Towards a Vision-Language Episodic Memory Framework: Large-scale Pretrained Model-Augmented Hippocampal Attractor Dynamics

Chong Li
lichong23@m.fudan.edu.cn
Fudan University

Taiping Zeng*
zengtaiping@fudan.edu.cn
Fudan University

Xiangyang Xue
xyxue@fudan.edu.cn
Fudan University

Jianfeng Feng
jffeng@fudan.edu.cn
Fudan University

Abstract

Modeling episodic memory (EM) remains a significant challenge in both neuroscience and AI, with existing models either lacking interpretability or struggling with practical applications. This paper proposes the Vision-Language Episodic Memory (***VLEM***) framework to address these challenges by integrating large-scale pretrained models with hippocampal attractor dynamics. VLEM leverages the strong semantic understanding of pretrained models to transform sensory input into semantic embeddings as the neocortex, while the hippocampus supports stable memory storage and retrieval through attractor dynamics. In addition, VLEM incorporates prefrontal working memory and the entorhinal gateway, allowing interaction between the neocortex and the hippocampus. To facilitate real-world applications, we introduce EpiGibson, a 3D simulation platform for generating episodic memory data. Experimental results demonstrate the VLEM framework's ability to efficiently learn high-level temporal representations from sensory input, showcasing its robustness, interpretability, and applicability in real-world scenarios.

Keywords: episodic memory; hippocampal attractor dynamics; vision-language model; cognitive framework

1. Introduction

The rapid progress in AI has led to the language models that can produce texts almost indistinguishable from human writing (Digutsch & Kosinski, 2023), representing a major step forward in realizing semantic memory functions (Kumar, 2020). However, modeling episodic memory—another key type of memory related to personal experiences—remains a significant challenge.

Episodic memory (EM) refers to the ability to store and consciously recall specific memories of past events (Tulving, 1972). It is characterized by: (i) **Egocentricity**. Episodic memory plays a crucial role in shaping our sense of self. Unlike semantic memory, which involves shared general knowledge, episodic memory is inherently self-referenced and unique to each individual (Penaud, Yeh, Gaston-Bellegarde, & Piolino, 2023). (ii) **Mental time travel**. Episodic memory helps us make decisions by allowing us to recall and relive moments from the past, guiding our choices in the present (Tulving, 2002; Nicholas, Daw, & Shohamy, 2022). (iii) **Real-world convergence**. In the real world, experiences are continuous and countless, making it impossible to retain all details (Neisser, 1992). However, episodic memory can always reliably store information about "what" happened, "where", and "when" (Yonelinas & Ritchey, 2015; Chandra, Sharma, Chaudhuri, & Fiete, 2025), thus condensing an infinite stream of observations into finite, discrete events. Overall, in the context of AI, episodic memory offers specific advantages: (i) **Robustness**. The properties of attractors make EM models resistant to noise. (ii) **Interpretability**. The attractor state space corresponds to the event space, so a change in state represent shifts between events.

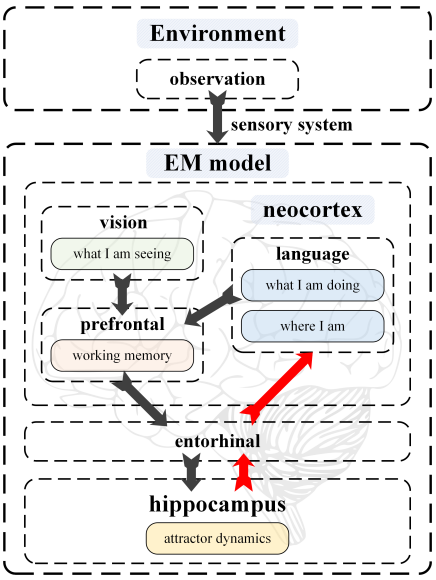

Figure 1: **Diagram of Vision-Language Episodic Memory framework.** This diagram illustrates the biologically inspired structure of our episodic memory model. The "vision" and "language" components handle the perception of visual inputs and self-state descriptions. "Working memory" processes sequential inputs for short-term storage, while the "entorhinal cortex" acts as a gateway between the neocortex and hippocampus. The "hippocampus" manages episodic memory through attractor dynamics.

Due to its unique characteristics, episodic memory has garnered significant attention across psychology, neuroscience, and artificial intelligence. Despite extensive research, the underlying mechanisms of episodic memory remain unclear, and there is no consensus on the optimal methods for modeling and effectively leveraging its capabilities. Originally proposed in psychology (Tulving, 1972), early EM models were primarily designed to explain findings from psycholog-

*Corresponding author.

