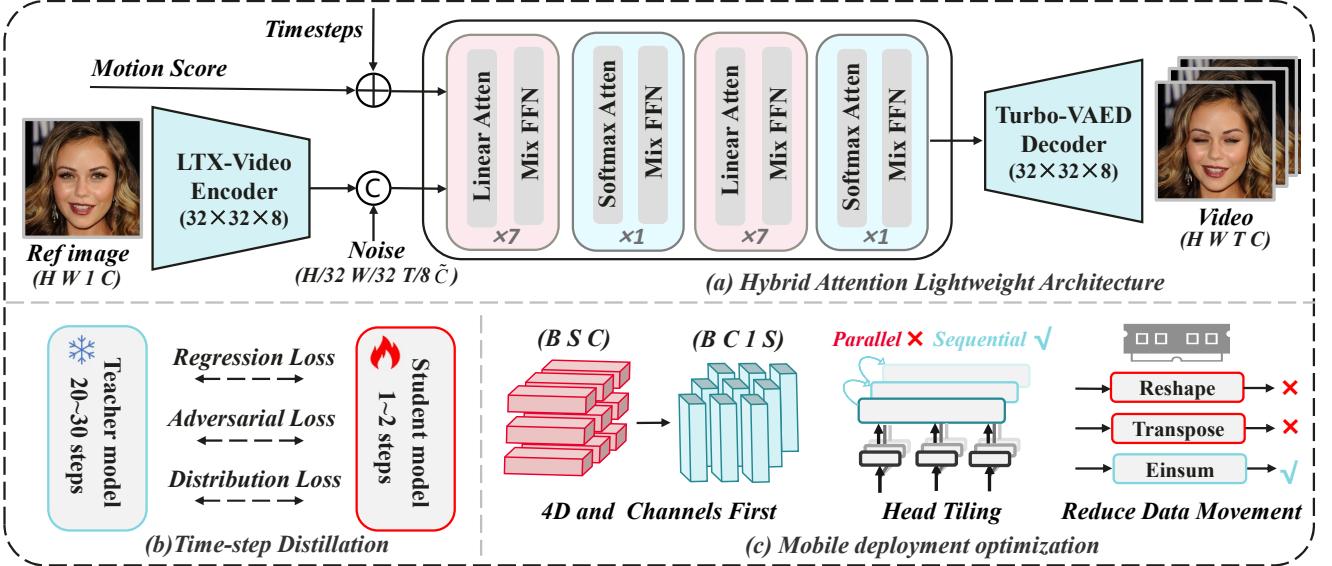


Figure 3. Our proposed MobileI2V framework employs a hybrid attention framework to support fast inference, utilizes time-step distillation to reduce the number of inference steps, and finally optimizes the model on mobile devices to further accelerate inference.

High Compression Ratio VAE. We use the LTX-Video [11] VAE encoder and the Turbo-VAED [60] decoder, with a downsampling factor of $32 \times 32 \times 8$. The number of channels is 128. For a 720p video, the latent size after VAE is $40 \times 23 \times 3$. Turbo-VAED distills the VAE of LTX-Video, reducing the size of the model decoder and significantly improving decoding speed. The decoding speed of the distilled VAE is $3 \times$ that of the LTX-Video VAE.

4.2. Design of Image to Video

The first frame of the video is used as the reference image in the I2V task. As shown in Fig. 3, the reference image, after VAE encoding, is used to replace the first frame in the noisy latent. We refer to the design of LTX-Video [11], setting time t as independent for each token and setting the reference frame’s t to 0. We extracted the optical flow scores from the training data and used them as conditions to be fed into denoiser, thereby controlling the motion magnitude of the objects. Conditioned on a motion score, the I2V model converges faster and allows users to intuitively control the generated video’s motion intensity.

4.3. Time-step Distillation

As shown in Fig. 4, we explore a timestep distillation strategy that combines multiple techniques, enabling video generation within only one or two inference steps. Specifically, inspired by successful practices in text-to-image (T2I) distillation [3, 33, 51], we employ three complementary losses: regression loss, adversarial loss, and distribution matching loss. To this end, the distillation process jointly maintains a Student Model G_S , a Teacher Model G_T , a Fake Score Model F , and a Discriminator D . Among them, G_S , G_T ,

and F are initialized from pretrained weights, while D is randomly initialized. During training, the parameters of G_S , F , and D (denoted as θ_S , θ_F , and ψ , respectively) are optimized, whereas the teacher G_T remains frozen. The three losses are designed as follows:

Regression Loss. A clean video x_0 is perturbed to a noisy state x_t at a given noise level. Both the student model G_S (single-step generation) and the teacher model G_T (multi-step generation) are applied to x_t . The mean squared error (MSE) between their predictions is used as the regression loss:

$$\mathcal{L}_{\text{reg}}(\theta_S) = \mathbb{E}_{x_0, t, \epsilon} \left[\|G_S(x_t, t) - \hat{x}_0^{\text{Teacher}}\|_2^2 \right]. \quad (4)$$

Here, $\hat{x}_0^{\text{Teacher}}$ denotes G_T ’s multi-step generated video starting from x_t .

Adversarial Loss. The predicted clean video \hat{x}_0 from G_S and the ground-truth clean video x_0 are treated as adversarial pairs. After adding noise of a certain magnitude, they are passed into the frozen teacher G_T to extract features, which are subsequently fed into the discriminator D . The adversarial loss is then computed as:

$$\begin{aligned} \mathcal{L}_{\text{adv}}^G(\theta_S) &= -\mathbb{E}_{x_0, t, \epsilon} \left[\sum_k D_{\psi, k} \left(G_T^{(k)}(\hat{x}_t, t) \right) \right], \\ \mathcal{L}_{\text{adv}}^D(\psi) &= \mathbb{E}_{x_0, t, \epsilon} \left[\sum_k \text{ReLU} \left(1 - D_{\psi, k} \left(G_T^{(k)}(x_t, t) \right) \right) \right. \\ &\quad \left. + \sum_k \text{ReLU} \left(1 + D_{\psi, k} \left(G_T^{(k)}(\hat{x}_t, t) \right) \right) \right]. \end{aligned} \quad (5)$$

