Progressive Autoregressive Video Diffusion Models

Supplementary Material

A. Summary

In this Appendix, we cover parallel works in Appendix B, related works in Appendix C, limitations and discussions in Appendix D, training details in Appendix E, evaluation details in Appendix F, additional qualitative results in Appendix G, and additional ablation study in Appendix H.

B. Parallel Works

The core idea of our PA-VDM is to 1. *assign progressively increasing noise levels to the F frames in the attention window* and 2. *autoregressively apply the video diffusion model on progressively noised frames to generate long videos*. The first part is inspired by Diffusion Forcing [4], which proposes to assign independent per-frame noise levels to some frames rather than a single noise level. We began developing our work right after July 1st, 2024, when Diffusion Forcing [4] was released on arXiv. Our manuscript was first submitted to peer review in early October, 2024. During this period, our work was developed independently, without the knowledge of two papers, Rolling Diffusion [38] and FIFO-Diffusion [19]. While Rolling Diffusion, FIFO-Diffusion, and PA-VDM arrive at a similar high-level idea in parallel, the three methods have different focuses, naming and framing of the idea, implementation details, experimental setups, and final result quality.

Compared to [6, 38], PA-VDM:

1. shows that it is possible to adapt a pre-trained video diffusion model to the progressive noise level schedule through finetuning, thus avoiding the otherwise immensely expensive computation cost of pre-training video diffusion models. [38] is trained from scratch and [19] is training-free.
2. achieves state-of-the-art 60-second long video generation at a quality comparable to frontier video diffusion models, demonstrating much longer video length and better quality than [19, 38].

We provide comparisons between our models (PA-*M* and PA-*O*) and [19] on our 60-second long video generation benchmark (Appendix F.2) in Sec. 4.2 and Tab. 1. Our method achieves substantially better qualitative and quantitative results than [19]. Notably, our method achieves FVD scores of 358.020 (PA-*M*, fine-tuned) and 548.117 (PA-*O*, training-free), significantly higher than [19]'s FVD of 975.459, which ranks the last among the methods we tested. [19] incorporates techniques that require inference cost that more than doubled the original cost, while our method only requires additional inference cost that is a fraction of the original cost (10% for PA-*M* and 16.66% for PA-*O*). We do not compare PA-VDM with [38] as there is no released code and it does not support text-conditioned open-domain generation.

C. Related Works

The field of long video generation has faced significant challenges due to the computational complexity and resource constraints associated with training models on longer videos. As a result, most existing text-to-video diffusion models [1, 10, 14, 15] have been limited to generating fixed-size video clips, which leads to noticeable degradation in quality when attempting to generate longer videos. Recent works are proposed to address these challenges through innovative approaches that either extend existing models or introduce novel architectures and fusion methods.

Freenoise [36] utilizes sliding window temporal attention to ensure smooth transitions between video clips but falls short in maintaining global consistency across long video sequences. Gen-L-video [49], on the other hand, decomposes long videos into multiple short segments, decodes them in parallel using short video generation models, and later applies an optimization step to align the overlapping regions for continuity. FreeLong [28] introduces a sophisticated approach which balances the frequency distribution of long video features in different frequency during the denoising process. Vid-GPT [7] introduces GPT-style autoregressive causal generation for long videos.

More recently, Short-to-Long (S2L) approaches are proposed, where correlated short videos are firstly generated and then smoothly transit in-between to form coherent long videos. StreamingT2V [12] adopts this strategy by introducing the conditional attention and appearance preservation modules to capture content information from previous frames, ensuring consistency with the starting frames. It further enhances the visual coherence by blending shared noisy frames in overlapping regions, similar to the approach used by SEINE [5]. NUWA-XL [56] leverages a hierarchical diffusion model to generate long videos using a coarse-to-fine approach, progressing from sparse key frames to denser intermediate frames. However, it has only been evaluated on a cartoon video dataset rather than natural videos. VideoTetris [46] introduces decomposing prompts temporally and leveraging a spatio-temporal composing module for compositional video generation.

Another line of research focuses on controllable video generation [16, 45, 60, 61] and has proposed solutions for long video generation using overlapped window frames. These approaches condition diffusion models using both frames from previous windows and signals from the current

window. While these methods demonstrate promising results in maintaining consistent appearances and motions, they are limited to their specific application domains which relies heavily on strong conditional inputs.