ical behavioral experiments (Criss & Howard, 2015), such as the word frequency mirror effect (Malmberg, Holden, & Shiffren, 2004). These models focused on pattern matching and associations within tasks (Humphreys, Bain, & Pike, 1989; Gillund & Shiffrin, 1984; Anderson, Silverstein, Ritz, & Jones, 1977). As neuroscience advanced, biologically inspired computational models became central to EM research, aiming to replicate its cognitive functions. Cognitive architectures, such as Soar (Laird, Newell, & Rosenbloom, 1987; Laird, 2008), incorporated episodic memory as a core component of overall cognitive processes, drawing insights from biological evidence (Langley, Laird, & Rogers, 2009). Further studies focused on the neural circuits involved in episodic memory, trying to build biologically realistic models that support its function (E. T. Rolls & Treves, 2024). Additionally, the key biological attractor dynamics in the CA3 region of the hippocampus have become widely accepted in neuroscience (E. T. Rolls & Treves, 2024; Allen & Fortin, 2013; Squire, Knowlton, & Musen, 1993; Jeong, Chung, & Kim, 2015; T. Rolls, 2018). Attractor networks, such as Hopfield network (Hopfield, 1982; Krotov, 2023), have been used to mimic the convergent properties and neuronal dynamics of episodic memory. Thus, a deeper understanding of episodic memory through interpretable computational models can provide a solid foundation for integrating biologically plausible mechanisms into AI frameworks. This approach could bring us closer to understanding how these mechanisms work and, for the first time, enable the application of biological episodic memory models in real-world scenarios.

Episodic memory is also considered crucial for AI, as it supports numerous critical high-level cognitive functions (Hassabis & Maguire, 2007; Eğilmez, 2015). Without the ability to remember past experiences, AI agents risk repeating previous mistakes and wasting valuable cognitive resources (Jockel, Weser, Westhoff, & Zhang, 2008). Additionally, the capacity to retrieve specific experiences is vital for fast learning in new or sparse-reward situations (Allen & Fortin, 2013; Boyle & Blomkvist, 2024). As a result, various AI research initiatives aim to glean insights from episodic memory to enrich AI agents (Eğilmez, 2015; Jockel et al., 2008). The Ego4d (Grauman et al., 2022) project released a large-scale egocentric video dataset and EM benchmark, treating episodic memory as a special video modality, followed by studies that frame memory retrieval as a video question-answering task (Datta et al., 2022; Bärmann & Waibel, 2022). However, "episodic memory" here primarily describes systems that possess some features of it but differ in significant ways (Boyle & Blomkvist, 2024). Therefore, it is crucial to introduce biological episodic memory mechanisms into AI models to improve their interpretability and robustness.

Clearly, there is a notable difference between EM models in neuroscience and AI. The former (Hopfield, 1982; E. T. Rolls & Treves, 2024) focuses on interpretable mechanism but faces challenges in practical application, while the latter (Datta et al., 2022; Bärmann & Waibel, 2022) prioritizes applicability but lacks interpretablity and robustness.

To bridge this gap, we propose a novel Vision-Language Episodic Memory (VLEM) framework, which augment hippocampal attractor dynamics with large-scale pretrained model to create a biologically plausible episodic memory system within AI framework. Specifically, as illustrated in Fig. 1, we integrate the powerful semantic understanding capability of large-scale pretrained models (Schuhmann et al., 2022) to mimic the semantic processing in the cortex, transforming sensory input into semantic embeddings. With an understanding of the current state, the hippocampus supports episodic memory through its attractor mechanism, enabling stable storage and retrieval of experiences. Additionally, our framework incorporates working memory to track short-term historical states, while the entorhinal cortex collects information from the cortex and projects it back after interacting with the hippocampus, acting as a gateway between the two. The EM model maps an unlimited number of observations to a finite set of stable states based on its attractor properties, with gradient descent optimization used to learn the attractor space through end-to-end training. To ensure the EM model performs effectively in real-world scenarios, we have further developed EpiGibson, the first high-fidelity EM synthesis platform within a 3D physical simulation, built on OmniGibson (Li et al., 2022). Through our VLEM framework, we explore the construction of an EM model for a human-like agent operating in a physically realistic environment.

In our experiments, we tested the model on both pattern-based and simulation-based datasets. We reported its prediction performance during simulation, its robustness under noisy conditions, and its interpretability based on memory retrieval. Finally, we visualized the data in the simulation-based dataset to demonstrate the feasibility of applying our model in real-world scenarios. These results show that our VLEM framework can efficiently and reliably learn high-level temporal representations from an agent's sensory input in the environment.

In summary, we have these contributions: 1) We introduce the VLEM framework, which combines large-scale pretrained models with hippocampal attractor dynamics, leveraging AI's strong semantic understanding and biologically plausible episodic memory. 2) We present EpiGibson, a 3D physical simulation platform for generating episodic memory data, capable of simulating daily life and recording memory-related data. 3) We validate the robustness, interpretability, and real-world applicability of our framework through carefully designed experiments.

