BELIEF BEYOND VISIBILITY: GROUNDING, UPDATING, AND PRESERVING PHYSICAL STATE FOR MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex, long-horizon manipulation requires robots to update their estimates when interactions change an object's state and retain supported information when objects become hidden and viewpoints shift. Memory-augmented policies exploit interaction histories, while persistent scene models retain objects and relations across observations. However, task success and average attribute accuracy do not reveal when, after a transition, the system jointly recovers the required physical attributes of the correct task object. We formulate *task-conditioned structured belief maintenance*, including causal estimation of object existence, visibility, object–container relations, spatial evidence, and container articulation, with physical state separated from task roles. We introduce CERB (Control-Evidence-Routed Belief), which combines a learned multi-view object-centric observer with action-conditioned temporal inference and measurement-based belief revision. In CERB, state candidates are inferred from visual history, executed actions, and proprioceptive feedback, corrected by accumulated observation evidence, and integrated with the previously accepted belief by a learned update. Thus, new evidence can rectify an erroneous estimate without requiring a physical transition. To evaluate these capabilities, we introduce PRIMARY35, a 35-task, three-view benchmark whose event-aligned criteria require applicable task binding and physical fields to recover together. Across three training repetitions, under the same frozen observer and complete output interface, CERB improves actionable recovery from 38.9% to 59.6% and complete-object recovery from 43.0% to 45.3% over the strongest adapted temporal baseline on in-distribution episodes held out from temporal training. The actionable-recovery gain persists when target-object assets are held out from temporal training.

1 INTRODUCTION

Consider a robot that places a jar into a refrigerator as part of a multi-step task. Once the refrigerator door is closed, the jar may no longer be visible, even though its physical state has not changed: it remains inside the refrigerator. A subsequent task query may then refer to a different object while the underlying scene itself remains unchanged. These cases call for distinct belief updates. The robot must revise state attributes when the world changes, preserve previously supported information when objects become hidden, and rebind task roles without altering the estimated scene.

The challenge is therefore not simply to retain more history, but to determine how different evidence should affect the current belief. A closing gripper signals an attempted grasp, and disappearance from view signals a change in observability. Neither alone determines whether an object is held, contained, or still present. Conversely, a failed action may reveal an incorrect prior estimate even when the world itself has not changed. A belief-maintenance system must therefore infer *what state the evidence supports* and *whether that evidence should revise the current belief*.

Memory-augmented policies and persistent scene models address important aspects of the above problem, but a key question remains: after a physical transition, when does the robot recover the jointly correct state of the correct task object? Average attribute accuracy can be dominated by unchanged intervals, while individually correct attributes may still fail to form a valid task-conditioned belief if

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.

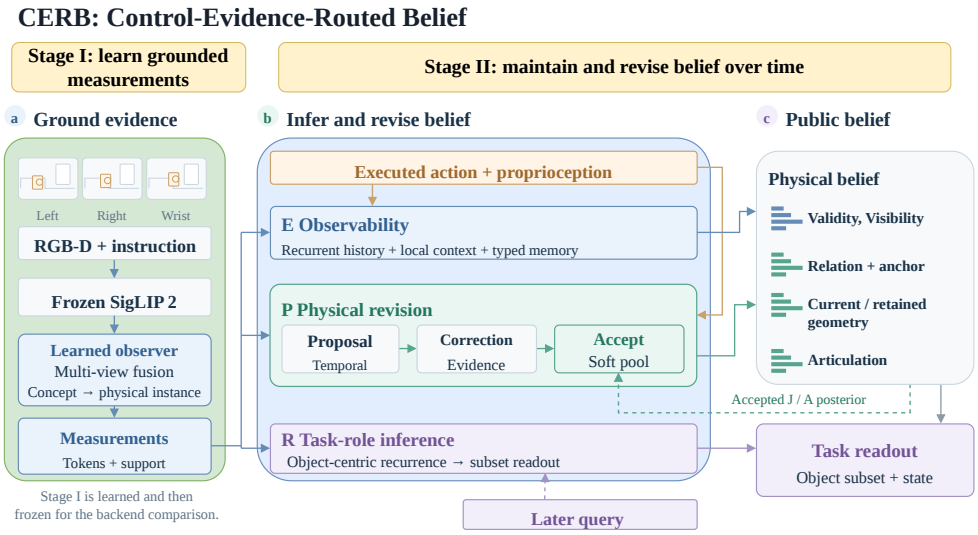


Figure 1: **CERB maintains a queryable physical belief.** A learned multi-view observer grounds current evidence. Three temporal branches estimate observability (E), revise physical fields (P), and read task roles (R). Completed feedback and measurement correction inform the accepted relation/articulation posterior, which becomes the next update's prior. Visibility and geometry retain field-specific evidence rules, and subsequent task queries do not alter physical inference.

they refer to different objects or become correct at different times. Task completion also depends on the planner and robot controller, so it cannot isolate the quality of state estimation.

Accordingly, we formulate *task-conditioned structured belief maintenance* and evaluate recovery by requiring the correct task binding and all applicable physical fields to become jointly correct within a finite post-change window. We introduce PRIMARY35, a 35-task, three-view benchmark that measures joint recovery, update timing, and hidden-state retention. We then propose CERB with a two-stage architecture (Figure 1). Stage I learns an object-centric observation interface from instruction-paired multi-view RGB-D data, grounding object and container descriptions, associating visual evidence with physical instances, and producing current-state measurements with field-specific support. Stage II learns from annotated interaction histories to maintain belief over time. Stage II separates three complementary functions: E tracks validity and visibility, P revises physical state from visual history, executed actions, and proprioceptive feedback, and R maintains task-object bindings. Within P, accumulated measurements correct temporal proposals, while a learned reliability model balances the corrected estimate, the action-conditioned prediction, and the previous belief. The accepted relation/articulation posterior becomes the prior for the next update, while later role queries do not alter physical inference.

The contributions of this study are threefold: **(1) A task and event-centered evaluation.** Primary35 measures joint post-change recovery, transition timing, and hidden-state retention. **(2) A learned object-centric observation interface.** The multi-view observer couples language grounding, physical-instance association, calibrated spatial measurements, and evidence support. **(3) A recursive evidence-guided belief model.** CERB combines action-conditioned proposals with measurement-based correction and recurs on the accepted public state, while separating physical estimation from task-role readout. We compare complete temporal backends under this common observer, examine their functional components, and test belief utility with a shared planner and executor.

2 RELATED WORK

Manipulation policies and memory. Memory-augmented manipulation policies condition actions on visual memories (Fang et al., 2025; Shi et al., 2025), retrieved interactions (Shah et al., 2026), recurrent queries or tokens (Li et al., 2026; Cherepanov et al., 2026), and object-centric temporal

slots (Chung et al., 2026). RB-VLA similarly maintains a compact recurrent belief for a diffusion policy (Bagaria et al., 2026). These methods establish the value of history for action selection. CERB makes the supported physical state explicit before planning. We evaluate adapted belief-estimation systems, not the original policies, using a shared learned observer and structured outputs.

Object-centric and persistent scene representations. Object-centric representations learn instance correspondence and dynamics (Locatello et al., 2020; Kipf et al., 2022; Elsayed et al., 2022; Wu et al., 2023). Persistent memories extend scene understanding to occluded objects (Huang et al., 2024), video with depth and pose (Fan et al., 2025), and relational reasoning for manipulation (Qian et al., 2026). Interactive scene graphs support exploration and state revision (Jiang et al., 2024; Yang et al., 2026; Bao et al., 2026). CERB builds on object persistence, but correspondence alone does not establish whether an object is held or contained. Its learned observer supplies field-specific evidence, and its temporal model maintains physical-state probabilities separately from task roles.

Belief estimation and its evaluation. Partially observable models and differentiable filters estimate hidden state (Kaelbling et al., 1998; Lee et al., 2020), while world models support latent imagination, physical expectations, and action-conditioned prediction (Hafner et al., 2020; Piloto et al., 2022; Zheng et al., 2026). CERB learns a discriminative update over named physical fields rather than an exact Bayesian filter or a future-video generator. Primary35 evaluates when those fields recover jointly for the correct task object, complementing the execution-based memory evaluations of LIBERO-Mem, RoboMME, and RoboDojo (Chung et al., 2026; Dai et al., 2026; Chen et al., 2026). Gated DeltaNet-2 and Mamba-3 (Hatamizadeh et al., 2026; Lahoti et al., 2026) supply additional recurrent operators for our system comparisons.

