MemPrism

Task-Conditioned Relational Memory Views for Long-Horizon Agents

Paper by Zhisheng Chen et al.

Independent paper-reading presentation

arXiv:2608.06745v1 | Read on 27 September 2026

Preserve interaction facts; choose their organization at decision time.

Sources: version-matched author TeX, 41-page PDF, and original figures. Results are reported by the paper, not independently reproduced.

Available evidence can still be hard to use

Storage	Retrieval	Representation
Was the needed fact preserved?	Was the relevant evidence accessed?	Are the needed relations easy to see?

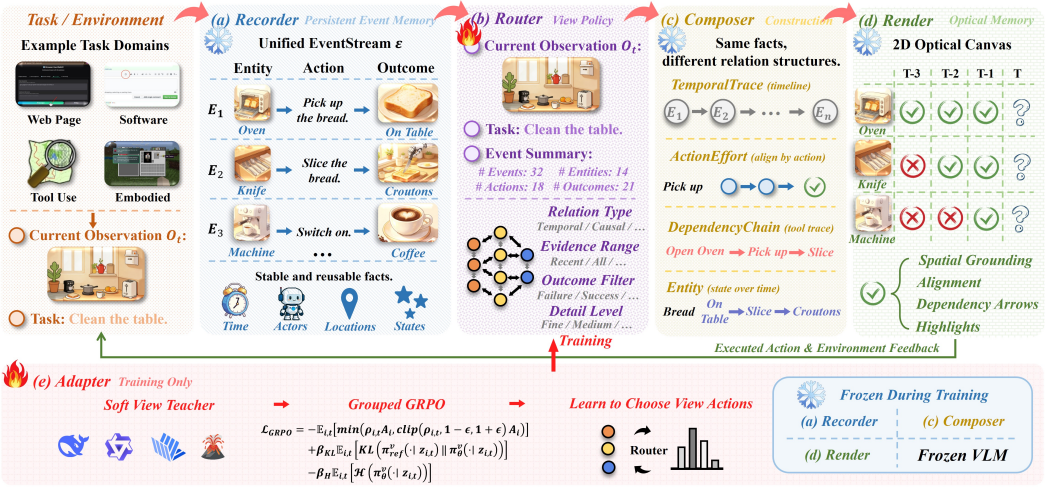
- Repeated failures call for aligned attempts and outcomes.
- Object tracking calls for a state history organized by entity.
- Progress tracking calls for temporal order.

Interpretation

The paper makes memory organization a decision variable.

Paper, Introduction, PDF pp.1–3. These failure categories can interact in the implementation.

Stable history supports temporary working views



Original Fig. 2, PDF p.5. Adapter trains view selection; Recorder, Composer, Render, and the task VLM remain frozen.

Recorder preserves what actually happened

$$e_t = (o_t^-, u_t, y_t^+, r_t, m_t) \xrightarrow{\Phi+\text{validation}} \tilde{e}_t = (t, e_t^{\text{ent}}, e_t^{\text{act}}, e_t^{\text{out}}, \Delta s_t, \xi_t)$$

- Raw transition: observation, executed action, feedback, reward, metadata.
- Structured event: entity, action, outcome, observed new state values, optional visual evidence.
- Outcomes: exception, state update, or no observed change.

An append-only boundary

$\mathcal{E}_{\leq t} = \mathcal{E}_{<t} \oplus \tilde{e}_t$. Generated views are discarded; they never overwrite history.

Eqs. (2)–(4), PDF pp.4–5; Appendix D.1, pp.22–23. Validation does not guarantee correct semantic extraction.

Router chooses one of 144 joint view actions

$$a_t^v = (\tau_t, w_t, c_t, \gamma_t), \qquad 4 \times 3 \times 4 \times 3 = 144$$

Factor	Choices
Relation τ	Temporal / Effect / State / Dependency
Range w	Last 6 / last 10 / all events
Outcome c	All / exception / update / no change
Granularity γ	Coarse / medium / fine

- One joint softmax retains dependencies among the factors.
- Sample during training; take argmax during evaluation.
- Router selects the memory view; the task VLM selects the action.