2. Methods

In daily life, humans receive sensory inputs from the environment, which the brain processes to form an understanding of the self-state. Here, we categorize the self-state into low-level action descriptions (e.g., putting vegetables in a pan) and high-level event descriptions (e.g., cooking in the kitchen

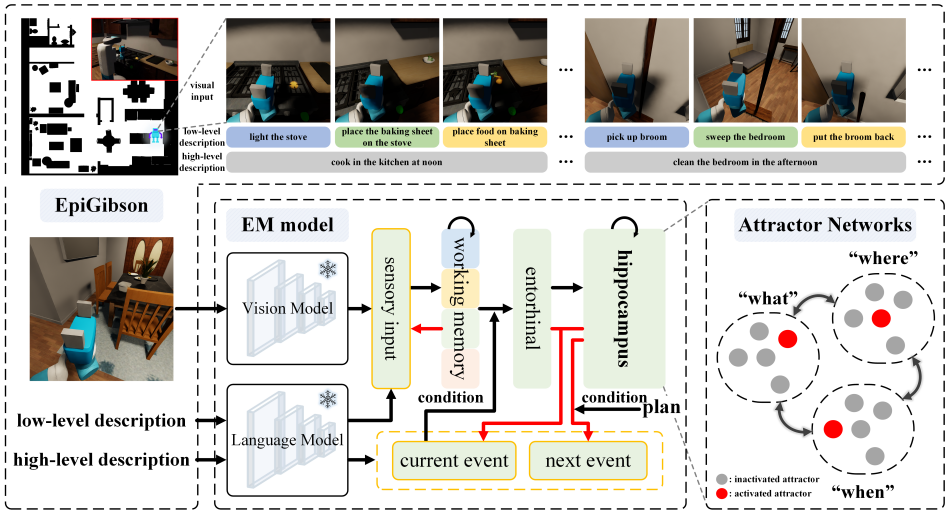

Figure 2: **Details in VLEM framework. 1) EpiGibson**: A simulation platform designed to evaluate the real-world applicability of our episodic memory model. It records continuous visual inputs and textual descriptions ("where", "what", "when", and action) by sequentially performing events and actions in a 3D virtual environment, creating realistic datasets for model evaluation. **2) EM Model**: Vision and language models process visual and text inputs from EpiGibson to generate sensory and event embeddings, which are stored in working memory. The most relevant slot is chosen to encode the current sensory input. This short-term state is then passed to the entorhinal cortex, which connects the neocortex and hippocampus, helping to maintain or predict events. Red arrows show the backward projection, and the yellow outline marks the target for loss calculation. **3) Attractor Networks**: The hippocampus is modeled as three interconnected attractor networks, representing the state of "where", "what" and "when" respectively, with event transitions represented by attractor state transitions.

at noon). More specifically, we present the Vision-Language Episodic Memory framework, which combines hippocampal attractor dynamics with large-scale pretrained models to create a biologically inspired episodic memory system within AI framework. This framework processes sensory input into semantic embeddings, incorporates working memory and the entorhinal cortex for short-term memory and information gateway respectively, and uses the hippocampus to store and retrieve experiences.

2.1 Data Synthesis

To better evaluate the robustness, interpretability, and real-world applicability of VLEM, we use two types of datasets: a pattern-based dataset, created with randomly generated patterns, and a simulation-based dataset, which records data from simulating events and actions with continuous visual input in our developed simulation environment, EpiGibson.

Dataset definition The episodic memory model continuously generates high-level event descriptions from visual input and low-level action descriptions. As shown in Fig 2, the episodic memory dataset is defined as a combination of continuous visual input and low-level/high-level descriptions. Specifically, the dataset consists of a sequence of L events $[E_1, E_2, \cdots, E_L]$, where each event E_i corresponds to a description of what happened, denoted as $\text{text}_{what,i}$. Each event E_i also contains an action sequence $\mathbf{A}_i = [A_{i,1}, A_{i,2}, ..., A_{i,na_i}]$, representing the execution process of event i. Each action $A_{i,j}$ is linked to a low-level description $\text{text}_{action,i,j}$, and includes a sensory sequence $\mathbf{S}_{i,j} = [S_{i,j,1}, S_{i,j,2}, ..., S_{i,j,ns_{ij}}]$, along with the corresponding "where" descriptions $\{\text{text}_{where,i,j,k} \mid 1 \le k \le ns_{ij}\}$ and "when" descriptions $\{\text{text}_{when,i,j,k} \mid 1 \le k \le ns_{ij}\}$. Overall, for each data sample $D_{i,j,k}$, sensory input is $S_{i,j,k}$, low-level description is $\text{text}_{action,i,j}$ and high-level description is combination of $\text{text}_{what,i}$, $\text{text}_{where,i,j,k}$ and $\text{text}_{when,i,j,k}$. Besides, there is a plan description text_{plan} summarizing the event sequence. These data samples are then flattened across the time dimension, with the data sample at time step t represented as: $\hat{D}_t = \{S_t, \text{text}_{action,t}, \text{text}_{where,t}, \text{text}_{what,t}, \text{text}_{when,t}\} = D_{i,j,k}$. The P_{plan} and $P_{when,t}$ are generated as random patterns that vary across different days and times of day, respectively. The other embeddings are then derived using pre-trained vision and language models:

$$P_{sensory,t} = \text{VisionModel}(S_t) \quad (1)$$

$$P_{X,t} = \text{LanguageModel}(\text{text}_{X,t}), \quad X \in \{where, what, action\} \quad (2)$$

2.2Vision-Languange Episodic Memory Framework

As shown in Fig. 2, our proposed EM framework consists of four modules: vision and language models, working memory, entorhinal cortex, and episodic memory. These modules represent key cognitive functions in the human brain. The modeling methods for each module are detailed below.