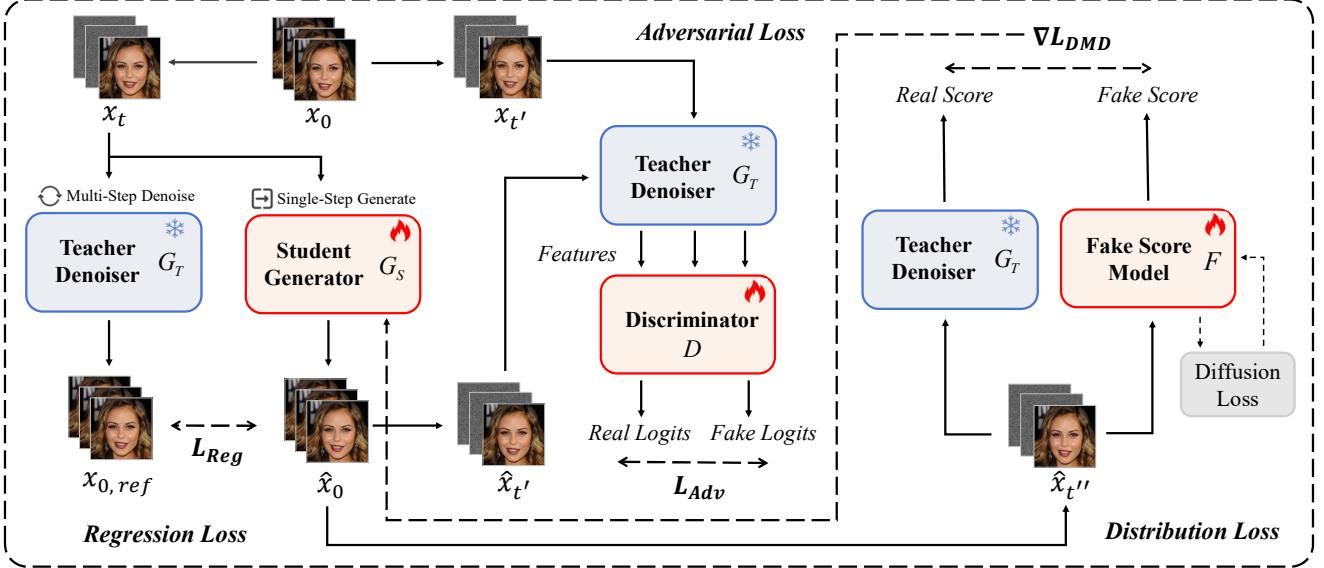


Figure 4. The proposed i2v time-step distillation framework. It includes three parts: regression loss, adversarial loss, and distribution loss.

Here, x_t and \hat{x}_t are obtained by adding noise to x_0 and \hat{x}_0 , respectively. $G_T^{(k)}(\cdot)$ denotes G_T 's output at the k -th feature branch.

Distribution Matching Loss. The predicted clean video \hat{x}_0 from G_S is first perturbed by noise. The resulting noisy sample is evaluated by both G_T and the fake score model F , producing a real score and a fake score, respectively. Following the formulation in DMD [51], we compute the distribution matching loss as:

$$\begin{aligned} \nabla \mathcal{L}_{DM}(\theta_S) &= \mathbb{E}_t \left[\nabla_{\theta_S} \text{KL}(p_{fake,t} \| p_{real,t}) \right] \\ &= -\mathbb{E}_t \left[\int (s_r(\hat{x}_t, t) - s_f(\hat{x}_t, t)) \frac{dG_S(x_t, t)}{d\theta_S} dx_t \right], \end{aligned} \quad (6)$$

$$\mathcal{L}_{DM}(\theta_F) = \mathbb{E}_{\hat{x}_0, t, \epsilon} \left[\|F(\hat{x}_t, t) - \hat{x}_0\|_2^2 \right]. \quad (7)$$

Here, s_r and s_f denote the scores computed by G_T and F , respectively. The noisy sample \hat{x}_t is obtained by adding noise to the \hat{x}_0 .

4.4. iPhone Mobile Model Deployment

Following Apple's guidelines in *Deploying Transformers on the Apple Neural Engine (ANE)*, we optimize all attention computations (softmax attention and linear attention) by aligning data layout and operator mappings with the ANE compiler's preferences. We apply three families of optimizations:

4D and Channels-First. We represent intermediate tensors as $(B, C, 1, S)$ and keep standard Linear projections. After projection we reshape with lightweight transpose so Q/K/V live in a 4D channels-first layout during attention. The ANE keeps the last axis unpacked: when sequence length occupies that axis, accesses along it are contiguous,

improving prefetch efficiency and minimizing the impact of padding. In contrast, making a channel dimension the last axis inflates buffers and hurts L2 residency.

Head Tiling. When computing on a GPU, we compute multi-head attention in parallel. However, when computing on a mobile device, we compute each head individually and serially. Multi-head attention is explicitly split into per-head Q , K , and V , computed on smaller tiles using a per-head function list, improving L2 locality and post-compilation multicore utilization.

Reduced Data Movement. Along the attention path, redundant reshape/transpose operations are eliminated, retaining only a single necessary transpose on K . The core products are implemented using an `einsum` form that maps directly to batched matrix multiplies on hardware, substantially lowering bandwidth pressure.

As shown in Fig 5, these optimizations speed up attention computation by more than $2\times$ for both linear and softmax attention modules. This further improves the inference speed of the model.

5. Experiment

5.1. Data Process and Training

Data Process. There are many open source video datasets, such as Openvid [29], VFHQ [42] and Celebav-text [52]. In terms of data, we collected face videos and landscape videos with resolutions of 512 and 720P. The 512 resolution data is approximately 160K (landscape 115K + face 45K), and the 720P data is approximately 170K (landscape 93K + face 88K). We filtered the data by aesthetic score, optical flow score, and reconstruction score to select a batch of high-quality data. We used DOVER [39] to evaluate the aesthetic and clarity scores of the videos and filtered

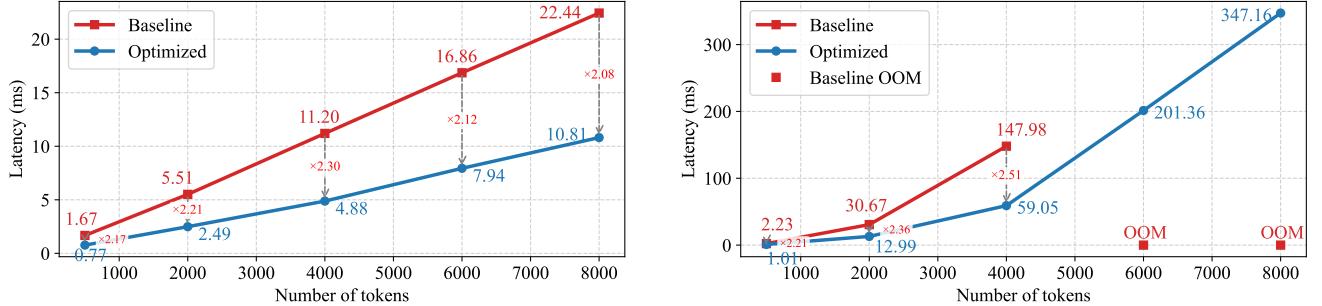


Figure 5. On-device latency on iPhone 16 Pro versus sequence length for linear attention (left) and softmax attention (right).

