TAP-T: Replacing Recurrence with Self-Attention for Temporal Latent Space Modeling

Joseph BOUIYODA
Department of Computer Engineering, ENSPY
University of Yaounde I
bouiyodajoseph@gmail.com

22 janvier 2026

Résumé

Recurrent neural networks form the backbone of many temporal latent space models for video generation, such as the Time-Aware Path (TAP) model. However, their sequential nature creates an information bottleneck, limiting their ability to model long-range dependencies effectively. Inspired by the success of the Transformer architecture in natural language processing, we propose TAP-T, a novel framework that replaces the recurrent temporal predictor of TAP with a Transformer-based module. TAP-T processes a sequence of past latent frame representations in parallel, using self-attention to directly model the relationships between distant points in time. This approach fundamentally overcomes the sequential bottleneck of RNNs. We hypothesize that this architectural change will lead to significant improvements in long-term consistency for video generation tasks, which we plan to validate on datasets like Moving MNIST.

1 Introduction

Video generation models aim to learn complex real-world dynamics to synthesize realistic future sequences. A prominent approach involves learning these dynamics in a compressed latent space, as exemplified by the Temporal Latent Space Modeling for Video Generation (TAP) framework [1]. The core of such models is often a recurrent neural network (RNN) that predicts the next latent state z_{t+1} based on the previous state z_t. This sequential processing, while intuitive, inherently struggles with long-term dependencies. Information from the distant past must be compressed and passed through every intermediate state, leading to information loss and the well-known "vanishing gradient" problem.

In parallel, the field of sequence modeling has been revolutionized by the Transformer architecture, introduced by Vaswani et al. [2], which completely dispenses with recurrence. By relying solely on self-attention mechanisms, Transformers can compute relationships between any two points in a sequence in constant time, regardless of their distance. This has led to state-of-the-art results in natural language processing.

This raises a compelling question : can the principles of the Transformer be applied to model the temporal dynamics of a video's latent space ? In this paper, we explore this question by proposing TAP-Transformer (TAP-T), an architecture that replaces the recurrent temporal predictor in the TAP model with a Transformer encoder block. Our main contribution is the adaptation of the self-attention mechanism for temporal prediction in latent space, with the goal of improving long-term consistency in generated videos.

2 Proposed Method : TAP-Transformer (TAP-T)

The TAP-T architecture replaces the sequential heart of the TAP model with a parallel, attention-based predictor. It maintains the same encoder-decoder structure for frame-level processing but redefines how temporal dynamics are learned.

2.1 Architecture Overview

Like the original model, an encoder E first maps each input frame x_t into a latent vector z_t. However, instead of processing these vectors one by one, our model operates on a sequence of the last N latent vectors, $[z_{t-N+1}, ..., z_t]$.

The process at each prediction step is as follows (Figure 1) :

1. A sequence of the N most recent latent vectors is collected.
2. **Positional Encodings** are added to this sequence to provide the model with information about the temporal order of the frames, which is otherwise lost in the non-recurrent architecture.
3. The resulting sequence is passed through a **Temporal Transformer** module. This module uses self-attention to compute an updated representation for each latent vector in the context of all others in the sequence.
4. The output representation corresponding to the last timestep, $\tilde{z}_t$, which now contains context from the entire window, is used to predict the next latent state $\hat{z}_{t+1}$.

Placeholder for Figure 1 : TAP-T Architecture Diagram

FIGURE 1 – Architecture of the proposed TAP-T model. A sequence of latent vectors is processed by a Temporal Transformer block, which uses self-attention to model long-range dependencies before predicting the next state.

2.2 Temporal Transformer Module

The core of our proposal is the Temporal Transformer module, which is analogous to a single layer of a Transformer encoder. It takes a sequence of latent vectors $Z = [z_1, ..., z_N]$ (after adding positional encodings) as input.

The self-attention mechanism computes a new representation for each z_i by attending to all other vectors in Z. This is done using scaled dot-product attention :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In this self-attention context :

— **Queries (Q), Keys (K), and Values (V) :** are all linear projections of the same input sequence Z. This allows every frame in the history window to interact with every other frame to build a rich, context-aware representation.

The output of this module is a new sequence $\tilde{Z} = [\tilde{z}_1, ..., \tilde{z}_N]$. The final vector, $\tilde{z}_N$, is then passed through a feed-forward network to produce the final prediction $\hat{z}_{N+1}$.

3 Experiments and Expected Results

We propose to evaluate TAP-T against the original recurrent TAP model on a task requiring long-term memory.

3.1 Dataset and Task

We will use the **Moving MNIST** dataset, with sequences of 50 frames. The task is frame prediction : given the first 10 frames, predict the next 40. This long prediction horizon will strongly penalize models that lack long-term consistency.

3.2 Evaluation Metrics

— **Mean Squared Error (MSE) :** To measure pixel-level prediction accuracy over the 40 predicted frames.
— **Object Consistency Score :** We will use a pre-trained MNIST classifier to identify the digit in the last predicted frame. The accuracy of this classification measures the model's ability to preserve the digit's identity over the long term.

3.3 Expected Results

We expect TAP-T to show a marked improvement over the baseline TAP model, especially in the consistency score. While the initial frame predictions might be of similar quality, the recurrent model's performance should degrade over time, while the Transformer's ability to access the initial frames should allow it to maintain the digit's identity.

TABLE 1 – Expected comparison on long-horizon Moving MNIST prediction.

Model	Prediction MSE (last 20 frames) ↓	Object Consistency @ frame 50 ↑
TAP (Baseline, Recurrent)	0.081	65%
TAP-T (Ours, Transformer)	**0.055**	**92%**

3.4 Ablation Study

To validate the contribution of the Transformer architecture, we will conduct an ablation study on the number of self-attention layers in the Temporal Transformer module. We will compare models with 1, 2, and 4 layers. We expect performance to increase with depth, as more layers allow for more complex temporal relationships to be learned, but with diminishing returns and increased computational cost.

4 Conclusion

We introduced TAP-T, a novel architecture for temporal latent space modeling that replaces the standard recurrent predictor with a Transformer-based module. By leveraging self-attention, TAP-T can directly model dependencies between distant frames, offering a principled solution to the problem of long-term consistency in video generation.

Our proposed experimental setup on Moving MNIST is designed to highlight the benefits of this approach. A primary limitation of TAP-T is the quadratic computational complexity of self-attention with respect to the sequence length. Future work should therefore investigate the integration of efficient, sparse attention mechanisms to allow the model to process even longer video sequences without prohibitive computational costs, further closing the gap on long-range temporal modeling.

Références

[1] A. Voufo, J. K. T. Feugmo, G. N. T. Tchapet, E. T. L. T. Tchiotsop, D. T. Tchiotsop, and C. F. F. T. Fotsing. Temporal latent space modeling for video generation. *ArXiv*, abs/2102.05095, 2021.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.