ADeltaM: An Exploratory Counterfactual Delta-Memory Interface for Egocentric Agents

Liyang Ruan
Department of Microelectronics
Southern University of Science and Technology
12311411@mail.sustech.edu.cn

Jiahao Cao
Department of Computer Science
Jilin University
caojh2423@mails.jlu.edu.cn

Abstract

Foundation-model embodied agents need short-horizon transition memories that estimate how turning, moving, or interacting changes the egocentric state. We study ADeltaM, an exploratory action-conditioned delta-memory interface for egocentric world modeling. ADeltaM stores latent state changes keyed by state and action, retrieves relevant deltas, and composes them with the current state for one-step and rollout prediction. We intentionally scope the contribution as a 4-page workshop diagnostic rather than a complete world model: our experiments use a procedural RGB-D gridworld, not frozen foundation-model features or standard embodied benchmarks. We evaluate persistence, linear residual, kNN delta retrieval, no-memory MLP, and GRU dynamics baselines, retain DANN/CORAL for continuity, and introduce a same-state different-action counterfactual protocol. Across three paired seeds, ADeltaM variants improve rollout and counterfactual delta error over most baselines; however, a linear residual model is close, and our study does not yet prove general embodied-agent utility. We present the result as a concrete memory interface and evaluation protocol for future foundation-model agent integration.

1. Introduction

Embodied foundation-model agents combine grounded perception, memory, planning, and action. A vision-language or vision-language-action model can propose subgoals, but it still needs a local transition model to score candidate actions. In egocentric interaction, useful counterfactual evidence is action specific: turning left rather than right changes target visibility, interacting can change object-state flags, and moving forward changes depth.

We study a compact transition-memory primitive for that layer. Instead of storing full future states, ADeltaM stores latent deltas associated with executed actions. At test time, a state-action query retrieves prior deltas and composes them with the current latent state. The intended role is not to replace RSSM/Dreamer-style world models [6–8], but to provide an inspectable episodic memory component that a foundation-model planner could call for short-horizon action scoring. In that pipeline, a VLM/VLA agent supplies candidate actions and semantic goals, while the memory module supplies local transition evidence.

We narrow the claim to match the evidence. The experiment is a small procedural RGB-D gridworld with engineered latent observations. It is aligned with the workshop theme of world models, memory, and interaction, but does not yet use VLM/VLA planners, frozen visual foundation features such as CLIP [11], DINOv2 [10], SigLIP [15], or V-JEPA [1], or benchmarks such as Habitat [12] and EmbodiedBench [14]. We frame the paper as an exploratory short-paper contribution: a delta-memory interface, stronger diagnostic baselines, and a same-state different-action counterfactual test.

2. Related Work

Latent world models learn predictive state dynamics for planning and control, from World Models and PlaNet to Dreamer-style agents [5–8]. These methods motivate stronger future comparisons. ADeltaM is narrower: it externalizes reusable transition fragments in episodic memory rather than learning a full recurrent latent simulator, making retrieval and inspection part of the modeling interface.

Action-conditioned visual prediction models study how controls affect future observations [3]. Retrieval and memory-based agents store experience for planning, long-horizon reasoning, or manipulation; recent embodied-agent systems increasingly combine foundation perception and explicit memory [2, 9]. ADeltaM differs in the unit of storage: it stores local transition deltas, not full frames, language memories, or value traces. This is also distinct from generative delta-token world-modeling directions: our focus is retrieval-based transition memory for action scoring, not high-fidelity video generation.

3. Method

Let an egocentric trajectory be (o_t, a_t, o_{t+1}), where o_t is an RGB-D observation and a_t is a discrete action. An encoder maps observations into latent states $z_t = f_\theta(o_t)$. In our diagnostic implementation, z_t is an engineered RGB-D latent vector; replacing this representation with frozen foundation-model features is a key next step.