D. Limitations and discussions

A limitation of our method is the demand of a well-trained base video diffusion model. Similar to the *replacement* methods [15, 58] and other approaches like StreamingT2V [12], our method autoregressively applies a video diffusion model to generate long videos. Such autoregressive video generation poses huge challenge on the base video diffusion model. Some slight errors remaining in the "clean" frames $\mathbf{x}^0$ may not be noticeable in a single video clip; however, in the autoregressive scenario, these error can be carried onto later frames, resulting in quality degradation. Further more, as the video diffusion model is only trained on denoising latent frames of real video data, it may poorly handle such distribution shift towards the generated erroneous frames [6, 54], resulting in more severe quality drop. This means that even after finetuning on our progressive noise levels, our method could still generate videos with some degree of quality degradation close to the ending, if the base video diffusion model is not well trained. Among the qualitative videos generated by our PA-*M*, in some cases, the video quality slightly degrades in the last 10 seconds.

Another limitation of our method is the subtle temporal flickering happening about every second in our PA-*M* results. It is caused by a flaw in the backbone video diffusion model *M*'s 3D VAE, as evident by the presence of such flickering in both PA-*M* and RW-*M* results while no such flickering is present in the PA-*O* results.

There are many promising future directions to extend this work. We only train on progressively increasing noise levels to reduce the space of noise levels for easier convergence. If sufficient computing resources are available, training on fully random, per-frame independent noise levels would enable a single model for various tasks with arbitrary lengths, including video extension, connection, temporal super-resolution. Another promising future application of the long video generation ability of our models is to use them as world simulators, useful for tasks in robotics and 3D vision. Being able to generate long videos without quality degradation is an substantial step towards this direction.

E. Training details

M is pre-trained on captioned image and video datasets, containing 1 million videos and 2.3 billion images. These data are licensed and have been filtered to remove low-quality content. We train PA-*M* on video clips of $16, 32, ..., 176$ raw frames that correspond to $F = 5, 10, ..., 55$ latent frames. The $F = 55$ attention window length is derived by setting $F = S + 5$, where $S = 50$ is the number of sampling steps in M ($S = 30$ in O) and 5 is the length of an additional chunk of latent frames, as described in Secs. 3.3 and 3.4. The shorter latent frame lengths $F = 5, 10, ..., 50$ are used for the variable length training, as discussed in Sec. 3.2. RW-*M* is trained on videos of 64 frames that corresponds to $F = 20$ frames.

E.1. Modification to the base model

To implement progressive autoregressive video diffusion models on top of their pre-trained foundation video diffusion models, we do not need to modify the base model architectures. Instead, we only need to modify the model's forward, training, and inference procedures. In the training and inference procedures, we replace the single noise level $\mathrm{t} \in [0, T)$ from regular diffusion model training [13, 15] with our per-frame noise level $\mathbf{t}_{0:F-1}$ and $\boldsymbol{\tau}'_{0:S-1}$ (Secs. 3.1 and 3.5). To accommodate this change, we only need to make a single modification to the the noise level embedding computation in the model's forward procedure. While the regular timestep only has the batch size dimension B, our progressive timesteps has two dimensions B, F. We first flatten them into the batch dimension of size $B \times F$, pass it to the timesteps embedding module, unflatten the two dimensions, and finally broadcast the timestep embedding to the same shape of the frames so they can be combined through either addition, concatenation, modulation, or cross-attention [32, 33, 48].

F. Evaluation details

F.1. Baselines

As discussed in Sec. 4, using our base models, we implement two baseline autoregressive video generation methods on three models, which are denoted as RW-*M*, RN-*O*-base, and RN-*O*. We also compare to Stable Video Diffusion (SVD) [1] and StreamingT2V [12] model families. Specifically, we consider the SVD-XT model from SVD, a image-to-video model that generates a short video clip of 25 frames at 576x1024 resolution given an conditioning image. We apply it autoregressively, using the last image of the previous clip as the condition for generating a new clip. This is equivalent to the *replacement-without-noise* method except that it only conditions on a single frame rather than a chunk of 17 frames as RN-*O*. We also consider the StreamingSVD model from StreamingT2V, a image-to-long-video generation model that uses SVD as the base model [12]; its autoregressive video generation is enabled by training additional modules that connect to the base model via cross-attention. Similar to our progressive autoregressive video diffusion models, StreamingSVD can autoregressively generate long videos at 720x1280 resolution with arbitrary lengths, which we set to 1440 frames. We also compare to a con-

current work FIFO-Diffusion [19] implemented on Open-Sora-Plan v1.0.0 [23], denoted as FIFO-OSP. It generates at 256x256 resolution with a context window of 65 latent frames. See Appendix B for a discussion on [19] and other concurrent works. See Appendix F for details on our testing set, quantitative metrics, and traditional video quality evaluation.

