Towards Continually-Evolving AI: Selective and Expandable Multimodal Memory

Jaehong Yoon
The College of Computing and Data Science
Nanyang Technological University
Block N3-02c-97, 50 Nanyang Avenue, Singapore
jaehong.yoon@ntu.edu.sg

A. Handling Long-Horizon Dynamics with Continual Multimodal Models. In a world where data, tasks, and environments are constantly changing, models trained on static datasets quickly fall out of sync with real-world needs. Continual learning in multimodal settings is essential, allowing AI systems to flexibly integrate diverse inputs, such as text, images, and video, while adapting to new contexts. These models must steadily expand their knowledge, balancing the retention of valuable information, refinement of outdated knowledge, and integration of new insights. This has been a key focus of my research (Yoon et al. 2018, 2020, 2022; Maharana et al. 2025). Given the continuous expansion of the instruction-tuning datasets, I led **Adapt-∞** (Maharana et al. 2025), a framework that dynamically selects informative data to balance continual learning and efficiency, enabling practical deployment in resource-constrained scenarios (See Figure 1). Additionally, models often need to acquire new information while discarding outdated knowledge. To evaluate such knowledge-refinement capabilities over time, **EvolvingQA** (Kim et al. 2024) designed a new benchmark assessing LLMs' temporal adaptation abilities. The existing methods struggle to remove outdated information, highlighting the challenges in rectifying and updating knowledge.

B. Dynamic Multimodal Memory Agent for Long Video Reasoning. Recent advances in video large language models have demonstrated strong capabilities in understanding short clips. However, scaling them to hours- or days-long videos remains highly challenging due to limited context capacity and the loss of critical visual details during abstraction. Existing memory-augmented methods mitigate this by leveraging textual summaries of video segments, yet they rely heavily on text and fail to leverage visual evidence when reasoning over complex scenes. Moreover, retrieving from fixed temporal scales further limits their ability to capture events that span variable durations. To address this, I introduce WorldMM (Yeo et al. 2025), a novel multimodal memory agent that constructs and retrieves from multiple complementary memories, encompassing both textual and visual representations. WorldMM comprises three types of memory: episodic memory, which indexes factual events

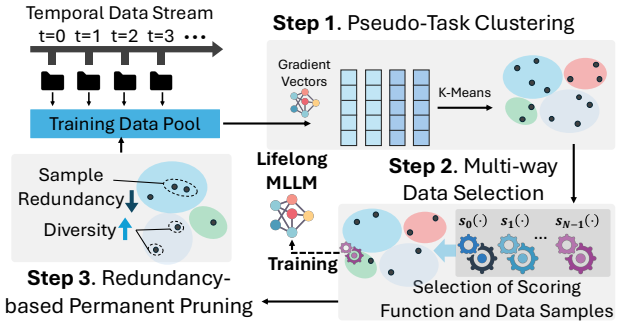

Figure 1: Lifelong Multimodal Learning via Selective Knowledge Accumulation and Expansion

across multiple temporal scales, semantic memory continuously updates high-level conceptual knowledge, and visual memory preserves detailed information about scenes. During inference, an adaptive retrieval agent iteratively selects the most relevant memory source and leverages multiple temporal granularities based on the query, continuing until it determines that sufficient information has been gathered.

C. Self-training to Adapt to Unseen Tasks using Auto-generated Training Sources. Beyond distributional gaps between training and test data, these systems often need to adapt to untrained or unfamiliar tasks after deployments. Self-training provides an effective solution by leveraging synthetic datasets to improve model capabilities in handling OOD tasks, addressing data scarcity while ensuring adaptability to evolving challenges across diverse environments. I introduced **SELMA** (Li et al. 2024) that enhances text-to-image (T2I) generation by fine-tuning automatically generated multi-skill image-text datasets. With skill-specific LoRA expert learning and merging, SELMA reduces errors in object relationships and missing details, improving text-image alignment across diverse prompts. In a similar vein, I proposed **EnvGen** (Zala et al. 2024) for continual embodied agent learning. This approach uses LLMs to dynamically generate and adapt to training environments, enabling embodied agents to learn and generalize in real-world settings more efficiently, especially for long-horizon tasks and open-world reasoning.

References

Kim, Y.; Yoon, J.; Ye, S.; Bae, S.; Ho, N.; Hwang, S. J.; and Yun, S.-Y. 2024. Carpe Diem: On the Evaluation of World Knowledge in Lifelong Language Models. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*.

Li, J.; Cho, J.; Sung, Y.-l.; Yoon, J.; and Bansal, M. 2024. SELMA: Learning and Merging Skill-Specific Text-to-Image Experts with Auto-Generated Data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Maharana, A.; Yoon, J.; Chen, T.; and Bansal, M. 2025. Adapt-: Scalable Lifelong Multimodal Instruction Tuning via Dynamic Data Selection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yeo, W.; Kim, K.; Yoon, J.; and Hwang, S. J. 2025. WorldMM: Dynamic Multimodal Memory Agent for Long Video Reasoning.

Yoon, J.; Kim, S.; Yang, E.; and Hwang, S. J. 2020. Scalable and Order-robust Continual Learning with Additive Parameter Decomposition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yoon, J.; Madaan, D.; Yang, E.; and Hwang, S. J. 2022. Online Coreset Selection for Rehearsal-based Continual Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yoon, J.; Yang, E.; Lee, J.; and Hwang, S. J. 2018. Lifelong Learning with Dynamically Expandable Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zala, A.; Cho, J.; Lin, H.; Yoon, J.; and Bansal, M. 2024. EnvGen: Generating and Adapting Environments via LLMs for Training Embodied Agents. In *Conference on Language Modeling (COLM)*.