AR-Diffusion: Asynchronous Video Generation with Auto-Regressive Diffusion
supplementary material

A. Discussions

Discussion: Why temporal causal attention over bidirectional temporal attention? We incorporate the temporal causal attention mechanism for three main reasons: **1) Reducing noise interference**: Subsequent video frames often contain more noise and less information than preceding frames. When preceding frames engage in cross-attention with subsequent frames, their information can be corrupted by the noise contained in subsequent frames. **2) Auto-Regressive flexibility**: Temporal causal attention enables models to behave similarly to an auto-regressive model, which is well-suited for generating videos of variable lengths. **3) Potential for image integration**: Temporal causal attention provides the capability to integrate image data as an initial video frame in future training, allowing for exclusive optimization of the starting frame.

Discussion: Why use x_0 prediction instead of ϵ prediction or v prediction? In many diffusion models, ϵ prediction [2] and v prediction [4] losses are more frequently employed. However, x_0 prediction is crucial for AR-Diffusion to effectively learn temporal correlations. In synchronous diffusion models, video frames are uniformly corrupted with equal timesteps, preserving most temporal correlations, allowing the model to directly learn temporal relationships from the input. In contrast, asynchronous diffusion disrupts these correlations due to varying levels of corruption across frames, making it challenging for AR-Diffusion to learn temporal dependencies from the inputs alone. The use of x_0 prediction forces the model to generate outputs that maintain strong temporal correlations across frames, leading to improved video consistency.

B. Limitation

Despite the promising results achieved by our proposed AR-Diffusion, there are several limitations that need to be addressed in future work. The primary limitation is that, while our model leverages video data for training, there is potential to further enhance its performance by incorporating image data. Images, being more readily available and diverse, can provide additional training signals that help improve the visual quality and diversity of generated frames. Integrating image data into the training process could also help in scenarios where video data is scarce or difficult to obtain. Future research should explore methods to effectively combine image and video data during training to further boost the performance of AR-Diffusion.

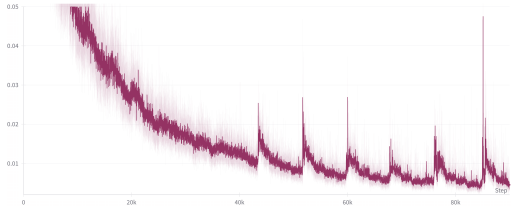


Figure 1. Loss curve of Diffusion Forcing [1] on UCF-101.

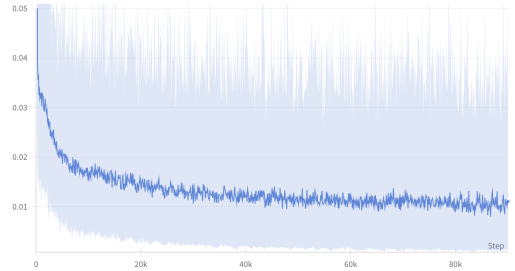


Figure 2. Loss curve of AR-Diffusion (ours) on UCF-101.

C. Training Stability

Training stability is a critical aspect of machine learning model performance. Stable training processes ensure consistent learning and convergence to optimal solutions. This paper explores the training stability of Diffusion Forcing [1] and our AR-Diffusion by analyzing their training loss curves. In particular, we visualize their training loss curves in Fig. 1 and Fig. 2, respectively. The training loss curves are plotted over a series of training steps. For Diffusion Forcing, the training loss decreases with noticeable fluctuations. These fluctuations suggest some instability in the training process. Significant spikes in the loss are observed around 40k and 80k steps, indicating moments of instability. Despite these spikes, the overall trend shows a decrease in loss, suggesting ongoing learning. For our AR-Diffusion, the training loss decreases with fewer and less pronounced fluctuations. This

Table 1. Ablation study on VAE using 4 A800 for 100 hours w/o the fine-tune stage.

VAE Model	Token Length	Sample Time	Taichi-HD		Sky-Timelapse	
			FVD_{rec}	FVD_{gen}	FVD_{rec}	FVD_{gen}
AR-VAE (ours)	16x32	3.1s	18.1	83.9	15.3	79.6
OpenSora-VAE-v1.2	4x32x32	30.7s	41.2	785.4	11.3	643.8

suggests better stability in the training process compared to Diffusion Forcing.