Vision and Language Models With the rapid development of AI, large-scale pre-trained models now achieve human-level semantic understanding in areas like vision and language by learning from vast amounts of data. While

these models differ from the brain's biological mechanisms, research shows they effectively map data to semantic space (Wang et al., 2022). As a result, we use pre-trained vision and language models to simulate how the brain encodes the semantic understanding of vision and language, aiding the learning of working memory and episodic memory.

Specifically, in our framework, we use the CLIP model (Schuhmann et al., 2022; Radford et al., 2021; Ilharco et al., 2021), which aligns images and text in a shared semantic space using contrastive learning. This model encodes the brain's semantic understanding of visual inputs and self-state descriptions, mirroring how the neocortex processes this information. As shown in Eq. 1-2, the vision and language models map image and text data into semantic space.

Working Memory Previous studies suggest that WM can be represented as controllable activity slots, and RNN slots have been used to model WM with gradient descent optimization(Whittington, Dorrell, Behrens, Ganguli, & El-Gaby, 2024). Building on this idea, we further improved the model by adding loss functions to better align with WM mechanisms and a cross-attention-based readout process.

We use N_{slots} unordered RNNs to model each slot, with $N_{slots} = 7$. Although some studies suggest a working memory capacity of 4 items (Cowan, 2001), we follow the more classical view of the "magical number seven" (Miller, 1956). These slots are connected to each other through fully connected layers. Let the states of working memory slots as $\mathrm{WM}_i \in \mathbb{R}^{N_{WM}}, 1 \le i \le N_{slots}$, where N_{WM} is dimension of each working memory slot. The working memory iteration is then defined as: $\mathrm{WM}_{i,t+1} = \tanh(W_{input,i}P_{input,t} + \sum_{j=1}^{N_{slots}} W_{i,j}\mathrm{WM}_{j,t} + b_i)$. Here, $P_{input,t} = \mathrm{concat}(P_{sensory,t}, P_{action,t}) \in \mathbb{R}^{N_S+N_A}$ represents the sensory input dimension to working memory, while $W_{input,i}, W_{i,j}$ are learnable weight matrices and b_i is the bias term.

Since all the RNN slots are identical in structure, it's important to prevent them from storing identical embeddings. We define the similarity between two slots as: $\mathrm{sim}_{WM}(i,j) = \frac{1}{\|WM_i\|_2 \|WM_j\|_2}\mathrm{WM}_i \cdot \mathrm{WM}_j^T$, where $\|\cdot\|_2$ denotes L2 norm. Because not all slots are activated at all times, we introduce an factor to represent the activation level of each slot, defined as: $a(i) = \| WM_i \|_2 / \sqrt{N_{WM}}$. To encourage diversity between slots while allowing some slots to remain inactive, we derive the working memory loss function as: $\mathcal{L}_{WM} = \frac{1}{N_{slots}(N_{slots}-1)} \sum_{i \neq j} a(i)a(j) \left|\mathrm{sim}_{WM}(i,j)\right| = \frac{1}{N_{WM}N_{slots}(N_{slots}-1)} \sum_{i \neq j} \left|\mathrm{WM}_i \cdot \mathrm{WM}_j^T\right|$.

Entorhinal Cortex The entorhinal cortex, acting as the gateway between the neocortex and hippocampus, collects information from working memory. While there are other pathways from specific brain regions to the entorhinal cortex, we simplify the model by ignoring these connections, as we assume that all necessary information can be encoded directly in working memory, theoretically from a modeling perspective. Therefore, the entorhinal state can be considered a readout of working memory, conditioned on the predicted current event embedding $\hat{P}_{curEvent,t} \in \mathbb{R}^{3N_P}$. The readout $\mathrm{Ento}_t \in \mathbb{R}^{N_{ento}}$ is calculated as: $\mathrm{Ento}_t = \mathrm{CrossAttn}(\hat{P}_{curEvent,t}, \mathbf{WM}_t)$. Here, $\mathrm{CrossAttn}(X_q, X_{kv}) = \mathrm{softmax}(c \cdot Q \cdot K^T) \cdot V$, $Q = X_q \cdot W_q$, $K = X_{kv} \cdot W_k$, $V = X_{kv} \cdot W_v$, and $c = N_{ento}^{-0.5}$ is a constant scale factor.

Attractor Networks In this study, we focus on modeling the CA3 region of the hippocampus. Based on the ability of episodic memory to stably recall the three attributes of an event—"where", "what", and "when"—we model each attribute with a separate attractor network and connect them to form a event attractor network. This allows us to build an attractor network that explicitly captures all the attributes of an event, offering a comprehensive model of the hippocampus.