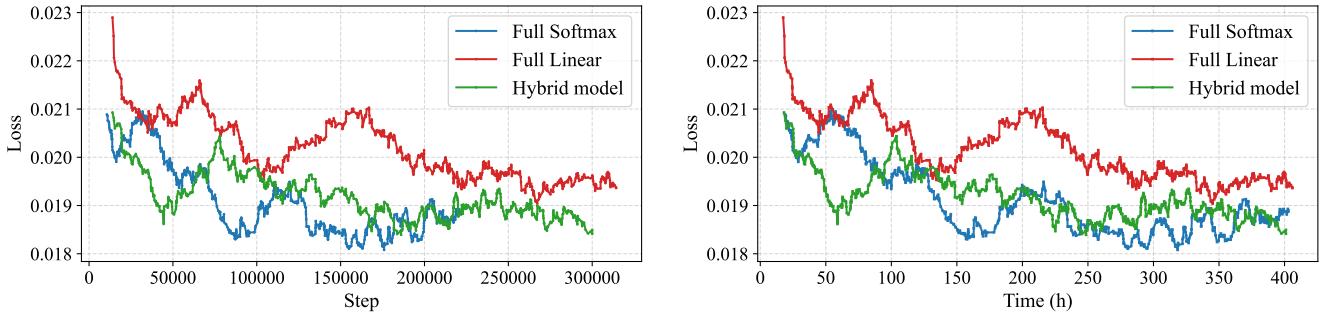


Figure 6. Loss curves for training steps (left) and training time (right) for different attention models.

Model	Type	Steps	Params	Resolution	Latency _{A100}	Latency _{Mobile}	FVD _{hum↓}	FVD _{scen↓}
DynamiCrafter [43]	U-Net	30	1.1B	1024×576	57.1 s	OOM	67.98	39.19
CogVideoX1.5 [46]	DiT	30	5.0B	720×480	73.1 s	OOM	60.83	41.03
SVD-XT [1]	U-Net	30	1.5B	1024×576	45.9 s	OOM	32.39	34.74
LTX-Video [11]	DiT	30	1.9B	1280×720	9.00 s	OOM	36.04	33.34
Ours	DiT	30	0.27B	1280×720	2.00 s	20.13 s	26.99	25.09
Ours (Distilled)	DiT	2	0.27B	1280×720	0.23 s	2.24 s	31.69	27.06

Table 1. Speed and quality comparison of different models performing 17-frame I2V tasks.

out data below 70. We refer to the optical flow evaluation in opensora [55] and obtain videos with scores greater than 0.1 and less than 6.5. The high compression rate VAE performs poorly in reconstructing some data, so we selected data with a reconstruction PSNR greater than 32 by calculating the PSNR before and after compression.

Training. Training the model on 24 Nvidia V100 32G GPUs would take approximately one week. We use the CAME [27] optimizer with a learning rate of $1e - 4$. We first train the model using low-resolution videos, and then we train it using high-resolution videos. This helps the model converge faster.

5.2. Evaluation

We adopt FVD [36] as the metric for video generation quality evaluation. We collected 1,348 facial data samples and 1,739 aerial landscape data samples to serve as the test set.

We compared the generation speeds of different models on the A100 and iPhone, as well as the generation results on two datasets, as shown in Tab. 1. The visualization re-

sults are shown in Fig. 7. Existing I2V models, such as CogVideoX1.5, DynamiCrafter, SVD, and LTX-video, cannot run on mobile devices due to their huge number of parameters. It can be observed that our method is significantly faster than existing I2V methods, is deployable on mobile devices, and achieves comparable performance to current I2V methods in both facial and landscape scenarios. After distillation, the generation speed of the two-step model is significantly improved, and there is no obvious decrease in the indicators.

6. Ablation Experiment

Hybrid Architecture. In order to verify the effectiveness of the proposed hybrid architecture model, we compared the hybrid architecture model with the full linear attention and full softmax attention models. As can be seen from the Fig. 6 and Tab. 2, the loss reduction curve of the hybrid model with only two attention layers is close to that of the full softmax model. The full softmax model demonstrates strong representational capabilities, as evidenced by

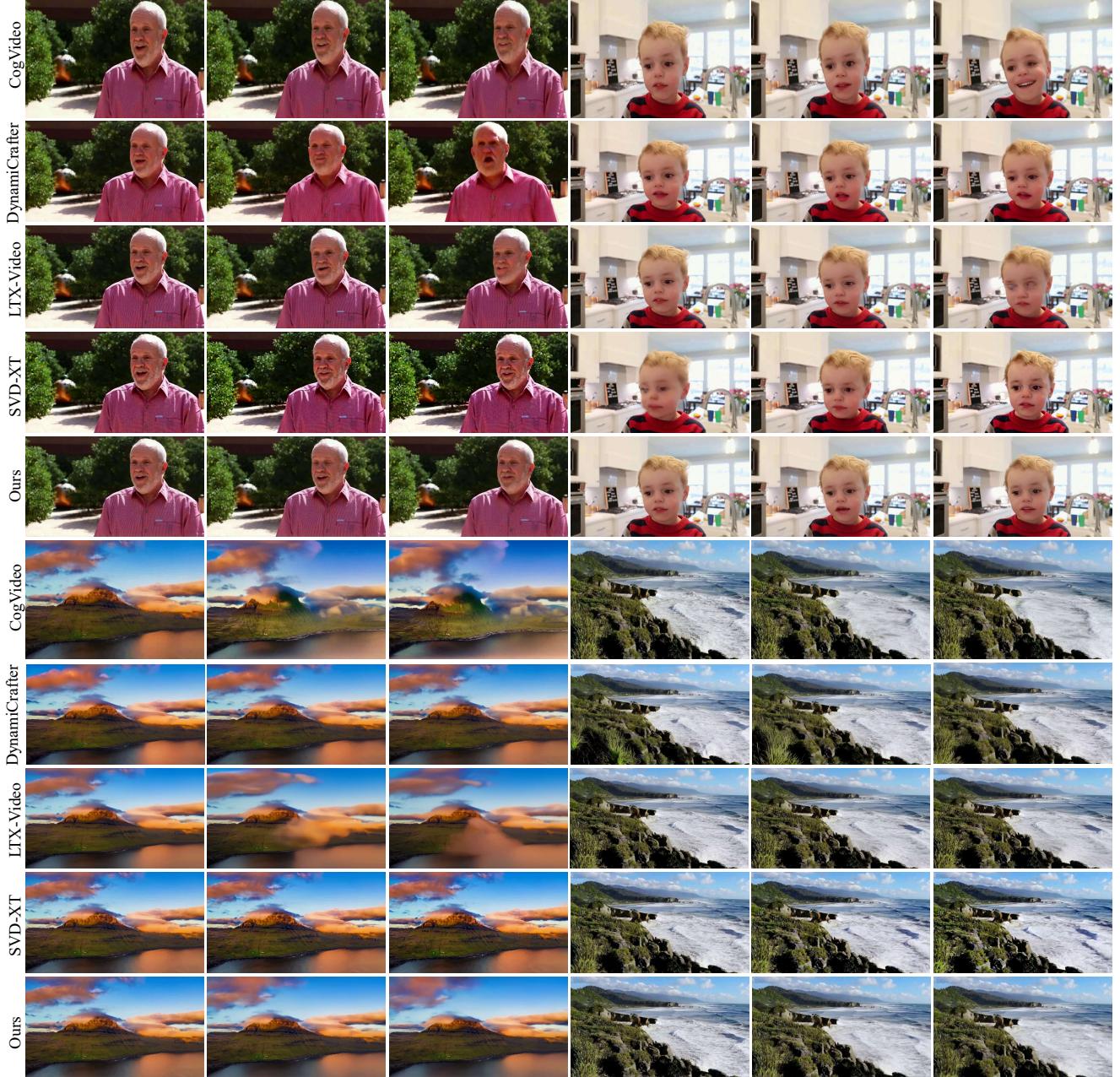


Figure 7. Visualization of 30-step Image-to-Video inference results across different models.

its best performance in the FVD_{scen} metric. However, its speed is much slower than the full linear model and the hybrid model. The hybrid model is slightly slower than the full linear model in terms of speed, but outperforms the full linear model in terms of metrics.

