Improving Long-Term Consistency in Temporal Latent Space Models with External Memory

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

22 janvier 2026

Résumé

Temporal latent space models, such as the Time-Aware Path (TAP) model, have shown promise in video generation by learning dynamics in a compressed latent space. However, they often struggle with long-term temporal dependencies, leading to inconsistencies in generated sequences. We propose an architectural modification, TAP-M (Time-Aware Path with Memory), which augments the TAP framework with an external memory module inspired by Neural Turing Machines. This module allows the model to store and retrieve latent representations of past frames, enabling it to access distant information and significantly improve long-range object persistence. We hypothesize that experiments on long-sequence datasets like Moving MNIST will show that TAP-M reduces reconstruction error and maintains object properties over longer time windows compared to the original TAP baseline.

1 Introduction

Learning to model and predict temporal sequences is a fundamental challenge in machine learning, with critical applications in video generation, forecasting, and reinforcement learning. State-of-the-art methods often rely on learning dynamics in a low-dimensional latent space, as proposed by models like the Temporal Latent Space Modeling for Video Generation (TAP) [1]. These models typically use a recurrent structure to predict the next latent state based on the previous one.

While effective for short-term dynamics, this recurrent dependency creates an information bottleneck. The model's "memory" is limited to the information passed through its hidden state, making it prone to forgetting crucial details from the distant past. For instance, in a long video sequence, an object that exits the frame and re-enters later may be generated with inconsistent properties (e.g., a different color or shape), as the model has "forgotten" its initial appearance. This limitation highlights a core problem : maintaining long-term consistency.

To address this challenge, we propose to augment the TAP architecture with an external memory module. Inspired by Neural Turing Machines (NTMs) [2], our modification, which we call TAP-M, equips the model with a large, addressable memory bank. At each timestep, the model writes the latent representation of the current frame into the memory. The temporal prediction module can then use an attention mechanism to read from the entire history of stored frames, not just the immediately preceding one. This allows the model to directly access past information, breaking the strict sequential dependency and enabling better handling of long-range temporal consistency.

2 Proposed Method : TAP with Memory (TAP-M)

Our proposed method, TAP-M, builds upon the original TAP architecture by introducing a memory-augmented temporal predictor. The core components are an encoder, a memory module, and a modified temporal predictor.

2.1 Architecture Overview

The original TAP model consists of an encoder E that maps an input frame x_t to a latent representation $z_t = E(x_t)$, and a temporal model T that predicts the next latent state $\hat{z}_{t+1} = T(z_t)$.

Our TAP-M architecture (Figure 1) modifies this process. At each timestep t :

1. The frame x_t is encoded into $z_t = E(x_t)$.
2. The latent vector z_t is written into the next available slot of an external memory matrix M.
3. The temporal model T_{mem} uses an attention mechanism to predict the next latent state. It formulates a query based on the current state z_t and attends over all past states $\{z_1, ..., z_t\}$ stored in the memory M.

Placeholder for Figure 1 : TAP-M Architecture Diagram

FIGURE 1 – High-level architecture of our proposed TAP-M model. The temporal predictor uses an attention mechanism to read from the entire memory of past latent states $(z_1, ..., z_t)$ to predict the next state $\hat{z}_{t+1}$.

2.2 Memory-Augmented Temporal Prediction

The key innovation lies in the temporal predictor T_{mem}. Instead of relying solely on z_t, it computes the next state $\hat{z}_{t+1}$ as a function of z_t and a context vector c_t, which summarizes relevant information from the past.

The context vector c_t is computed using scaled dot-product attention [3] :

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

In our case :

— **Query (Q) :** A projection of the current latent state z_t. It represents the question : "What information from the past is relevant to predict the next frame, given what I see now ?"

— **Keys (K) and Values (V) :** Projections of all latent states currently stored in the memory matrix $M = [z_1, z_2, ..., z_t]$.

The temporal model then predicts the next state as $\hat{z}_{t+1} = f(z_t, c_t)$, where f can be a simple feed-forward network. This mechanism allows the model to directly access, for instance, the latent state z_1 of the first frame when predicting frame z_{t+1}, overcoming the limitations of a purely recurrent path.

3 Experiments and Expected Results

To validate our approach, we propose a comparative study between the original TAP model (baseline) and our TAP-M on a long-sequence video dataset.

3.1 Dataset and Task

We will use the **Moving MNIST** dataset. This dataset consists of sequences of frames showing digits moving and bouncing within a frame. We will use long sequences (e.g., 50-100 frames) to specifically challenge the models' long-term memory. The task is to predict the next frame given a sequence of previous frames.

3.2 Evaluation Metrics

We will evaluate the models on two primary metrics :

— **Mean Squared Error (MSE) :** To measure the per-pixel reconstruction accuracy of the predicted frames.

— **Object Consistency Score :** A custom metric to measure long-term persistence. For Moving MNIST, this could be the classification accuracy of the digit in the final predicted frame compared to the ground truth. A model that "forgets" will produce a blurry, unclassifiable digit.

3.3 Expected Results

We expect TAP-M to significantly outperform the TAP baseline, particularly on the Object Consistency Score for long sequences. The direct access to past frames provided by the memory module should prevent the model from forgetting the identity of the digit over time.

TABLE 1 – Expected comparison on long-sequence Moving MNIST.

Model	Frame Prediction MSE ↓	Object Consistency ↑
TAP (Baseline)	0.045	78%
TAP-M (Ours)	**0.032**	**96%**

3.4 Ablation Study

To demonstrate the impact of our memory module, we will conduct an ablation study by varying the size of the memory. We hypothesize that performance will improve as the memory size increases, up to a certain point where it can hold the entire relevant history. This will show that the performance gain is directly attributable to the external memory and not other confounding factors.

4 Conclusion

In this work, we proposed TAP-M, an extension of the Temporal Latent Space Model that incorporates an external, attention-based memory module to address the problem of long-term consistency in video generation. By allowing the temporal predictor to directly access all past latent states, our model is designed to overcome the information bottleneck of purely recurrent architectures.

While the implementation and full-scale experiments are left for future work, our architectural proposal presents a clear path toward improving long-range temporal modeling. Limitations include the increased computational cost due to the attention mechanism, which scales with sequence length. Future research could explore more efficient, sparse attention mechanisms to mitigate this cost while retaining the benefits of long-term memory.

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] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014.

[3] 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.