3 Problem Formulation and Evaluation

3.1 Causal, task-conditioned physical belief

Let $t = 0, \ldots, T-1$ index visual observations in an episode of length T. At boundary t, the estimator receives three synchronized RGB-D views I_t with calibrated camera geometry and the original instruction ι. For $t \geq 1$, it also receives the elapsed action–proprioception interval:

$$\mathcal{U}_t = \{(a_k, p_{k+1}^{\text{prop}})\}_{k=n_{t-1}}^{n_t-1} \tag{1}$$

where k indexes robot action steps, a_k is the action command applied at step k, p_{k+1}^{prop} is the proprioceptive measurement recorded after that action, and n_t is the index of the latest available proprioceptive measurement at observation t. At the initial observation, $\mathcal{U}_0 = \varnothing$, while the initial proprioceptive measurement $p_{n_0}^{\text{prop}}$ is available. Each pair is available only after the action has been applied and its proprioceptive feedback recorded. Executing the command does not imply that its intended physical effect occurred. Actions and post-action proprioception are recorded at 20 Hz, while multi-view observations arrive at a nominal 4 Hz. At each visual step, CERB processes the complete action–proprioception sequence between the preceding and current observations. The estimator may therefore use $I_{0:t}$ and all executed actions and proprioceptive feedback through n_t, but not I_{t+1} or a future action outcome.

Object and container registers index the predicted state variables. We use i for a register and f for one of its physical fields. Each physical belief comprises object validity, five-state visibility V, a joint primary relation–anchor variable J, current and last-supported coarse position distributions G, and container validity and articulation state A. The 13-state relation space includes UNKNOWN, NONE, HELD, and IN/ON relations with respect to each of five anchors. Rather than encoding all simultaneously valid relations, J records a single primary relation according to a predefined priority rule. The spatial fields preserve the most recently supported geometric evidence when direct observations are unavailable and are also included in recovery evaluation whenever applicable.

A subsequent query q_t selects a role-dependent subset of objects from the already estimated scene. Let $\mathcal{B}_t^{\text{phys}}$ denote the physical part of the belief. With the original instruction, sensory and action history, register interface, and initialization held fixed, any two subsequent queries q_t and q_t' leave it unchanged:

$$\mathcal{B}_t^{\text{phys}}(q_t) = \mathcal{B}_t^{\text{phys}}(q_t') \tag{2}$$

Only the role answer may change. This distinction enables the same estimated scene to support a new task target without treating rebinding as physical motion. Primary35 instantiates these causal inputs and state variables in manipulation streams, where reference annotations identify the physical changes to be evaluated.

3.2 Instantiating the Problem on Manipulation Streams

Primary35 pairs multi-view RoboCasa365 manipulation streams (Nasiriany et al., 2024; 2026) with object-centric state annotations. Natural trajectories and articulation examples are complemented by paired interventions: from the same initial state and action sequence, the attempted action either changes or leaves unchanged the target state, whose subsequent observations may differ. Related episodes remain together when forming training and development partitions.

The *in-distribution* (IID) test uses new episodes from the task and object distribution used for temporal learning. The *target-asset* test instead uses target objects whose 3D assets were entirely withheld from temporal learning. Both assess temporal generalization under a shared frozen observer. Table 5 summarizes the partitions, and Appendix H analyzes observer-source overlap and task composition.

Candidate information. All declared object registers, container anchors, role queries, and role–object candidates remain available during temporal training, validation, and testing. Episode metadata does not restrict these candidate sets. Appendix D.1 distinguishes this input setting from the validity indicators used for supervision and scoring.

Reference states and events. Reference states use prediction-independent grasp adjudication based on contact and relative motion, with the same rules for training and evaluation (Appendix E). Ambiguous cases retain their original labels and are examined separately. A scored transition requires three-step reference dwell on each side, with onset at the first destination step. These label-defined events anchor the evaluation of whether, and how promptly, belief responds to a change.

3.3 Event-Centered Recovery, Timing, and Persistence

Evaluation proceeds from instantaneous joint correctness $C_{e,t}$, through event-level recovery $H_{e,h}$, to aggregate ABR and Whole scores. For an object–role event e, the binary indicator $C_{e,t}$ equals one when role binding and all applicable physical fields are correct at the same time t. For a task-addressed container event, it instead checks container validity and articulation. Applicable geometry checks use a coarse 0.20 m tolerance.

Recovery asks whether this joint correctness is reached within a finite post-change window. Let τ_e denote the event onset, τ_e^+ the onset of the next event in the same stream, T the episode length, and h the maximum allowed recovery offset in visual steps. If no later event exists, set $\tau_e^+ = +\infty$. We define

$$H_{e,h} = \mathbf{1}\big[\exists t \in [\tau_e, \min(\tau_e + h, \tau_e^+ - 1, T - 1)] : C_{e,t} = 1\big] \tag{3}$$

Here $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. Thus, $H_{e,h}$ records whether a jointly correct estimate occurs at least once within the window. *Actionable Belief Recovery* (ABR) computes a recovery rate over eligible object–role and container events within each episode, then averages episode scores within and across source families. ABR@8 permits a two-second delay at 4 Hz, including onset and at most eight subsequent observations. *Whole@8* applies the same windowed-recovery logic without role matching to a broader object-event population. The scores therefore differ in both correctness requirements and event population, so subtracting them does not isolate binding errors.

A recovery hit alone does not establish correct update timing, accurate probabilities, or continued correctness. We therefore report complementary diagnostics. *Transition macro-F1* matches predicted and reference changes one-to-one in causal windows. *Event negative log-likelihood* (Event NLL) scores relation/articulation probabilities around events rather than across all frames. *Hidden retention* checks relation–anchor and applicable articulation at specified endpoints of physically stable hidden intervals. The *reference-stable write rate* measures prediction changes while the reference remains unchanged, including changes that repair an earlier error. Appendix D.2 defines all predicates and reductions.

4 CERB: FROM GROUNDED EVIDENCE TO PERSISTENT BELIEF

CERB addresses these recovery and persistence requirements by separating learning to read the scene from learning to revise its state. Stage I supplies object-centered measurements, and Stage II turns the causal measurement and action history into an explicit physical belief and a query-dependent role answer. The distinction is functional: a powerful latent history is useful only if its information reaches the public state at the appropriate time.

4.1 LEARNING AN EVIDENCE-BEARING OBSERVATION INTERFACE

Stage I converts each synchronized multiview frame into typed measurements and field-specific evidence support. It separates what a field predicts from whether the current frame directly supports that prediction. The observer resets between frames, so temporal history is handled entirely by Stage II. Figures A1 and A2 detail the observation pathway.

Language specifies the entities, and vision supplies the evidence. Learned instruction proposals and multiview fusion adapt frozen SigLIP 2 image/text features (Tschannen et al., 2025). A typed query decoder produces object and container hypotheses from instruction tokens and spans. Instruction-conditioned latents read all three views with calibrated geometry, then write fused context back to spatially indexed patches for instance association and geometric readout. A separate instruction-conditioned query produces the scene token.

Foreground-guided physical-instance association. The observer deliberately factorizes *whether task-relevant evidence exists* from *which register owns it*. Let n index patches across the three views and i index an available object register in $\mathcal{R}_o$. The foreground probability γ_n and conditional ownership probability π_{in} define the foreground mass Γ_{in} assigned to that register. A separate occupancy score w_{in} and valid-depth confidence $d_n \geq 0$ weight the calibrated 3D point p_n^{3D}:

$$\begin{aligned} &\Gamma_{in} = \gamma_n \pi_{in}, && \sum_{i \in \mathcal{R}_o} \Gamma_{in} = \gamma_n, \\ &\omega_{in} = \frac{d_n w_{in}^2}{\sum_{n'} d_{n'} w_{in'}^2}, && \mu_i = \sum_n \omega_{in} p_n^{\mathrm{3D}}, \qquad v_i = \sum_n \omega_{in} (p_n^{\mathrm{3D}} - \mu_i)^{\odot 2} \end{aligned} \tag{4}$$

Here ω_{in} is the normalized pooling weight, $\mu_i \in \mathbb{R}^3$ the mean position, and $v_i \in \mathbb{R}^3$ the coordinate-wise variance of the weighted depth points. The factor w_{in}^2 emphasizes confident patches, while $\odot$ denotes elementwise multiplication and $\odot 2$ squares each coordinate of the deviation. Ownership probabilities $\pi_{\cdot n}$ normalize over the nonempty register set $\mathcal{R}_o$. Geometry is unavailable when total depth support is zero. A learned identity correction redistributes ownership without changing foreground mass, and detached association scores condition the occupancy branch. Thus additional same-class hypotheses change foreground allocation, not its total mass. Containers use an analogous binder with a NULL alternative.

A typed contract for temporal inference. For every object, Stage I exposes appearance, direct observability, relation–anchor evidence, coarse current position and uncertainty, and field-specific support. Containers expose appearance, articulation, geometry, uncertainty, and support. These outputs form object tokens o_t^{obj}, container tokens o_t^{ctr}, scene context g_t, and support values $\epsilon_t^{i,f} \in [0, 1]$ for field f of register i. Measurement distributions and support are separate: a field may have a well-formed prediction while carrying no direct evidence for a public write. Direct observability distinguishes observed from unsupported evidence, whereas the temporal model uses history to infer the five-state visibility belief.

Implementation and stage boundary. Executed actions and proprioception in $\mathcal{U}_t$ bypass the observer and enter Stage II, where later queries affect only role readout. Stage I learns an instruction-to-role binder after instance association, distinct from the temporal role posterior learned in Stage II. All compared systems share the frozen observer. Appendices A and C.1 specify its architecture, supervision, and container-capacity transfer.

