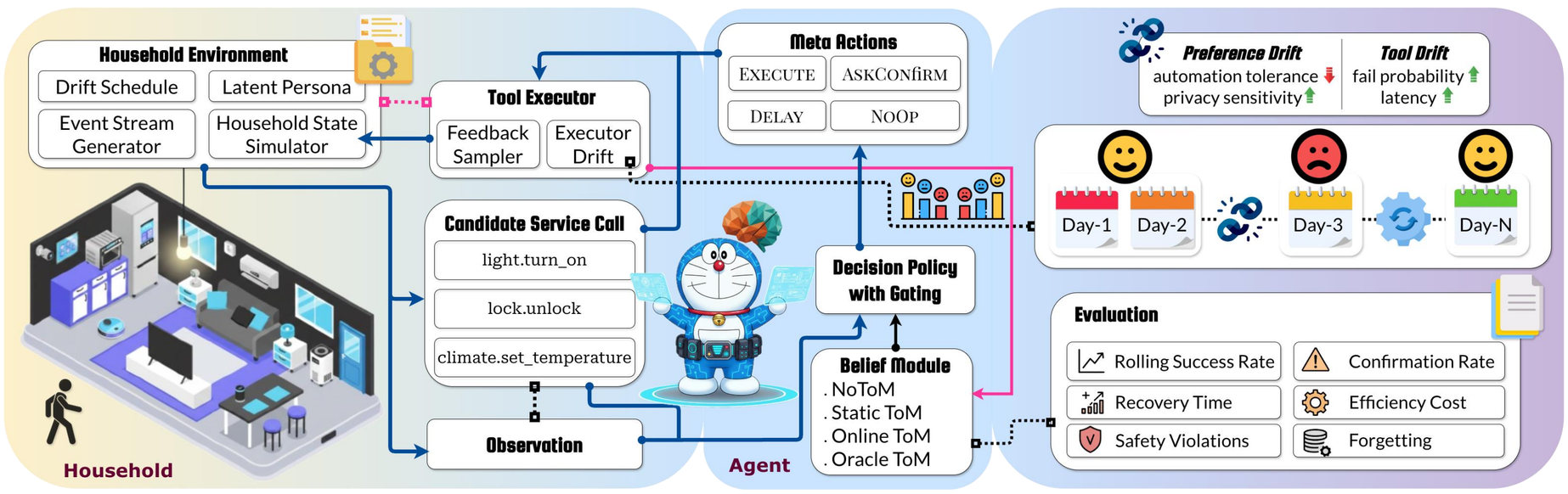
DomusMind: A Benchmark For Evaluating Lifelong Smart Home Agents Under Drift

Rong Xu[1], Yinxin Wan[2], Xiaochan Xue[3]

[1]Stevens Institute of Technology, [2]University of Massachusetts Boston,[3]University of Hawaii at Mānoa

TL;DR: Smart-home agents must operate continuously under evolving user preferences and device reliability, yet most evaluations remain episodic and reset-based. **DomusMind** introduces a persistent benchmark with preference drift and tool drift to measure long-horizon success, safety, burden, and recovery.

Benchmark Overview



SCAN ME

DomusMind evaluates agents in a persistent smart-home loop. At each step, the agent decides whether to execute, confirm, delay, or abstain under changing user preferences and tool reliability.

Why Lifelong Smart-Home Evaluation?

- Real smart homes evolve over time in user preference and device reliability.
- Reset-based benchmarks fail to capture persistent post-drift misalignment.
- Lifelong evaluation should measure both performance and burden during adaptation.

Two Drift Types

Preference drift
- **User autonomy tolerance and privacy preference may change over time.**
- **An action that was once helpful can later become intrusive.**

Tool drift
- **Device reliability, latency, and failure rate may change over time.**
- **Execution can fail even when the agent's intent is aligned.**

Decision Space

- At each step, the environment provides an observation and at most one candidate service call.
- The agent selects one meta-action: EXECUTE, ASKCONFIRM, DELAY, or NOOP.
- ASKCONFIRM executes the call only after explicit approval.
- DELAY / NOOP skip execution and allow the situation to reappear later.

- **EXECUTE** – attempt the proposed service call
- **ASKCONFIRM** – query the user, execute only if approved
- **DELAY** – postpone the decision
- **NOOP** – abstain from acting

Baselines and Metrics

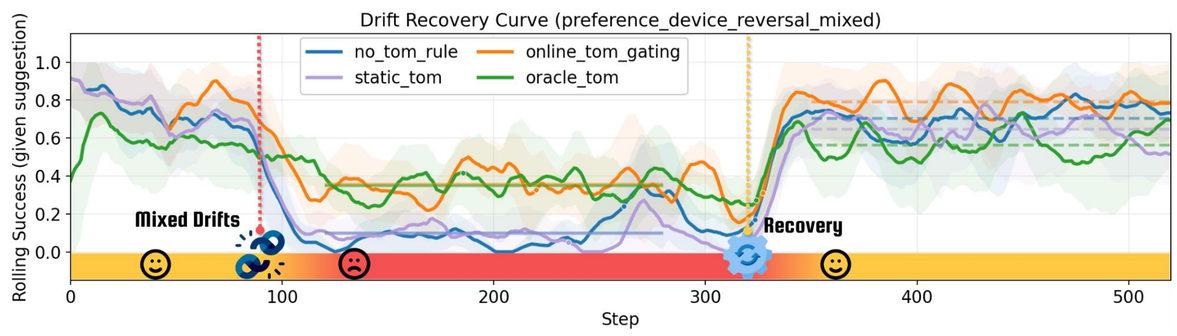
Baselines

- **NoToM**: no user modeling.
- **STATIC**: one-time personalization only.
- **ONLINE**: continual belief update + confirmation when uncertain.
- **ORACLE**: true persona access but still exposed to tool drift.