Model	$FVD_{hum} \downarrow$	$FVD_{scen} \downarrow$	$Speed_{v100} \downarrow$	$Speed_{mobile} \downarrow$
Linear	27.47	25.75	3.72s	13.74s
Softmax	28.89	23.66	7.88s	22.97s
Hybrid	26.99	25.09	4.64s	14.85s

Table 2. Ablation study on the effect of hybrid architecture.

Time-step Distillation. To validate the effectiveness of timestep distillation, we compare two-step inference outputs of the model before and after distillation, as illustrated in Fig. 8. In videos generated with only 2 inference steps, the non-distilled model exhibits noticeable blurring in later frames, while the distilled model effectively eliminates this blurring and produces outputs that closely approximate those obtained with 20 steps. As shown in Tab. 3, the metrics for the model’s one-step generation improve markedly after distillation.

Distillation Loss Function. To evaluate the contributions



Figure 8. Visualization results of time-step distillation ablation experiment. Original_20: 20-step results of the undistilled model. Distilled_2: 2-step results of the model after distillation. The first frame is the reference image input to the model.

of each loss in the time-step distillation, we conducted ablation studies, with the results summarized in Tab. 3. The key findings are as follows: (1) Enabling QK normalization in all attention layers and applying timestep normalization significantly improves the visual quality of single-step generation. (2) Adding the adversarial loss on top of the regression loss further enhances the metrics, but the performance still falls short of the full scheme. We attribute this to the limited representation capacity of the lightweight I2V backbone, which restricts the discriminator from providing sufficiently strong supervision. (3) Incorporating only the distribution matching loss, in contrast, degrades performance. Visualization shows that although this loss increases motion magnitude, it introduces noticeable blurriness, thereby impairing overall perceptual quality. (4) Jointly training with all three losses achieves the best results, demonstrating complementary benefits. The single-step generation output exhibits substantially improved sharpness and temporal consistency compared to any individual or pairwise combination of losses.

Model	NFE	Norm	L_{reg}	L_{adv}	L_{dm}	FVD _{h↓}	FVD _{s↓}
Teacher	1	-	-	-	-	144.85	110.13
Teacher	1	✓	-	-	-	122.05	104.95
Distilled	1	✓	✓	-	-	53.08	40.20
Distilled	1	✓	✓	✓	-	51.93	42.32
Distilled	1	✓	✓	-	✓	115.45	56.07
Distilled	1	✓	✓	✓	✓	43.62	32.68
Distilled	2	✓	✓	✓	✓	31.69	27.06

Table 3. Ablation study on time-step distillation loss function.

Mobile Optimization. We ablate three implementation optimizations—4DC (4D channels-first), RDM (reduced data movement), and HT (head tiling)—on Core ML (iPhone 16 Pro). The ablation study uses linear attention and processes 2,760 tokens. The results are in Fig. 9. Applying all three optimizations together reduces the latency from 7.69 ms to 3.38 ms, yielding a 2.27× speed-up. Among the three optimizations, RDM yields the greatest

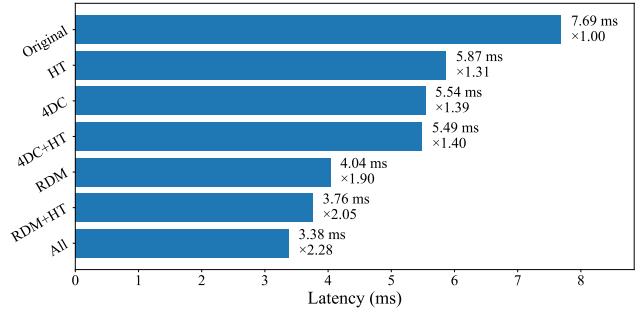


Figure 9. Ablation study of linear attention optimizations on Core ML (iPhone 16 Pro).

improvement: latency drops to 4.04 ms, a 1.9× speed-up. Applying the 4DC and HT optimizations on top of RDM shaves off another 0.66 ms. The experiment demonstrates the effectiveness of all three optimizations and shows that they can be stacked for maximum gain.

7. Limitations and Future Work

Although the proposed model demonstrates a clear speed advantage over existing ones, it still has several limitations. First, due to limited training resources and data availability, the proposed I2V model is not universal. Subsequent large-scale data training is needed to make it more general. Furthermore, due to the high compression rate of VAE, some areas of faces and landscapes may be blurry. Increasing the motion amplitude of the video exacerbates this problem. To address these issues, it is crucial to research higher-quality high-compression VAEs. Furthermore, further research on stronger linear attention models [10, 45] and their downstream applications will be a promising research direction.

8. Conclusion

In this paper, we propose a lightweight diffusion model for fast I2V on mobile devices. We have discovered that a hybrid architecture combining linear and softmax layers can effectively balance generation speed and quality, proving

highly efficient on mobile platforms. By employing high compression rates, we significantly reduce the length of tokens and utilize time-step distillation to decrease the number of sampling steps to 1-2, thereby substantially enhancing video generation speed. In addition, we also proposed three optimization methods for the linear attention on the mobile side. Under the condition of one-step inference, the fastest mobile video generation speed can reach 96ms per frame. MobileI2V enables fast, high-resolution image-to-video generation on mobile devices, poised to power a broad range of future on-device applications.