4.2 INFERRING STATE FROM HISTORY AND EXECUTED ACTIONS

Temporal inference estimates what remains true and what has changed. CERB separates observability (E), physical revision (P), and task-role inference (R) into branches with separate histories. Below, E and P form state candidates, and Section 4.3 explains how these revise the public belief.

An attempted action is not an observed outcome. Closing the gripper may grasp an object or leave it unmoved. P therefore uses not only the executed action but also the resulting proprioceptive response. An interval encoder summarizes all valid action–proprioception pairs in $\mathcal{U}_t$, and a continuous recurrent encoder additionally carries their history and proprioceptive differences across intervals. Its output enters the scene context *before* temporal inference, so action feedback can change the proposed state itself. Both encoders stop at the current visual boundary: no later image or feedback is available.

Memory must support both persistence and change. An occluded object's state may depend on distant observations, whereas evidence of a new transition can be concentrated in the last few frames. In addition, visibility can change with viewpoint while containment remains unchanged. E and P address these different temporal demands with three complementary paths: a two-layer GRU carries long-range context s_t^i for register i, while causal local attention summarizes recent observations in ℓ_t^i, and typed memory $m_t^{i,f}$ lets each field f retain and revise evidence at its own learned rates. The current register input x_t^i is projected from its observer token and scene context to the same 304-dimensional space as s_t^i, ℓ_t^i, and $m_t^{i,f}$. Learned residual coefficients combine these paths:

$$z_t^{i,f} = s_t^i + \eta_o^{i,f}\phi_o(x_t^i - s_t^i) + \eta_\ell^{i,f}\ell_t^i + \eta_m^{i,f}\phi_m(m_t^{i,f} - s_t^i), \quad \eta_{\{o,\ell,m\}}^{i,f} \in [-1,1] \tag{5}$$

The recurrent state provides continuity, and the residual terms let current evidence, recent context, and field-specific memory adjust it. Here ϕ_o and ϕ_m are learned residual MLPs, and the η coefficients are input-dependent signed gates from the field router, with their time index omitted. Unlike the probability weights below, these coefficients can add or subtract a residual correction. Typed decoders map the combined field representation $z_t^{i,f}$ to a current-state distribution p_t^f, suppressing the register index below. For relation and articulation, a separate action-conditioned head estimates the state at the end of the elapsed action interval, yielding $p_t^{f,\mathrm{act}}$. Both are estimates, not facts to be written automatically. Appendix B.2 gives the memory updates and operator dimensions.

4.3 REVISING THE BELIEF EXPOSED TO THE ROBOT

The public belief supplies state estimates to queries and planning. Revision must correct stale estimates as well as represent physical changes. CERB separates *whether evidence is available* from *which state it supports*, so an attempted action need not become an asserted outcome.

Three alternatives, one accepted belief. For relation and articulation ($f \in \{J, A\}$), the update weighs three alternatives: retaining the previously accepted belief b_{t-1}^f, adopting a measurement-corrected state candidate $\widehat{p}_t^f$, or adopting the action-conditioned candidate $p_t^{f,\mathrm{act}}$. Let ν_t indicate that a completed interval is available and set $e_t^f = \mathbf{1}[\epsilon_t^f > 0 \text{ or } \nu_t = 1]$. A field-specific reliability network r_f produces weights $w_t^f = \mathrm{softmax}(r_f(\cdot))$ over retention (K), the measurement-corrected candidate (O), and the action candidate (A). These three source labels index nonnegative weights summing to one; the source label A is distinct from the articulation field $f = A$. The accepted field distribution is

$$b_t^f = \begin{cases} w_{t,K}^f b_{t-1}^f + w_{t,O}^f \widehat{p}_t^f + w_{t,A}^f p_t^{f,\mathrm{act}}, & e_t^f = 1 \\ b_{t-1}^f, & e_t^f = 0 \end{cases} \tag{6}$$

When neither source of evidence is available, the previous distribution is copied unchanged. Otherwise, the reliability weights determine the extent of revision from field support, candidate uncertainty and disagreement, and temporal context. The predicted probability of physical change contributes to this decision, but does not by itself authorize an update. As a result, a failed grasp may weaken an incorrect HELD belief even without a detected transition, while an executed action cannot force its expected outcome into the public state. If no completed interval is available, its candidate in the

probability pool is replaced by b^f_{t-1}. The accepted result is emitted directly and becomes the prior for the next update.

Measurements can correct a plausible but wrong proposal. The remaining question is how to obtain $\widehat{p}^f_t$. A recurrent measurement encoder compares current and preceding observer tokens, their differences, field support, and proprioceptive response, exchanging evidence across registers. Unlike the command-conditioned proposal path, this encoder does not directly read command vectors. The accumulated measurement evidence provides a state-wise correction δ^f_t to the temporal candidate:

$$\widehat{p}^f_t = \mathrm{softmax}_{\mathcal{S}_f}\left(\log p^f_t + \delta^f_t\right) \tag{7}$$

Here $\mathcal{S}_f$ is the finite legal state set for field f, and p^f_t, $\widehat{p}^f_t$, and δ^f_t are vectors indexed by its states. Logarithm and addition act elementwise: δ^f_t adjusts candidate log scores using measurement evidence, and softmax normalizes them. Equation 6 then pools the corrected candidate with prior belief and the action-conditioned estimate. Appendix B.3 specifies the correction network, fixed output scale, and reliability features.

Field-specific persistence and task readout. Different fields require different update rules. E provides validity and support-weighted visibility, while P supplies relation, articulation, and current/last-supported geometry. Geometry follows a separate support-based carry rule because it has no action-effect head. A detached object-centric temporal readout in R answers task queries without changing physical inference. Appendix B.4 specifies field assembly and Appendix J establishes normalization, valid support, and no-evidence retention.

4.4 Learning the emitted belief

With the observer frozen, temporal learning supervises physical fields, event recovery, hidden retention, role subsets, and action outcomes. Proper-score losses train P's temporal, action, corrected, and public predictions. A command–outcome-balanced term discourages treating an attempted grasp as a successful one, using labels only during training. Learning proceeds from recurrent initialization through typed proposals to physical correction, whose final stage updates all active P parameters while holding the other branches fixed. Appendix C gives the objectives and schedules, and Appendix B specifies the implementation.

5 Experiments

We ask whether CERB improves structured-belief recovery, how its branches and public update contribute, and whether the findings withstand reference sensitivity checks and support downstream action.

5.1 Complete-system comparison

We compare CERB with four complete baselines built around GRU, SlotSSM, Gated DeltaNet-2, and Mamba-3. Each baseline uses one temporal estimator to predict the full belief interface. All systems share the learned frozen observer, causal inputs, register capacities, temporal-learning partition, references, and scorer. CERB instead assigns fields to specialized branches and revises their public state using evidence. CERB itself uses GRU recurrence and SlotSSM-based role inference as building blocks. The comparison therefore evaluates complete belief estimators, not isolated recurrent cells. Appendix C specifies their training schedules.

Joint recovery improves across both test populations. CERB raises ABR@8 from 38.90% to 59.60% on IID episodes and from 49.10% to 61.44% on held-out target assets relative to Gated DeltaNet-2, the strongest recovery baseline (Table 1). Whole@8 gains are smaller, and the asset contrast remains uncertain: $+1.66$ points with a paired 95% interval of $[-.52, 3.94]$. Transition F1 also improves over the strongest timing baseline, SlotSSM, from .1702 to .3048 on IID and .2761 to .4584 on target assets. Event NLL is lower than for all four baselines. CERB and the SlotSSM baseline use the same trained role readout, so their joint-recovery difference reflects physical estimation rather than stronger task binding.

Table 1: **Complete belief systems under a common learned observer.** Baseline names identify their temporal estimator, not the original policy model. GDN-2 denotes Gated DeltaNet-2. Entries are mean ± sample standard deviation over three training repetitions, with source-family aggregation. ABR and Whole are percentages.

Belief system	ABR@8 ↑	Whole@8 ↑	Trans. F1 ↑	Event NLL ↓
In-distribution (IID)				
GRU-based	30.48±2.75	21.30±4.82	0.1381±0.0069	2.5288±0.2186
SlotSSM-based	37.91±0.40	27.61±4.40	0.1702±0.0108	2.1323±0.1073
GDN-2-based	38.90±1.33	43.01±0.68	0.1538±0.0142	1.5311±0.0309
Mamba-3-based	35.19±3.02	40.61±1.69	0.1409±0.0087	1.9071±0.1237
CERB	**59.60**±0.79	**45.26**±2.16	**0.3048**±0.0239	**0.4155**±0.0048
Target-asset shift				
GRU-based	26.10±7.00	22.18±5.92	0.2161±0.0214	3.5624±0.5690
SlotSSM-based	32.98±3.36	32.38±4.06	0.2761±0.0045	2.7691±0.3012
GDN-2-based	49.10±4.23	50.03±3.03	0.1936±0.0217	1.1414±0.0991
Mamba-3-based	46.68±3.09	49.12±1.94	0.1983±0.0097	1.5204±0.2058
CERB	**61.44**±2.11	**51.70**±2.35	**0.4584**±0.0546	**0.3598**±0.0108

The recovery gains are larger for container events than for object events. High recovery can also coexist with early articulation predictions and increased visibility writes, motivating the timing and persistence diagnostics in Appendix H. That appendix separates these behaviors and explains the higher raw asset scores through task composition.