D. Reconstruction Performance

In this section, we qualitatively analyze the performance of the Auto-Regressive Video Auto-Encoder (AR-VAE) on reconstructing video frames across different datasets: FaceForensics, UCF-101, TaiChi-HD, and Sky-Timelapse. The results are reported in Fig. 3. On the FaceForensics dataset, the reconstructed frames closely resemble the real frames, maintaining the overall structure and details of the scenes. The colors and shapes are well-preserved, indicating that AR-VAE effectively captures the essential features of the video frames. On UCF-101, the AR-VAE demonstrates strong performance in reconstructing dynamic actions, such as the movements of the individuals in the video. The reconstructed frames retain the motion and spatial details, ensuring temporal consistency and clarity. On the TaiChi-HD dataset, the reconstructed frames accurately reflect the subject movements. The AR-VAE maintains the continuity and fluidity of the actions, preserving the intricate details and background elements. On Sky-Timelapse, the AR-VAE effectively reconstructs the changing sky scenes, capturing the variations in cloud formations and lighting conditions. The reconstructed frames exhibit high fidelity to the real frames, maintaining the temporal progression and visual consistency. Overall, the AR-VAE shows impressive reconstruction capabilities across different types of video content, preserving both spatial and temporal features with high accuracy.

E. Ablation study on AR-VAE

We conduct ablation study on AR-VAE and report the results in Table **??**. FVD_{gen} is obtained by training AR-Diffusion with different VAEs. Inference time involves both generation and decoding processes. We train AR-Diffusion with different VAE on 4 A800 GPUs for 100 hours. As reported in Table 1, current SOTA VAE, i.e. Open-Sora-VAE, utilizes 8 times more tokens to represent a video than our AR-VAE, resulting in a smaller batch size (2 vs 16), slower training and infer speed, and much poorer generation performance.

Table 2. Sampling a F-frame video on an A800. * denotes 128^2 reso.

	Latte	FIFO	TATS	VIDM	Ours
Max Mem.	14.7GB	8.5GB	4.1GB	35.5GB	8.6GB
F=16	6.8s	595.8s	14.8s	115.9s	3.1s
F=128	52.0s	1639.8s	49.9s*	380.0s*	45.8s

F. Efficiency Comparison

As reported in Table 2, our method shows superior efficiency compared to others. We use official codes&ckpts. Different from reported in paper, here we include both generation and decoding time.

G. Settings of Hyper Parameters

The detailed settings of model hyper parameters are presented in Table 3.

Table 3. Hyper-parameters of the AR-VAE and the DiT backbone of AR-Diffusion.

AR-VAE	
Token Length L	32
Token Dimension D	4
Model Width	1024
Num Layers	24
Num Heads	16
MLP Ratio	4.0
AR-Diffusion	
Scale Factor	0.5
Hidden Size	1152
Depth	28
Num Heads	16
β Linear Start	0.0001
β Linear End	0.002
Num Timesteps	1000

H. Samples on Long Video Generation

In Fig. 4, we present the generated 128-frame long videos by our AR-Diffusion on four diverse datasets: TaiChi-HD, Sky-Timelapse, UCF-101, and FaceForensics. More displayable samples can be found in https://anonymouss765.github.io/AR-Diffusion. The qualitative results of our AR-Diffusion model demonstrate its capability to generate visually realistic and temporally coherent video frames. On the TaiChi-HD dataset, ehe generated frames exhibit smooth and natural transitions, capturing the fluid motion characteristic of Tai Chi exercises. The model maintains the consistency of the subject's movements and the background details, ensuring a coherent visual experience. On