Let the attractor states of "where", "what" and "when" be denoted as $\mathrm{EM}_{where}, \mathrm{EM}_{what}, \mathrm{EM}_{when} \in \mathbb{R}^{N_{EM}}$, where N_{EM} is the dimension of each attractor. Then attractor state of event is concatenation of them: $\mathrm{EM}_{event} = \mathrm{concat}(\mathrm{EM}_{where}, \mathrm{EM}_{what}, \mathrm{EM}_{when}) \in \mathbb{R}^{3N_{EM}}$. Following the principles of the Hopfield network (Hopfield, 1984), we treat an RNN with symmetric recurrent weights as an attractor network. Let $\mathrm{attrs} = \{where, what, when\}$, the iteration for episodic memory is then given by:

$$\mathrm{EM}_{X,t,0} = \tanh\Big(\sum_{X' \in \mathrm{attrs}} W_{X,X'}\mathrm{EM}_{X',t} + W_{Ento}\mathrm{Ento}_t\Big)$$
$$\mathrm{EM}_{X,t,k+1} = \tanh\Big(\sum_{X' \in \mathrm{attrs}} W_{X,X'}\mathrm{EM}_{X',t,k}\Big)$$
$$\mathrm{EM}_{X,t+1} = \mathrm{EM}_{X,t,K}, \quad X \in \mathrm{attrs}$$

where $W_{X,X'}$ is a symmetric matrix, $W_{X,X'} = W_{X',X}$, and K is the number of additional self-iteration steps for the attractor network.

Unlike previous approaches that rely on Hebbian learning, our framework uses gradient descent for model training, which means there is no need to predefine the attractor states. Instead, they can be learned directly from the data. In this case, we assume that the target attractor states can be predicted directly from the event states as: $\hat{P}_{EM,X,t} = \tanh(W_{event,X}P_{event,t}), X \in \mathrm{attrs}$. In order to automatically learn these states as attractors, we define the attractor loss function as: $\mathcal{L}_{EM1} = \sum_{X \in \mathrm{attrs}} \| \hat{P}_{EM,X} - \tanh(W_{X,X}\hat{P}_{EM,X}) \|_1 + \| \hat{P}_{EM,event} - \tanh(W_{event,event}\hat{P}_{EM,event}) \|_1$, where $\hat{P}_{EM,event}$ represents the combined state of the three attractor states, and $W_{event,event}$ represents the recurrent connection weights after combining the three RNNs into one. In addition, to prevent all attractors from converging to the same state, we introduce an episodic memory contrastive loss as: $\mathcal{L}_{EM2} = \frac{1}{N_{event}(N_{event}-1)} \sum_{i \neq j} \mathrm{sim}_{EM}(i,j)$, where N_{event} denote the number of unique events and $\mathrm{sim}_{EM}(i,j) = \hat{P}_{EM,event,i} \cdot \hat{P}_{EM,event,j}^T / (\| \hat{P}_{EM,event,i} \|_2 \cdot \| \hat{P}_{EM,event,j} \|_2)$. Further, in order to ensure that neurons in attractor states are either fully firing or not firing at all, meaning their activation values are 1 or -1, thus enhancing the capacity of the network, we also introduce the following loss constraint: $\mathcal{L}_{EM3} = - \| \hat{P}_{EM,event} \|_1$.

Backward Projection 1) From working memory to sensory input. To ensure that semantic information is encoded in working memory, we project the state of working memory back onto sensory input for prediction. Since working memory has multiple slots, with each slot encoding different semantic information, we designate the nearest slot as the one encoding the semantic information of the current sensory input. Therefore, the loss of encoding accuracy in working memory is as: $\mathcal{L}_{input} = \| \tanh(W_{input\_WM} P_{input}) - \mathrm{WM}_k \|_2^2 + \| W_{WM\_input} \mathrm{WM}_k - P_{input} \|_2^2$, where $k = \arg\min_k \| \tanh(W_{input\_WM} P_{input}) - \mathrm{WM}_k \|_2$. 2) From hippocampus to events. To understand the current state of an event and predict future events, we project from the hippocampus back to the encoding of the event. Since the attractor states in the hippocampus do not directly encode semantic information, we predict the current understanding of the event by using the condition entorhinal state from the hippocampus state, and we predict the next event using the condition plan embedding. Thus we have: $\hat{P}_{event} = W_{EM\_event}\mathrm{EM}_{event} + W_{ento\_event}\mathrm{Ento}$ and $\hat{P}_{nxtEvent} = W_{EM\_nxtEvent}\mathrm{EM}_{event} + W_{plan\_event}P_{plan}$. Then prediction loss of events is derived as: $\mathcal{L}_{event} = \| P_{event} - \hat{P}_{event} \|_2^2 + \| P_{nxtEvent} - \hat{P}_{nxtEvent} \|_2^2$.