Table 1 and Eqs. (6)–(9), PDF pp.5–6.

Four views expose different relations

TemporalTrace

Order events by step.

What happened recently, and in what order?

ActionEffect

Group action–entity pairs.

Did repeated attempts work?

EntityState

Group observed state updates.

Where is the object now?

DependencyChain

Link the latest event to related history.

Which earlier events provide context?

DependencyChain is a heuristic local trace, not a verified causal graph.

Composer, PDF p.6; Appendix D.3, p.24; teacher specifications, pp.31–37.

Scope selection can restore task-relevant evidence

$$\widehat{\mathcal{E}}_t = \mathrm{Merge}\big(\sigma_{w_t,c_t}(\mathcal{E}_{<t}),\ \mathrm{TaskRelevant}(\mathcal{E}_{<t}, g),\ \mathrm{RecentContext}(\mathcal{E}_{<t}, \tau_t)\big).$$

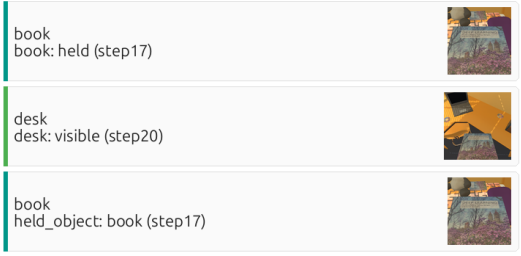
- Start with the chosen window and outcome filter.
- Restore events whose normalized terms match the task goal.
- Preserve recent anchor context for dependency views.

Interpretation

"Last 6" and "exception only" are selection rules, not strict final content limits.

Appendix D.3, Eq. (D.3), PDF p.24. Granularity changes the display, not stored events.

The image makes relations directly readable



- Cards group entity states.
- Alignment, arrows, and highlights express relationships.
- The VLM also receives current input and short text context.

Original memory panel, Fig. B.2, PDF p.19.

$$V_t = \mathcal{R}_{\tau_t,\gamma_t}(S_t), \qquad u_t \sim \pi_{\text{task}}(\cdot \mid X_t, g, H_t^{\text{short}}, V_t)$$

Eqs. (12)–(13), PDF p.7. The case reports one successful placement and progress 0 to 1/2, not full task completion.

A teacher initializes decision-conditioned routing

$$q_T(a \mid t) = q_T(a \mid o_t, g, \mathcal{E}_{<t}, u_t^{\mathrm{ref}})$$

$$\mathcal{L}_{\mathrm{SFT}} = \mathrm{KL}(q_T \| \pi_\theta^v) + \lambda_C \sum_{a \in \mathcal{A}^v} \pi_\theta^v(a \mid z_t) \widehat{C}_t(a)$$

- The teacher sees the reference next action and returns soft view labels.
- The deployed Router receives no reference action.
- The cost term discourages unnecessarily broad or detailed views.
- D1 trains the fusion encoder and Planner; MiniLM stays frozen.

Eqs. (14)–(15), PDF p.7; Appendix A.3, p.16. Structural cost is not a measured latency objective.

GRPO aligns view choices with downstream outcomes

$$A_i = \begin{cases} \dfrac{R_i - \mu_g}{\sigma_g + \varepsilon}, & \sigma_g > \varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

- Sample 16 trajectories per task group; normalize return within the group.
- All view decisions in an online trajectory share its advantage.
- A clipped objective uses a frozen SFT reference and optional entropy bonus.
- D2 updates only the residual Planner and 144-way head; the fusion encoder is frozen too.

Eqs. (16)–(18), PDF p.7; Appendix A.3–A.4, p.16. Mind2Web instead uses offline replay and action-correctness rewards.

The three benchmarks test different capabilities

Benchmark	Evaluation setting	Main reported metric
ALFWorld	140 episodes; up to 50 steps	Task SR
EB-ALFRED	300 tasks; up to 30 steps	Task SR
Mind2Web	3 offline test splits	Step action accuracy

- Shared memory-reader backbone: frozen Qwen2.5-VL-7B-Instruct.
- EB-ALFRED uses EmbodiedBench, not the official ALFRED leaderboard.
- Mind2Web advances with expert actions regardless of model predictions.