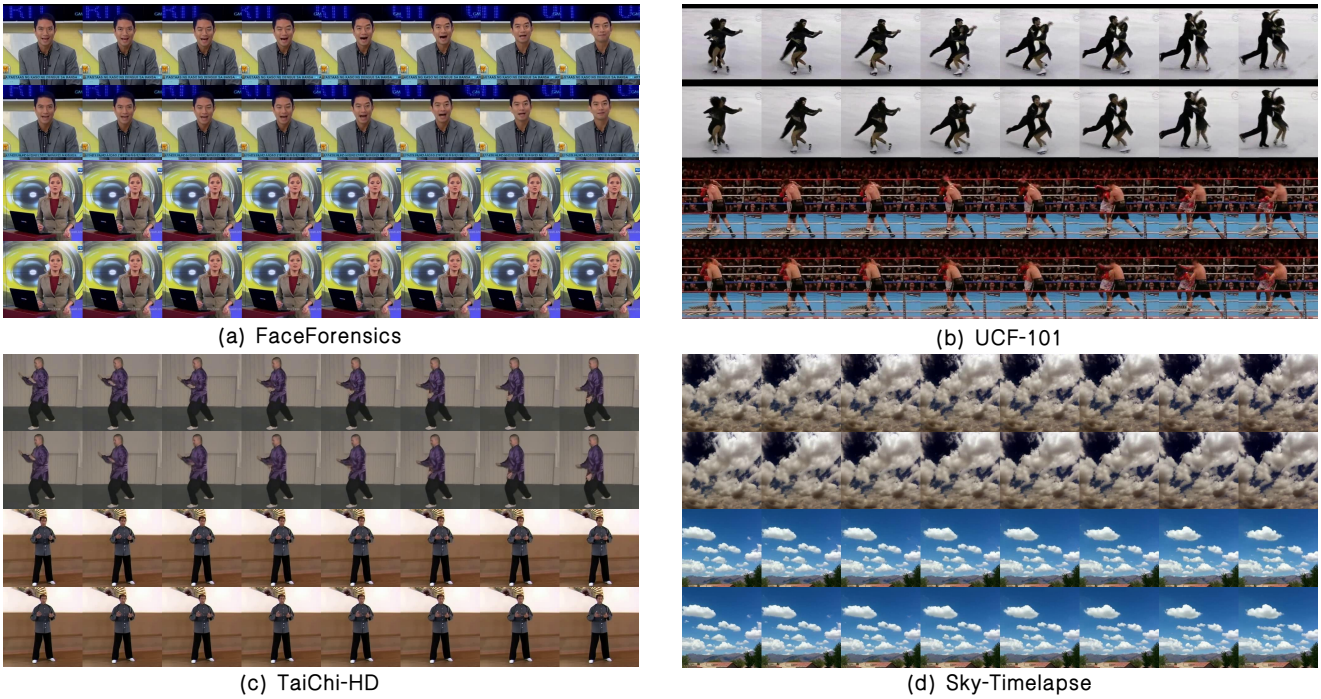


Figure 3. Real (first row) and reconstructed (second row) video frames using our AR-VAE on the (a) FaceForensics [3], (b) UCF-101 [6], (c) TaiChi-HD [5], and (d) Sky-Timelapse [7] datasets.

the Sky-Timelapse dataset, ehe model effectively synthesizes the gradual changes in the sky, including cloud movements and lighting variations. The temporal coherence is well-preserved, with the transitions between frames appearing seamless and natural. On the UCF-101 dataset, which includes various human actions, the AR-Diffusion model successfully generates frames that depict continuous and realistic motion. The actions are rendered with high fidelity, and the temporal progression of the activities is smooth and coherent. On the FaceForensics dataset, the generated video frames show the model's ability to handle complex facial movements and expressions. The transitions between frames are smooth, and the facial details are consistently maintained, demonstrating the model's robustness in generating temporally coherent video sequences. Overall, the AR-Diffusion model excels in producing high-quality video frames that are both visually realistic and temporally coherent, outperforming existing methods in handling diverse and challenging video generation tasks.

