KNOWING WHAT YOU NO LONGER KNOW: CALIBRATED BELIEF VALIDITY FOR PERSISTENT PERCEPTUAL WORLD MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent world memories let embodied agents answer questions about objects they saw minutes or hours ago. Every such system that reports a confidence reports *descriptive* uncertainty, how sure it is what an object is, and none reports *validity* uncertainty, how sure it is that a belief is still true, as a fitted or evaluated quantity. The consequence is systematic and invisible to current benchmarks: asked where an object is after it moved off-camera, models from 14B to 235B answer with the stale location on every such real-kitchen item and never once flag it as stale, at confidently-wrong rates of 1.000 to 0.970 and still 0.879 when told that declining is acceptable, while correctly declining about objects they never saw. We prove that any confidence which is a function of object class and elapsed time is at exactly chance on an elapsed-time-matched pair, and we introduce STILL-Bench, a benchmark with *derived* epistemic ground truth over real changing kitchens and controlled ones, built from such pairs. Fitting an interval-censored hazard to 9,120 real observation gaps, we find that validity *is* predictable, at AUROC 0.765 against 0.587 for fitted recency, and that the signal is almost entirely *where the object was last seen*: expiry runs from 0.029 on a counter to 0.614 in a drawer. A place-conditioned validity model separates stale from fresh beliefs where every published time-based heuristic is provably at chance, at PVD 0.710 ($p = 0.029$), answers 98.0% of queries at 5% risk or less where recency answers 61.3%, and ranks places well enough to be the best budget-limited re-observation policy we tested on procedurally generated houses. Three cautions follow: descriptive uncertainty does not predict validity and is *anti*-predictive on pooled pairs; a simulator whose injected changes lack the real place regularity inverts a validity evaluation; and the fitted hazard does not transfer between two real datasets, because a belief formed by *seeing* an object in an enclosure and one formed by *putting* it there carry opposite hazards.

1 INTRODUCTION

A robot that had walked the same office every day was asked, in a real deployment, to find a desk that had been vacant for several days. It chose an occupied one, because on every walk that desk had been vacant and nothing in its memory said that a belief can stop being true (Ginting et al., 2025, §8). The failure is the default behaviour of every persistent perceptual memory we tested (Fig. 1). Asked where an object is after it moved while the camera was elsewhere, models from 14B to 235B answer with the old location on every such real-kitchen item and never once flag it as stale, at confidently-wrong rates of 1.000 to 0.970; a 20B model that deliberates does no better; sanctioning abstention lifts the largest to 0.091 on stale items at a cost of six points on fresh ones; and all of them decline, unprompted, about objects they have *never* seen. The machinery for expressing uncertainty works; what is missing is a *quantity*: every spatial memory that reports a confidence reports descriptive uncertainty, estimated from disagreement among concurrent views, and none reports validity uncertainty, which grows with elapsed non-observation. Existing invalidation mechanisms cannot supply it: bitemporal stores (Rasmussen et al., 2025; Qian et al., 2026) invalidate on a *written* contradiction and decay heuristics (Tang et al., 2026) prune on a *positive* absence check, and neither fires for the object that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **A memory whose beliefs are cooling off.** One STILL-Bench item in the kitchen where it was recorded. **(A)** The frames the agent recorded of the fork, at the times they were taken, and the fifteen minutes in which nothing confirmed or contradicted the belief. **(B)** The question, the true answer now, and what the memory holds instead. *Centre:* the kitchen as the simulator renders it, with all 18 beliefs as boxes at the places their records name, coloured by estimated validity, a ray from each of the 120 recorded camera poses to every belief it produced, and the fork's true location in red. **(C)** The two systems' actual outputs. A persistent memory with typed status and valid time has no quantity for whether a record still holds, so it asserts the stored place at confidence 0.90 and is wrong; ours reports $p_{\text{valid}} = 0.29$, below its risk-controlled threshold, and declines. **(D)** Every belief on the validity axis: the 5 last seen inside a cabinet are already written off, while those on open surfaces sit in the assert band. **(E)** The same construction on every substrate.

moved while no one was looking. Proposition 1 proves that any confidence which is a function of object class and elapsed time is at exactly chance on a pair of questions matched on both.

To make the failure measurable we build STILL-Bench, whose epistemic labels are derived from observation and change timelines over real egocentric kitchen video (Perrett et al., 2025) and a controlled simulator (Kolve et al., 2017), around that matched pair. Fitting an interval-censored hazard to 9,120 real gaps shows that validity *is* predictable (AUROC 0.765 against 0.587 for fitted recency) and that the signal is where the object was last seen, not what it is (Fig. 3c). A place-conditioned validity model reaches PVD 0.710 ($p = 0.029$) where Proposition 1 pins every time-based heuristic at 0.5, answers 98.0% of queries at 5% risk or less against recency's 61.3%, and its estimate is sharp enough to decide where to look next (Fig. 5, appendix).

The three cautions in the abstract fell out of doing this carefully, and each carries a number: descriptive uncertainty is *anti*-predictive on pooled pairs (PVD 0.407, $p = 0.009$) and at chance on real ones, a simulator without the real place regularity inverts the evaluation (PVD 0.300, $p = 0.008$), and the hazard does not transfer between two real datasets because *seeing* and *putting* carry inverted per-place hazards.

Contributions. The descriptive and validity distinction with Proposition 1 and the matched-pair instrument; STILL-Bench, with derived epistemic ground truth, fifteen structural gates and a two-distribution evaluation rule; a fitted, calibrated, actionable validity model, and the finding that placement context carries the signal; and the three cautions. The memory architecture itself is prior art, ported and cited in Sec. 2.

2 Related work

Persistent spatial memory. Object-centric memories that outlive an episode span dense 4D maps (Schmid et al., 2024), dynamic voxel stores (Liu et al., 2024), caption databases indexed by position and time (Anwar et al., 2024), snapshot memories (Yang et al., 2024), persistent object stores (Fan et al., 2024), scene graphs for planning (Wang et al., 2024; Yan et al., 2024) and hierarchical 4D scene graphs that answer spatio-temporal questions (Gorlo et al., 2025; Ma et al., 2026b; Bartoli et al., 2026). DAAAM (Gorlo et al., 2025) appends descriptions to a history with no supersession, R4DSG (Ma et al., 2026b) is by its authors' account retrospective and offline, and Mind Palace (Ginting et al., 2025) always commits to an answer and reports the desk failure we opened with. Event graphs (Nguyen et al., 2025) and cross-session instance association (Huang et al., 2024; Behrens et al., 2024; Zhao et al., 2023) track *what changed*; none carries a per-belief confidence that a stored fact still holds.

Uncertainty over spatial memory. UQ-DAAAM (Zhang et al., 2026a), the closest system and the one whose protocol we adopt, scores multi-view caption scatter, which has no time term, and is calibrated for descriptive correctness on static questions. Variability attributes (Looper et al., 2022), belief scene graphs (Saucedo et al., 2024), calibrated functional graphs (Fu et al., 2026), predictive scene graphs (Saavedra-Ruiz et al., 2026), landmark evanescence (Fuentes et al., 2025) and map-space belief prediction (Marques et al., 2025) model that scenes change, as priors or planning quantities, not as a per-belief validity that is fitted, reported and scored; a position paper (Ma et al., 2026a) asks when episodic memory should be trusted and calls for revalidation under shift, proposing neither.

Temporal decay and typed memory. HitMem (Tang et al., 2026) is the only spatial memory with a time-dependent score, $\exp(-\lambda(v)\Delta t)$ with hand-set λ; it is never evaluated as a prediction and Proposition 1 places it at chance on matched pairs. Text-agent memories carry the machinery we port: bitemporal facts (Rasmussen et al., 2025), typed status with provenance and a resolver that abstains on conflict (Qian et al., 2026), evidence kept separate from defeasible belief (Xu et al., 2026), supersession over bitemporal ledgers (Yadav, 2026; Wang, 2026), immutable revisions (Park, 2026), probabilistic belief updates (Liao et al., 2026b; Singh, 2026), evidence-first provenance (Joshi, 2026), decaying reliability (Lu et al., 2026b) and paging (Packer et al., 2023). All invalidate on a *written* contradiction, so an unchallenged stale state resolves as active.

Staleness, abstention and unanswerability. STALE (Chao et al., 2026) and FreshQA (Vu et al., 2023) test stale facts in text, where staleness is announced. Abstention has benchmarks and surveys (Wen et al., 2024; Kirichenko et al., 2025; Miyai et al., 2024) for robots (Yeke et al., 2026), embodied question answering (Wu et al., 2025), occlusion (Zhang et al., 2026d), unanswerable video (Yoon et al., 2025) and matched physics pairs (Pramono et al., 2026). Language models can report what they know (Kadavath et al., 2022) but verbalise overconfidently (Xiong et al., 2023), and semantic entropy (Kuhn et al., 2023) targets confabulation rather than expiry. Long-horizon egocentric, streaming and world-model benchmarks (Yang et al., 2025a; Yan et al., 2025; Chen et al., 2026; Yang et al., 2025b; Zhang et al., 2026c; Liao et al., 2026a; Lu et al., 2026a; Cai et al., 2026) report that persistence fails, and PACT (Jin et al., 2026) shows that re-inferring over one observation adds agreement without evidence; none supplies epistemic ground truth over real changing scenes.

Active re-observation. Where to look next is studied for exploration (Cheng et al., 2024; Zhai et al., 2025), drift detection (Yin & Du, 2026), search in changed environments (Zhang et al., 2026b), persistent monitoring (Allu et al., 2024) and active SLAM (Placed et al., 2022); none carries a per-belief probability of having expired (Sec. 6.4).

3 PROBLEM: DESCRIPTIVE VERSUS VALIDITY UNCERTAINTY

A persistent perceptual memory holds beliefs $b = (k, v, t_{\text{obs}})$: object k was at place v when last observed at t_{obs}. *Descriptive* uncertainty, $p_{\text{desc}} = P(v \text{ correct at } t_{\text{obs}})$, asks whether the memory recorded the world correctly and shrinks with more views of that moment, which is what caption scatter measures (Zhang et al., 2026a). *Validity* uncertainty, $p_{\text{valid}}(\Delta t) = P(v \text{ still true at } t_{\text{obs}} + \Delta t)$, asks whether the world has since changed; it grows with elapsed non-observation, and no additional viewing at t_{obs} bears on it. Every spatial memory we know of that reports a confidence reports the first quantity.

Proposition 1 (matched pairs) *Let a fresh and a stale item share object class c and observation gap Δt. For any confidence $f(c, \Delta t)$ both members receive identical values, so the pairwise validity discrimination*

$$\text{PVD} = P(f_{\textit{FRESH}} > f_{\textit{STALE}}) + \tfrac{1}{2} P(f_{\textit{FRESH}} = f_{\textit{STALE}}) \tag{1}$$

equals $\frac{1}{2}$ exactly, covering recency, decay of the form $\exp(-\lambda(c)\Delta t)$ and any time-free score.

Proof. Both members carry the same c and the same Δt and f reads nothing else, so $f_{\text{FRESH}} = f_{\text{STALE}}$ on every pair and Eq. (1) is evaluated at a tie throughout: the first probability is 0, the second 1, and $\text{PVD} = \frac{1}{2}$. The bound constrains the score's arguments, not its accuracy, so it holds however well f tracks validity marginally and under any monotone rescaling. □

In practice Δt is matched to a tolerance, so the residual mismatch must be sign-balanced across pairs; we enforce a skew of at most 0.10 and measure between 0.006 and 0.031.