Interpretation

Mind2Web action accuracy is not autonomous web-task success.

Experiment Setup, PDF p.8; Appendix C.1, p.20; Fig. D.1, p.24. CFG-SFT is a separate task-model training baseline.

ALFWorld gains depend substantially on GRPO

Method	SR (%) ↑
Full History	31.43
LangMem	38.27
Mem0	37.49
MemPrism SFT	34.29
MemPrism SFT+GRPO	**40.71**

Absolute differences

+2.44 pp vs. LangMem

+9.28 pp vs. Full History

+6.42 pp vs. SFT

SFT alone does not beat the strongest external baseline here.

Selected rows of Table 2, PDF p.9. Point estimates; numerical confidence intervals are not shown in the PDF.

Embodied gains leave a clear visual failure boundary

Method	Base	Com.	Cplx.	Vis.	Spat.	Long	Avg.
Full History	16	10	14	0	10	12	10.3
Mem0	16	16	14	0	10	12	11.7
MemPrism SFT	20	20	22	0	16	16	15.7
SFT+GRPO	24	22	26	0	16	18	**17.7**

- Average SR improves by 7.4 pp over Full History and 6.0 pp over Mem0.
- Visual-subset SR remains 0%; memory organization does not solve all perception problems.
- CFG-SFT reaches 8% on Visual, but changes the task model's parameters.

Selected rows of Table 3, PDF p.9. All entries are SR percentages. Six subsets contain 50 tasks each.

Web gains hold across the three offline splits

Method	Task	Website	Domain	Overall
Full History	12.57	7.95	7.91	8.79
LangMem	13.20	8.85	8.42	9.45
MemPrism SFT	12.95	10.05	10.85	10.81
SFT+GRPO	**14.23**	**11.13**	**12.83**	**12.87**

- Overall improvement: +4.08 pp over Full History; +3.42 pp over LangMem.
- Predictions are scored against human actions; errors do not redirect the next page state.
- Episode success is mentioned in the setup but no value is supplied in Table 2.

Selected rows of Table 2, PDF p.9; protocol on pp.8, 24, 40–41. Values are step action accuracy (%).

Dynamic organization also helps in text

Organization	Text SR	Optical SR	Optical gain
Fixed	30.00%	35.71%	+5.71 pp
Dynamic	37.14%	**40.71%**	+3.57 pp
Dynamic gain	+7.14 pp	+5.00 pp	

- Dynamic organization helps with either presentation medium.
- Optical presentation helps under either organization setting.
- These are conditional differences, not two independent additive effects.

Re-tabulated from Fig. 3, panel II, PDF p.10. Fixed = TemporalTrace + last 10 + medium + all outcomes.

Always showing fine detail hurts the most here

Fixed relation views

View	SR (%)
ActionEffect	30.00
DependencyChain	32.14
EntityState	32.85
TemporalTrace	35.71

One constrained factor

Fixed setting	SR (%)
Type = TemporalTrace	35.71
Filter = all	40.00
Granularity = fine	32.86
Range = last 10	38.57
Full planner	40.71

Fine-only loses 7.85 pp relative to the plotted full planner.

Fig. 3, panels I and III, PDF p.10. Left: last 10 + medium + all. The prose gives 40.70%, while the figure gives 40.71%; differences here use the figure.

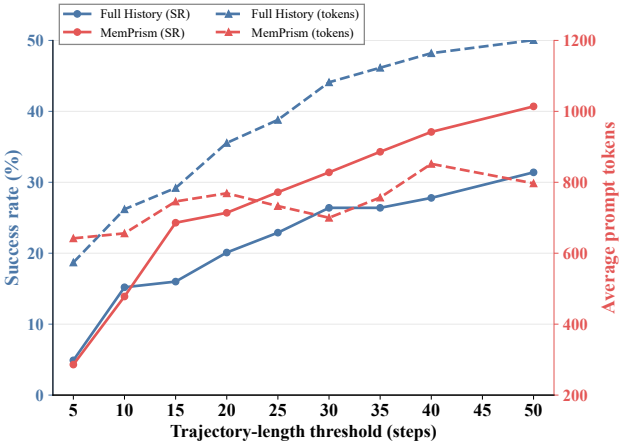
Global view frequencies do not explain the gain

View-selection strategy	ALFWorld SR (%)
Frequency-matched random	30.71
Context-conditioned SFT	34.29
Context-conditioned SFT+GRPO	**40.71**

- Random selection matches overall configuration frequencies but ignores the current state.
- Conditioning and outcome-guided learning both improve the reported result.
- Online trajectories can diverge: this is an episode-level comparison, not identical evidence at every decision.