References

- [1] 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](#) [6](#)
- [2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3:1, 2024. [1](#)
- [3] Clément Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 15686–15695. AAAI Press, 2025. [4](#)
- [4] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [2](#)
- [5] Jierun Chen, Dongting Hu, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7997–8008, 2025. [3](#)
- [6] Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025. [3](#)
- [7] Geoffrey Cideron, Andrea Agostinelli, Johan Ferret, Sertan Girgin, Romuald Elie, Olivier Bachem, Sarah Perrin, and Alexandre Ramé. Diversity-rewarded CFG distillation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [3](#)
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [3](#)
- [9] Yu Gao, Jiancheng Huang, Xiaopeng Sun, Zequn Jie, Yujie Zhong, and Lin Ma. Matten: Video generation with mamba-attention. *arXiv preprint arXiv:2405.03025*, 2024. [3](#)
- [10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. [8](#)
- [11] 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](#), [4](#), [6](#)
- [12] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 5938–5948. IEEE, 2023. [3](#)
- [13] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5961–5971, 2023. [3](#)
- [14] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *European conference on computer vision*, pages 124–140. Springer, 2024. [3](#)
- [15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. [3](#)
- [16] Yushi Huang, Xingtong Ge, Ruihao Gong, Chengtao Lv, and Jun Zhang. Linvideo: A post-training framework towards o (n) attention in efficient video generation. *arXiv preprint arXiv:2510.08318*, 2025. [3](#)
- [17] Mude Hui, Rui-Jie Zhu, Songlin Yang, Yu Zhang, Zirui Wang, Yuyin Zhou, Jason Eshraghian, and Cihang Xie. Arflow: Autgressive flow with hybrid linear attention. *arXiv preprint arXiv:2501.16085*, 2025. [3](#)
- [18] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 5156–5165. PMLR, 2020. [2](#)
- [19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xincheng Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng,

- Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zun-nan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanyvideo: A systematic framework for large video generative models, 2025. 2
- [20] Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025. 3
- [21] Xiaohui Li, Shaobin Zhuang, Shuo Cao, Yang Yang, Yuan-dong Pu, Qi Qin, Siqi Luo, Bin Fu, and Yihao Liu. Lin-earsr: Unlocking linear attention for stable and efficient image super-resolution, 2025. 3
- [22] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Luhuan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model, 2024. 1, 2
- [23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3
- [24] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 3
- [25] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *CoRR*, abs/2310.04378, 2023. 2, 3
- [26] Weijian Luo, Zemin Huang, Zhengyang Geng, J. Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 3
- [27] Yang Luo, Xiaozhe Ren, Zangwei Zheng, Zhuo Jiang, Xin Jiang, and Yang You. CAME: confidence-guided adaptive memory efficient optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4442–4453. Association for Computational Linguistics, 2023. 6
- [28] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025. 2
- [29] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhen-heng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 5
- [30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. 1, 2
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 2
- [32] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3
- [33] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 106:1–106:11. ACM, 2024. 2, 3, 4
- [34] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, pages 87–103. Springer, 2024. 3
- [35] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 32211–32252. PMLR, 2023. 3
- [36] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. 6
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1
- [38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2
- [39] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20087–20097. IEEE, 2023. 5
- [40] Yushu Wu, Zhixing Zhang, Yanyu Li, Yanwu Xu, Anil Kag, Yang Sui, Huseyin Coskun, Ke Ma, Aleksei Lebedev, Ju Hu,

- et al. Snapgen-v: Generating a five-second video within five seconds on a mobile device. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2479–2490, 2025. 3
- [41] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2, 3
- [42] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 656–665. IEEE, 2022. 5
- [43] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 6
- [44] Haitam Ben Yahia, Denis Korzhenkov, Ioannis Lelekas, Amir Ghodrati, and Amirhossein Habibian. Mobile video diffusion. *arXiv preprint arXiv:2412.07583*, 2024. 3
- [45] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 8
- [46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*. 6
- [47] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihang Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2
- [48] Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in Neural Information Processing Systems*, 37:56166–56189, 2024. 3
- [49] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 3
- [50] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédéric Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. 2024. 3
- [51] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédéric Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6613–6623. IEEE, 2024. 2, 3, 4, 5
- [52] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-text: A large-scale facial text-video dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14805–14814. IEEE, 2023. 5
- [53] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris N. Metaxas, Sergey Tulyakov, and Jian Ren. SF-V: single forward video generation model. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. 2024. 2, 3
- [54] Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXII*, pages 225–242. Springer, 2024. 3
- [55] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 2, 6
- [56] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. 1
- [57] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 3
- [58] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model, 2024. 3
- [59] Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, and Xinggang Wang. Dig: Scalable and efficient diffusion models with gated linear attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7664–7674, 2025. 3
- [60] Ya Zou, Jingfeng Yao, Siyuan Yu, Shuai Zhang, Wenyu Liu, and Xinggang Wang. Turbo-vaed: Fast and stable transfer of video-vaes to mobile devices, 2025. 2, 4

MobileI2V: Fast and High-Resolution Image-to-Video on Mobile Devices

Supplementary Material

In this appendix, we provide more experimental results in Section 9 and mobile test results in Section 10. Section 11 provides the limitations of the paper and future work.

9. More Experimental Results

9.1. VbenchI2V Results

We tested the VbenchI2V indicators of different models, and the results are shown in Tab. 4. We offer different model versions, including those before and after distillation. It can be observed that adjusting the motion score significantly affects the dynamic degree index, which in turn influences the overall score. For example, when inferring a landscape video, increasing the motion score from 2 to 5 causes the dynamic degree to change from 0.157 to 0.495. The overall indicator improves, while other indicators decrease slightly. Overall, the proposed model can achieve comparable performance metrics to existing models such as DynamiCraft, CogVideoX1.5, SVD, and LTX-Video. We provide further visual comparison results, as shown in Fig. 15 and Fig. 16. For methods that require prompts, we provide standardized prompts, such as “Driving people in the image to move, such as shaking their heads, talking or smiling.”, “Move the image perspective and convert it into a video.”

9.2. More Resolution Results

We trained a model with a 960×960 resolution version on the facial video dataset. The visualization results are shown in Fig. 11. From the figure, it can be observed that at this resolution, the model is capable of effectively driving facial expressions, enabling basic actions such as blinking and opening the mouth. Compared with 720P data, the 960-resolution training video content is cleaner and has better generation effect.

9.3. Ablation Experiment Supplement

In order to verify the effectiveness of the proposed hybrid attention model, we provide the training loss curves of different types of attention mechanisms, as shown in the Fig. 10. As can be seen from the figure, with the same number of training steps and training time, both the full softmax model and the hybrid model significantly outperform the full linear model. At the same training time, the full softmax model is significantly slower than both the full linear model and the hybrid model. This figure demonstrates that the our proposed hybrid model achieves superior results in both speed and performance.

10. Mobile Results

Model conversion and deployment. We convert PyTorch models to Core ML for on-device inference on the iPhone 16 Pro. Specifically, we first export the model to TorchScript and then use `coremltools` to generate a Core ML model, which we package in the `.mlpackage` format. All automatic optimization passes provided by `coremltools` are enabled (e.g., fusing `batch_norm` into `conv` or `conv_transpose`). During inference, FP16 is used for both activations and weights by default, while numerically sensitive operations—such as selected RMSNorm layers—are kept in FP32. To ensure functional equivalence before and after conversion, we compare model outputs on identical inputs and report the peak signal-to-noise ratio (PSNR) between the PyTorch and Core ML results. All benchmarks are obtained via the Core ML inference APIs on-device. We tested the speed of the model on mobile devices, as shown in Fig. 12.

Mobile speed. We tested the speed of the model on mobile devices, as shown in Fig. 12. To generate a $1280 \times 720 \times 17$ frame video, the inference times for the VAE encoder, decoder, and denoiser on a mobile devices are 337ms, 503ms, and 412ms, respectively. In actual I2V end-to-end testing, the module’s time will be approximately 10% slower compared to when it is tested individually.