Corollary 1 *A validity model can beat chance on such pairs only through information that varies within a pair: evidence gathered during the gap, or the placement context of the last observation.*

4 STILL-BENCH

Substrates. The real tier uses HD-EPIC's public movement annotations (Perrett et al., 2025): 156 videos over nine kitchens, 8,382 objects and 36,760 sightings, 36,617 of which carry a 3D location and a symbolic fixture. The gap between two sightings of an object is when its belief can expire unseen. The controlled tier uses AI2-THOR (Kolve et al., 2017) sessions in which a privileged agent relocates objects between visits without notice, giving exact coverage ground truth, gaps out to 26 hours and cue-bearing changes; it supplies pairs and horizons only and is never used to fit the deployed hazard (Sec. F); the leave-one-house-out fit of Sec. 6.4 is a separate transfer analysis.

Derived labels and matched pairs. Each item's epistemic class is computed from its observation and change timelines by one function, re-derived as a validation gate behind timeline invariants. Three filters are measured: an in-hand annotation value is excluded; fixture changes have median displacement 0.48 m against 0.05 m for jitter, so 0.20 m separates relocation from noise; and answer granularity equals change granularity, since 41.5% of relocations move an object between two instances of the same fixture kind. A fresh and a stale item about the same object instance share template, wording, evidence budget and gap to within 15% or ten seconds, with residual mismatch sign-balanced (Proposition 1); relative matching yields 33 real pairs and 90 pooled in v0.9. Stale items are blind (declining is the only correct behaviour) or cued (a caveat is correct).

Composition, gates, rule. Version 1.0 holds 560 items: 90 pairs in three question forms plus never-observed, coverage and stale-premise probes. Fifteen structural gates pass on every version; three shortcut gates, run on v0.9, leave 0.5%, 2.5% and 5.0% of its 200 items solvable against limits of 30%, 40% and 40%, and cutting those 14 items with their 4 partners gives the certified set of 182 items and 86 pairs. A human annotator reproduces the derived labels at $\kappa = 0.897$ over 50 items, 95% CI $[0.762, 1.000]$; a model labelling blind reaches 0.828 on the same items and the two agree at 0.930. One annotator is not the pre-registered gate, which asks for two, so it stays open. A fifth class for contradiction reached 0.20 and was withdrawn: the disagreement gap inside one such item spans 0.267 s to 16,291 s, and that gap decides the honest reading (Sec. B). Discrimination is measured on the pairs, whose balanced prior removes the confound; calibration and selective risk are measured on



Figure 2: **Overview of EWM.** Perception writes place-valued observations from the recorded frames (left, with their times); the store keeps one record per object with its place, valid time and typed status, and STALE, the status text-agent memories cannot represent, is what an unconfirmed but uncontradicted record receives. A hazard $p_{\text{valid}} = \exp(-\lambda(x)\Delta t^{k})$ with $\log \lambda(x) = \beta^{\top} x$ over place, coarse class and observable context is fitted by interval-censored maximum likelihood to 9,120 real away gaps, and gives each record a validity at query time; for the queried fork, 0.29. The gate panel is the calibration itself: risk among answered gaps against the threshold, the Hoeffding-inflated curve above it, and $\tau = 0.976$ the smallest threshold whose inflated risk clears the 2% target; the queried record falls below it and the system declines in the words it actually produced. Every number and string shown is the fitted model's own output.

the natural distribution of all 9,120 real gaps, of which 6.0% expired, since a calibration metric on the pair set is guaranteed to indict a correctly calibrated model (Sec. 6.4).

5 METHOD: EWM

EWM adds one layer to an otherwise off-the-shelf pipeline, so that the belief layer is the single independent variable in every experiment (Fig. 2). Perception, retrieval and answer rendering are unchanged; what is new is that each belief carries a fitted, calibrated probability of still being true, and that this probability, not a phrasing heuristic, decides whether the system answers.

The store. A belief is a record $b = (k, v, [t_{\text{obs}}, \cdot), \sigma, \pi)$: object k was at place v, confirmed at t_{obs}, with typed status σ and provenance π, the frame identifiers that support it. The ledger is append-only; a new sighting of k opens a new record and links back to its predecessor rather than overwriting it, so the valid-time history survives. Typed status, valid time, provenance closure and abstention on conflict are ported from text-agent memories (Qian et al., 2026; Rasmussen et al., 2025) and are not our contribution. What those designs cannot express is a record that is admissible, never contradicted, and too old to trust; we add exactly one status, STALE, for it, and the rest of this section is how a record earns it.

The validity hazard. For a record last confirmed Δt ago in placement context x we model the probability that it still holds as a Weibull survival function,

$$p_{\text{valid}}(\Delta t \mid x) = \exp\!\left(-\lambda(x)\,\Delta t^{\,k}\right), \qquad \log \lambda(x) = \beta^{\top} x, \tag{2}$$

Table 1: **Discrimination on Δt-matched pairs, certified v0.9.** Every item survived the question-only, last-sighting-only and current-view-only shortcut baselines. CDD is the decision-level analogue of PVD: a system that hedges every item and one that asserts every item both score exactly 0.5, which is why behavioural accuracy on stale items is never read alone (Tab. 5).

system	PVD ↑ all	Tier R PVD ↑	Tier R p	Tier C PVD ↑	CDD ↑ Tier R	behav. STALE ↑
B0 constant	0.500	0.500	n/a	0.500	0.500	0.000
B9 Graphiti+ECHO	0.500	0.500	n/a	0.500	0.500	0.000
B12 recency	0.471	0.484	1.000	0.464	0.500	0.000
B14 HitMem decay	0.477	0.500	1.000	0.464	0.500	0.000
B11 descriptive UQ	0.407	0.532	0.727	0.336	0.500	0.000
EWM (ours)	0.558	**0.710**	0.029	0.473	0.613	0.977

86 pairs (Tier R $n = 31$, Tier C $n = 55$).

with a shared shape k and a log-linear rate. The covariates x are the canonical place of the last sighting under a shared ontology, a coarse object class, and the agent's observable context during the gap; Sec. 6.2 reports what each contributes, and place carries almost all of it.

The training signal is interval-censored. For an away gap we know that the belief held when the object was last seen and what its state was when the object was seen again, but never the instant it failed. Writing $s_i = p_{\text{valid}}(\Delta t_i \mid x_i)$, the objective is therefore

$$-\frac{1}{n}\sum_{i=1}^{n}\Big[y_i \log s_i + (1-y_i)\log(1-s_i)\Big] + \eta\|\beta_{1:}\|_2^2, \tag{3}$$

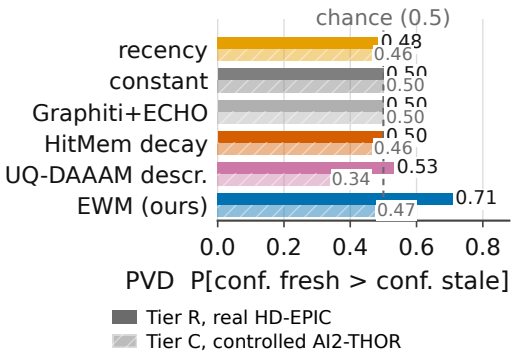
where $y_i = 1$ if the belief survived the gap. The first term is the survival contribution and the second the probability of failure somewhere inside the interval. We minimise Eq. (3) with L-BFGS-B, ridge $\eta = 10^{-4}$ on the non-intercept coefficients, and the intercept initialised at the marginal failure rate. A free k replaces the pre-registered constant hazard, which proved misspecified: it is worth $+0.041$ AUROC and moves ECE from 0.015 to 0.003. All fitting is leave-one-kitchen-out, so no query is ever scored by a model that saw its own kitchen.

From a probability to an action. A probability is not yet a decision, and a fixed cut-off on it carries no guarantee. We calibrate a threshold by split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) on held-out gaps: for a target risk r and confidence $1-\alpha$, τ is the smallest score whose empirical risk among answered items, inflated by a Hoeffding slack $\sqrt{\log(1/\alpha)/2n}$, stays below r. The guarantee is distribution-free and does not assume that p_{valid} is itself calibrated, which matters because the base expiry rate is 6% and any target above it is satisfied by answering everything. We use $r = 2\%$, and $\tau = 0.976$ on the certified set.

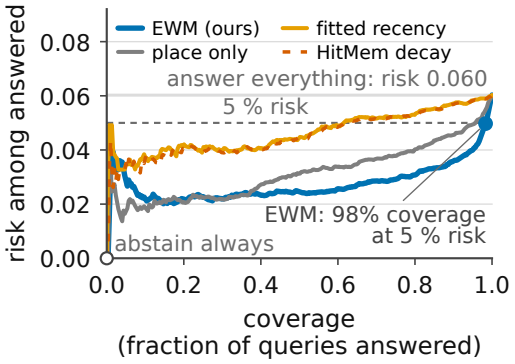
The policy then has three outcomes: assert when $p_{\text{valid}} \geq \tau$; answer with a caveat that states the age of the evidence when p_{valid} falls in $[0.8, \tau)$; and decline below 0.8. The order of the two tests is not cosmetic. Testing the fixed floor first asserts confidently on everything above 0.8 whenever $\tau > 0.8$, which is precisely the failure this paper measures in other systems, and it also empties the hedged band, because the conformal test then only ever sees scores under 0.8; getting it wrong turns 143 caveats into confident assertions on the certified set. Finally, the number the system reports as its confidence *is* p_{valid}, so the reported confidence is a prediction about the world rather than a statement about a caption, and beliefs are rendered in the question's own form, so a belief that nothing changed answers "yes, still there" rather than restating a place.

6 Experiments

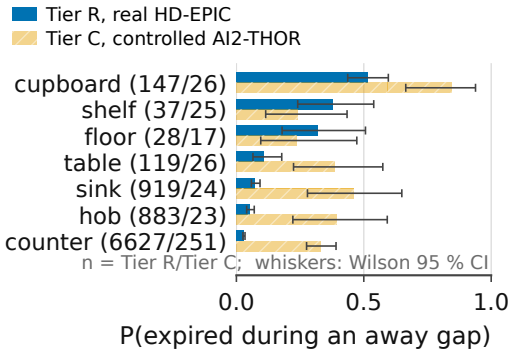
Setup. Benchmark v0.9 (200 items, 90 matched pairs) for the main comparison v1.0 (560 items, three question forms) for phrasing, and v1.1 (555 items, v1.0 without the withdrawn class) for the closed-form re-run, under the oracle-perception condition in which the annotated timeline is the perception, so every error is attributable to memory or reasoning. Baselines are the closed-form systems of Tab. 1, a constant reader, the Graphiti and ECHO bitemporal transplant (Rasmussen



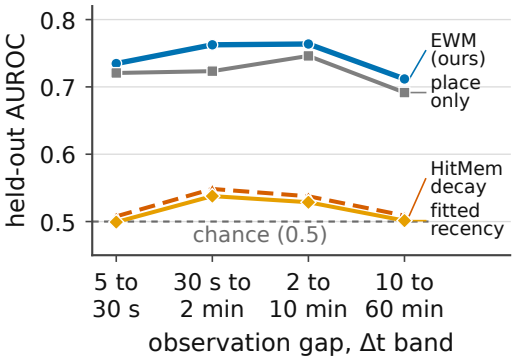
(a) discrimination on matched pairs



(b) selective risk on the natural distribution



(c) expiry by last-seen place, real and controlled



(d) discrimination within elapsed-time bands

Figure 3: **Results.** (a) PVD per system and tier on the 86 certified pairs of v0.9: constant and bitemporal readers tie every pair, recency and decay sit at chance, descriptive uncertainty falls below it, and only the place-conditioned validity model separates stale from fresh on real pairs. (b) Risk–coverage on 9,120 real away gaps: EWM answers 98.0% of queries at 5% risk where recency answers 61.3%. (c) Among places both tiers share, real expiry runs from 0.029 on a counter to 0.517 in a cupboard (0.614 in a drawer, real tier only); the controlled tier reorders them (Spearman $\rho = 0.04$). (d) Within a band elapsed time has nothing left to rank on, and place still does.

et al., 2025; Qian et al., 2026), fitted recency, HitMem's decay (Tang et al., 2026), UQ-DAAAM's descriptive uncertainty under its own protocol (Zhang et al., 2026a), together with a 14B language model over the same memory and a hosted 122B model on real pairs. Discrimination is measured on pairs, calibration and selective risk on the natural distribution (Guo et al., 2017; Naeini et al., 2015; Brier, 1950; Geifman & El-Yaniv, 2017).