Fig. 3, panel IV, PDF pp.10–11; shared evaluation protocol, Appendix C.1, p.20.

At 50 steps, the prompt is about one-third smaller



Full History → MemPrism

SR:
31.4% → 40.7%

Average prompt tokens:
1,201 → 797

About 33.6% fewer tokens

Original Fig. 4 and reported endpoints, PDF p.11. Solid circles: SR, left axis. Dashed triangles: tokens, right axis. X-axis uses cumulative length thresholds, not disjoint task bins.

The same Router transfers to three unseen VLMs

Task VLM	Full History SR	MemPrism SR	Token reduction
Qwen2.5-VL-7B	31.43%	40.71%	∼33.6%
Qwen3-VL-4B	39.17%	41.43%	30.64%
Qwen2.5-VL-3B	16.30%	21.42%	∼43.9%
MiniCPM-V-2.6	14.17%	23.57%	48.15%

- Router was trained with Qwen2.5-VL-7B; no per-model adaptation for the other rows.
- All three new readers improve; the gain ranges from 2.26 to 9.40 pp.
- Token savings are comparable only within each model.

Re-tabulated from Fig. 5 and its discussion, PDF p.12. Cross-reader transfer does not establish cross-environment zero-shot transfer.

Larger optical views are not automatically better

Max side (px)	SR (%)	Avg. prompt tokens	Invalid actions (%)
448	38.57	703	0.043
672	**40.71**	797	0.064
896	40.00	917	0.097

- Same checkpoint and rendering rules; only the resolution cap changes.
- The default 672-pixel cap has the highest reported SR among these three settings.
- Token savings alone do not measure total latency, extraction cost, or storage.

Table C.2, PDF p.22; separate cost categories, Appendix D.5, p.25. Intervals are not shown for this comparison.

Evidence supports the interface within a limited scope

- **Statistical precision:** the PDF gives point estimates. It refers to archived confidence intervals without printing their values.
- **Task scope:** online horizons are 50 or 30 steps; web evaluation is offline; the Visual subset remains at 0% SR.
- **System cost:** full-history scans and persistent storage grow with history, despite compact task prompts.
- **Representation limits:** extraction errors, four fixed relation types, and short Router summaries remain bottlenecks.

Tables 2–3, p.9; Appendix A.2, p.15; C.1, p.20; D.5, p.25. Bottleneck assessment is reading analysis.

Learn how history should support the next decision

1. Preserve reusable interaction records and construct working memory on demand.
2. Learn the relation, scope, and detail that help a frozen reader act.
3. Evaluate organization, medium, and full system cost separately.

What the experiments establish

Task-conditioned views improve the reported benchmarks and transfer across three new VLM readers; broader autonomy and end-to-end efficiency remain open.

Zhisheng Chen et al., MemPrism: Task-Conditioned Relational Memory Views for Long-Horizon Agents. arXiv:2608.06745v1, 2026. Main discussion ends here; backup follows.

Backup: the full GRPO objective

$$\rho_{i,t} = \frac{\pi_\theta^v(a_{i,t}^v \mid z_{i,t})}{\pi_{\text{old}}^v(a_{i,t}^v \mid z_{i,t})}$$

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = & -\mathbb{E}_{i,t}[\min\left(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i\right)] \\ & + \beta_{\text{KL}}\mathbb{E}_{i,t}[\text{KL}\left(\pi_{\text{ref}}^v(\cdot \mid z_{i,t}) \,\|\, \pi_\theta^v(\cdot \mid z_{i,t})\right)] \\ & - \beta_H\mathbb{E}_{i,t}[\mathcal{H}(\pi_\theta^v(\cdot \mid z_{i,t}))]\,.\end{aligned}$$

- π_{old}^v: rollout policy; π_{ref}^v: frozen SFT policy.
- Zero group variance gives $A_i = 0$; KL and entropy terms can still act.
- Mind2Web uses asymmetric clipping and different coefficients.