2.3 Training Strategy

We perform end-to-end training to optimize the EM model. The pretrained vision and language models are frozen, while all other weights are learnable. The loss used to optimize the entire model is combination of predefined losses:

$$\mathcal{L} = \mathcal{L}_{WM} + \mathcal{L}_{input} + \mathcal{L}_{event} + \alpha(\mathcal{L}_{EM1} + \mathcal{L}_{EM2} + \mathcal{L}_{EM3})$$

where α is the scale of the EM-related loss. In order to enhance the learning performance of episodic memory, α will gradually increase as the training progresses.

3. Experiments

To more thoroughly assess the VLEM model's performance, we tested three attractor models in our experiments: (1) the Hopfield Network (Hopfield, 1984), (2) VLEM(merged), a version of VLEM with a single combined attractor, and (3) VLEM, the full VLEM model. In VLEM(merged), the three separate attractors for "where", "what", and "when" are combined into one, removing the explicit distinction between these categories. The Hopfield Network further extends VLEM(merged) by replacing the learned attractor weights with new ones through Hebbian learning.

3.1 Datasets and Metrics

Pattern-based Synthesis In pattern-based synthesis, we construct a random tree graph where each node represents a unique location ("where"). A virtual agent completes its action list by navigating through the tree to execute actions at corresponding locations. The agent begins at the location of the first action. Once the action is completed, it moves along a path to the next location to perform the following action. Transitioning to the next location and completing an action takes random time. For each time step, patterns for "where", "what" and "when" are recorded as P_{where}, P_{what} and P_{when}, respectively, where $P_{where}, P_{what}, P_{when} \in \mathbb{R}^{T \times N_P}$, with T being the total number of time points and N_P being the dimension of each pattern. The embeddings for sensory input $P_{sensory,t} \in \mathbb{R}^{N_S}$ and low-level action description $P_{action,t} \in \mathbb{R}^{N_A}$ are calculated using two randomly initialized fully connected layers by passing $P_{event,t}$. The plan embedding is the average of all event embeddings: $P_{plan} = \mathrm{mean}(P_{event}) \in \mathbb{R}^{3N_P}$.

Let $N_{what}, N_{where}, N_{when}, N_{action}$ represent the number of unique patterns for "what", "where", "when" and actions, respectively. We create datasets with various settings for $(N_{what}, N_{where}, N_{when}, N_{action})$: "large" (50,20,10,100), "medium" (20,10,5,50) and "small" (10,5,3,20).

Simulation-based Synthesis To further assess the real-world applicability of our model, we developed **EpiGibson**, the first episodic memory physical simulation platform, based on OmniGibson (Li et al., 2022). Like the pattern-based dataset, the simulation-based dataset is created by having the agent perform actions sequentially in each event, with each action's code manually programmed. During these interactions, the robot's visual inputs, along with the corresponding low-level and high-level textual descriptions, are recorded at each time step. As a result, the data format from the simulation-based dataset is the same as that of the pattern-based dataset. The key difference is that in the simulation-based synthesis, the agent continuously interacts with a 3D virtual environment within a physical simulation, providing a high-fidelity reproduction of human daily life while capturing the required data.

Specifically, as shown in Fig. 2, at each time step t, the data sample includes visual input $S_t \in \mathbb{R}^{H \times W \times 3}$, and text descriptions for "where", "what", "when" and action, denoted as $\mathrm{text}_{where,t}$, $\mathrm{text}_{what,t}$, $\mathrm{text}_{when,t}$ and $\mathrm{text}_{action,t}$ respectively. Furthermore, the event description is defined as a combination of "where", "what" and "when" descriptions. The action descriptions are then summarized to produce the plan description text_{plan} by ChatGPT-4o. Finally, all patterns are derived using the equations in Eq. 1-2.

Metrics We used two metrics, MSE and correlation, to evaluate our predictions. We tested the predictions for the current event, the next event, and the sensory input separately.

Implementation Details The learning rate starts at 2e-4, decaying every 500 steps, with training limited to 5,000 steps. All training and inference were done on a NVIDIA A800 GPU. The code is available at: `https://github.com/fudan-birlab/VLEM`.

3.2 In-simulation Accuracy

To evaluate how well our model understands the state during agent simulation, we first assess its ability to predict sensory input, checking if the working memory has correctly stored the current sensory information. Further, we also evaluate the model's predictions of the current and next events to see if the

		pattern-based dataset (large)						simulation-based dataset					
		VLEM		VLEM(merged)		Hopfield		VLEM		VLEM(merged)		Hopfield	
	σ	corr	MSE	corr	MSE	corr	MSE	corr	MSE	corr	MSE	corr	MSE
curEvent	0	**0.999**	**0.001**	0.922	0.136	0.443	0.851	**0.657**	**0.670**	0.558	0.804	0.366	1.044
	1	**0.961**	**0.077**	0.743	0.440	0.387	0.897	**0.674**	**0.618**	0.570	0.746	0.377	1.030
nxtEvent	0	**0.736**	**0.450**	0.393	0.926	0.116	1.194	**0.366**	**1.044**	0.102	1.179	-0.007	1.341
	1	**0.729**	**0.461**	0.177	1.314	0.105	1.267	**0.377**	**1.030**	0.100	1.181	-0.028	1.365
sensory	0	0.939	0.117	**0.941**	**0.113**	0.622	0.649	**0.944**	**0.108**	0.940	0.115	0.763	0.418
	1	0.871	0.240	**0.882**	**0.221**	0.577	0.718	0.909	0.174	**0.915**	**0.162**	0.738	0.456

Table 1: **Evaluation results for accuracy across synthetic datasets.** We evaluated our models on both pattern-based and simulation-based datasets, using metrics to assess predictions for the current event, next event, and sensory input. Our VLEM significantly outperforms Hopfield and demonstrates robustness even under Gaussian noise with a standard deviation of $\sigma = 1$. Bold values indicate the best results.