5.2 Functional specialization and public readout

We expose alternative outputs of the trained CERB branches without further optimization (Table 2). $\mathsf{E}+\mathsf{R}$ and $\mathsf{P}+\mathsf{R}$ use all physical fields from one branch with the same role readout. CERB uses E for validity/visibility, P for relation/articulation, and support-based retention of P's geometry.

Table 2: **Recovery and timing under alternative CERB readouts.** All rows reuse the trained branches and the same R role predictions, without further optimization. Means over three training repetitions, with ABR and Whole in percent. Appendix F reports variability.

Readout	ABR@8 ↑	Whole@8 ↑	Trans. F1 ↑	Event NLL ↓
In-distribution episodes				
$\mathsf{E}+\mathsf{R}$	60.58	46.44	.140	.321
$\mathsf{P}+\mathsf{R}$	61.33	43.24	.314	.416
CERB	59.60	45.26	.305	.416
Held-out target assets				
$\mathsf{E}+\mathsf{R}$	64.44	53.29	.178	.229
$\mathsf{P}+\mathsf{R}$	60.53	50.24	.476	.360
CERB	61.44	51.70	.458	.360

Recovery and transition timing favor different readouts. $\mathsf{E}+\mathsf{R}$ attains higher recovery and lower Event NLL than CERB, but its transition F1 is .140 versus .305 on IID and .178 versus .458 on target assets. Recovery asks whether a correct state occurs within a window, whereas transition F1 requires the predicted change itself to match a reference transition. The gap demonstrates why correct state estimates and correctly timed changes must both be evaluated. CERB retains P's relation/articulation predictions while its field assembly raises Whole@8 over $\mathsf{P}+\mathsf{R}$ by 2.02 and 1.46 points, at a small cost in aggregate F1. The resulting trade-off favors physical-transition fidelity over $\mathsf{E}+\mathsf{R}$ and complete-state recovery over $\mathsf{P}+\mathsf{R}$. Geometry retention also changes in the latter contrast, so the gain is not attributable to validity/visibility substitution alone.

Dedicated role inference improves task-object binding. Using R rather than the native E or P role head raises IID role-set exactness by 14.67 and 8.56 points without changing physical predictions.

Appendix F reports variability and field-level comparisons, and Appendix G examines the temporal-proposal components.

5.3 Robustness and downstream utility

Source-exclusion and grasp-resolved analyses preserve CERB's higher mean recovery, although some Whole@8 intervals include zero (Appendix H). These checks separate sensitivity to known observer-source exposure and ambiguous grasp references from the full-cohort comparison.

Table 3: **Belief-guided recovery with the same planner.** Task Success counts successes / rollouts on 18 shared development starts across seven scenes, repeated for each of three trained models per system.

Belief system	Task Success	Hidden	Failed open	Failed grasp
GRU-based	7/54	2/21	3/21	2/12
SlotSSM-based	22/54	8/21	4/21	10/12
GDN-2-based	17/54	7/21	3/21	7/12
Mamba-3-based	13/54	4/21	2/21	7/12
CERB	**27/54**	**10/21**	**7/21**	**10/12**

A shared planner turns field probabilities into OPEN, SEARCH, GRASP, PLACE, VERIFY, and DONE decisions, executed by the same skill library. CERB completes 27/54 development recoveries versus 22/54 for SlotSSM (Table 3), with gains on hidden-target and failed-open starts. Success requires independently measured placement to persist for the final second. This proof of concept connects belief to execution, while the paired scene-cluster interval includes zero. Appendix I gives the planning rule and rollout analysis.

6 Discussion

The results distinguish task-object binding, physical-state estimation, and the timing of belief revision. Separate role inference improves binding, while the physical branches contribute differently to probabilistic accuracy and transition timing. A useful belief must bring these capabilities together at the moment a robot needs to act.

Primary35 studies these capabilities with calibrated simulation observations, fixed-capacity registers, and one primary relation per object. Real sensors, changing object inventories, and simultaneous relations are natural extensions. The shared-planner development study connects belief recovery to physical execution. Independently sampled closed-loop tasks would test this connection more broadly, including the robot's ability to recognize completion and stop.

7 Conclusion

CERB treats manipulation memory as the maintenance of a queryable physical state, not only the retention of latent history. Its learned multi-view observer supplies grounded evidence, while temporal inference integrates visual history, executed actions, and proprioceptive feedback, corrects state candidates, and recurs on the belief exposed to the robot. Primary35 makes the timing and joint correctness of that state measurable. The results support evidence-conditioned belief revision for object-centered recovery, while separating physical estimation, task binding, and downstream execution as complementary capabilities.

REFERENCES

Vaidehi Bagaria, Bijo Sebastian, and Nirav Kumar Patel. Recursive belief vision language action models. *arXiv preprint arXiv:2602.20659*, 2026. URL https://arxiv.org/abs/2602.20659.

Shuo Bao, Wei Dong, Shuyue Zhang, Ming Shang, Yuchen Huang, Han Yu, Chengjie Xu, Yiheng Bi, Kai Sun, Fuchun Sun, and Xinzhou Wang. PBD-AG: Persistent baseline-delta active graphs with uncertainty-aware inspection for long-horizon service robots. *arXiv preprint arXiv:2608.10449*, 2026. URL https://arxiv.org/abs/2608.10449.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, Honghao Su, Zhiyang Dou, Kaixuan Wang, Dandan Zhang, Yunze Liu, Yan Qin, Qiwei Liang, Qiwei Wu, Zijian Lin, Wenwei Lin, Yuran Wang, Minghua He, Tianshu Wu, Ruihai Wu, Jingquan Zhou, Kai-Chong Lei, Haibao Yu, Yuanfeng Ji, Weiyang Jin, Guanyu Lin, Xiaofan Li, Qi Xiong, Renjing Xu, Zhongyu Li, Wenhao Chai, Enze Xie, Ziwei Wang, Yao Mu, Hao Dong, Wojciech Matusik, Mingyu Ding, Wenbo Ding, Ping Luo, and Masayoshi Tomizuka. RoboDojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026. URL https://arxiv.org/abs/2607.04434.

Egor Cherepanov, Nikita Kachaev, Daniil Zelezetsky, Aydar Bulatov, Artem Pshenitsyn, Yuri Kuratov, Alexey Skrynnik, Aleksandr I. Panov, and Alexey K. Kovalev. μVLA: On recurrent memory for partially observable manipulation in VLA models. *arXiv preprint arXiv:2606.12497*, 2026. URL https://arxiv.org/abs/2606.12497.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Nhat Chung, Taisei Hanyu, Toan Nguyen, Huy Le, Frederick Bumgarner, Duy Minh Ho Nguyen, Khoa Vo, Kashu Yamazaki, Chase Rainwater, Tung Kieu, Anh Nguyen, and Ngan Le. Rethinking progression of memory state in robotic manipulation: An object-centric perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3407–3415, 2026. doi: 10.1609/aaai.v40i5.37337. URL https://doi.org/10.1609/aaai.v40i5.37337.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. RoboMME: Benchmarking and understanding memory for robotic generalist policies. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.04639.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, volume 235, pp. 10041–10071, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.07764.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6342–6352, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fan_Embodied_VideoAgent_Persistent_Memory_from_Egocentric_Videos_and_Embodied_Sensors_ICCV_2025_paper.html.

Haoquan Fang, Markus Grotz, Wilbert Pumacay, Yi Ru Wang, Dieter Fox, Ranjay Krishna, and Jiafei Duan. SAM2Act: Integrating visual foundation model with a memory architecture for robotic manipulation. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=anSWDvJm8v.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated DeltaNet-2: Decoupling erase and write in linear attention. *arXiv preprint arXiv:2605.22791*, 2026. URL https://arxiv.org/abs/2605.22791.

Yixuan Huang, Jialin Yuan, Chanho Kim, Pupul Pradhan, Bryan Chen, Li Fuxin, and Tucker Hermans. Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models. In *IEEE International Conference on Robotics and Automation*, 2024. URL https://arxiv.org/abs/2309.15278.

Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. RoboEXP: Action-conditioned scene graph via interactive exploration for robotic manipulation. In *Conference on Robot Learning*, 2024. URL https://jianghanxiao.github.io/roboexp-web/.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.12594.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. *arXiv preprint arXiv:2603.15569*, 2026. URL https://arxiv.org/abs/2603.15569.

Michelle A. Lee, Brent Yi, Roberto Martin-Martin, Silvio Savarese, and Jeannette Bohg. Multimodal sensor fusion with differentiable filters. *arXiv preprint arXiv:2010.13021*, 2020. URL https://arxiv.org/abs/2010.13021.