6.1 No published confidence separates stale from fresh

On the certified set (Tab. 1, Fig. 3a) no system's confidence distinguishes a still-true belief from an expired one: constant and bitemporal readers tie every pair, recency and decay reach 0.484 and 0.500 on real pairs, the 14B model 0.475 on the 100-item real pilot with 66 answers at confidence 1.00, and a 20B reasoning model 0.515 ($p = 1.000$) at a confidently-wrong rate of 0.848. None caveats a stale item, and the 14B declines on all five objects it never saw and none of the 40 that moved. Two hosted models, one item per call over all 33 real pairs, land *below* chance at PVD 0.364 ($p = 0.022$, 122B) and 0.409 ($p = 0.070$, 235B), both at a confidently-wrong rate of 0.970: each is on average *more* sure of an expired belief than of one that holds. Sanctioning abstention moves the 235B to 0.455 ($p = 0.581$) and its stale behaviour to 0.091, costing six points on fresh items and still flagging none of the 33 (Tabs. 7 and 9). Behavioural accuracy on stale items is 0.000 for every baseline and 0.977 for EWM, but that is an operating point: at a 2% target risk it hedges 0.988 of fresh items too. The decision-level statistic that survives is CDD (Tab. 1), where hedging everything and asserting everything both score 0.5: EWM reaches 0.613 on the real tier ($p = 0.016$) against 0.500 for every baseline. Figure 4 in the appendix shows one such pair. Sweeping the target risk (Tab. 5, appendix)

moves behavioural accuracy from 0.977 to 0.209 while PVD and CDD do not move, and across the three question forms of v1.0 the 14B model's PVD stays between 0.572 and 0.589: neither the operating point nor the phrasing is the story.

Outside the certified set, and across question forms. On 31 certified real pairs the real-tier result carries $p = 0.029$; on all 33, certified or not, it is PVD 0.727 at $p = 0.014$ with a cluster bootstrap of $[0.576, 0.879]$, against 0.500 to 0.530 for every baseline ($p \geq 0.86$), and 0.578 pooled over both tiers, which is not significant ($p = 0.161$) because 57 of the 90 situations are controlled ones at chance for the reason Sec. F gives (Tabs. 2 and 4, appendix). Asking every situation in three question forms, as v1.1 does, changes none of those numbers and cannot, since a closed-form confidence never reads the wording; its 270 pairs carry the information of 90 situations, and a test over 270 draws would be anti-conservative. Over-abstention on fresh items is 0.000 on the real tier and 0.144 pooled on this uncertified set, where a hazard fitted on a 6% base rate meets a substrate whose base rate is 37%. Nor does size close it: four models spanning $17\times$ in parameters, on the same 33 pairs under identical prompts, none above chance and none ever flagging a stale belief, with the confidently-wrong rate *rising* from 0.848 to 0.970 (Tab. 10, Sec. F).

Handed rendered frames instead of the timeline, a 7B and a 235B model behave alike (Tab. 8): from frames alone each is diffident, declining on 0.400 and 0.525 of fresh items with almost every pair tied, and adding the text memory raises fresh behaviour *and* the confidently-wrong rate together, to 0.900 and 0.750. Resolving what the object *is* supplies nothing about whether the belief holds; it converts abstention into confident error, at both ends of a 33-fold size range.

6.2 Validity is predictable, and the signal is place

Leave-one-kitchen-out on 9,120 real gaps (Tab. 18, appendix): a fitted global exponential reaches AUROC 0.587, class alone 0.593, place alone 0.705 and the full observable Weibull model 0.765 at Brier 0.047 and ECE 0.003. P1 passes at $+0.137$; P3 fails at $+0.011$, with the oracle activity feature adding $+0.006$, so the activity signal is absent, not unexploited. Expiry spans 0.029 on a counter to 0.614 in a drawer while object class spans 0.04 to 0.10 (Fig. 3c), and the nested ablation attributes $+0.118$ to place, $+0.041$ to the Weibull shape, $+0.011$ to context and $+0.008$ to class. Trained on the other kitchens the model reaches PVD 0.710 ($p = 0.029$) on the 31 certified real pairs and 0.700 ($p = 0.017$) on the independent 40-pair pilot: 0.889 on the 18 pairs whose last-seen places differ and 0.545, chance, on the 22 that share one, where a place-only model reaches 0.725, its ceiling. Within each elapsed-time band EWM holds AUROC 0.71 to 0.76 while recency and decay stay at 0.50 to 0.55 (Fig. 3d).

6.3 The hazard does not transfer, for the reason that makes it a hazard

A second real substrate comes from EPIC-KITCHENS-100's public narrations (Damen et al., 2020): a narrated placement starts a belief and the noun's manipulation timeline gives its state at fixed horizons from 30 s to 20 min (7,726 rows, 26 participants, none in HD-EPIC). Fixing the horizons in advance matters (Sec. F): at fixed horizons expiry rises from 0.100 to 0.435. P1 replicates (Tab. 11, appendix): place 0.690 against global 0.645, observable 0.747, and within horizons place alone 0.574 to 0.652. The fitted hazard does not carry across: HD-EPIC's place model is worse than a global decay on EK-100 (0.612 against 0.661) and vice versa (0.553 against 0.590), because the per-place orderings are inverted (Tab. 3, appendix): an object *seen* in a fridge expires at 0.958, one *put* there at 0.096; on a counter, 0.031 and 0.291. You see into an enclosure only while retrieving from it, and you put things into one to leave them: validity is a property of place *and* provenance, so the store types how each belief was formed and the hazard must be fitted on beliefs formed the way the deployed system forms them.

6.4 The estimate is calibrated enough to act on, and to act with

On the natural distribution (Fig. 3b; Tab. 17, appendix) EWM reaches AURC 0.0275 against 0.0463 for recency and 0.0455 for HitMem, and answers 98.0% of queries at 5% risk or less where recency answers 61.3%; over 2,000 bootstraps the AURC gap is $+0.0185$ $[+0.0133, +0.0241]$ and the coverage gap $+0.338$ $[+0.141, +0.485]$. Measured first on the pair set the same model appeared to

fail (confidently-wrong rate 0.788): a model calibrated to a 6% expiry rate must look overconfident on a set half stale by construction, which the two-distribution rule prevents.

Re-observation. On 40 procedurally generated houses (Tier P) (Deitke et al., 2022) under the same nine-session schedule, 38 of which reach a session boundary holding a stale belief, and with the hazard still fitted on HD-EPIC, the agent may re-look at k places before answering; a re-look refreshes the beliefs about objects found there and refutes those that named it for objects now gone. Ranking places by expected stale mass $\sum_{b \in P}(1 - p_{\text{valid}}(b))$ is the best budget-limited policy at every k (Fig. 5, appendix): after one look it leaves 0.190 of beliefs silently stale against 0.264 for refreshing the least recently confirmed place, 0.201 for the densest place and 0.280 at random, a paired difference of -0.074 $[-0.098, -0.052]$ against recency. On the hand-built kitchens a belief-count heuristic does slightly better (0.145 against 0.160), because relocations there are drawn from the objects present, and a greedy oracle leaves 0.052, so the headroom is large. Leave-one-house-out, observable context reaches AUROC 0.792 against 0.599 for place and 0.577 for a global decay, while HD-EPIC's place model scores 0.559 here, below the global 0.602: place coefficients are substrate-specific; context travels.

Section F reports the rest: the simulator inversion (PVD 0.300, $p = 0.008$), the failure composition, in which every baseline fails through one code and EWM is the only system that ever declines on a fresh item, and the label-legibility and shortcut-gate results.

7 Limitations

Scope. One real domain, nine kitchens, and object identity that does not persist across videos, so real gaps top out near an hour and no pair spans two videos. The 33 real pairs are near what the substrate admits, not a sampling choice: a pair needs a fresh and a stale situation on one object in one session at matched elapsed time, and the ceiling is 46 pairs with no cued quota and 18 at a 5% one, so growing the real tier costs the cued-versus-blind split we report. Two to 26 hours come from a controlled substrate whose change process we generated and stopped tuning once it matched the real regularity qualitatively but not quantitatively (37% against 6% base expiry).

Where the method stops. All of EWM's discrimination comes from pairs whose last-seen places differ (0.889); on same-place pairs it is at chance (0.545), and context adds $+0.011$, a limit of the signal rather than the estimator, which is where re-observation enters (Sec. 6.4).

Models and annotation. No *proprietary* frontier system was run; the largest reachable was a hosted 235B model, on all 33 real pairs in both conditions with every confidence parsed, and its perception arm uses rendered frames, the HD-EPIC video agreement being unfiled. The real tier is therefore read from annotated timelines, which is also why perception failures are absent there. Labels are derived, not authored, and the pre-registered two-annotator check is *half* complete: one annotator, neither of us, reproduces them blind at $\kappa = 0.897$, 95% CI $[0.762, 1.000]$ (Sec. B). One reader cannot separate a legible taxonomy from a lucky one, so the gate asks for two, stands open, and we report no pairwise human κ.

8 Conclusion

A memory that knows what it saw is not a memory that knows what is still true. The confidence current spatial memories report is calibrated for the first and silent about the second; the gap is structural in every published invalidation mechanism; the second quantity is predictable from real data, from where an object was rather than what it was; and a model that estimates it declines the questions it should and knows where to look next. Where the place gives nothing away, the honest answer is to go and look.

AI use statement

AI assistants were used as tools for coding assistance and for editing text under the authors' direction. The research questions, the hypotheses, the pre-registered protocol, the experimental design, the analyses and the conclusions are the authors' own. No generative model produced any figure element,

number, label or result: every quantity in the paper is regenerated from the results directory by a single script, and all code was reviewed by the authors and verified by the test suite and that reproduction. A language model serves as a measurement instrument in three places, each labelled as such wherever it appears: the label-legibility proxy at $\kappa = 0.828$, the same proxy run over the withdrawn contradiction class, and a proxy pass over the human-ceiling task used to debug the annotation instrument before any person was asked to spend time on it. None of the three is reported as a human result, and the analysis script excludes proxy annotators from the pre-registered gate by construction rather than by convention.

