GLOBER: Coherent Non-autoregressive Video Generation via Global Guided Video Decoder

Anonymous Author(s)
Affiliation
Address
email

A Appendix

A.1 Broader Impact

The goal of this work is to advance research on video generation methods. Our method has the potential to facilitate the workflow of film production and animation, thereby exhibiting a positive influence on creative video applications. Since our method is trained mainly on domain-specific datasets, the potential deleterious consequences of exploiting our model for malicious purposes, such as spreading misinformation or producing fake videos, seem to be insignificant. Nevertheless, it remains crucial to apply an abundance of caution and implement strict and secure regulations.

A.2 Experimental Results on Long Video Generation Tasks

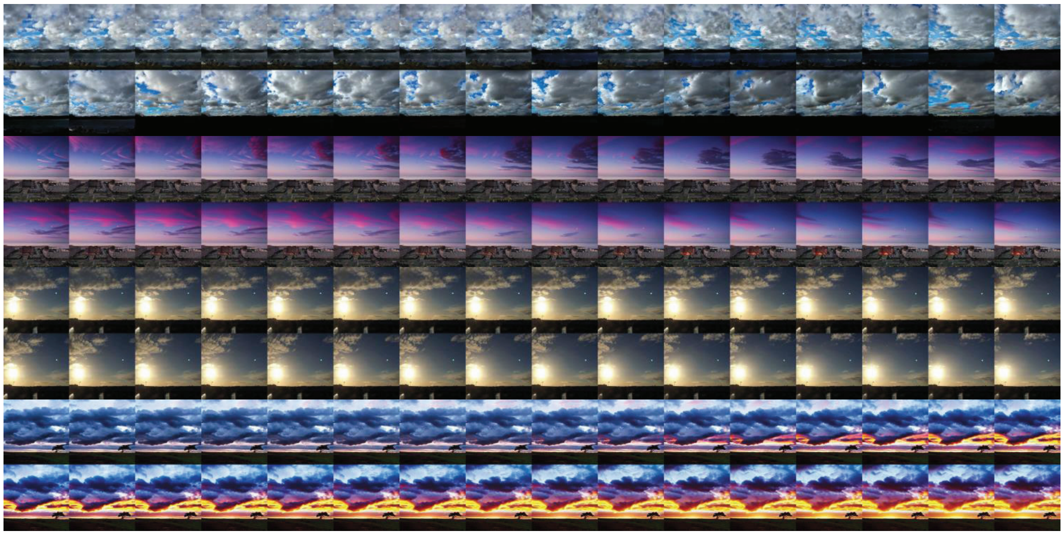
We obtain new state-of-the-art results on the SkyTimelapse and UCF-101 datasets for long video generation tasks. All experiments are conducted without conditional inputs. The quantitative results are reported in Table 1. MoCoGAN, MoCoGAN-HD, DIGAN, and StyleGAN-V are GAN-based methods, which dominate the field of vision generation until 2022. Based on diffusion probabilistic models, VIDM outperforms these GAN-based methods by a large margin. However, VIDM employs the auto-regression generation strategy to generate long videos, which lacks global guidance and suffers from error accumulation. Our method, GLOBER, outperforms VIDM significantly due to its incorporation of global features and non-autoregression generation strategy. We present several video samples in Fig. 1, which demonstrate that our method can generate long videos of remarkable quality.

Table 1: Quantitative Results of FVD comparison on the SkyTimelapse and UCF-101 datasets for 128-frame long video generation.