Hang Li, Fengyi Shen, Dong Chen, Liudi Yang, Xudong Wang, Jinkui Shi, Zhenshan Bing, Ziyuan Liu, and Alois Knoll. ReMem-VLA: Empowering vision-language-action model with memory via dual-level recurrent queries. *arXiv preprint arXiv:2603.12942*, 2026. URL https://arxiv.org/abs/2603.12942.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024. URL https://arxiv.org/abs/2406.02523.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.04356.

Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6:1257–1267, 2022. doi: 10.1038/s41562-022-01394-8.

Jianing Qian, Qinhe Peng, Emmanuel Panov, Leonor Fermoselle, Dinesh Jayaraman, Bernadette Bucher, and Tarik Kelestemur. Expanding spatial and temporal context for robotic imitation learning with scene graphs. *arXiv preprint arXiv:2606.01072*, 2026. URL https://arxiv.org/abs/2606.01072.

Rutav Shah, Yisu Li, Femi Bello, Yuke Zhu, and Roberto Martín-Martín. Memory retrieval in visuomotor policies for long-horizon robot control. In *Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2606.25136.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025. URL https://arxiv.org/abs/2508.19236.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05861.

Heqing Yang, Ziyuan Jiao, Shu Wang, Yida Niu, Si Liu, and Hangxin Liu. Integrated exploration and sequential manipulation on scene graph with LLM-based situated replanning. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2602.04419.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Jun Shi, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, and Xu Jia. Mem-World: Memory-augmented action-conditioned world models for persistent robot manipulation. *arXiv preprint arXiv:2606.18960*, 2026. URL https://arxiv.org/abs/2606.18960.

APPENDIX: ARCHITECTURE, LEARNING, AND SUPPORTING EVIDENCE

The supplement first specifies the architecture, learning objectives, benchmark, and reference semantics. It then examines the complete system and its temporal-proposal components, followed by field-level diagnostics, sensitivity analyses, and a planner proof of concept.

Reproducibility statement. The supplement specifies causal alignment, model modules, initialization and optimization stages, splits, scoring predicates, and aggregation. We will release code, configurations, and checkpoint provenance.

AI-use statement. Generative AI tools assisted with manuscript organization, language editing, editable figure preparation. The authors defined the method, experiments, and adjudication rules, made all final reference decisions, and verified the reported results, manuscript, and figures.

A LEARNING THE OBJECT-CENTRIC OBSERVATION INTERFACE

The observation model attributes current visual evidence to physical entities and estimates their state. It separates concept-level foreground prediction from instance association, then combines appearance with calibrated depth to produce measurements and field-specific support. Figure A1 shows the complete architecture, and Figure A2 details its operators. Observation learning precedes temporal training, as described in Appendix C.1.

A.1 EXTERNAL REPRESENTATIONS AND LEARNED INSTRUCTION PROPOSALS

At visual step t, the observer receives the RGB components of the three synchronized RGB-D views $I_t^{1:3}$, the original instruction ι with token/span representations, calibrated sensor geometry, and input-availability masks. Frozen SigLIP 2 image/text encoders (Tschannen et al., 2025) provide visual features $F_t \in \mathbb{R}^{3\times196\times768}$, instruction-token features $E_\iota \in \mathbb{R}^{L\times768}$, a pooled instruction embedding $e_\iota \in \mathbb{R}^{768}$, and encoded entity-span candidates, where L is the number of instruction tokens. Both encoders have 12 layers, 12 heads, and width 768. The image encoder uses 16×16 patches at 224×224 resolution, yielding 196 patches per view. Their representations are precomputed and remain fixed during observation learning.

The model allocates six object-query slots and five container-query slots. Semantic embeddings and proposal-validity scores are predicted from language, while externally supplied entity embeddings and entity-active masks are not passed to the instruction proposer. Soft proposal-validity scores guide grounding. Predicted support indicates available measurement evidence and remains distinct from sensor availability and supervision validity.

Affine projections with GELU and layer normalization map instruction tokens and the pooled embedding to width 256. Separate learned object and container queries receive their respective type embeddings and the pooled instruction:

$$Q^o_{0,i} = q^o_i + e^o_{\text{type}} + P_{\text{pool}} e_\iota, \qquad Q^c_{0,j} = q^c_j + e^c_{\text{type}} + P_{\text{pool}} e_\iota$$

Here i and j index object and container queries. The vectors q^o_i, q^c_j are learned base queries, $e^o_{\text{type}}, e^c_{\text{type}}$ are their type embeddings, and P_{pool} projects the pooled instruction. The resulting $Q^o_{0,i}, Q^c_{0,j}$ initialize the decoder. A two-layer Transformer decoder uses the concatenated 11×256 queries and the projected instruction tokens as memory. It has eight attention heads, feed-forward width 1,024, and GELU activations. The first six outputs correspond to object queries and the remaining five to container queries. Training matches predicted and target entity sets within each type, allowing query positions to represent different semantic entities across instructions.

Object semantic readout attends to encoded entity spans, whereas container readout attends to instruction tokens. The resulting representations have dimensions 6×768 and 5×768, accompanied by separate validity predictions. To form visual queries, their projections are combined with register addresses, identity representations, task relevance, and instruction/scene context in a 256-dimensional space. Association uses the current-frame initialization and excludes mutable physical content from the temporal estimators.

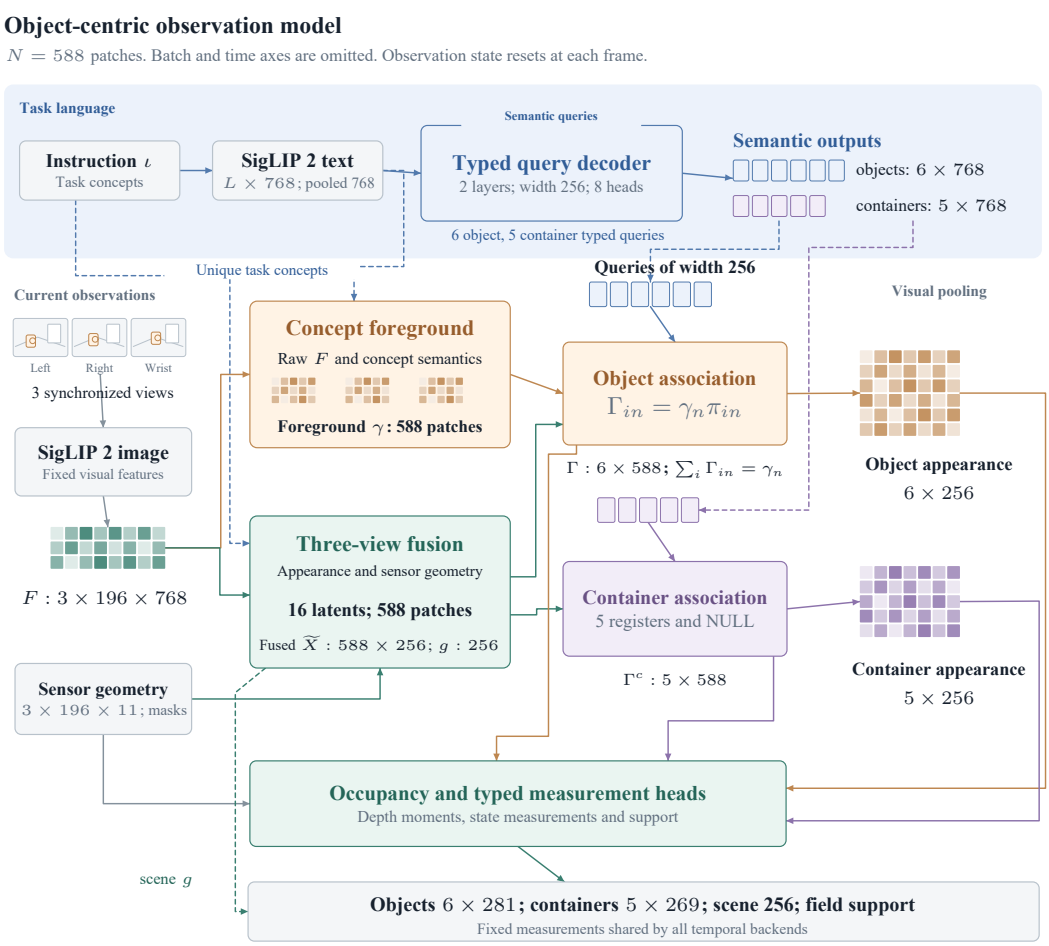


Figure A1: Observation-model architecture. Typed language proposals condition visual association. Raw SigLIP 2 patches enter parallel foreground and multiview-fusion branches. Object association combines fused patch features with foreground probabilities, whereas container association includes a NULL alternative. Register-specific appearance, depth occupancy, and measurement heads produce current physical observations and their support. The scene token is passed directly to the output. Dashed connections denote conditioning or a direct output connection. Dimensions omit batch and time axes. Images and feature maps are schematic. Figure A2 details the operators.

