# Composition of Memory Experts For Diffusion World Models

**Sebastian Stapf, Pablo Acuaviva, Aram Davtyan, Paolo Favaro**

{sebastian.stapf, pablo.acuavivahuertos, aram.davtyan, paolo.favaro}@unibe.ch - Computer Vision Group @ University of Bern

## 1. Building Consistent World Models

We aim to build world models that generate videos that are **consistent** with past observations in an **efficient and scalable** manner


## 2. Different Approaches to Memory

### Transformer

**Context window** acts as memory

$$\mathcal{M}_t = \{K_{1:t}, V_{1:t}\}$$
$$O_t = \mathrm{softmax}\left(q_t \frac{K_{1:t}^\top}{\sqrt{d}}\right) V_{1:t}$$
$$K_{t+1} = \mathrm{concat}(K_{1:t}, k_{t+1})$$
$$V_{t+1} = \mathrm{concat}(V_{1:t}, v_{t+1})$$

- ✅ Enables highly detailed, accurate predictions
- ❌ Expensive: scales data, model size, and test-time compute

### Recurrent Models

Information stored in some **state vector**

$$\mathcal{M}_t = h_t$$
$$O_t = C\mathcal{M}_t$$
$$\mathcal{M}_{t+1} = A\mathcal{M}_t + Bx_t$$

- ✅ Cheap reading and writing into hidden states
- ❌ Limited by state capacity

### TT-Training

Information stored in **temporary parameters**

$$\mathcal{M}_t = \theta_t$$
$$O_t = \mathrm{f}(\theta_t, x_t)$$
$$\mathcal{M}_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}$$

- ✅ Higher capacity than fixed hidden states
- ❌ More expensive update mechanism

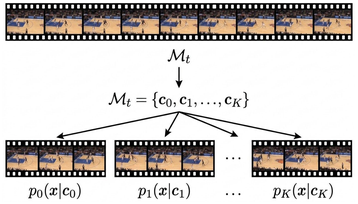
### Explicit Memory

Observations stored in an **external database**

$$\mathcal{M}_t = \{(e_i, d_i)\}_{i=1}^N$$
$$r_t = \mathrm{TopK}(\mathrm{sim}(q_t, E))$$
$$O_t = f(x_t, d_{r_t})$$
$$\mathcal{M}_{t+1} = \begin{cases} \mathcal{M}_t \cup (e_{t+1}, d_{t+1} \\ \mathcal{M}_t \end{cases}$$

- ✅ Cheap reading and exact retrieval
- ❌ Memory grows, search grows

## 3. Can We Combine Them?




**Memory could be distributed, not centralized**

- Do **not** force one backbone to solve all temporal scales
- Use a **composition of specialized memory experts** for different parts of the history

## 4. Memory as Product of Experts

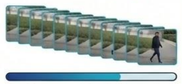
**IMMEDIATE PAST EXPERT (I2V)**

**Immediate** (1-3 Frames)

Pretrained I2V diffusion model for high-quality generation conditioned on immediate past.

$p_\theta(\mathbf{x} \mid c)$

**SHORT-TERM EXPERT (ST)**

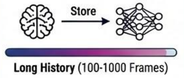


**Short-Term** (10-100 Frames)

Models local dynamics and fine details over a moderate context, using architectures like Mamba.

$p_\phi(\mathbf{x} \mid c_{\mathrm{ST}})$

**LONG-TERM MEMORY EXPERT (LT)**




**Long History** (100-1000 Frames)

**Test-Time** Finetuning

Stores episodic knowledge directly in weights via test-time finetuning for constant-time reuse.

$p_\psi(c_{\mathrm{LT}})(\mathbf{x} \mid c)$

**SPATIAL PRIOR EXPERT (SP)**

**Spatial Signals** (Poses, Maps)

Uses auxiliary spatial signals ($\mathcal{S}$) extracted from history to disambiguate locations and reduce drift.