FIFO-OSP FIFO-Diffusion [19] is a parallel work that adopts a similar high-level idea as our method on pre-trained video diffusion models without any fine-tuning (see more discussion in Appendix B). It provides training-free implementations on VideoCrafter2 and Open-Sora-Plan v1.1.0 [23]. We choose its Open-Sora-Plan implementation since our method is also implemented on DiT-base [32] models, *M* and Open-Sora (*O*) [58]. Open-Sora-Plan v1.1.0 generate videos at 512x512 resolution. Since there is no distributed inference support in the released code of FIFO-Diffusion, we adopt Open-Sora-Plan v1.0.0 in our reproduced FIFO-Diffusion results in order to saving computation costs by inferencing at the 256x256 resolution instead of the original 512x512 resolution.

F.2. Testing set

Text prompts and real videos Our testing set consists of 40 text prompts and the corresponding real videos, sampled from Sora [58] demo videos, MiraData [18], UCF-101 [44], and LOVEU [52, 53]. For each text prompt, we generate two videos with 1440 frames, 60 seconds long at 24 FPS, resulting in a total of 80 videos. We use these 80 videos from each model for both quantitative and qualitative results, unless specified otherwise. Due to computation resource limitations of sampling 1-minute long videos, we only obtained partial results from *M*-PA, StreamingSVD and FIFO-OSP, including 48, 40, 40 videos from 24, 40, 40 text prompts respectively. This testing set measures the zero-shot long video generation ability of the models, since none of them are specifically trained on any of the above datasets.

Real video initialization Since our focus is on long video generation, we focus on the video extension capability of the models rather than the text-to-short-video generation capability. Thus, we use the initial frames of the videos as the condition for all models, similar to the setting in [12]. *M*, *O* [58], StreamingSVD [12], SVD-XT [1], and FIFO-OSP [19, 23] use 16, 17, 1, 1, and 65 frames from the real video as the initial condition. Note that our PA-*M* and PA-*O* only require one chunk of frames (16 and 17 for *M* and *O* respectively), which is substantially less than the full context window of 65 frames required by FIFO-Diffusion [19]. This advantage is obtained from our variable-length autoregressive generation design as described in Sec. 3.2.

S	FVD↓
50	358.20
100	**339.59**
150	399.91

Table 2. Ablation on the number of sampling steps S of the PA-*M* model.

G. Additional Qualitative Results

We provide additional quailtative results in Fig. 7.

H. Additional Ablation Study

In our project webpage, we show an ablation study on our *Variable Length* design (Sec. 3.2). We compare Variable Length inference results of PA-*M* models trained with and without Variable Length. Without Variable Length training, the second video shows temporal jittering and abrupt scene change at the 1st and 59th seconds. This is because the model is not trained to generate the first/last chunk of latent frames to be consistent with the prior chunks. With Variable Length training, the first video avoids the jittering and abrupt scene change at the 1st and 59th seconds, and the video is temporally smooth. Furthermore, Variable Length inference enables the model to generate precisely 1440 frames, whereas without this technique the model would need to discard the noisy chunks remaining in the context window, which correspond to the 1441-1584th frames, when it reaches the 1440th frame. Being able to stop the autoregressive video denoising at a precise ending frame allows our model to generate a proper ending to the video, e.g. the woman exits the camera view in the first video, which is not possible without the Variable Length technique.

Additionally, we ablate the number of sampling steps S of the PA-*M*. Note that our progressive video denoising can work with arbitrary S; when the *chunked frames* technique is used, S only needs to be divisible by C. We compute FVD scores in the same way as described in Sec. 4.2. As shown in Tab. 2, further increasing S from 50 to 100 provides marginal benefits despite doubling the inference compute cost, while increasing S to 150 leads to slightly worse results.

Figure 7. Qualitative comparison of PA-*M* (ours), RW-*M*, PA-*O*-base (ours), RN-*O*-base, StreamingSVD from StreamingT2V [12], SVD-XT from Stable Video Diffusion [1], and FIFO-Diffusion [19]. Frames are evenly sampled from 1 minute long generated video, i.e. at 10, 20, 30, 40, 50, and 60 seconds. Our models can autoregressively generate 60-second, 1440-frame videos without quality degradation.