Eqs. (16)–(18), PDF p.7; benchmark-specific settings in Table A.1, p.16.

Backup: Router uses five input tokens

Input	Dimension	Scope
Observation text	384	MiniLM; at most 256 tokens
Goal text	384	MiniLM; at most 256 tokens
Event-window features	128	Last 8 events
Global statistics	8	At most last 64 events

$$[\mathrm{CLS}, o, g, w, s] \longrightarrow \text{4-layer Transformer} \longrightarrow z_t \in \mathbb{R}^{512}$$

- Separate projections to 512; 8 attention heads; FFN width 1024.
- Three residual MLP blocks produce joint 144-way logits.
- No CLIP embedding in the reported Router path; images go to the task VLM.

Appendix A.2, PDF p.15; D.4, pp.24–25. View windows are 6/10/all, distinct from summary windows.

Backup: training differs across benchmarks

Setting	ALFWorld	ALFRED path	Mind2Web
D1 epochs	50	20	50
D1 cost weight	0.02	0	0.02
D2 learning rate	3×10^{-6}	3×10^{-6}	10^{-6}
D2 iterations/steps	100	100	150
D2 KL weight	0.2	0.2	0.01
Training time	24 h	48 h	24 h

- D1: AdamW, learning rate 10^{-4}, batch 256, seed 7.
- D2: 8 task groups $\times$ 16 trajectories per iteration.
- Reported runs use eight H200 GPUs; rollout and model serving remain expensive.

Table A.1 and Appendix A.3–A.4, PDF pp.16–17. ALFRED training uses official training annotations; EB-ALFRED is reserved for evaluation.

Backup: the policy does not use every view equally

View type	ALFWorld	EB-ALFRED	Mind2Web
TemporalTrace	55.15	42.28	42.26
ActionEffect	31.24	21.98	7.62
EntityState	13.61	20.86	18.48
DependencyChain	0.00	14.88	31.64

- ALFWorld also assigns 0% of greedy selections to medium granularity.
- DependencyChain remains available; zero usage is policy concentration.
- Usage frequencies describe behavior, not the causal value of each view.

Selected rows of Table C.3, PDF p.23; explanation in C.6, p.22. Values are marginal selection frequencies (%).

Backup: AgentOCR token counts use a different scope

Method	ALFWorld SR	Reported tokens per step
AgentOCR (w/o RL)	25.30%	0.47k memory-context tokens
MemPrism	40.71%	797 complete prompt tokens

- Both conditions use a Qwen2.5-VL-7B task policy without task-policy RL.
- MemPrism trains its separate Router.
- The reported +15.41 pp SR gap is not an exactly budget-matched comparison.
- This table cannot establish that MemPrism uses fewer tokens than AgentOCR.

Appendix C.3–C.4, PDF pp.21–22; Table C.1, p.22. The AgentOCR value is reported from the cited setting.

Backup: evidence and reproducibility checks

- **Numerical consistency:** Fig. 3 uses 40.71%; its discussion uses 40.70%. This deck uses the plotted value.
- **Denominators:** some baseline SRs do not map to integer counts out of 140. Aggregation requires the external run records.
- **Uncertainty:** Appendix C.1 describes archived Wilson and paired bootstrap intervals, but no values appear in the PDF.
- **Reading scope:** all author TeX sections and appendices were read; key figures and formulas were checked against the PDF. No experimental rerun or external artifact audit was performed.

Paper: Zhisheng Chen, Bingfan Zeng, Bangde Cao, Zhengwei Xie, Yuxuan Li, Jinhan Li, Zheng Lu, Xiangchen Guan, Zikai Xiao, Rui Qian, Jingwei Song. arXiv:2608.06745v1.