References

[1] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024. 1

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

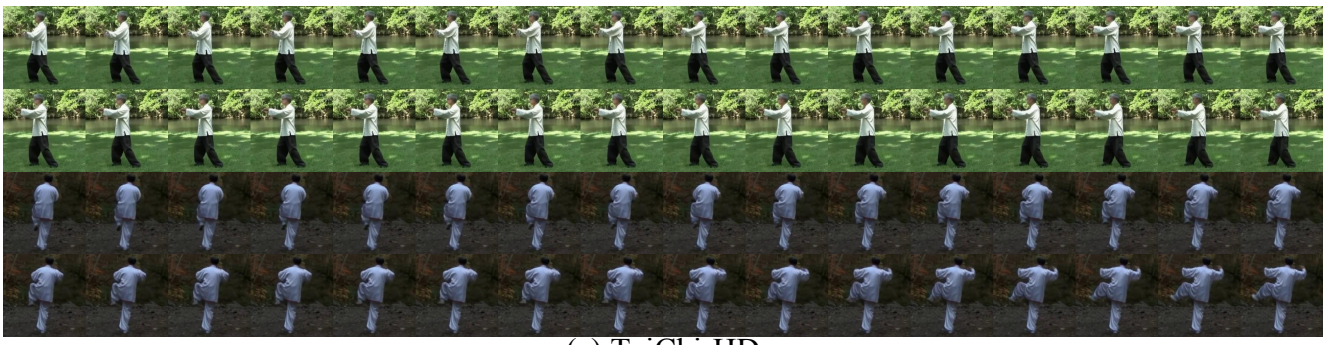
[3] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 3

[4] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[5] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 3

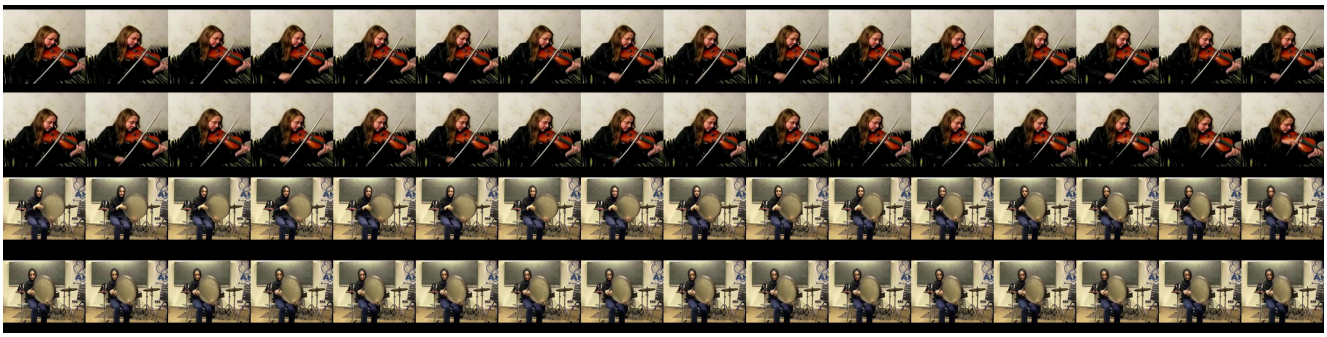
[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[7] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018. 3

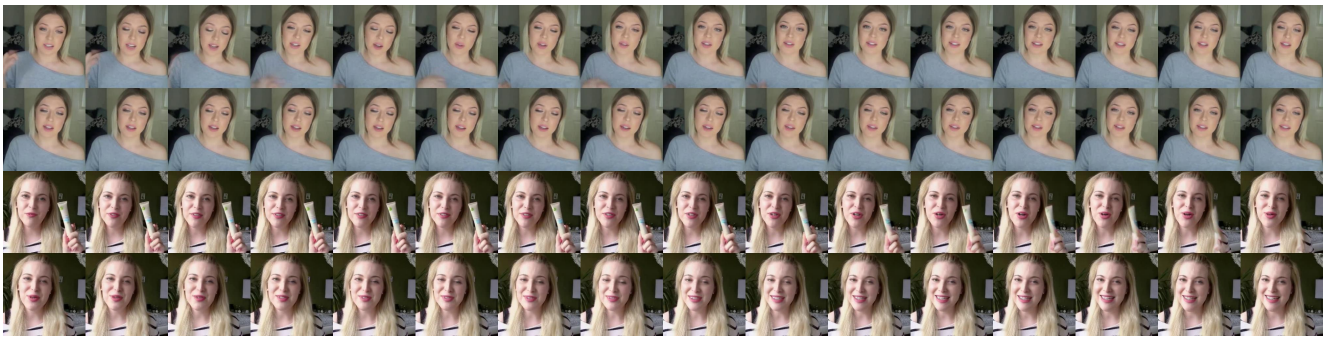


(a) TaiChi-HD

(b) Sky-Timelapse



(c) UCF-101



(d) FaceForensics

Figure 4. Generated 128-frame videos using our AR-Diffusion on four datasets. Each video is displayed with 4 skipped frames.