ADeltaM stores latent transition deltas,

$$\Delta z_t = z_{t+1} - z_t, \tag{1}$$

as entries $m_i = (k_i, a_i, \Delta z_i, c_i)$ containing a state key, action, delta, and metadata. Given query (z_t, a_t), the memory retrieves action-compatible neighbors and predicts

$$\hat{z}_{t+1} = z_t + \sum_{i \in \mathcal{N}_K(z_t, a_t)} w_i \Delta z_i. \tag{2}$$

Weights are computed from state similarity with an action-match bias. The default diagnostic configuration uses a 48-dimensional latent, $K = 14$ retrieval neighbors, memory capacity 768, action bias 6.0, and softmax temperature 0.04. A write policy commits a new delta when novelty or prediction error is high. We evaluate basic, tuple-entry, learned-write, and retrieval-focused variants.

4. Experimental Protocol

The environment is a 10×10 procedural egocentric RGB-D gridworld with local depth rays, RGB context, orientation, goal/object offsets, and interaction flags. The action set is forward, turn-left, turn-right, and interact. We train on 1152 transitions and evaluate 32 rollouts of horizon 16 over seeds $\{0, 1, 2\}$.

We compare against two core ablations and seven diagnostic baselines. Ablations remove action conditioning or store full states rather than deltas. Baselines include persistence, an action-conditioned linear residual model, kNN delta retrieval, a no-memory MLP transition predictor, a GRU history model, DANN [4], and CORAL [13]. DANN/CORAL remain weak world-model baselines, but are retained for continuity and are reported separately from the compact-memory argument; the added linear, kNN, MLP, and GRU baselines are the fairer transition comparisons.

All parametric baselines use the same small-data diagnostic budget and are not exhaustively tuned state-of-the-art dynamics models. Our goal is to test whether the memory interface is competitive in a constrained exploratory setting, not to claim superiority over tuned RSSM, transformer, or Dreamer-style systems.

The primary metric is open-loop rollout MSE. To test the counterfactual motivation directly, we also evaluate same-state different-action prediction on 48 held-out states: from the same state, each candidate action is rolled once in the environment and by the model. We report counterfactual delta MSE and action-match accuracy, where each predicted action delta is matched to the closest true action delta.

Table 1. Main diagnostic results. Mean ± std over three paired seeds. "Slots/params" denotes stored memory slots for memory methods and learned parameters for parametric models.

Method	Rollout ↓	CF-Δ ↓	Act. acc. ↑	Slots/params
ADeltaM-Tuple	**0.0207±0.0011**	**0.0073±0.0013**	0.807±0.019	832
ADeltaM	0.0208±0.0012	0.0074±0.0013	0.807±0.019	768
Linear residual	0.0215±0.0016	0.0073±0.0014	**0.821±0.025**	9408
kNN delta	0.0300±0.0008	0.0148±0.0020	0.748±0.010	768
Full-state memory	0.0325±0.0012	0.0135±0.0010	0.807±0.028	768
DirectMLP	0.0339±0.0009	0.0128±0.0014	0.733±0.019	24016
No-action delta	0.0388±0.0016	0.0298±0.0012	0.250±0.000	768
CORAL	0.0306±0.0008	0.0110±0.0011	0.750±0.019	24016
DANN	0.0315±0.0010	0.0112±0.0012	0.750±0.009	36898
GRU	0.0490±0.0005	0.0240±0.0015	0.399±0.020	61200
Persistence	0.0639±0.0028	0.0438±0.0014	0.250±0.000	0

5. Results

Table 1 and Fig. 1 support a narrower claim than a full embodied world-model paper. ADeltaM variants achieve the best rollout and counterfactual delta MSE among the compact memory methods and outperform kNN retrieval, full-state memory, MLP, GRU, DANN/CORAL, persistence, and the no-action ablation. The no-action ablation collapses to chance-level action matching, indicating that the counterfactual protocol is sensitive to action information. The main evidence is therefore interface-level: ADeltaM is a retrieval-based compact memory with competitive error under a small-slot budget.

The most important caveat is the linear residual baseline. It is close on rollout MSE (0.0215 vs. 0.0208 for the basic ADeltaM) and slightly better on action-match accuracy. These results do not prove that retrieval memory is always superior to simple action-conditioned residual prediction. The contribution should be read as an interface and protocol result—competitive error with compact retrieval, inspectability, and same-state action testing—rather than a decisive modeling win over simple residual dynamics.