Mobile UI. We designed the mobile image to video UI interface to facilitate the effect display, as shown in Fig. 13. Using this application, you can read images from your phone and then convert them into videos. The entire I2V process is carried out on the mobile devicce.

11. Limitations and Future Work

Limitations. Since we use a $32 \times 32 \times 8$ high compression rate VAE, blurring may occur in complex dynamic scenes such as faces. As shown in the Fig. 14, complex dynamic areas such as the mouth and eyes of a person’s face are prone to blurring. In addition, due to limited data resources, our model is mainly trained on faces and aerial scenery, resulting in poor performance on other scenes.

Future work. From the perspective of model architecture, a high compression VAE is crucial for accelerating generation. Currently, the capabilities of a $32 \times 32 \times 8$ compression VAE need to be improved. In addition, in this paper we have seen the potential of linear architecture on mobile devices, and adopting a more powerful linear architecture may be a means to improve the effect.

Model	SC	BC	MS	AQ	IQ	DD	Q_{score}	$i2v_{sub}$	$i2v_{back}$	$i2v_{score}$	T_{score}
DynamiCrafter	0.985	0.979	0.988	0.603	0.695	0.172	0.783	0.987	0.987	0.937	0.860
CogVideoX1.5	0.964	0.953	0.989	0.559	0.674	0.370	0.779	0.982	0.988	0.935	0.857
SVD	0.986	0.975	0.992	0.557	0.699	0.136	0.774	0.983	0.988	0.935	0.855
LTX-Video	0.983	0.979	0.994	0.542	0.709	0.509	0.808	0.985	0.988	0.936	0.872
Ours(U.2.5)	0.984	0.971	0.993	0.550	0.710	0.264	0.785	0.989	0.989	0.940	0.862
Ours(D.2.2)	0.989	0.982	0.994	0.552	0.703	0.157	0.779	0.986	0.985	0.935	0.857
Ours(D.2.5)	0.984	0.978	0.993	0.554	0.695	0.495	0.808	0.984	0.983	0.933	0.869

Table 4. Comparison of various models generate 17 frame video. (SC: subject consistency , BC: background consistency, MS: motion smoothness, AQ: aesthetic quality, IQ: imaging quality, DD: dynamic degree, Q_{score} : Quality Score, $i2v_{sub}$: i2v subject, $i2v_{back}$: i2v background, $i2v_{score}$: I2V Score, T_{score} : Total Score. U: 30-step results without time-step distillation. D: two-step results after time-step distillation. 2.2/2.5: motion score for human data and scenery data.)

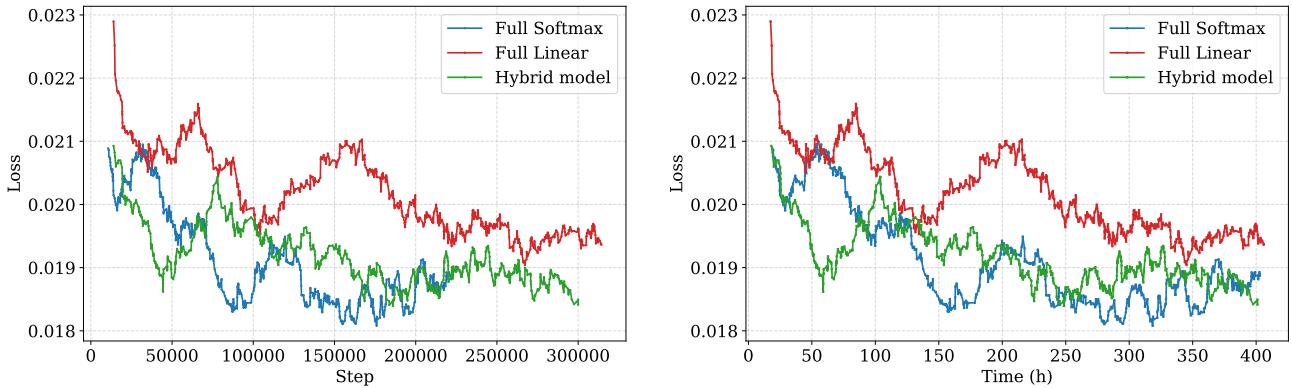


Figure 10. Training step number and loss curves and training time and loss curves for different types of attention modules.

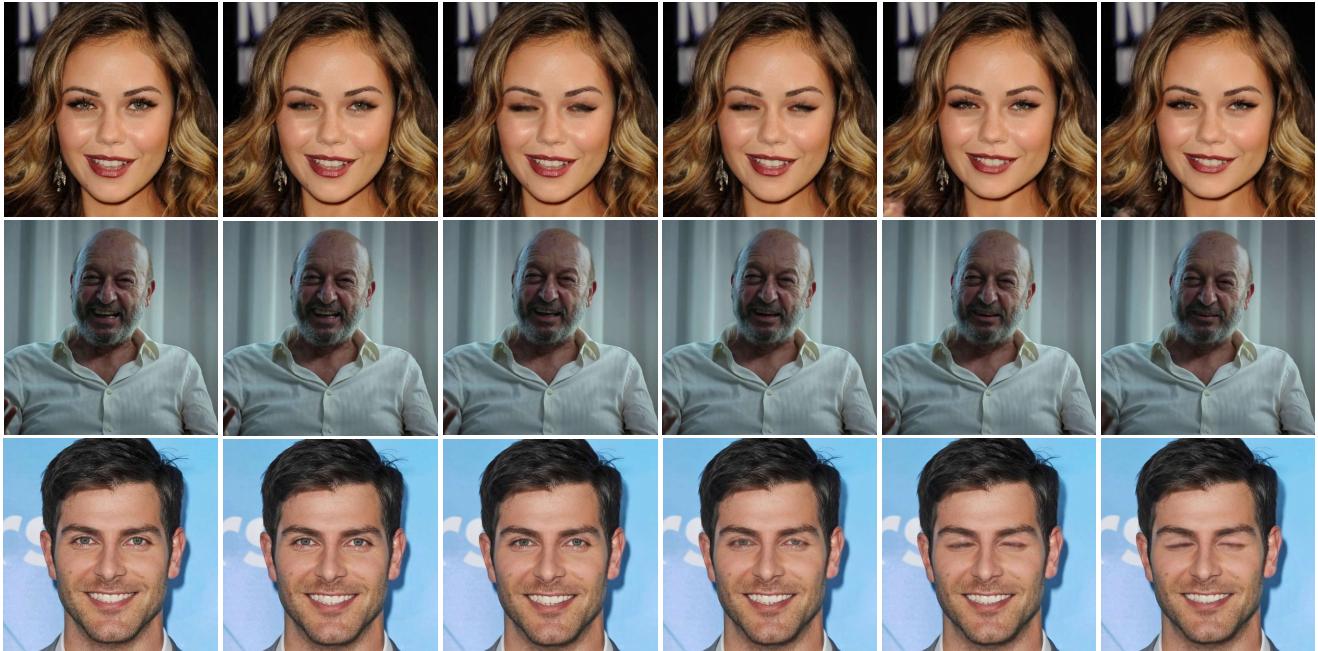


Figure 11. 17 frame 960×960 Image to Video Visualization Results (Inference steps: 28).