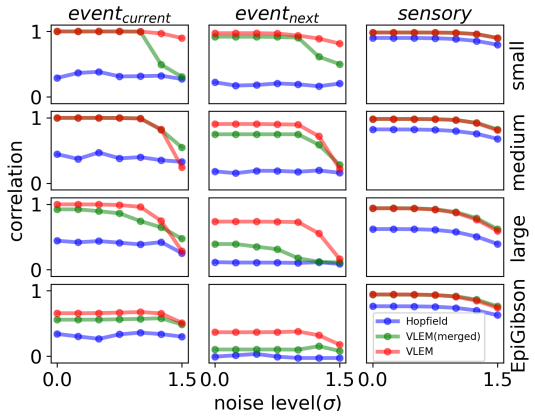

Figure 3: **Evaluation on various datasets and noise levels.**

episodic memory retains the event details based on the input stimuli. Specifically, at each time step, the model receives an input $P_{sensory,t}$, and we test whether its predictions $\hat{P}_{sensory,t}$, $\hat{P}_{event,t}$ and $\hat{P}_{nxtEvent,t}$ are accurate. The results are shown in Tab. 1. VLEM significantly outperforms the Hopfield Network and surpasses VLEM(merged) on most metrics.

3.3 Robustness

To test the model's robustness, we evaluate its performance under varying levels of dataset and noise. As shown in Fig. 3, for event prediction, as the number of events increases (from "small" to "large"), VLEM's advantage over competitors becomes more evident, especially in predicting the next event. This suggests that explicitly splitting event elements ("where", "what" and "when") and forming separate attractor networks increases the overall memory capacity, making the system more robust to larger data scales and capable of adapting to more complex environments. Additionally, we tested the model's robustness to input noise by adding varying levels of Gaussian noise ($\sigma : 0 \rightarrow 1.5$). The results show that VLEM is highly robust to input noise. Even with $\sigma = 1$, VLEM maintains small accuracy losses (see tab. 1). The stability in sensory prediction accuracy highlights the robustness of the working memory model, with only limited impact from changes in episodic memory modeling. Overall, the experimental results demonstrate that VLEM is more robust to complex events and noisy inputs.

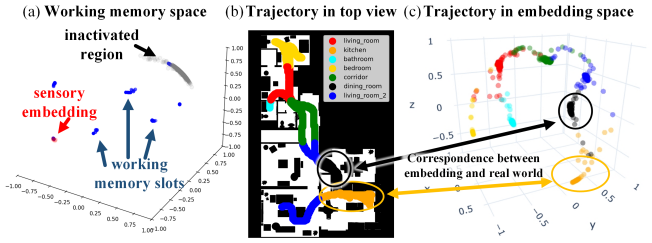

Figure 4: **Visualization for working memory and episodic memory.** (a) Working memory slots either encode distinct semantics or remain inactive, with one specific slot effectively capturing the sensory input. (b) Real trajectory of agent in simulation environment. (c) The event ("where") embeddings, derived from episodic memory, shows the agent's trace, with different colors representing different locations. This structure closely matches the real-world map.

3.4 Real-world Applicability and Interpretability

To test our model's performance in real-world applications, we collected episodic memory data from a physical simulation using EpiGibson. In this setup, our visual input consists of real images, and the self-state descriptions are text labels corresponding to the agent's state. We report the model's basic metrics on this dataset and visualize the results, comparing them with the actual physical labels. Results are shown in Tab. 1. Moreover, we use CEBRA (Schneider, Lee, & Mathis, 2023) for unsupervised encoding of neuronal dynamics into 3D space, as shown in Fig. 4. The results show that VLEM accurately learns spatial relationships consistent with the real world, effectively demonstrating the model's interpretability.

4. Conclusion

In this paper, we propose the Vision-Language Episodic Memory (VLEM) framework, which combines large-scale pretrained models with hippocampal attractor dynamics. The framework leverages AI's semantic understanding alongside the stability and interpretability of hippocampal dynamics, enabling reliable storage and retrieval of episodic experiences. We also present EpiGibson, a 3D simulation platform for generating episodic memory data, and show how the framework is robust, interpretable, and applicable in real-world scenarios. Our work advances biologically inspired memory models and their integration into AI systems.

References

Allen, T. A., & Fortin, N. J. (2013). The evolution of episodic memory. *Proceedings of the National Academy of Sciences*, *110*(supplement_2), 10379–10386.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological review*, *84*(5), 413.