Figure 2 visualizes compactness, but the x-axis mixes stored slots for memory methods with parameter counts for neural predictors. We treat it as a footprint sanity check only, not as a byte-level memory, FLOP, latency, or deployment-efficiency measurement.

Figure 3 illustrates why the interface is useful even when the visual proxy is simple. The retrieved-delta prediction can be inspected as a local transition explanation: the model exposes which state components are expected to move, which retrieved memories supported the move, and where absolute error accumulates after composition. This is different from treating rollout error as a single scalar. For a workshop setting, this inspectability is part of the motivation for studying delta memory before scaling to richer simulators.

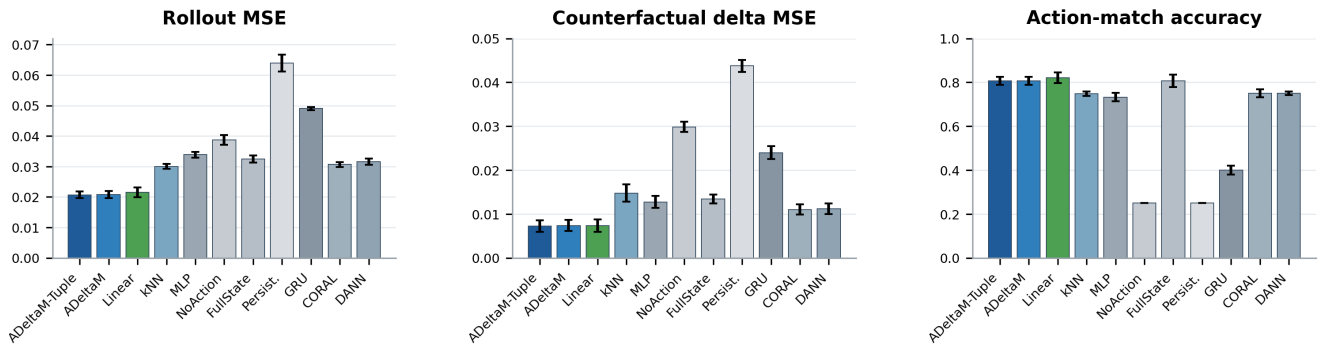

Figure 1. Clean diagnostic comparison across rollout error, same-state counterfactual delta error, and action-delta matching accuracy. The main signal is family-level: ADeltaM variants are strong among compact memory methods, while the linear residual baseline is close and prevents an overbroad novelty claim.

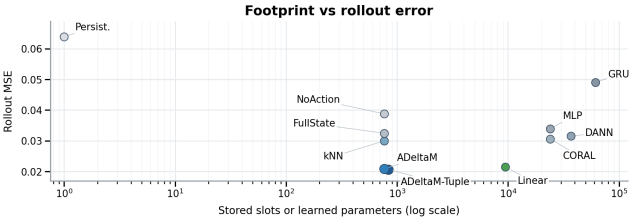

Figure 2. Footprint/error diagnostic. ADeltaM uses hundreds of slots, while parametric MLP/GRU baselines use tens of thousands of parameters in this implementation. This is a diagnostic footprint, not a hardware-normalized efficiency claim.

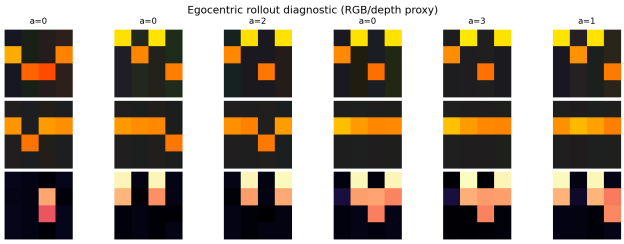

Figure 3. Qualitative rollout proxy. Top: target observation proxy. Middle: ADeltaM prediction. Bottom: absolute error. The figure is meant for failure inspection, not photorealistic world modeling.

6. Foundation-Feature Extension

The current implementation does not yet use a foundation model. We view the relevance as architectural: ADeltaM defines a small transition-memory API that a VLM/VLA planner could query when scoring candidate actions. A planner may propose actions in language or action tokens; the memory module estimates short-horizon physical consequences from egocentric state. This role is complementary to semantic reasoning and to memory-augmented embodied-agent systems [2, 9].