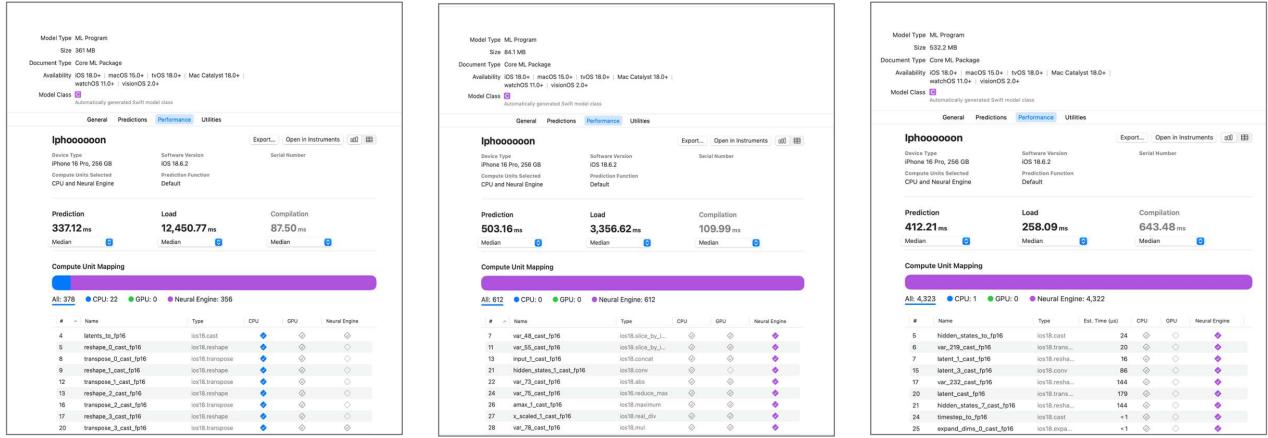


Figure 12. VAE’s encoder (left), decoder (middle) and denoiser (right) runtime on mobile devices.

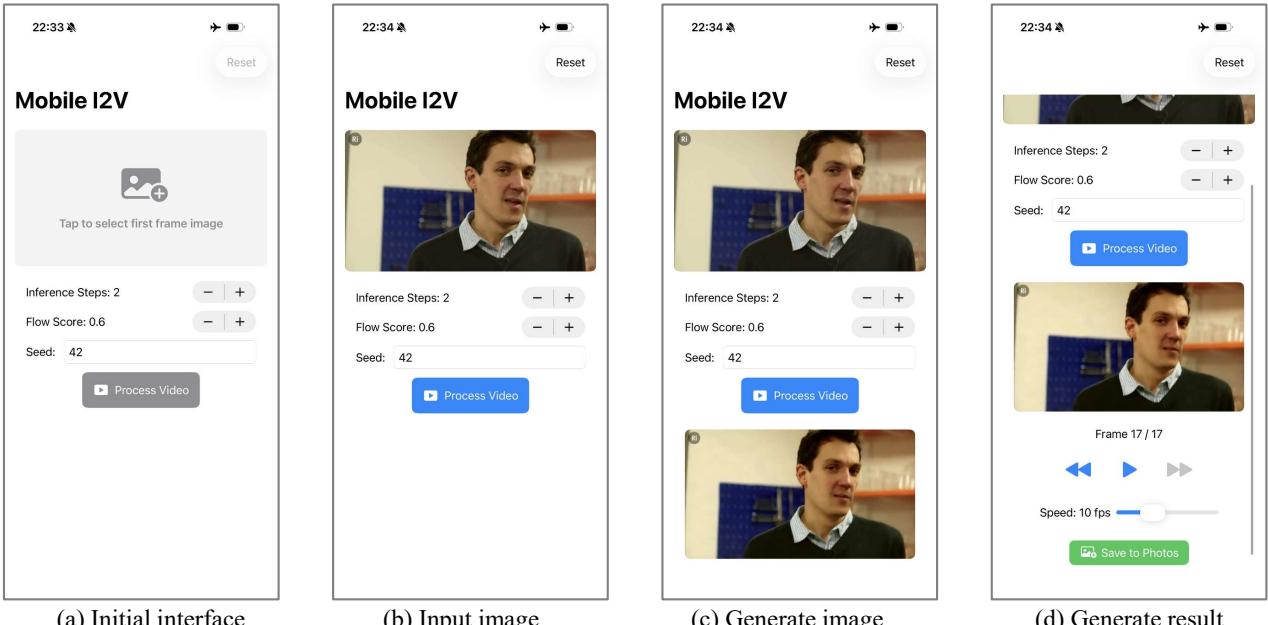


Figure 13. Screenshot of the UI interface on mobile devices.

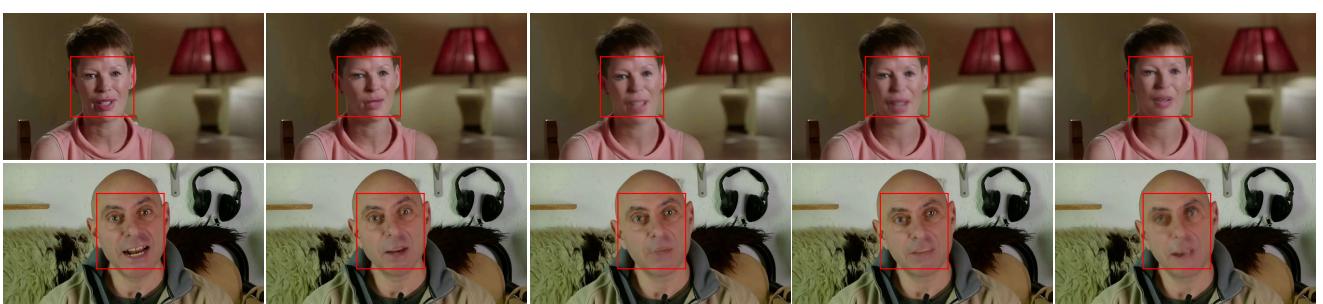


Figure 14. Some generated failure cases.