Ethics statement

The real-world substrate is HD-EPIC's publicly released movement annotations and EPIC-KITCHENS-100's public narrations; no video frames from either dataset are redistributed, and no participant is identifiable from the derived timelines used here. The controlled substrates are simulator scenes. The benchmark labels are derived from annotations, not collected from human subjects; the blind-labelling check reported as a legibility proxy was performed by a language model shown only the timelines and the class definitions. One human annotator has since completed the same blinded task, and the pre-registered two-annotator check therefore remains incomplete, which the paper states. That annotator consented to an unpaid, uncredited hour of labelling for research use, was told what the study measures before starting, and is identified in the released data by a given name only; no other personal data was collected, and their responses stayed on their own device until they chose to return them. The instrument and the protocol governing it are in the repository; the protocol forbids the paper's authors from annotating either task, and an author's own pass is retained in the released data marked as such and excluded from the gate by the analysis script rather than by convention. Every result of the study is written by that script and not by hand. The work is intended to make embodied memory systems more honest about what they no longer know; we see no direct path to harm from the released artefacts.

Reproducibility statement

Every number in the paper is regenerable from the repository with one command, `scripts/reproduce.sh`, which rebuilds the benchmark from the public annotations, refits the validity models, re-runs every closed-form baseline and the method, regenerates every table and figure, and diffs the headline statistics against the committed values; it reports an exact match in about three minutes on a laptop, and the parts that need a local language model, metered API access or the simulator are skipped and named. The regenerated artefacts are byte-identical when nothing substantive has changed, figures included, so a modified file after a reproduction means a number moved rather than a timestamp. The arms that do need metered access ship the rows they produced, so their statistics are recomputable from the repository without spending anything; every such row records the model that produced it, its elicitation, and whether its confidence parsed. The benchmark items, derived labels, fitted-model outputs and all results files are included. Section C gives the fitting details a reimplementation would otherwise have to guess, Sec. B the label derivation and gates, Sec. D the baseline reimplementations, and the proof of Proposition 1 is in Sec. 3.

References

Sai Haneesh Allu, Itay Kadosh, Tyler Summers, and Yu Xiang. Build once, monitor continuously: Persistent semantic mapping via autonomous exploration and open-vocabulary object updates, 2024. URL https://arxiv.org/abs/2409.15493.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation, 2024. URL https://arxiv.org/abs/2409.13682. ICRA 2025.

Ermanno Bartoli, Buwei He, Dennis Rotondi, Sebastian Koch, Federico Tombari, Kai O. Arras, Patric Jensfelt, Yixi Cai, and Iolanda Leite. GESTO: Human-centric spatio-temporal memory for reasoning in dynamic scenes, 2026. URL https://arxiv.org/abs/2608.10886.

Tjark Behrens, René Zurbrügg, Marc Pollefeys, Zuria Bauer, and Hermann Blum. Lost & found: Tracking changes from egocentric observations in 3d dynamic scene graphs, 2024. URL https://arxiv.org/abs/2411.19162. RA-L 2025.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Zikui Cai, Kaushal Janga, Tan Dat Dao, Seungjae Lee, Shivin Dass, Mingyo Seo, Kaiyu Yue, Mintong Kang, Nandhu Pillai, Monte Hoover, Aadi Palnitkar, Ruchit Rawal, Ruijie Zheng, Bo Li, Yuke Zhu, Roberto Martín-Martín, Tom Goldstein, and Furong Huang. Beyond episodic evaluation: Memory architectural bottlenecks in sequential embodied question answering, 2026. URL https://arxiv.org/abs/2607.21571.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid?, 2026. URL https://arxiv.org/abs/2605.06527.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A temporal object consistency benchmark for video large language models, 2026. URL https://arxiv.org/abs/2605.09904.

Kai Cheng, Zhengyuan Li, Xingpeng Sun, Byung-Cheol Min, Amrit Singh Bedi, and Aniket Bera. EfficientEQA: An efficient approach to open-vocabulary embodied question answering, 2024. URL https://arxiv.org/abs/2410.20263.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision, 2020. URL https://arxiv.org/abs/2006.13256. IJCV 2022.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation, 2022. URL https://arxiv.org/abs/2206.06994. NeurIPS 2022.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied VideoAgent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding, 2024. URL https://arxiv.org/abs/2501.00358. ICCV 2025.

Zhengyu Fu, René Zurbrügg, Kaixian Qu, Marc Pollefeys, Marco Hutter, Hermann Blum, and Zuria Bauer. FunFact: Building probabilistic functional 3d scene graphs via factor-graph reasoning, 2026. URL https://arxiv.org/abs/2604.03696.

Erick Fuentes, Jared Strader, Ethan Fahnestock, and Nicholas Roy. Belief roadmaps with uncertain landmark evanescence, 2025. URL https://arxiv.org/abs/2501.17982.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL https://arxiv.org/abs/1705.08500. NeurIPS 2017.

Muhammad Fadhil Ginting, Dong-Ki Kim, Xiangyun Meng, Andrzej Reinke, Bandi Jai Krishna, Navid Kayhani, Oriana Peltzer, David D. Fan, Amirreza Shaban, Sung-Kyun Kim, Mykel J. Kochenderfer, Ali akbar Aghamohammadi, and Shayegan Omidshafiei. Enter the mind palace: Reasoning and planning for long-term active embodied question answering, 2025. URL https://arxiv.org/abs/2507.12846.

Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Describe anything anywhere at any moment, 2025. URL https://arxiv.org/abs/2512.00565. CVPR 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL https://arxiv.org/abs/1706.04599. ICML 2017.

Chenguang Huang, Shengchao Yan, and Wolfram Burgard. BYE: Build your encoder with one sequence of exploration data for long-term dynamic scene understanding, 2024. URL https://arxiv.org/abs/2412.02449.

Zekai Jin, Hanrong Zhang, Yihong Tang, Fei Hu, Zhen Dong, and Yi Shao. Not all agreement counts as corroboration: Provenance-conserving multi-view fusion for typed action admission in human-robot collaboration, 2026. URL https://arxiv.org/abs/2609.01662.

Resham Joshi. Eywa: Provenance-grounded long-term memory for AI agents, 2026. URL https://arxiv.org/abs/2605.30771.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions, 2025. URL https://arxiv.org/abs/2506.09038.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3d environment for visual AI, 2017. URL https://arxiv.org/abs/1712.05474.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664. ICLR 2023.

Chih-Ting Liao, Xi Xiao, Chunlei Meng, Zhangquan Chen, Yitong Qiao, Weilin Zhou, Tianyang Wang, Xu Zheng, and Xin Cao. SpaMEM: Benchmarking dynamic spatial reasoning via perception-memory integration in embodied environments, 2026a. URL https://arxiv.org/abs/2604.22409.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability, 2026b. URL https://arxiv.org/abs/2605.05583.

Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. DynaMem: Online dynamic spatio-semantic memory for open world mobile manipulation, 2024. URL https://arxiv.org/abs/2411.04999. ICRA 2025.

Samuel Looper, Javier Rodriguez-Puigvert, Roland Siegwart, Cesar Cadena, and Lukas Schmid. 3d VSG: Long-term semantic scene change prediction through 3d variable scene graphs, 2022. URL https://arxiv.org/abs/2209.07896. ICRA 2023.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, and Xiaozhu Ju. Current world models lack a persistent state core, 2026a. URL https://arxiv.org/abs/2606.20545.

Yihao Lu, Wanru Cheng, Zeyu Zhang, and Hao Tang. MMA: Multimodal memory agent, 2026b. URL https://arxiv.org/abs/2602.16493.

Kailang Ma, Heye Huang, Inhi Kim, and Kitae Jang. Rethinking world models for safety-critical embodied systems, 2026a. URL https://arxiv.org/abs/2609.03774.

Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. R4DSG: Relative 4d scene graph memory for object-centric question answering in long egocentric video, 2026b. URL https://arxiv.org/abs/2608.11017. ACM MM 2026.

Joao Marcos Correia Marques, Nils Dengler, Tobias Zaenker, Jesper Mucke, Shenlong Wang, Maren Bennewitz, and Kris Hauser. Map space belief prediction for manipulation-enhanced mapping, 2025. URL https://arxiv.org/abs/2502.20606. RSS 2025.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Robust understanding evaluation for large multimodal models, 2024. URL https://arxiv.org/abs/2403.20331. ACL 2025.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Phuoc Nguyen, Francesco Verdoja, and Ville Kyrki. Event-grounding graph: Unified spatio-temporal scene graph from robotic observations, 2025. URL https://arxiv.org/abs/2510.18697. RA-L 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Young Bin Park. Graph-native cognitive memory for AI agents: Formal belief revision semantics for versioned memory architectures, 2026. URL https://arxiv.org/abs/2603.17244.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset, 2025. URL https://arxiv.org/abs/2502.04144. CVPR 2025.

Julio A. Placed, Jared Strader, Henry Carrillo, Nikolay Atanasov, Vadim Indelman, Luca Carlone, and José A. Castellanos. A survey on active simultaneous localization and mapping: State of the art and new frontiers, 2022. URL https://arxiv.org/abs/2207.00254.

Fnu Pramono, John Cai, and Sourabh Kulkarni. TRAPSBench: Vision-language models encode but fail to express epistemic restraint, 2026. URL https://arxiv.org/abs/2608.13167.

Yu Qian, Hong Miao, Boyang Guo, Tingyi Jiang, Shan Zhao, Tianxing Le, Lintian Li, and Meng Liu. ECHO: A cognitively inspired, auditable memory plane for long-horizon agents, 2026. URL https://arxiv.org/abs/2608.21755.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Miguel Saavedra-Ruiz, Charlie Gauthier, Kumaraditya Gupta, Shima Shahfar, Kirsty Ellis, Steven Parkison, and Liam Paull. Predictive spatio-temporal scene graphs for semi-static scenes, 2026. URL https://arxiv.org/abs/2605.00121.

Mario A. V. Saucedo, Akash Patel, Akshit Saradagi, Christoforos Kanellakis, and George Nikolakopoulos. Belief scene graphs: Expanding partial scenes with objects through computation of expectation, 2024. URL https://arxiv.org/abs/2402.03840. ICRA 2024.

Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic SLAM in dynamic environments, 2024. URL https://arxiv.org/abs/2402.13817. RSS 2024.

Pranav Singh. When does belief-based agent memory help? reliability-conditional updating and provenance-capped poisoning defense, 2026. URL https://arxiv.org/abs/2606.22030.

Ruijie Tang, Chenye Zou, Guoquan Wu, Jun Wei, Wei Chen, and Jiaxin Zhu. HitMem: Hierarchical temporal 3d memory with multi-modal context-aware retrieval for dynamic environments, 2026. URL https://arxiv.org/abs/2609.00950. ECCV 2026.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. ISBN 978-0-387-00152-4.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation, 2023. URL https://arxiv.org/abs/2310.03214.

Ziming Wang. TOKI: A bitemporal operator algebra for contradiction resolution in LLM-Agent persistent memory, 2026. URL https://arxiv.org/abs/2606.06240.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. KARMA: Augmenting embodied AI agents with long-and-short term memory systems, 2024. URL https://arxiv.org/abs/2409.14908. ICRA 2025.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2024. URL https://arxiv.org/abs/2407.18418.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say "i don't know": Benchmarking abstention in embodied question answering, 2025. URL https://arxiv.org/abs/2512.04597. CVPR 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs, 2023. URL https://arxiv.org/abs/2306.13063. ICLR 2024.