The instruction model is trained with six object and four container queries. Its adaptation to five containers is described in Appendix C.1. The original instruction conditions grounding, and subsequent role queries q_t are restricted to the role readout. Executed actions and their subsequent proprioceptive measurements in $\mathcal{U}_t$ enter the temporal interval encoder, not this frame-local observer. Reference segmentations, physical states, and event outcomes supply supervision rather than observation features. Calibrated depth is an explicit input.

A.2 Frame-Local Multiview Fusion

Each patch combines projected appearance, a camera-role embedding, and a learned projection of an 11-component geometry descriptor: world-frame ray origin (3), world-frame ray direction (3), calibrated world-frame depth point (3), log-transformed depth $\log(1+d)$ (1), where d is metric depth, and geometry confidence (1). Flattening the view and patch axes produces $X_t \in \mathbb{R}^{588\times256}$. Let $\bar{e}_\iota \in \mathbb{R}^{256}$ denote the projected instruction and $Z_0^{\text{fuse}} \in \mathbb{R}^{16\times256}$ the learned fusion queries. An eight-head cross-attention block is

$$\mathrm{CA}(Q,K) = \mathrm{LN}\big(Q + \mathrm{MHA}(Q,K,K)\big),$$

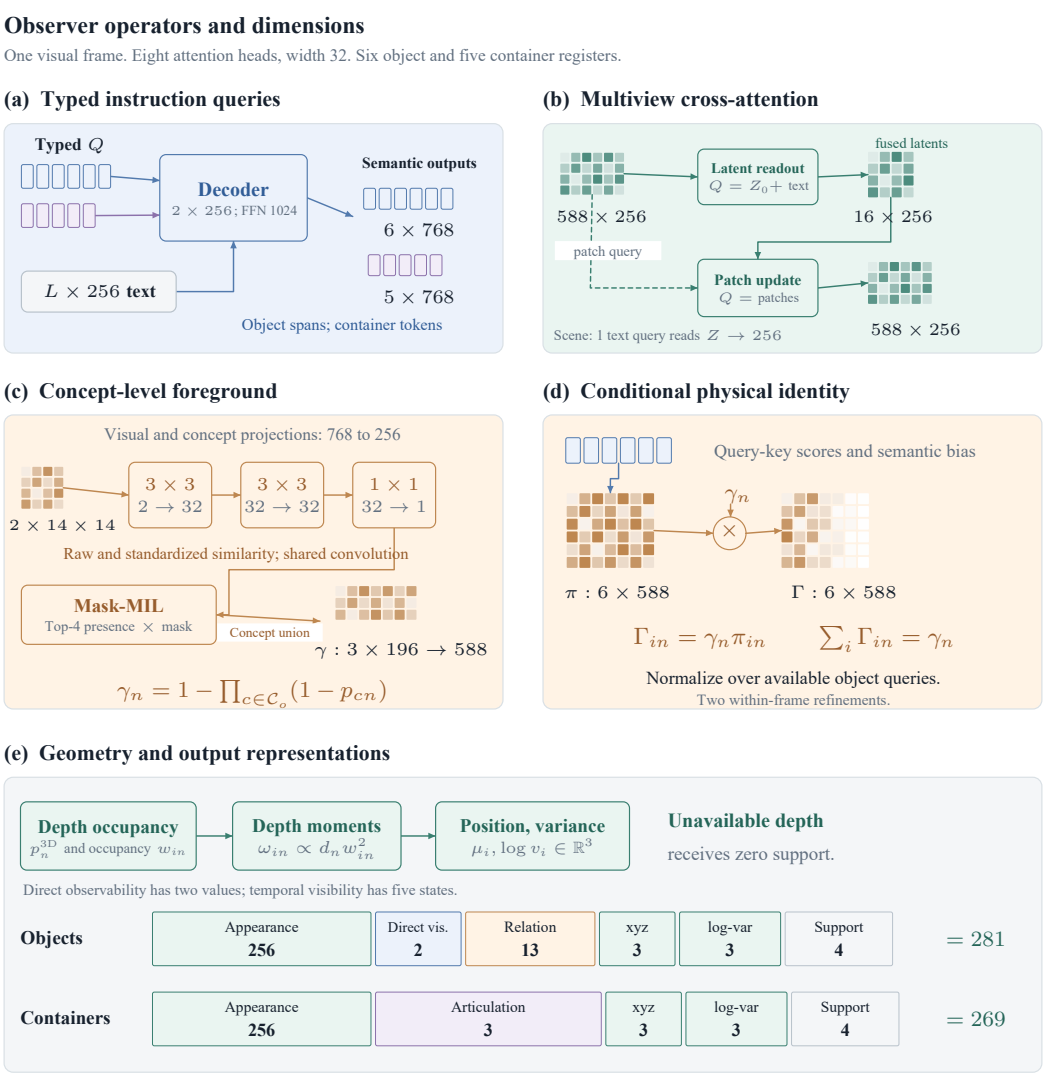


Figure A2: Observation operators and output dimensions. (a) Typed queries extract semantic representations from L instruction tokens and their entity spans, producing six object and five container slots. (b) Fusion latents read the camera patches and return fused information to them. (c) A shared convolutional network predicts concept masks and presence. (d) Normalized instance ownership π_{in} allocates foreground evidence γ_n across object slots. (e) Depth occupancy w_{in} and valid-depth confidence d_n define normalized weights ω_{in} for the calibrated points p_n^{3D}, yielding mean position μ_i and diagonal uncertainty. Missing depth gives zero geometry support. Direct vis. denotes the observer's two-way direct-observability measurement, distinct from the temporal visibility posterior. Log-var. denotes position log-variance. Box sizes are schematic.

where Q contains queries, K supplies keys and values, LN is layer normalization, and MHA is multihead attention with 32 dimensions per head. With learned instruction projection W_ι, three independently parameterized blocks compute

$$Z_t^{\text{fuse}} = \text{CA}_{\text{read}}(Z_0^{\text{fuse}} + W_\iota \bar{e}_\iota, X_t), \qquad Z_t^{\text{fuse}} \in \mathbb{R}^{16\times 256}, \tag{8}$$

$$\widetilde{X}_t = \text{CA}_{\text{write}}(X_t, Z_t^{\text{fuse}}), \qquad \widetilde{X}_t \in \mathbb{R}^{588\times 256}, \tag{9}$$

$$g_t = \text{CA}_{\text{scene}}(\bar{e}_\iota, Z_t^{\text{fuse}}), \qquad g_t \in \mathbb{R}^{256} \tag{10}$$

The instruction projection is broadcast over the 16 fusion queries, and the singleton query axis of the scene summary is omitted. The first block lets each latent read all available patches across

cameras. Its per-head attention matrix has shape 16×588. The second returns the fused information to the patch representations, preserving the patch-level features required for association. A single instruction-conditioned query reads the scene summary in the third block. Camera identity and geometry are embedded before concatenation, and missing patches are masked. The learned queries are reinitialized to Z_0^{fuse} at every frame. Cross-view correspondence is learned through these spatial interactions.

A.3 Foreground Evidence and Instance Assignment

Concept foreground. A parallel branch uses the raw SigLIP features to estimate foreground for each task-object concept. Learned visual and text projections map their inputs from 768 to 256 dimensions and compute normalized cosine similarity. For each concept and view, the raw and spatially standardized similarity maps form a two-channel 14×14 grid. A shared network applies two 3×3 convolutions with 32 hidden channels, concept and camera conditioning, and a 1×1 output convolution. Its residual combines with scaled raw similarity and a learned bias to produce mask logits.

A mask-based multiple-instance readout estimates concept presence from the four largest valid patch logits in each view. The local patch probability is multiplied by the corresponding presence probability and the valid concept/patch masks. Let p_{cn} denote the probability for concept c at patch n. Writing $\mathcal{C}_o$ for the set of unique active object concepts, their foreground union is

$$\gamma_n = 1 - \prod_{c \in \mathcal{C}_o} (1 - p_{cn}), \qquad \gamma \in [0,1]^{3\times 196}$$

The product combines the per-concept background probabilities, and its complement yields a soft foreground union. The same convolutional weights process each concept–view grid. Foreground is computed before expanding concepts into physical-instance hypotheses, so unused instance slots do not increase the measured foreground mass.