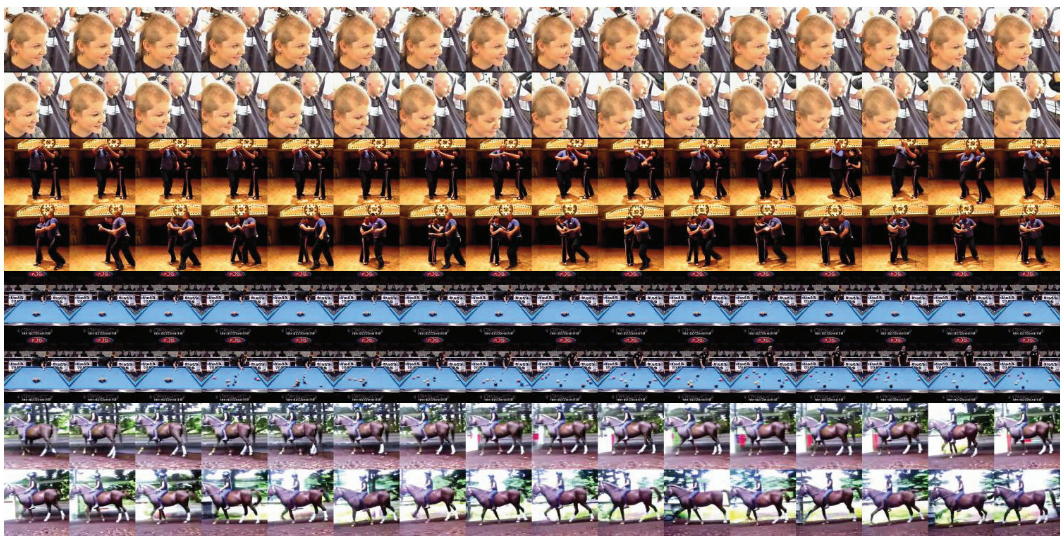
Method	UCF-101	Sky Time-lapse
MoCoGAN [CVPR18]	3679.0	575.9
+StyleGAN2 backbone	2311.3	272.8
MoCoGAN-HD [ICLR21]	2606.5	878.1
DIGAN [ICLR22]	2293.7	196.7
StyleGAN-V [CVPR22]	1773.4	197.0
VIDM [AAAI23]	1531.9	140.9
GLOBER (ours)	**1177.4**	**125.5**

A.3 More Qualitative Results

We present more qualitative results on the UCF-101, Sky Time-lapse, and TaiChi-HD datasets in the link: `https://anonymouss765.github.io/GLOBER`.



(a) Sky Time-lapse



(b) UCF-101

Figure 1: Genetated long videos with 128 frames on the Sky Time-lapse and UCF-101 datasets (4 frames skipped).

A.4 Sensitivity Analysis of Unconditional Guidance Scale

We investigate the effectiveness of the unconditional guidance scale μ that is used when employing class-condition constraints. Table 2 presents the influence of μ on the FVD score of videos conditionally decoded by the video decoder on the UCF-101 256^2 benchmark. Table 3 presents the influence of μ on the FVD score of videos conditionally sampled by the video generator on the UCF-101 256^2 dataset. It is evident that appropriate selection of the unconditional guidance scale is important in ensuring the quality of videos decoded or sampled with class conditions.

Table 2: Sensitivity analysis of the unconditional guidance scale μ for video reconstruction on the UCF-101 dataset.

μ	0	3	6	9	12	15
FVD	211.4	114.3	**106.7**	133.2	281.1	670.8

Table 3: Sensitivity analysis of the unconditional guidance scale μ for video generation on the UCF-101 dataset.

μ	0	3	6	9	12	15
FVD	575.6	173.0	172.6	171.5	**168.9**	224.3

A.5 Settings of Hyper Parameters

The detailed settings of model hyper parameters are presented in Table 4.

Table 4: Hyper-parameters of the video auto-encoder and the quantitative results on video reconstruction. Experimental settings on the UCF-101 dataset are the same for both conditional and unconditional video generation except given video descriptions.

	UCF-101 256^2	Sky Time-lapse 256^2	TaiChi-HD 128^2
Batch Size	40	32	32
Learning Rate	1e-5	1e-4	5e-5
		KL-VAE	
f_{frame}	8	8	4
		Video Encoder	
f_{video}		2	
Input Shape		32	
Input Channels		4	
Output Channels		16	
Model Channels		320	
Num Res. Blocks		2	
Num Head Channels		64	
Attention Resolutions		[16, 8]	
Channel Multiplies		[1, 2]	
		Video Decoder (UNet)	
Input Shape		32	
Input Channels		4	
Output Channels		4	
Model Channels		320	
Num Res. Blocks		2	
Num Head		8	
Attention Resolutions		[32, 16, 8]	
Channel Multiplies		[1, 2, 4, 4]	
		Video Generator (DiT)	
Input Shape	16	16	16
Input Channels	16	16	16
Model Channels	1152	1024	1024
Num Head	16	16	16
Depth	28	20	20
Mlp Ratio	4	4	4