Wenzhuo Xu, Yuchen Zhu, Chongjian Ge, Xuan Shen, Jing Shi, Jason Kuen, Yongxin Chen, Molei Tao, Christopher McComb, Noelia Grande Gutiérrez, and Jiuxiang Gu. VeriPhy: Agentic physical reasoning for world model evaluation and refinement, 2026. URL https://arxiv.org/abs/2609.03153.

Neeraj Yadav. Temporal validity in retrieval memory: Eliminating stale-fact errors for AI agents over evolving knowledge, 2026. URL https://arxiv.org/abs/2606.26511.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, Changzhi Sun, Jixiang Luo, Dell Zhang, Hao Sun, Chi Zhang, and Xuelong Li. TeleEgo: Benchmarking egocentric AI assistants in the wild, 2025. URL https://arxiv.org/abs/2510.23981.

Zhijie Yan, Shufei Li, Zuoxu Wang, Lixiu Wu, Han Wang, Jun Zhu, Lijiang Chen, and Jihong Liu. Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation, 2024. URL https://arxiv.org/abs/2410.11989. RA-L 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant, 2025a. URL https://arxiv.org/abs/2503.03803. CVPR 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video, 2025b. URL https://arxiv.org/abs/2511.04670.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning, 2024. URL https://arxiv.org/abs/2411.17735. CVPR 2025.

Doguhan Yeke, Elif Su Temirel, Ananth Shreekumar, Brandon Lee, Dongyan Xu, and Z Berkay Celik. The yes-man syndrome: Benchmarking abstention in embodied robotic agents, 2026. URL https://arxiv.org/abs/2605.20544.

Xingkun Yin and Hongyang Du. GLOVE: Global verifier for LLM memory-environment realignment, 2026. URL https://arxiv.org/abs/2601.19249.

Eunseop Yoon, Hee Suk Yoon, Mark A. Hasegawa-Johnson, and Chang D. Yoo. Can video LLMs refuse to answer? alignment for answerability in video large language models, 2025. URL https://arxiv.org/abs/2507.04976.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-centric embodied question answering, 2025. URL https://arxiv.org/abs/2505.13948.

Harry Zhang, Nicolas Gorlo, and Luca Carlone. Remember with confidence: Uncertainty quantification for spatio-temporal memory with probabilistic guarantees, 2026a. URL https://arxiv.org/abs/2606.08277.

Wei Zhang, Ping Gong, Yujie Wang, Leilei Yao, Minghui Bai, Rongfeng Ye, Yinchuan Wang, Yachao Wang, Chen Sun, and Chaoqun Wang. IGV-RRT: Prior-real-time observation fusion for active object search in changing environments, 2026b. URL https://arxiv.org/abs/2603.21887.

Yehang Zhang, Jianchong Su, Haojian Huang, Yifan Chang, Tianhao Zhou, Xinli Xu, Yingjie Xu, Yinchuan Li, Zexi Li, and Ying-Cong Chen. WorldLines: Benchmarking and modeling long-horizon stateful embodied agents, 2026c. URL https://arxiv.org/abs/2606.18847. EMNLP 2026.

Yue Zhang, Zun Wang, Han Lin, Yonatan Bitton, Idan Szpektor, and Mohit Bansal. Seeing isn't knowing: Do VLMs know when not to answer spatial questions (and why)?, 2026d. URL https://arxiv.org/abs/2605.30557.

Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3d scenes from egocentric videos, 2023. URL https://arxiv.org/abs/2312.04117. CVPR 2024.

A PRE-REGISTRATION AND WHAT IT DECIDED

Three predictions were registered before any system was run, together with the numeric gate that would falsify each. **P1: validity is predictable from observable context.** Gate: a place-conditioned hazard must beat a global decay by at least 0.05 AUROC under leave-one-kitchen-out evaluation. Outcome: passed, $+0.118$ for place alone and $+0.137$ with observable context on HD-EPIC; on a second dataset with no shared participants place alone reaches $+0.045$, just under the margin, while observable context reaches 0.747 against the global 0.645 (Sec. 6.3). **P2: the gap-cue signal is usable.** Gate: cued and blind stale items must separate. Outcome: untested on real data, because the cue requires wearer pose and gaze from a download we do not hold; the controlled tier supplies cued items by construction and is reported as such. **P3: scene activity predicts expiry beyond place.** Gate: $+0.05$ AUROC from activity features. Outcome: failed, $+0.011$ observable and $+0.006$ with the oracle activity count, so the paper claims place and not activity.

Two pre-registered escape hatches were exercised. The discrimination statistic was fixed as pairwise validity discrimination after the originally registered statistic proved degenerate on a matched design, and the ambiguity band that triggered a 90-pair extension was declared before the pilot was scored. Both changes, and the twenty-three others that shaped this work, are recorded with their evidence and rejected alternatives in the decision log released with the code, including six scoring bugs we caught in our own favour and reverted: an evidence-count asymmetry between pair members, a gap definition that inverted the time effect, an object-identity key that dissolved every relocation, a batching artefact that manufactured discrimination, a private place vocabulary in the scorer, and a yes/no negation-scope error.

B BENCHMARK CONSTRUCTION

Labels are derived, never authored. Each item carries a full timeline of observations and change events. The epistemic class is computed from that timeline by a single function, so no annotator opinion enters the label, and eight timeline invariants are checked before an item is admitted: observations must be ordered, a change must lie between the sightings it separates, a change marked unobserved must have no frame covering it, and so on. A violation raises rather than silently relabels.

Structural gates. Fifteen gates run on every build. Five check that labels are re-derivable and invariants hold; four check pair integrity, that both members exist, that their elapsed times match within tolerance, that their wording is identical except where a template requires a location phrase, and that roles are balanced; three check composition, item count, stratum balance and the cued fraction; three check leakage, that no item's question reveals its own answer, that no evidence frame post-dates the query, and that no pair spans two scenes.

Matching on elapsed time. Fixed elapsed-time bands were abandoned after measurement: the real substrate's gap distribution yielded seven usable pairs. Matching is therefore relative, within 15% or ten seconds, whichever is larger, and the residual mismatch is sign-balanced so that no systematic direction survives. Measured skew is at most 0.031 on every shipped file, against a declared limit of 0.10.

The class we withdrew, and the repair that failed. A fifth class for co-temporal contradiction was built and then removed. The model labeller agreed with the derived label on 1.00 of stale and unobserved items but only 0.20 of contradicted ones, and the bitemporal baseline resolved every such item as supersession rather than conflict, which is the correct reading of sequentially ordered evidence from one agent. Measuring the class rather than arguing about it shows why it cannot be patched. Across the ten such items the project holds, the elapsed time between the disagreeing sightings inside a *single* item ranges from 0.267 s to $16{,}291$ s: a factor of sixty thousand in the one quantity that decides whether the honest reading is a contradiction or a superseded belief. At the bottom, two labels a quarter of a second apart disagree and no history orders them. At the top, two sightings four and a half hours apart disagree because something moved and nobody recorded it. One definition covered both. The obvious repair, narrowing contradiction to co-temporal or unordered evidence and giving sequential disagreement a class of its own, was pre-registered and then tested through the same blinded instrument. Two labellers disagree about it, and the disagreement is instructive.

A language model never used the new class at all, answering fresh or stale instead, which lowered its agreement. A human annotator used it on *eight of the nine* items she reached, agreeing with the original class on all nine once supersession is folded back into it. So the disagreement axis is real to a reader, and the redefinition points the right way: a reader does say that the later sighting wins. What she does not reproduce is our boundary. She called disagreements 0.2 s and 1.0 s apart superseded, where a one-second co-temporal rule calls them contradictions, and called a 1.5 s item a contradiction. A reader takes the later sighting even when two sightings are a fifth of a second apart, which suggests that *all* ordered disagreement is supersession and that contradiction belongs only to genuinely unordered evidence, of which this substrate contains none. On one annotator and nine items that refutes the boundary rather than establishing a replacement, so the class stays withdrawn from v1.1 and returns only with a multi-agent or out-of-order-evidence tier, where it means something and where the boundary can be declared from the substrate rather than guessed.

The two-annotator check: the instrument, and why it is reported as open. The pre-registered gate is Cohen's $\kappa \geq 0.75$ between two annotators, neither of whom may be an author. **One** annotator has completed it, reaching $\kappa = 0.897$ against the derived labels with a 95% interval of $[0.762, 1.000]$ over 50 items at a median of 12.3 s each, and agreeing with the model labeller at $\kappa = 0.930$ with two disagreements in fifty. The gate asks for two, so it remains open and this paper reports no pairwise human κ. What the repository does contain is the whole instrument, so that the missing ingredient is a person's afternoon rather than a build. Two tasks, kept separate because they answer different questions. In the labelling task an annotator sees the observable timeline and the world's change ledger and assigns one epistemic class; that ledger is exactly what the derived label is computed from, so a labeller who cannot see it would be guessing rather than reproducing. In the ceiling task an annotator answers the question itself from the *sightings only*, standing where a system stands: a system never sees the change ledger, and showing it would hand over the answer on every stale item and return a ceiling near 1.0 by construction. Both tasks strip the derived label, the ground-truth answer, the required behaviour and the pair role, and a test asserts those fields absent from the files an annotator receives, because a blinding that is only intended is not one. Three refusals are encoded rather than remembered. A task with fewer than thirty completed items reports as incomplete and emits no statistic. Fewer than two *human* annotators reports the gate as open and can never report it as passed; a model annotator is admitted, is identified as a proxy, and is excluded from every gate computation by construction. An author's identifier among the responses is a hard error. Running a model through the instrument first paid for itself. It found that the first build showed the change ledger in both tasks, which would have made the ceiling meaningless, and it did so before any person had been asked for their time.

C Validity model and fitting

Section 5 gives the model, its likelihood and its decision rule; this section records what a reimplementation would otherwise have to guess. The design matrix holds an indicator per canonical place, an indicator per coarse object class and the context covariates, with an intercept; k is a single shared shape parameter, estimated jointly rather than profiled. The optimiser is L-BFGS-B with at most 3,000 iterations, ridge 10^{-4} on every coefficient except the intercept, and the intercept initialised at $\log\big((1-\bar{y})/\overline{\Delta t}\big)$, the constant-hazard solution, which keeps the first step in a sensible region. The conformal threshold is chosen by search rather than by a quantile formula: the candidate set is the observed scores together with 0 and 1, each candidate's risk is the failure rate among the gaps it would answer, and τ is the smallest candidate whose risk plus $\sqrt{\log(1/\alpha)/2n}$ at $\alpha = 0.10$ stays under the target. With a target above the base expiry rate every candidate qualifies and the search returns 0, which is why the target must sit below it. The nested ablation behind Tab. 18 adds one feature family at a time in a fixed order, so each row is the marginal contribution of that family given the ones above it, and the Weibull shape is the last addition rather than the first.

D Baseline reimplementations

Bitemporal store. The typed-status and valid-time design of published text-memory systems is transplanted onto perceptual beliefs: statuses are active, superseded, revoked and unresolved, with valid and transaction time and a provenance link per record. The transplant is faithful in one respect

that matters for our claim: it resolves a later disagreeing sighting as supersession, which is right, and has no representation for a belief that is unconfirmed but uncontradicted, which is the gap we fill.

Exponential decay. The published decay baseline scores a memory by $\exp(-\lambda \Delta t)$ with λ set per functional role and zero for carriers. No code or data was released, so it is reimplemented from the paper's equations and constants, and it appears here as an upper bound on what any class-and-time confidence can achieve, which Proposition 1 fixes at exactly 0.5 on a matched pair.