Metrics

- **Success**: aligned execution or safe abstention.
- **Confirmation rate**: user interruption frequency.
- **Safety violations**: unsafe executions.
- **Recovery time**: return to pre-drift performance.
- **Aggregate cost**: burden from confirmation and failures.

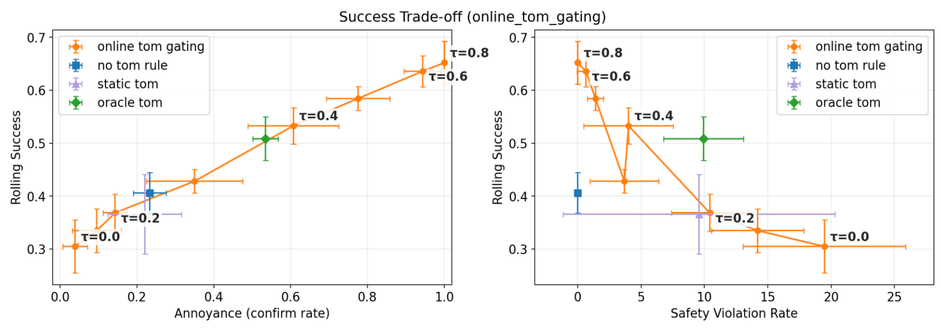
Evaluation Results



- **ONLINE** achieves the most robust overall adaptation under preference, tool, and mixed drift.
- After drift, **NoToM** and **STATIC** remain misaligned until the later regime revisit.
- **ORACLE** access to user preferences does not remove failures under tool drift.
- Sweeping the confirmation threshold reveals a tunable **success–annoyance frontier**.

	Policy	Success	r_{safevio}	Cost	PreDrift	PostDrift	F	T_{recover1}	T_{recover2}	r_{confirm}
Pref-R	NoToM	0.45±0.02	0.00±0.00	13.05±0.89	0.65±0.25	0.36±0.05	0.29±0.29	143.60±117.07	3.40±5.64	0.25±0.03
	STATIC	0.42±0.10	10.57±14.54	43.38±41.91	0.69±0.22	0.26±0.11	0.44±0.29	115.20±129.76	0.00±0.00	0.19±0.14
	ONLINE	0.64±0.03	1.04±0.93	19.82±2.98	0.80±0.15	0.80±0.24	-0.00±0.26	82.60±91.43	3.60±7.50	0.76±0.04
	ORACLE	0.53±0.02	14.48±3.34	54.03±9.36	0.55±0.27	0.43±0.21	0.12±0.38	38.00±45.99	19.00±23.72	0.44±0.04
Tool-R	NoToM	0.55±0.04	0.00±0.00	16.91±0.91	0.59±0.31	0.54±0.23	0.06±0.46	15.40±31.71	2.80±6.26	0.25±0.03
	STATIC	0.50±0.08	10.57±14.54	46.91±41.83	0.58±0.27	0.48±0.27	0.10±0.35	27.20±37.28	6.80±15.21	0.19±0.14
	ONLINE	0.63±0.03	0.87±0.61	23.77±1.25	0.78±0.14	0.57±0.28	0.21±0.37	75.20±90.18	3.40±7.60	0.79±0.09
	ORACLE	0.43±0.04	25.43±2.81	85.46±11.21	0.51±0.24	0.26±0.11	0.25±0.35	23.20±44.90	26.40±20.17	0.00±0.00
Mixed-R	NoToM	0.44±0.03	0.00±0.00	16.23±0.58	0.65±0.25	0.40±0.12	0.25±0.35	171.40±101.28	3.40±5.64	0.25±0.03
	STATIC	0.41±0.09	10.57±14.54	46.83±41.92	0.69±0.22	0.25±0.06	0.44±0.26	145.20±140.53	0.00±0.00	0.19±0.14
	ONLINE	0.60±0.04	0.51±0.47	21.73±1.41	0.80±0.15	0.45±0.21	0.35±0.34	130.00±122.08	7.00±6.28	0.90±0.06
	ORACLE	0.49±0.03	14.48±3.34	56.19±9.11	0.55±0.27	0.43±0.17	0.12±0.26	95.60±101.99	4.00±5.34	0.44±0.04

Table 1: Performance (mean±std) under strong preference (Pref-R), tool (Tool-R), and mixed (Mixed-R) drift. "-R" denotes a later regime revisit. Baselines: NoToM (rule), Static (one-time personalization), Online (uncertainty-gated confirmation, $\tau = 0.5$), and Oracle (true persona).



Conclusions

- **DomusMind** enables no-reset evaluation of lifelong smart-home agents under controlled drift.
- Separating preference drift from tool drift reveals execution reliability as a distinct bottleneck.
- **ONLINE** ToM with uncertainty-gated confirmation offers the best robustness–burden trade-off.

For more information: https://abby-xu.github.io/