Long Video Generation with Memory Efficient Bidirectional Transformers

Jaehoon Yoo[1], Semin Kim[1], Jinwoo Kim[1], Seunghoon Hong[1]

[1]Korea Advanced Institute of Science and Technology(KAIST)

Recent weather forecasting models built upon deep-learning based video prediction methods to improve the accuracy. Moreover, transformers with discrete tokens have shown rapid improvement in synthesizing complex videos. Yet, the quadratic memory and computational cost of standard transformers make it hard to model long-range and high-resolution videos. To address this issue, we propose an efficient video generative transformer with latent bottlenecks. Based on recent advances in bidirectional mask Transformers, our method learns to decode the entire spatio-temporal volume of a video in parallel from partially observed patches. The proposed transformer achieves a linear memory complexity in both encoding and decoding, by projecting observable context tokens into a fixed number of latent tokens and conditioning them to decode the masked tokens through the cross-attention. We also present optimization challenges in training bidirectional transformers with long videos and propose short-term guided curriculum learning that alleviates the issue. Empowered by linear complexity and stable optimization scheduling, our method demonstrates superior performance over the standard transformer-based generative models in long-term video synthesis.

Key words: Weather Forecast, Bidirectional Transformers, Memory Efficient Transformers, Data Driven Models.