Descriptive uncertainty. The strongest published uncertainty baseline scores a belief by the spread of its own descriptive embeddings, calibrated and thresholded on a held-out split following its authors' protocol. Run unchanged on our items it is anti-predictive, which is the cleanest evidence we have that descriptive and validity uncertainty are different quantities rather than two estimates of one.

Elicitation. Every language-model arm whose PVD we report answers one item per call. Batching several items into one call is prohibited for any within-pair statistic, not merely discouraged: with a shared context the two members of a matched pair are no longer independent, and a batch-level shift in caution produces apparent discrimination. We measured that artefact once, at a correlation of 0.781 between the batch mean and the within-pair difference, retracted the affected numbers and reran unbatched. The rerun is worth stating because it settles what the retraction was worth. Batched, the 122B appeared to reach PVD 0.650 ($p = 0.002$). Unbatched over all 33 real pairs it measures 0.364 ($p = 0.022$), which is not merely smaller but on the *other side of chance*: the batched artefact pointed the opposite way. A cost optimisation that touches how a judgement is elicited can therefore invert a paired result, and the only defence is to elicit one item per call and pay for it (Tab. 9). The 190-item batched run is retained in the results directory for the record; no statistic in this paper rests on it.

Does size close it? The arm we could not run is a proprietary frontier system, and the implication behind that gap is that one would have done better. That cannot be settled by a model we have no access to, but it can be asked whether the failure is capacity-limited at all. Four models spanning $17\times$ in parameters, 14B to 235B, answered the *same* 33 real pairs one item per call under identical prompts (Tab. 10). None separates stale from fresh above chance: PVD runs 0.364 to 0.515. The confidently-wrong rate rises with size rather than falling, 0.848 to 0.970. Behavioural accuracy on stale items is 0.000 at every size, and staleness is flagged on *zero* of 33 stale items at every size. Four points from four model families are a description and not a scaling law, and we read no slope from them. The claims that need no fit are the ones we rest on: nothing in this range does better than chance, and nothing in it ever says a belief might have expired. That is what one expects of a quantity that is absent from the prompt rather than of a capability that more parameters supply, and it is the reason we do not think the missing arm changes the conclusion.

The real-perception arm in full. The main comparison hands systems an annotated timeline so that perception failures cannot be mistaken for memory failures. Lifting that protection is what Tab. 8 reports, on the 95 frame-bearing controlled items whose frames are rendered at the recorded camera poses, for a 7B local model and a hosted 235B one on identical items. Given only the frames of the object's own sightings, neither model's caution is about validity at all. Each declines on a quarter to a half of the fresh items, which lifts stale-item behaviour to 0.425 while discriminating not at all: 29 and 40 of the 40 pairs tie on confidence, the confidently-wrong rate is 0.000 because nothing is ever asserted above 0.8, and CDD is at or below indifference. Four frames instead of one does not change the character of it. Supplying the text memory alongside the frames does, and in the same direction for both: fresh behaviour rises to 0.625 and 0.850, and the confidently-wrong rate rises with it, to 0.900 and 0.750. The reading is that these are two different uncertainties and only one of them was ever being measured. Diffidence from frames is *descriptive* uncertainty, about what the object is and where the frame was taken; remove it by supplying the memory and no validity signal appears in its place, only confidence. That this holds at 7B and at 235B is the part worth keeping: it is not an artefact of one model's calibration. Across both models and all three conditions, staleness is flagged on 1 of 240 stale judgements. One number here is deliberately not interpreted. The frames-plus-memory PVDs are 0.537 and 0.600 at $p = 0.487$ and $p = 0.134$, and 40 pairs cannot separate 0.55 from 0.65; the pre-registered power rule for this design says so, and the arm's claim therefore rests on the per-item behavioural columns, which need no pair power.

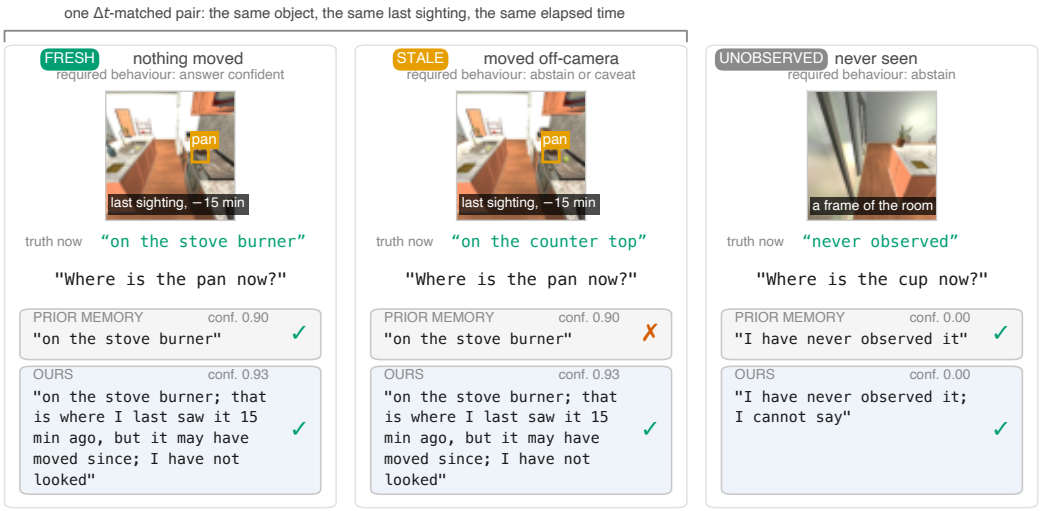


Figure 4: **The two members of one matched pair, and a never-observed probe.** Rendered frames from the controlled tier, real system outputs, and a tick or cross assigned by the benchmark's own required-behaviour rule. The pair is chosen by stated criteria, not by hand: both items hold the *same* belief, formed at the *same* place after the *same* 15 minutes, so the columns differ only in whether the world moved. A memory without a validity model therefore answers both identically and confidently, and is wrong on the stale one; EWM attaches 0.93 to the belief, which is below τ, and hedges both; the rule accepts a caveat on a fresh item and scores over-abstention separately. The hedges are identical because the covariates are: on same-place pairs EWM cannot discriminate (Sec. 6.2). In the third column both systems correctly report never having seen the object.

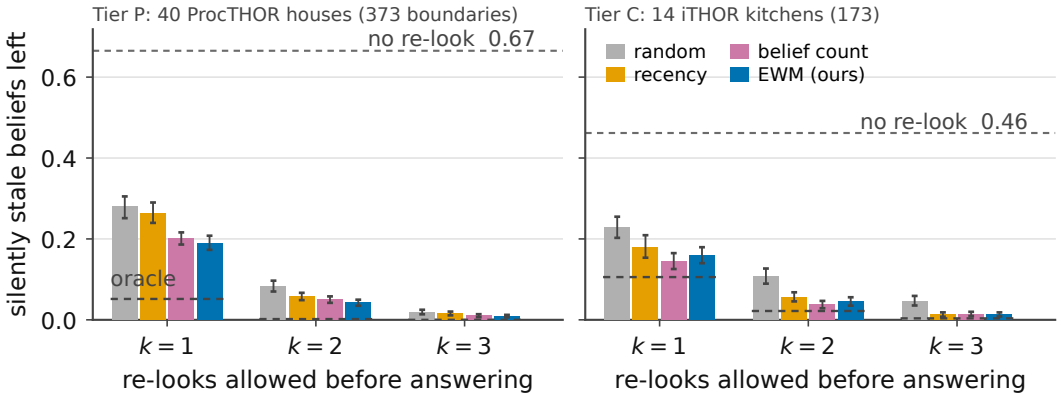


Figure 5: **Active re-observation.** Beliefs left *silently* stale after k re-looks chosen by each policy, with 95% bootstrap intervals over session boundaries. The hazard is fitted on HD-EPIC and nothing is fitted on either substrate. Ranking places by expected stale mass is the best budget-limited policy on procedurally generated houses at every budget; on the hand-built kitchens a belief-count heuristic matches it, because relocations there are drawn from the objects present. The dashed rule is a greedy oracle with ground truth.

E ADDITIONAL FIGURES

Three figures: the re-observation result that Sec. 6.4 refers to, the matched pair as two belief timelines, and the composition of behavioural failures.

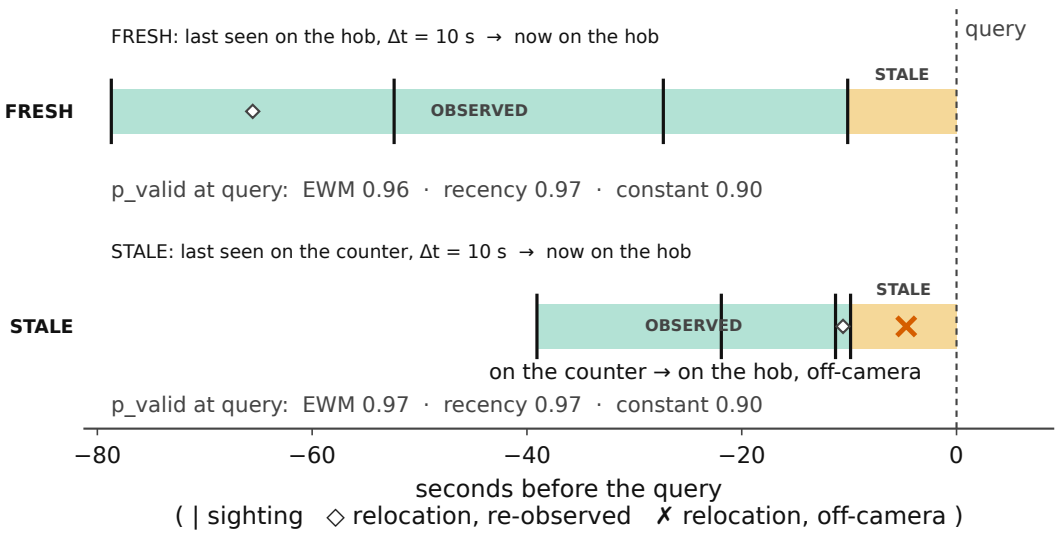


Figure 6: **One Δt-matched pair** (same object instance, a serving spoon, HD-EPIC kitchen P09). Both items share template, wording, evidence budget and observation gap ($\Delta t = 10$ s). In the FRESH member the object is still where it was last seen (the hob); in the STALE member it was relocated off-camera (counter $\rightarrow$ hob, $\times$) between the last sighting and the query while the agent was not looking. Ticks are sightings; the dashed line is the query time. No monotone function of Δt can separate the two, and this pair also defeats EWM: the stale belief was formed on a counter, where real expiry is 0.029, and the fresh one on a hob, so the hazard ranks them the wrong way round, 0.970 against 0.958. Discrimination is a rate and not a per-pair guarantee: 0.889 on pairs whose last-seen places differ and 0.545 on those that share one (Tab. 12).