Concretely, for a candidate action set $\mathcal{A}_t$, the planner can call ADeltaM once per action and receive $(\hat{z}^a_{t+1}, u^a, \mathcal{R}^a)$, where $\mathcal{R}^a$ is the supporting memory neighborhood and u^a is computed as the weighted neighbor-distance spread plus the weighted variance of retrieved deltas. A lightweight action score can combine semantic desirability from the foundation model with transition evidence,

$$S(a) = S_{\mathrm{VLM}}(a) - \lambda u^a - \gamma d(g, \hat{z}^a_{t+1}), \tag{3}$$

where g is a goal or subgoal representation and $d(\cdot,\cdot)$ is a task-specific distance. This formulation is intentionally modular: the foundation model remains responsible for perception and goal semantics, while ADeltaM supplies a short-horizon physical prior over egocentric state changes.

A direct next-step extension is to replace engineered latents with frozen visual foundation features while keeping the same transition-memory protocol. Each gridworld observation can be rendered as a small egocentric RGB panel, optionally concatenated with a depth colormap, then embedded by a frozen CLIP, DINOv2, or SigLIP image encoder [10, 11, 15]. The memory would store deltas in that frozen feature space:

$$\Delta\phi_t = \phi(o_{t+1}) - \phi(o_t), \tag{4}$$

where ϕ is fixed and only the memory/retrieval components change. The same rollout and counterfactual metrics would then test whether ADeltaM remains competitive when the latent state is no longer hand-engineered. This experiment is intentionally compatible with the current code: we would change the state extractor, not the memory algorithm, and report a direct side-by-side table for engineered, CLIP, DINOv2, and SigLIP latents.

We would not reduce this analysis to a single headline number. Instead, we would test whether action conditioning still separates turn-left, turn-right, forward, and interact deltas in frozen feature space; whether retrieved neighbors remain semantically coherent under viewpoint changes; and

whether a linear residual model still nearly matches memory when features are fixed. We would expect this extension to be most informative on three comparisons. First, ADeltaM should retain an advantage over no-action delta memory if action labels help disambiguate feature changes. Second, the gap to linear residual prediction may shrink or widen depending on whether foundation features linearize local dynamics. Third, compact memory may become more attractive because frozen features can be high dimensional and expensive to model parametrically. Until that experiment is run, the present claim remains a diagnostic interface result rather than evidence that ADeltaM improves foundation-model embodied agents.

7. Limitations

The study remains preliminary. We use a small procedural gridworld rather than Habitat, AI2-THOR, ProcTHOR, ENACT-style egocentric world-modeling data, or EmbodiedBench-style agent evaluation. We report only three seeds and no downstream navigation/manipulation success. RSSM/Dreamer and transformer dynamics baselines are not implemented. The reported slot/parameter counts are implementation-level footprint diagnostics, not latency, FLOP, or memory-byte measurements. Future work needs harder interaction outcomes, frozen foundation-feature latents, and real RGB-D or egocentric video rollouts.

8. Conclusion

ADeltaM is best understood as an exploratory counterfactual delta-memory interface for egocentric agents. In a small RGB-D gridworld, we observe improved rollout and same-state counterfactual delta prediction over several diagnostic baselines, while a linear residual model remains close and the foundation-model connection is not yet experimentally realized. We therefore view the work as a short workshop discussion seed: the mechanism is concrete, the evaluation protocol is inspectable, and the next steps toward foundation-model embodied agents are explicit.

References

[1] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

[2] Haoquan Fang, Markus Grotz, Wilbert Pumacay, Yi Ru Wang, Dieter Fox, Ranjay Krishna, and Jiafei Duan. Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation. In *International Conference on Machine Learning*, 2025.

[3] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, 2016.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2016.

[5] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[6] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

[7] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[8] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[9] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *arXiv preprint arXiv:2408.03615*, 2024.

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[12] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *International Conference on Computer Vision*, 2019.

[13] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, 2016.

[14] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *International Conference on Machine Learning*, 2025.

[15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, 2023.