Bärmann, L., & Waibel, A. (2022, June). Where did i leave my keys? - episodic-memory-based question answering on egocentric videos. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops* (p. 1560-1568).

Boyle, A., & Blomkvist, A. (2024). Elements of episodic memory: insights from artificial agents. *Philosophical Transactions B*, *379*(1913), 20230416.

Chandra, S., Sharma, S., Chaudhuri, R., & Fiete, I. (2025, Jan 15). Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*. Retrieved from https://doi.org/10.1038/s41586-024-08392-y doi: 10.1038/s41586-024-08392-y

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. doi: 10.1017/S0140525X01003922

Criss, A. H., & Howard, M. W. (2015). Models of episodic memory. *The Oxford handbook of computational and mathematical psychology*, *399*, 165–183.

Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., & Parikh, D. (2022, June). Episodic memory question answering. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 19119-19128).

Digutsch, J., & Kosinski, M. (2023, March). Overlap in meaning is a stronger predictor of semantic activation in gpt-3 than in humans. *Scientific reports*, *13*(1), 5035. Retrieved from https://europepmc.org/articles/PMC10050205 doi: 10.1038/s41598-023-32248-6

Eğilmez, E. (2015). *The role of episodic memory in artificial intelligence*. Unpublished master's thesis, Middle East Technical University.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological review*, *91*(1), 1.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., ... Malik, J. (2022, June). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 18995-19012).

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, *11*(7), 299–306.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088-3092. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.81.10.3088 doi: 10.1073/pnas.81.10.3088

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., ... Schmidt, L. (2021, July). *Openclip*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5143773 (If you use this software, please cite it as below.) doi: 10.5281/zenodo.5143773

Jeong, W., Chung, C. K., & Kim, J. S. (2015). Episodic memory in aspects of large-scale brain networks. *Frontiers in human neuroscience*, *9*, 454.

Jockel, S., Weser, M., Westhoff, D., & Zhang, J. (2008). Towards an episodic memory for cognitive robots. In *Proc. of 6th cognitive robotics workshop at 18th european conf. on artificial intelligence (ecai)* (pp. 68–74).

Krotov, D. (2023, 05). A new frontier for hopfield networks. *Nature Reviews Physics*, *5*. doi: 10.1038/s42254-023-00595-y

Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*, 40 - 80. Retrieved from https://api.semanticscholar.org/CorpusID:221495897

Laird, J. E. (2008). Extending the soar cognitive architecture. In *Agi* (Vol. 171, pp. 224–235).

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), 1–64.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160.

Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., ... Fei-Fei, L. (2022). BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In *6th annual conference on robot learning*. Retrieved from https://openreview.net/forum?id=_8DoIe8G3t

Malmberg, K. J., Holden, J. E., & Shiffren, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 319.

Miller, G. A. (1956, March). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*(2), 81-97. Retrieved from http://www.musanim.com/miller1956/

Neisser, U. (1992). *Phantom flashbulbs: False recollections*

of hearing the news about challenger. Affect and accuracy in recall/Cambridge University Press.

Nicholas, J., Daw, N. D., & Shohamy, D. (2022). Uncertainty alters the balance between incremental learning and episodic memory. *Elife*, *11*, e81679.

Penaud, S., Yeh, D., Gaston-Bellegarde, A., & Piolino, P. (2023). The role of bodily self-consciousness in episodic memory of naturalistic events: an immersive virtual reality study. *Scientific Reports*, *13*(1), 17013.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Icml.*

Rolls, E. T., & Treves, A. (2024). A theory of hippocampal function: new developments. *Progress in Neurobiology*, 102636.

Rolls, T. (2018). The storage and recall of memories in the hippocampo-cortical system. *Cell and tissue research*, *373*(3), 577–604.

Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*. doi: https://doi.org/10.1038/s41586-023-06031-6

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., ... Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth conference on neural information processing systems datasets and benchmarks track.* Retrieved from https://openreview.net/forum?id=M3Y74vmsMcY

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual review of psychology*, *44*(1), 453–495.

Tulving, E. (1972). Episodic and semantic memory. *Organization of memory/Academic Press*.

Tulving, E. (2002). Episodic memory: From mind to brain [Journal Article]. *Annual Review of Psychology*, *53*(Volume 53, 2002), 1-25. Retrieved from https://www.annualreviews.org/content/journals/10.1146/annurev.psych.53.100901.135114 doi: https://doi.org/10.1146/annurev.psych.53.100901.135114

Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., ... Gao, W. (2022). Large-scale multi-modal pre-trained models: A comprehensive survey. Retrieved from https://github.com/wangxiao5791509/MultiModal_BigModels_Survey

Whittington, J., Dorrell, W., Behrens, T., Ganguli, S., & El-Gaby, M. (2024, 11). A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. *Neuron*. doi: 10.1016/j.neuron.2024.10.017

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: an emotional binding account. *Trends in cognitive sciences*, *19*(5), 259–267.