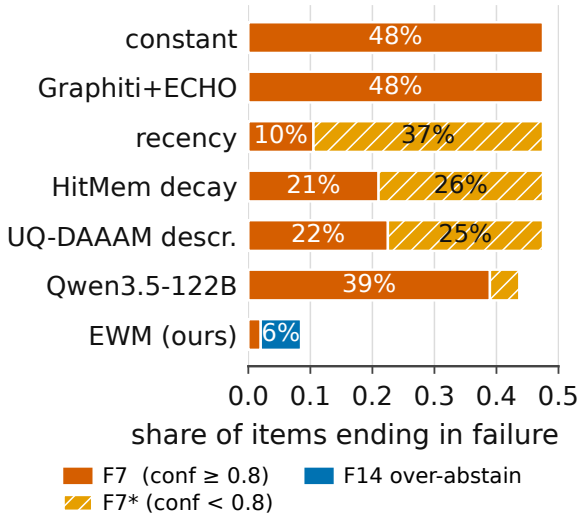


Figure 7: **Failure composition** (fraction of items ending in each behavioural error code, v0.9, including the hosted 122B VLM; counts in Tab. 7). Baselines fail entirely by F7, the stale value asserted at confidence 0.8 or above. Recency and HitMem shift mass to F7* (asserted below 0.8) only on Tier C's hour-long gaps, by being uniformly diffident. EWM is the only system with F14 (declining on a fresh item), the price of being the only one that declines at all. Oracle-perception condition, so perception failures F1–F4 are absent by construction.

F FURTHER ANALYSES

Descriptive uncertainty is anti-predictive. UQ-DAAAM's $\log\det S_j$ score (Zhang et al., 2026a), with its held-out normalisation and its top-20% abstention threshold, scores PVD 0.400 overall and 0.342 on the controlled tier at $p = 0.001$, which is significantly below chance. The mechanism is visible in the score itself: an object sighted in several places yields high caption scatter and therefore low confidence, while being unusually well observed. The metric conflates having been seen in several places with being hard to identify. Used as a validity proxy it is worse than using nothing.

A simulator can invert a validity evaluation. Our first controlled tier injected relocations with uniform-random destinations. That reversed the real regularity, giving cupboard expiry 0.13 against counter 0.33, and a model fitted on real data then scored PVD 0.300 at $p = 0.008$ on that substrate, significantly below chance. A retrieval-driven policy restores the enclosure and surface mechanism qualitatively, at cupboard 0.846, but not the magnitudes, with a base rate of 37% against the real 6%, so we stopped there: tuning an evaluation substrate until it agrees with the model under test is circular. The controlled tier is therefore used for pair supply and long horizons only, never for fitting, and a validity model fitted on real data is at or below chance on it: evaluated as pair sets distinct from the Tier C pairs of Tab. 1, 40 pairs over 25 ProcTHOR houses and 40 over the frame-bearing iTHOR kitchens give the HD-EPIC-fitted hazard PVD 0.475 ($p = 0.875$) and 0.350 ($p = 0.058$), while the closed-form baselines sit between 0.500 and 0.562 on the same pairs. Any staleness benchmark built from injected relocations inherits this hazard.

Why the horizons are fixed in advance. The first build of the EPIC-KITCHENS-100 tier ran each surviving row to the end of its video and each expired row to the next manipulation, which makes elapsed time protective by construction, since the long rows are the censored ones, and scored a global exponential at AUROC 0.363. Reading the object's narrated state at fixed horizons from 30 s to 20 min makes Δt independent of the outcome, and expiry then rises monotonically with horizon.

Failure composition. Baselines fail entirely through one code, asserting the stale value at confidence 0.8 or above. Recency and HitMem shift mass to the same error below that threshold only on the controlled tier's hour-long gaps, by being uniformly diffident rather than by discriminating. EWM is the only system that ever declines on a fresh item, at 0.151 of them on the certified set, which is the price of being the only system that declines at all. The language-model arms never flag a stale belief in any condition. Over the stale items they answer wrongly, the 14B's reported confidence runs 0.85 to 1.00 with the memory in the prompt, a confidently-wrong rate of 1.000 at the 0.8 threshold; the two hosted models, measured over all 33 real pairs rather than the 20 cleanly parsed ones an earlier run reached, sit at 0.970, which is 32 of 33 (Tab. 9). Offering abstention explicitly in the system prompt leaves the 14B at 0.800 and still produces no caveat on any stale item; on the 235B it lowers the rate to 0.879 and produces a caveat or a decline on 3 of 33, while costing 0.061 of fresh-item accuracy. No arm, in any condition, at any scale we could reach, ever says that a belief might have expired. A third arm answers the obvious objection that a model which deliberates might do better. A 20B reasoning model, run one item per call on the 33 real pairs with its answer parsed from every response, scores PVD 0.515 ($p = 1.000$) and AUROC 0.498, flags no stale belief at all, and asserts the expired value above 0.8 confidence on 28 of the 33 it gets wrong. Its reported confidence runs 0.70 to 1.00. Sanctioning abstention does not rescue it either. Told in the system prompt that declining is acceptable, the same model answers wrongly on 22 of the 33 stale items rather than on all 33, and its confidently-wrong rate falls from 0.848 to 0.606; but it still flags no belief as stale, still answers no stale item correctly, and the extra refusals are untargeted, taking fresh-item accuracy from 1.000 to 0.485. It is not identifying expiry, only declining more often. Deliberation does not supply the missing quantity, because the quantity is not in the prompt: nothing the model can read says how long a belief of this kind survives unobserved. Getting that measurement right needed one correction worth recording. This model emits its reasoning in a separate channel that is nonetheless drawn from the same generation budget, so under the 160-token cap we first used, 23 of 66 answers arrived empty and scored as wrong: the arm measured the cap. At 512 tokens no answer is truncated and none fails to parse.

Label legibility and shortcut gates. A human annotator and a language model, each shown only timelines and the class definitions and never the derived label, reproduce those labels at $\kappa = 0.897$

Table 2: **The same on the uncertified v0.9, for reference.** Before the shortcut cuts; the certified set is the one the paper cites.

system	PVD ↑ all	Tier R PVD ↑	Tier R p	Tier C PVD ↑	CDD ↑ Tier R	behav. STALE ↑
B0 constant	0.500	0.500	n/a	0.500	0.500	0.000
B9 Graphiti+ECHO	0.500	0.500	n/a	0.500	0.500	0.000
B12 recency	0.494	0.515	1.000	0.482	0.500	0.000
B14 HitMem decay	0.500	0.530	0.860	0.482	0.500	0.000
B11 descriptive UQ	0.400	0.500	1.000	0.342	0.500	0.000
EWM (ours)	0.578	**0.727**	0.014	0.491	0.606	0.967

90 pairs (Tier R $n = 33$, Tier C $n = 57$).

Table 3: **The same place, the opposite hazard.** Expiry rate by the place a belief names, horizon-matched to 30 to 300 s (n = 4,525 HD-EPIC, n = 6,702 EK-100). A sighting inside an enclosure means the agent was retrieving from it; a placement inside one means it was putting something away. The two orderings correlate at Spearman $\rho = -0.43$.

place	sighting-formed ↓	placement-formed ↓
fridge	0.958	0.096
drawer	0.690	0.212
cupboard	0.607	0.155
counter	0.031	0.291
sink	0.075	0.272
hob	0.049	0.178

(95% CI $[0.762, 1.000]$, raw 0.94, median 12.3 s per item) and $\kappa = 0.828$ (95% CI $[0.664, 0.963]$, raw 0.90) over the same 50 items; per class the human reaches 1.000 on never-observed, 0.885 on stale and 0.864 on fresh. They agree with each other at $\kappa = 0.930$, with two disagreements in fifty. The model figure is a legibility proxy and is labelled as one throughout. The human figure is one annotator, and the pre-registered gate asks for two, so no pairwise human κ is reported and the gate is open. That check is a built instrument rather than an intention: a blinded task builder, an offline annotator needing no server, an agreement analysis that refuses to emit a gate verdict from fewer than two human annotators, and an adjudication pass over the disagreements. About an hour of one non-author's time completes it. The value of running a model through the same instrument first was practical: it caught a leak that would have made the human ceiling meaningless, before anyone was asked for their time. A third pass, by an author, is retained and excluded by protocol; at 8.6 s per item against the human's 12.3 it reached $\kappa = 0.126$, which is what a skimmed pass looks like and is why pace is now reported beside every human κ. The withdrawn contradiction class deserves its own sentence, because we tested the obvious repair and the two labellers split over it. The proposal was to narrow contradiction to co-temporal or unordered disagreement and give sequential disagreement its own class. A language model never used the new class at all, calling the items fresh or stale, so its agreement fell. A human annotator used it on eight of the nine items she reached and agreed with the original class on all nine. The axis is therefore real to a reader and the redefinition points the right way; what a reader does not reproduce is our one-second boundary, since she treats a disagreement a fifth of a second apart as supersession too. On one annotator and nine items that unseats the boundary without replacing it, so the class stays withdrawn. The three shortcut gates, which try to answer from the question alone, from the last sighting alone and from the current view alone, leave 0.5%, 2.5% and 5.0% of v0.9 solvable against pre-registered limits of 30%, 40% and 40%. The 14 solvable items and their 4 pair partners are cut, which defines the certified set of 182 items and 86 pairs.

Table 4: **Discrimination on v1.1, all items and all three question forms.** Every situation, certified or not, asked in three question forms. The PVD column is identical to Tab. 2 to three decimals and must be: a closed-form confidence reads the belief and the gap and never the wording, so the 270 pairs carry exactly the information of their 90 situations. The pair-level p printed here therefore counts each situation three times and is anti-conservative; the situation-level sign test and a cluster bootstrap over situations give 0.727, $p = 0.014$, $[0.576, 0.879]$ on the 33 real situations and 0.578 at $p = 0.161$ pooled (Sec. F).

system	PVD ↑	Tier R		Tier C	CDD ↑	behav.
	all	PVD ↑	p	PVD ↑	Tier R	STALE ↑
B0 constant	0.500	0.500	n/a	0.500	0.500	0.000
B9 Graphiti+ECHO	0.500	0.500	n/a	0.500	0.500	0.000
B12 recency	0.494	0.515	0.841	0.482	0.500	0.000
B14 HitMem decay	0.500	0.530	0.610	0.482	0.500	0.000
B11 descriptive UQ	0.400	0.500	1.000	0.342	0.500	0.000
EWM (ours)	0.578	**0.727**	<0.001	0.491	0.606	0.967

270 pairs (Tier R $n = 99$, Tier C $n = 171$).

Table 5: **The conformal target risk sets the decision split, not the discrimination** (certified v0.9, 182 items). Behavioural accuracy on stale items spans its whole range as the knob moves; PVD and Tier R CDD do not move at all. The base expiry rate is 6%, so a target risk below that drives τ towards 1 and the hedged band swallows nearly every item.

risk	τ	assert/caveat/abstain	behav. F	behav. S	PVD	CDD	CDD-R
0.02	0.976	6/143/33	0.849	0.977	0.558	0.529	0.613
0.05	0.849	148/1/33	0.849	0.221	0.558	0.535	0.613
0.10	0.800	149/0/33	0.849	0.209	0.558	0.529	0.613
0.20	0.800	149/0/33	0.849	0.209	0.558	0.529	0.613

Table 6: **Active re-observation.** Fraction of beliefs left *silently* stale after k re-looks chosen by each policy, at session boundaries with at least one stale belief (Tier P $n = 373$, Tier C $n = 173$). The hazard is fitted on HD-EPIC; nothing is fitted on either substrate. On Tier C a belief-count heuristic matches the hazard because relocations there are drawn from the objects present (Sec. 7).