Physical-instance association. Object queries and fused patch features produce eight-head association scores of shape $8 \times 6 \times 588$. Suppressing the head and time indices, let ℓ_{in} be the raw score, r_{in} its learned identity correction, and $\mathcal{R}_o$ the available object-query set. For each head, define

$$c_n = \log \sum_{j \in \mathcal{R}_o} e^{\ell_{jn} + r_{jn}} - \log \sum_{j \in \mathcal{R}_o} e^{\ell_{jn}}, \tag{11}$$

$$\ell'_{in} = \ell_{in} + r_{in} - c_n, \qquad \pi_{in} = \frac{e^{\ell'_{in}}}{\sum_{j \in \mathcal{R}_o} e^{\ell'_{jn}}},$$

$$\Gamma_{in} = \gamma_n \pi_{in}, \qquad \Gamma_{\varnothing n} = 1 - \gamma_n \tag{12}$$

The shared offset c_n preserves the pre-correction log-partition. It cancels in the object-ownership softmax. Relative ownership changes through r_{in}, while the separate foreground path fixes the total mass. This offset is not an additional evidence term. In the temporal experiments, $\mathcal{R}_o$ contains all six declared object registers, with no episode-specific candidate restriction. Ownership normalization gives $\sum_{i \in \mathcal{R}_o} \Gamma_{in} = \gamma_n$. The background term $\Gamma_{\varnothing n}$ accounts for mass outside predicted foreground. Two within-frame refinements update ownership before each register pools a 256-dimensional appearance token.

Containers use a separate competitive attention module with five queries and an explicit NULL/background alternative, independently of the object foreground probabilities. Averaging over heads gives object and container assignment maps of shapes 6×588 and 5×588. Their pooled appearance representations have shapes 6×256 and 5×256.

A.4 Geometry and Typed Measurement Support

A separate occupancy branch adds a learned geometry residual to the logit of the detached, head-averaged assignment probability, then applies a sigmoid. It therefore uses normalized assignment, not the raw attention score. Detachment prevents geometry gradients from changing assignment through

this path. Let $p_n^{3D} \in \mathbb{R}^3$ be the calibrated depth point, w_{in} its predicted occupancy for register i, and $d_n \geq 0$ its valid-depth confidence. For positive total support,

$$\omega_{in} = \frac{d_n w_{in}^2}{\sum_{n'} d_{n'} w_{in'}^2}, \qquad \mu_i = \sum_n \omega_{in} p_n^{3D} \tag{13}$$

The normalized weights ω_{in} yield the weighted mean position μ_i. Squaring occupancy emphasizes patches with high assignment confidence. Weighted point dispersion provides diagonal uncertainty with numerical floors and ceilings. Geometry without depth support is marked unavailable. Additional heads predict direct observability, relation–anchor probabilities, and articulation, each with field-specific support.

An object token concatenates appearance (256 dimensions), direct observability (2), relation–anchor probabilities (13), position (3), position log-variance (3), and support indicators (4), for a total of 281 dimensions. The relation categories are UNKNOWN, NONE, HELD, and IN/ON for each of five container anchors. All five anchors are available in the main temporal experiments. The four indicators encode object confidence, relation support, geometry support, and sensor/register availability. All declared registers remain enabled in the temporal experiments. Direct observability supplies a current measurement distinct from the five-state visibility posterior of the temporal decoder.

A container token concatenates appearance (256), articulation probabilities (3), position (3), position log-variance (3), and support indicators (4), giving 269 dimensions. For the open-versus-closed probability $p_{\text{open}} \in [0, 1]$ and articulation support $\epsilon^A \in [0, 1]$, the UNKNOWN/CLOSED/OPEN measurement is $[1 - \epsilon^A,\ \epsilon^A(1 - p_{\text{open}}),\ \epsilon^A p_{\text{open}}]$. Its indicators encode container confidence, articulation support, geometry support, and availability. Field support is also supplied separately to the temporal model. The 256-dimensional scene token bypasses entity pooling.

For batch size B and sequence length T, the observation tensors have shapes $[B, T, 6, 281]$, $[B, T, 5, 269]$, and $[B, T, 256]$. Pixel-assignment maps are omitted from the temporal inputs.

A.5 From Visual Grounding to Task-Role Queries

Stage I learns both image-to-slot association and task-role binding. The Stage-I role binder is trained with role/subset supervision and refined after the instance-association update (Appendix C.1). The binder predicts membership, cardinality, and empty/unknown probabilities from learned representations.

For the temporal comparison, the learned observer is used as a frame-local measurement model and reinitialized at every visual frame. The public role-subset posterior is produced by a temporal readout rather than the observation-stage role head. In CERB, this readout uses the SlotSSM-based role branch R. The shared interface contains learned measurements and support, constructed without preselecting task candidates. Subsequent role queries act on detached temporal features, without feedback to image association or physical recurrence. The comparison holds the learned frontend fixed while evaluating the complete temporal belief backends.

B Complete Temporal Belief Architecture

This appendix expands the computation in Section 4, from the elapsed action–proprioception interval to the field distributions exposed to a query. E, P, and R are neural estimation branches with separate parameters and histories. Their field predictions form the public belief through the updates defined below.

B.1 Causal alignment and action encoding

Equation 1 ends at $(a_{n_t-1}, p_{n_t}^{\text{prop}})$. The source-clock subscript $k + 1$ denotes feedback following action a_k, not a future visual frame. For image indices $0, 5, 10$, the second visual step consumes $(a_0, p_1^{\text{prop}}), \ldots, (a_4, p_5^{\text{prop}})$, and the third consumes $(a_5, p_6^{\text{prop}}), \ldots, (a_9, p_{10}^{\text{prop}})$. The first visual step has no completed interval. Variable-length final intervals use all available tuples in order.

Interval summary. Actions have shape $[B, T, W, 12]$, proprioception $[B, T, W, 16]$, and valid-sample masks $[B, T, W]$, where B is batch size, T the number of visual steps, and W the padded

number of action–proprioception samples per interval. Each valid tuple is augmented with its normalized ordinal and inverse interval length. LayerNorm and a $30 \to 128$ GRUCell (Cho et al., 2014) produce the interval context. Its hidden state resets at each visual boundary, advances only on valid samples, and is normalized at output, and empty intervals are explicitly zeroed. Each branch has its own interval encoder.

Continuous response in P. The physical branch also maintains a width-64 GRU across all available action–proprioception samples. Its 44-component input concatenates the 12-dimensional action, scaled 16-dimensional proprioception, and scaled causal proprioceptive difference. Quaternion differences account for equivalent signs. Differences are zero at the first valid sample. An MLP with LayerNorm and SiLU precedes the GRU, and output LayerNorm follows it. The state selected at a visual boundary is the state after the last completed sample. A zero-initialized $64 \to 256$ projection adds the response to the observer's scene context before both temporal proposal and effect decoding. An empty current interval contributes zero through this bridge.

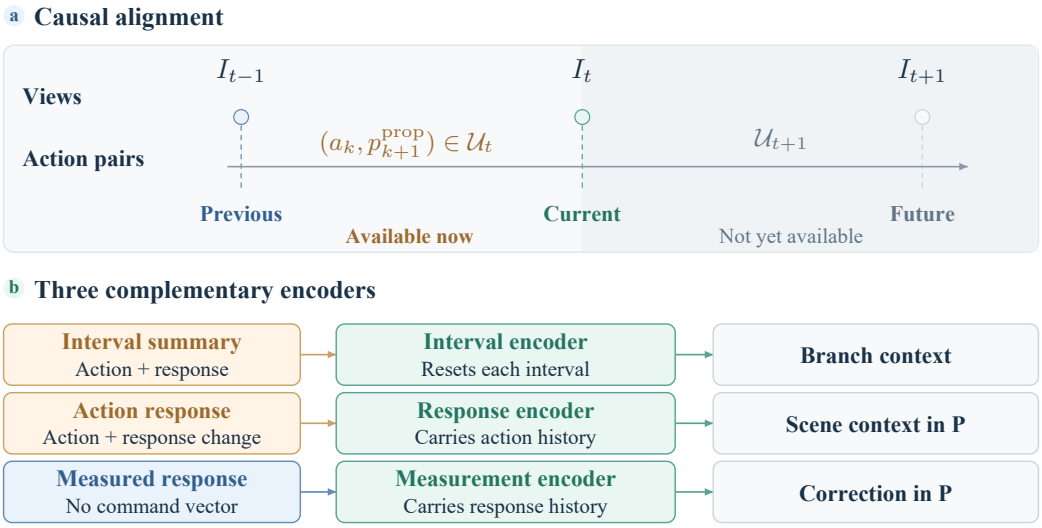


Figure B1: **Completed feedback, not future observations.** (a) At visual step t, $\mathcal{U}_t$ contains the completed pairs $(a_k, p^{\text{prop}}_{k+1})$ between the previous and current views. Each pair associates an executed action with its subsequent proprioceptive measurement. The next interval $\mathcal{U}_{t+1}$ is unavailable. (b) The interval encoder resets at each visual boundary, whereas P's action-response and measurement encoders carry history across intervals. The measurement path reads proprioceptive feedback without command vectors. The first visual step has no completed interval.

B.2 Recurrent proposals and typed change memory

E and P use width-304 register states with separate learned parameters. Type-specific projections align object/container tokens of width 281/269, and the projected tokens combine with scene context, register address, and type embeddings to form

$$x_t^i = \mathrm{LN}\big(P_{\mathrm{type}(i)}\,\mathrm{LN}(o_t^i) + P_g\,\mathrm{LN}(g_t) + e^i_{\mathrm{register}} + e_{\mathrm{type}(i)}\big) \tag{14}$$

Here o_t^i is the observer token of register i, and $\mathrm{type}(i)$ distinguishes object from container. The matrices $P_{\mathrm{type}(i)}$ and P_g are learned projections, and e^i_{register} and $e_{\mathrm{type}(i)}$ are learned address and type embeddings. The context g_t includes the response bridge in P. A two-layer GRU evolves each register independently using weights shared across registers. Local attention reads up to 16 encoded tokens of the same register, including the present token, with projected width 32 and eight heads. Early missing entries are masked.