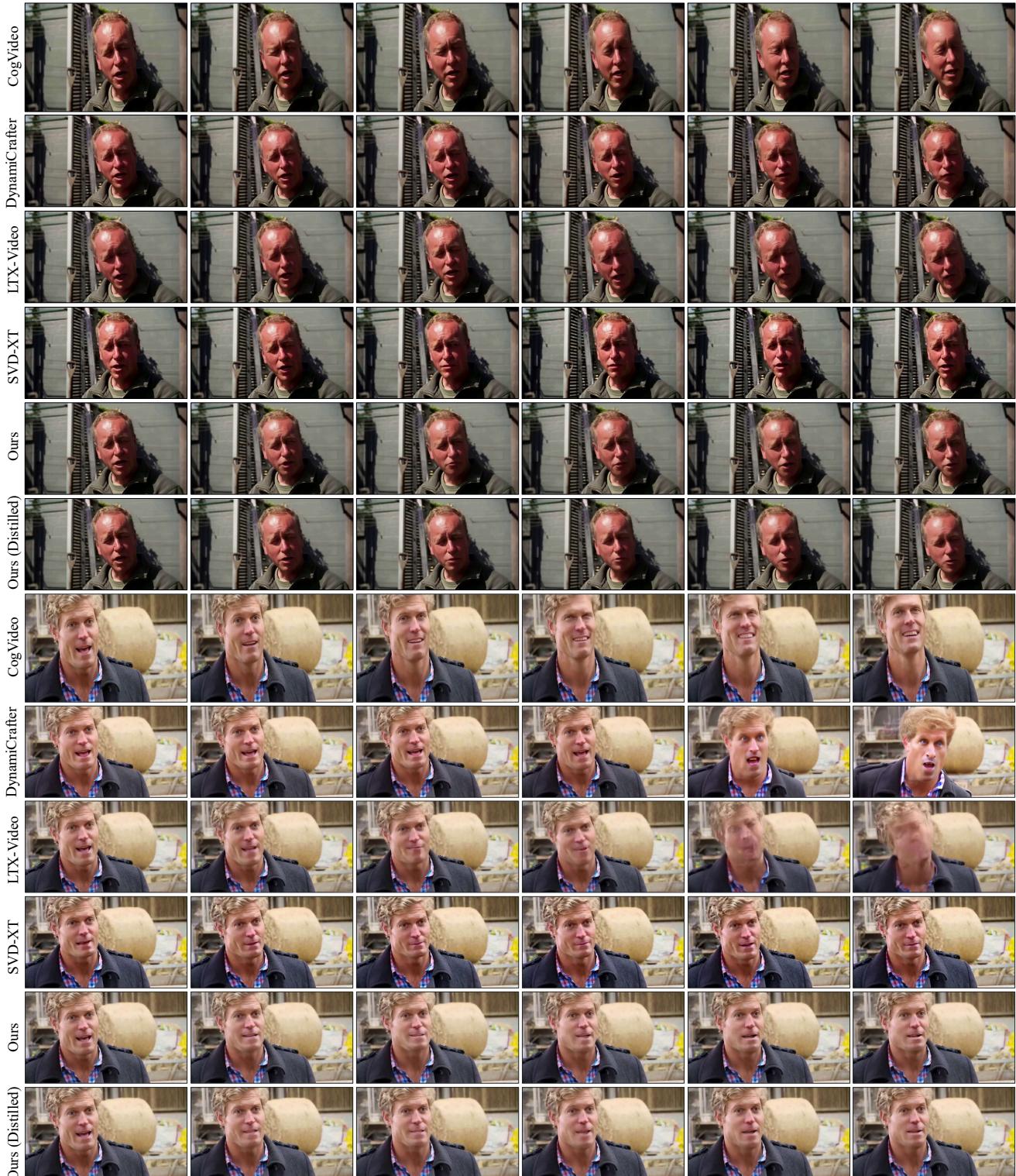


Figure 15. More visual comparison results.

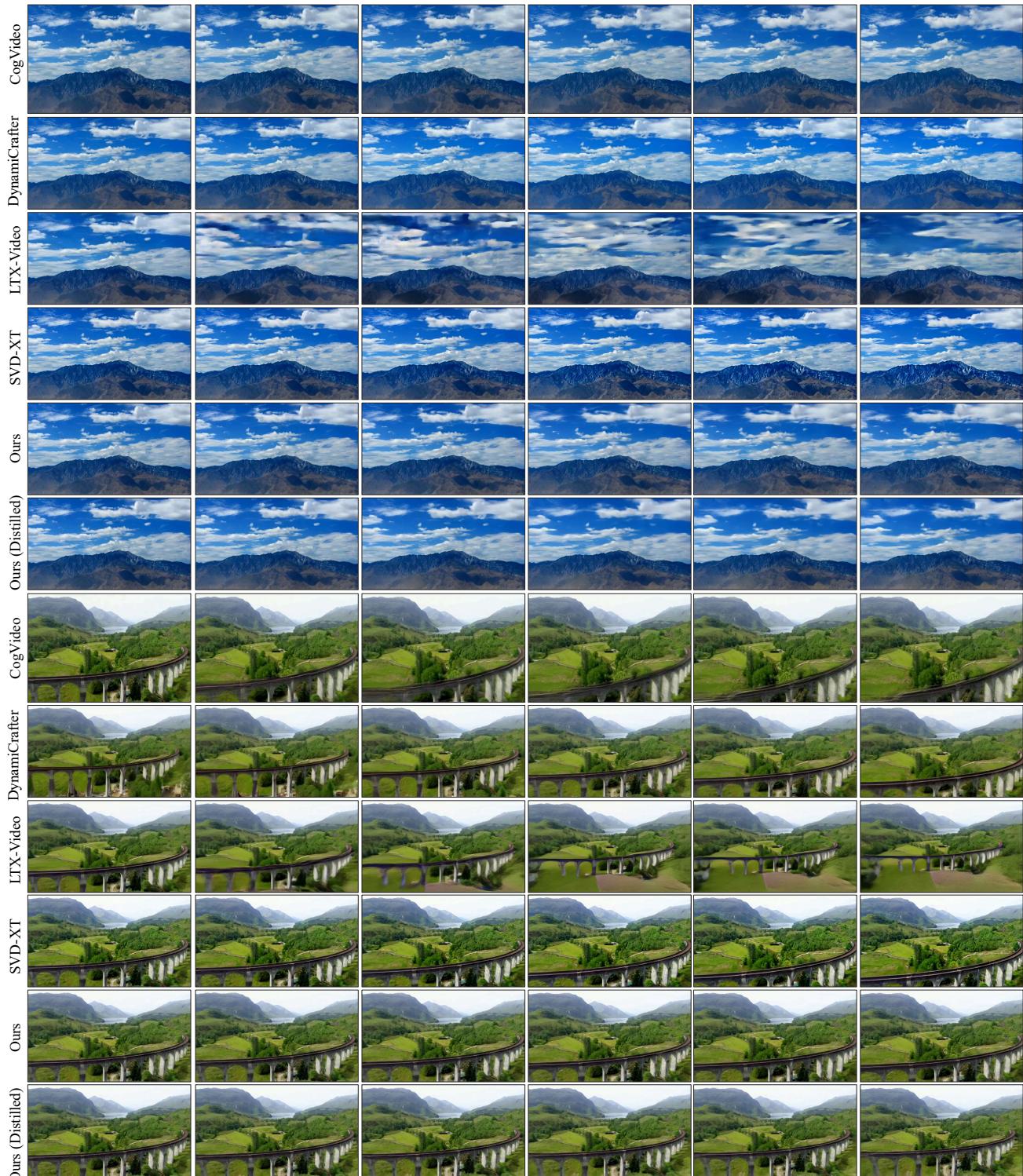


Figure 16. More visual comparison results.