	Tier P (ProcTHOR)			Tier C (iTHOR)		
policy	k=1	k=2	k=3	k=1	k=2	k=3
random	0.280	0.083	0.019	0.229	0.108	0.047
recency	0.264	0.058	0.015	0.180	0.056	0.013
belief count	0.201	0.050	0.010	0.145	0.037	0.014
EWM	0.190	0.042	0.009	0.160	0.045	0.013
oracle (greedy, GT)	0.052	0.002	0.000	0.106	0.022	0.004

Table 7: **Failure taxonomy** (behavioural errors). Counts of items. F7: the stale value asserted at ≥ 0.8 confidence; F7*: asserted below 0.8, the only benign way to be wrong on a stale item; F11: a reasoning error; F14: over-abstention on a fresh item; F16: a parse failure. The first block is v0.9 (200 items); the hosted- and local-model blocks are the 66 real-tier items of Tab. 9, elicited one item per call; the perception blocks are the 95 frame-bearing controlled items of Tab. 8. The 122B's four parse failures are all in the 56 rows a 2026-09-03 pass produced before per-item checkpointing existed; every row elicited since parses. Regenerated from `results/e9_*.json`.

system	items	errors	F7	F7*	F11	F14	F16
constant	200	95	95	0	0	0	0
descriptive UQ	200	95	45	50	0	0	0
recency	200	95	21	74	0	0	0
HitMem decay	200	95	42	53	0	0	0
Graphiti+ECHO	200	95	95	0	0	0	0
EWM (ours)	200	17	4	0	0	13	0
gpt-oss-20B, real tier	66	33	28	5	0	0	0
gpt-oss-20B, abstention offered	66	39	20	2	2	15	0
Qwen3-14B, real tier	66	34	29	3	1	0	0
Qwen3.5-122B, real tier	66	36	32	0	0	0	4
Qwen3-VL-235B, real tier	66	33	32	1	0	0	0
Qwen3-VL-235B, abstention offered	66	32	29	1	0	2	0
Qwen2.5-VL-7B, frames only	95	55	0	15	24	16	0
Qwen2.5-VL-7B, frames + memory	95	61	17	1	26	3	0
Qwen3-VL-235B, frames only	95	57	0	17	19	21	0
Qwen3-VL-235B, frames + memory	95	43	29	4	4	6	0

Table 8: **Shown the frames** (real-perception arm F1–F3 on the 95 frame-bearing controlled items, 40 pairs, frames rendered at the recorded camera poses; a 7B local model and a hosted 235B one on identical items). The pattern is the same at both ends of a 33-fold range in size. From frames alone each model is uniformly diffident: it declines on a quarter to a half of the fresh items, which raises stale-item behaviour without discriminating, and almost every pair ties on confidence. Add the text memory and the diffidence goes, taking the confidently-wrong rate from 0.000 to 0.750 and 0.900: resolving what the object *is* converts abstention into confident error rather than supplying any signal about whether the belief still holds. Staleness is flagged on 1 of 240 stale judgements across both models and all three conditions. The F3 PVDs are reported with their p and are *not* interpreted: at 40 pairs the design cannot separate 0.55 from 0.65.

condition	model	behavioural ↑		CWR ↓	over-abst. ↓	flags ↑	PVD
		FRESH	STALE		FRESH	STALE	
F1 last frame only	7B	0.300	0.425	0.000	0.400	0.025	0.463 (29 ties)
	235B	0.250	0.425	0.000	0.525	0.000	0.500 (40 ties)
F2 four frames	7B	0.275	0.300	0.000	0.525	0.000	0.425 (30 ties)
	235B	0.375	0.125	0.000	0.250	0.000	0.512 (39 ties)
F3 frames + memory	7B	0.625	0.075	0.900	0.075	0.000	0.537 (27 ties)
	235B	0.850	0.150	0.750	0.150	0.000	0.600 (18 ties)

Table 9: **The hosted models on real pairs, one item per call** (33 Δt-matched real-kitchen pairs, 66 items, every confidence parsed). Both models sit *below* chance: each is on average more sure of a belief that has expired than of one that still holds. Sanctioning abstention in the system prompt is the fair test of the objection that the models were never told they could decline; it moves the operating point and adds no discrimination. Elicitation is one item per call throughout, which a batched run cannot substitute for on any within-pair statistic (Sec. D).

arm	PVD	p	behavioural ↑		CWR ↓
			FRESH	STALE	
122B, no abstention offered	0.364	0.022	0.909	0.000	0.970
235B, no abstention offered	0.409	0.070	1.000	0.000	0.970
235B, abstention sanctioned	0.455	0.581	0.939	0.091	0.879

Table 10: **Does size close it?** The same 33 real Δt-matched pairs, one item per call, condition A, identical prompts, over a 17× span in parameters. No size separates stale from fresh above chance, the confidently-wrong rate rises rather than falls, and staleness is flagged on zero of 33 stale items at every size. Four points from four model families describe rather than fit: we quote no slope, and the claims that need none are the last two columns.

model	B	PVD	CWR ↓	AUROC	STALE ↑	flags ↑
Qwen3-14B	14	0.424	0.879	0.429	0.000	0.000
gpt-oss-20B	20	0.515	0.848	0.498	0.000	0.000
Qwen3.5-122B	122	0.364	0.970	0.450	0.000	0.000
Qwen3-VL-235B	235	0.409	0.970	0.417	0.000	0.000

Table 11: **Does the validity hazard transfer across real datasets?** AUROC of the fitted hazards. HD-EPIC: 9,120 away gaps, beliefs formed by *sighting*; EK-100: 7,726 fixed-horizon rows, beliefs formed by *placing*. Place beats a global decay within either dataset and is worse than one across them.

protocol	global ↑	place ↑	observ. ↑
EK-100 alone, LOO participant (26)	0.645	0.690	0.747
HD-EPIC → EK-100	0.661	0.612	0.518
EK-100 → HD-EPIC	0.590	0.553	0.623

Table 12: **PVD of fitted validity models on the 40-pair Tier R pilot** (trained on other kitchens). 40 pairs: 18 with differing last-seen places (separable), 22 same-place. A place-only model's ceiling is **0.725** (all separable pairs right + chance on the rest).

model	PVD ↑	p	separable pairs (n=18) ↑	same-place pairs (n=22) ↑
M1-global-exp	0.525	0.8746	0.444	**0.591**
M2-class	0.525	0.8746	0.444	**0.591**
M3-place	**0.725**	0.0064	**0.889**	**0.591**
M4-cls+place	**0.725**	0.0064	**0.889**	**0.591**
M5-observ.	**0.725**	0.0064	**0.889**	**0.591**
M5w-obs.-weibull (ours)	0.700	0.0166	**0.889**	0.545

Table 13: **Per-template behaviour** (90 pairs × 3 question forms; correlated, never pooled). The closed-form rows are v1.1, the item set with the withdrawn class removed; the 14B rows are v1.0, the only run in which a language model was asked all three forms. A closed-form confidence does not read the wording, so its columns are identical across forms by construction; the language model's are not.

system	template	FRESH behav. ↑	STALE behav. ↑	CWR (STALE) ↓	PVD ↑	p
HitMem decay	T1_where_now	**1.000**	0.000	**0.444**	0.500	1.000
HitMem decay	T4_what_changed	**1.000**	0.000	**0.444**	0.500	1.000
HitMem decay	T7_still_in	**1.000**	0.000	**0.444**	0.500	1.000
Qwen2.5-14B	T1_where_now	0.978	0.000	0.967	0.572	0.079
Qwen2.5-14B	T4_what_changed	0.944	0.011	0.922	0.572	0.047
Qwen2.5-14B	T7_still_in	0.744	0.078	0.678	**0.589**	0.072
EWM (ours)	T1_where_now	0.856	**0.967**	0.800	0.578	0.161
EWM (ours)	T4_what_changed	0.856	**0.967**	0.800	0.578	0.161
EWM (ours)	T7_still_in	0.856	**0.967**	0.800	0.578	0.161

Table 14: **Expiry rate by the place the object was last seen** (away gaps). Tier R base rate 0.060; Tier C base rate 0.374. Tier C is engineered (retrieval-driven policy). Across the seven places both tiers carry, the two orderings are uncorrelated (Spearman $\rho = 0.04$) and the magnitudes do not carry over; only the coarse contrast reproduces, the enclosure ranking above every surface in both tiers. All but one Tier C place rests on $n \leq 26$. "n/a": place absent from Tier C.

	Tier R (real)		Tier C (controlled)	
place	P(expired)	n	P(expired)	n
counter	0.029	6627	0.331	251
hob	0.053	883	0.391	23
sink	0.074	919	0.458	24
table	0.109	119	0.385	26
bin	0.111	18	n/a	n/a
microwave	0.121	33	n/a	n/a
dishwasher	0.145	62	n/a	n/a
oven	0.200	50	n/a	n/a
floor	0.321	28	0.235	17
shelf	0.378	37	0.240	25
cupboard	0.517	147	0.846	26
fridge	0.545	66	n/a	n/a
drawer	0.614	88	n/a	n/a

Table 15: **Label legibility: labelling blind vs the derived labels** (50 published-set items, one item per call; each labeller saw timelines and class definitions only, never the derived label). A human annotator reaches $\boldsymbol{\kappa = 0.897}$, 95% CI $[0.762, 1.000]$, raw 0.94; a language model reaches $\kappa = 0.828$, 95% CI $[0.664, 0.963]$, raw 0.90; the two agree with each other at $\kappa = 0.930$ with two disagreements in fifty. The model figure is a legibility proxy. One human is *not* the pre-registered gate, which asks for two and remains open. The contradiction class is shown separately because it is not in the published set: it was withdrawn in v1.1, and the ten items the project holds are supplied from earlier versions and excluded from κ. Narrowing it and adding a class for sequential disagreement divides the labellers: the model never used the new class, while the human used it on eight of nine items and agreed with the original class on all nine. The class stays withdrawn, but the reader who reaches for supersession is the human one.

class	human	model	n
UNOBSERVED	1.000	1.000	5
STALE	0.885	0.821	23
FRESH	0.864	0.773	22
CONTRADICTED (withdrawn, excluded)	n/a	0.30	10

Table 16: **Shortcut gates** on benchmark v0.9, run with a local 14B language model ($n = 200$). Items flagged for cutting: 14, plus their 4 pair partners.

gate	result	detail
question only	PASS	0.5% solvable (limit 30%); 1 item to cut
last sighting only	PASS	2.5% solvable (limit 40%); 5 items to cut
current view only	PASS	5.0% solvable (limit 40%); 10 items to cut

Table 17: **Selective risk, natural distribution**. Bootstrap (2,000): AURC(recency) − AURC(EWM) = **+0.0185** [+0.0133, +0.0241]; cov@5%(EWM) − cov@5%(recency) = **+0.338** [+0.141, +0.485].

model	AURC ↓	cov. @5% risk ↑
M1-global-exp	0.0463	0.613
HitMem	0.0455	0.618
M3-place	0.0315	0.955
M5w-observable-weibull (ours)	**0.0275**	**0.980**

Table 18: **Validity prediction, natural distribution** (9,120 held-out away gaps, LOO-kitchen). M5w is the validity model used by EWM.

model	AUROC ↑	Brier ↓	ECE ↓
M1-global-exp	0.587	0.0609	0.046
M2-class	0.593	0.0612	0.045
M3-place	0.705	0.0566	0.031
M4-class+place	0.713	0.0562	0.031
M5-observable	0.724	0.0505	0.015
M6-plus-oracle	0.730	0.0506	0.016
M5w-observable-weibull (ours)	**0.765**	**0.0467**	**0.003**