$p_\lambda(\mathbf{x} \mid \mathcal{S})$

- Memory state composed of multiple contexts:
  - Short-term past / Long-term past
  - Possibly other structured memory formats
- Each context is handled by a **specialized diffusion model**

$$\Longrightarrow \quad p(\mathbf{x} \mid \mathcal{M}) = p(\mathbf{x} \mid c_1, \ldots, c_K) \propto \prod_{i=1}^{K} p_i(\mathbf{x}|c_i)$$

## 5. Key Enabler

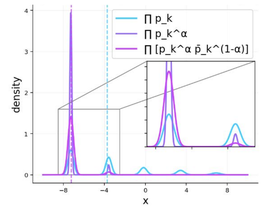
Combine independent diffusion models **at sampling time** by adding up the scores to sample from the product distribution

$$x_0 \sim p_\theta(x_0)$$
$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \nabla_{x_t} \log p_\theta(\mathbf{x}_t, t)\right) + \sigma_t \eta$$

$$x_0 \sim p_\theta(x_0) p_\phi(x_0) / Z$$
$$\Longrightarrow \quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - [\nabla_{x_t} \log p_\theta(\mathbf{x}_t, t) + \nabla_{x_t} \log p_\phi(\mathbf{x}_t, t)]\right)$$

## 6. Spurious Modes

Spurious likelihood modes get amplified if simultaneously present in different experts $\Longrightarrow$ We can remove them in the CFG way:




$$p_{\mathrm{CoME}}(\mathbf{x} \mid c, c_{\mathrm{ST}}, c_{\mathrm{LT}}, \mathcal{S}) \propto \left[p_\theta(\mathbf{x} \mid \varnothing)^{1-\alpha_0} p_\theta(\mathbf{x} \mid c)^{\alpha_0}\right] \times \left[p_\phi(\mathbf{x} \mid \varnothing)^{1-\alpha_1} p_\phi(\mathbf{x} \mid c_{\mathrm{ST}})^{\alpha_1}\right] \times \left[p_{\psi(\varnothing)}(\mathbf{x} \mid c)^{1-\alpha_2} p_{\psi(c_{\mathrm{LT}})}(\mathbf{x} \mid c)^{\alpha_2}\right] \times \left[p_\lambda(\mathbf{x} \mid \varnothing)^{1-\alpha_3} p_\lambda(\mathbf{x} \mid \mathcal{S})^{\alpha_3}\right]$$

## 7. Experiments

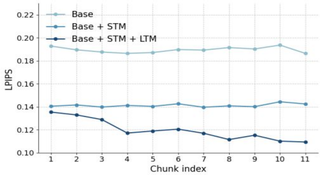
### Accurate predictions

Can the composition of experts make more accurate predictions?

| Method | LPIPS ↓ | SSIM ↑ | PSNR ↑ |
|---|---|---|---|
| Base | 0.209 | 0.771 | 19.16 |
| + STM | 0.156 | 0.820 | 21.29 |
| + LTM | 0.171 | 0.805 | 19.98 |
| + SLTM | 0.150 | 0.833 | 20.65 |
| + STM+LTM | 0.114 | 0.862 | 22.32 |
| CoME | 0.097 | 0.892 | 23.07 |
| Sliding | 0.183 | 0.753 | 19.02 |
| SSM | 0.158 | 0.828 | 20.62 |
| Full | 0.113 | 0.859 | 22.78 |

### Long-Term Memory

Does the TT-Learning of the long term memory contribute?




### Planning Capability

Does the composition help with planning? We investigate the NWM setting:

| | GNM | NOMAD | NWM | CoME |
|---|---|---|---|---|
| ATE (↓) | 1.87 | 1.93 | 1.13 | 0.96 |
| RPE (↓) | 0.73 | 0.52 | 0.35 | 0.28 |

| STM | LTM | NWM+STM | NWM+LTM |
|---|---|---|---|
| 1.05 | 1.10 | 0.98 | 1.07 |
| 0.32 | 0.32 | 0.30 | 0.33 |