Five field types—validity, visibility, relation, geometry, articulation—each carry three width-304 change-memory cells. Let $c_t \in \mathbb{R}^{128}$ be the interval context and $\bar{\epsilon}_t^{i,f} = \mathbf{1}[\epsilon_t^{i,f} > 0]$ indicate direct support. A shared rank-64 feature and a field-specific projection form the proposed memory content:

$$\varphi_t^i = \mathrm{GELU}\big(W_v\,\mathrm{LN}([x_t^i; s_t^i; c_t]) + b_v\big)\,, \qquad \Delta_t^{i,f} = W_f^v \varphi_t^i + b_f^v \tag{15}$$

Brackets $[\cdot;\cdot]$ denote feature concatenation. The learned weights W_v, W_f^v and biases b_v, b_f^v map the shared feature φ_t^i to proposed memory content $\Delta_t^{i,f}$. A separate rank-64 projection produces visual and action-conditioned write logits $\ell_t^{i,f,\mathrm{vis}}$ and $\ell_t^{i,f,\mathrm{act}}$. With ν_t indicating an available completed interval, the write strength is

$$\alpha_t^{i,f} = \sigma(\ell_t^{i,f,\mathrm{vis}})\bar{\epsilon}_t^{i,f} + \sigma(\ell_t^{i,f,\mathrm{act}})(1-\bar{\epsilon}_t^{i,f})\nu_t \mathbf{1}[f \in \{J,G,A\}] \tag{16}$$

Here σ is the logistic sigmoid and $\alpha_t^{i,f} \in [0,1]$ the write strength. The two terms are mutually exclusive. Action-conditioned memory writes apply to relation, geometry, and articulation. Each field maintains three memory cells, updated as

$$h_{t,k}^{i,f} = (1-\alpha_t^{i,f}\rho_k)h_{t-1,k}^{i,f} + \alpha_t^{i,f}\rho_k \Delta_t^{i,f}, \tag{17}$$

where $k \in \{1,2,3\}$ indexes the initially zero memory cells $h_{t,k}^{i,f}$. Their rates $\rho_k \in (0,1)$ are learned and shared across fields within each branch, starting from distinct initializations for fast, intermediate, and slow replacement. Each cell's effective update rate is $\alpha_t^{i,f}\rho_k$, combining its learned rate with the evidence-dependent write strength. The readout is $m_t^{i,f} = \sum_{k=1}^3 \lambda_{t,k}^{i,f} h_{t,k}^{i,f}$, where $\lambda_{t,k}^{i,f} \geq 0$ and $\sum_{k=1}^3 \lambda_{t,k}^{i,f} = 1$. The convex mixture selects among memories with different replacement rates. A softmax selector computes these weights from φ_t^i, a learned field embedding, and normalized log support age, and age resets on direct observation. The field router concatenates the slow, current, local, and memory vectors, interval context, and four indicators: direct support, normalized support age, reappearance after an unsupported interval, and completed-interval availability. A $1348 \to 64 \to 3$ network maps these features to the signed gates in Equation 5. Both residual maps ϕ_o and ϕ_m apply LayerNorm, a $304 \to 64$ linear layer, GELU, and a $64 \to 304$ linear layer. Their weights are shared across fields, while E and P have separate parameter sets.

Observability branch. E retains a causal outcome-correction module in addition to this proposal. Current/previous tokens and their differences are combined with private latents, interval context, and interval proprioceptive statistics (mean, endpoint, endpoint-minus-start). Width-128 per-register recurrence produces residual relation/articulation logits for its internal estimates and effect heads. These estimates support the geometry consistency rule below. E's final visibility readout is defined separately, but it does not use P's continuous-response bridge or measurement-memory updater.

Role branch. R uses a width-256 SlotSSM backend. Input projection and register/type embeddings feed observation cross-attention, per-slot Mamba2 recurrence (Dao & Gu, 2024), then cross-slot self-attention and a feed-forward block. The implementation uses one block, four attention heads, SSM state width 64, convolution width 4, expansion 2, and head width 64. The public role posterior is taken from this branch's complete typed decoder. R reuses the trained SlotSSM-based baseline for role inference.

The encoder input dimensions follow their feature concatenations: $12 + 16 + 2 = 30$ for interval summaries, $12 + 16 + 16 = 44$ for continuous action response, and $16 + 16 = 32$ for proprioceptive response. The field router concatenates four width-304 vectors, a width-128 interval context, and four indicators, yielding 1,348 inputs.

B.3 Measurement innovation and accepted-posterior recursion

P's measurement path is distinct from its command-conditioned proposal. A continuous proprioceptive GRU reads scaled proprioception and causal differences, giving 32 input components and 64 hidden components. It receives no command vector. For each register, the current token, previous token, token difference, response state, and scalar field support are projected to width 128. Four-head attention mixes registers within the current frame, and a per-register GRU then carries the resulting evidence across time. Neither attention nor recurrence reads future observations.

Suppressing the register index, let $m_t^{\text{meas}} \in \mathbb{R}^{128}$ denote the measurement-memory state after within-frame attention and the recurrent update. For field f, the correction head receives this state and the detached context $\xi_t^f = [b_{t-1}^f, p_t^f, p_t^{f,\text{act}}, \epsilon_t^f, \nu_t]$: previous belief, temporal candidate, action-conditioned candidate, direct field support, and completed-interval availability. Its output is

$$\delta_t^f = B_\delta \tanh\Big(g_f(m_t^{\text{meas}}, \operatorname{sg} \xi_t^f)/B_\delta\Big) \tag{18}$$

Here $B_\delta > 0$ is a fixed correction scale, $\tanh$ acts on each state component, and sg is the identity in the forward pass but stops gradients through its argument. For a field with $C_f = |\mathcal{S}_f|$ legal states, the correction MLP g_f has dimensions $(128 + 3C_f + 2) \to 128 \to C_f$. The elementwise mapping bounds each correction to $(-B_\delta, B_\delta)$ and is approximately linear near zero. The correction scale and memory-rate initializations are specified in the accompanying configuration files. Equation 7 converts the adjusted scores to probabilities. The correction head's final layer is initialized to zero. The reliability MLP has dimensions $(304 + 128 + 3C_f + 10) \to 128 \to 3$ and reads the private latent, interval context, the three candidate distributions, and ten scalars: observation support, interval availability, predicted physical-change probability, three entropies, three pairwise total-variation distances, and log evidence age. Its features are detached, while candidate distributions themselves remain differentiable through the public proper score.

Relation and articulation start from normalized UNKNOWN-dominated priors with nonzero probability on the other legal states. The implemented update uses the corrected current candidate *inside* the probability pool. With no direct support and no completed interval, it copies the previous accepted probability tensor. Within the probability pool, an unavailable action candidate is replaced by the previous belief. The same soft pool is used at training and inference for relation and articulation. These public fields are emitted directly. Other fields retain their own update rules below.

B.4 Typed output schema and complete field assembly

The system exposes six object registers and five container registers. Object visibility takes VISIBLE, PARTIAL, OCCLUDED, OUT-OF-VIEW, or UNAVAILABLE. The primary relation is

$$J_t^i \in \{\text{UNKNOWN}, \text{NONE}, \text{HELD}\} \cup \{\text{IN}(j), \text{ON}(j) : j = 1, \dots, 5\} \tag{19}$$

Separate diagonal Gaussians describe current and last-supported positions. Containers predict validity and articulation in UNKNOWN/CLOSED/OPEN. Let $b_t^{i,f}$ denote a physical field distribution and $p_t(Z^r \mid q_t)$ the posterior over object subsets Z^r for role-query index r. Then

$$\mathcal{B}_t = \Big((b_t^{i,f})_{i \in \mathcal{R}_o,\, f \in \mathcal{F}_o},\ (b_t^{j,f})_{j \in \mathcal{R}_c,\, f \in \mathcal{F}_c},\ (p_t(Z^r \mid q_t))_r\Big) \tag{20}$$

where $\mathcal{R}_o$ and $\mathcal{R}_c$ index the six object and five container registers, $\mathcal{F}_o = \{\text{valid}, V, J, G^{\text{cur}}, G^{\text{rel}}\}$ and $\mathcal{F}_c = \{\text{valid}, A\}$. Here G^{cur} and G^{rel} denote current and last-supported geometry, respectively. This factorization makes field uncertainty accessible without asserting a fully joint Bayesian scene posterior. In particular, per-object relation decoding does not enforce cross-object exclusivity such as at most one directly held object.

Typed heads. For hidden width d, let z_i and z_j denote the current object and anchor representations, with time and field indices suppressed. Three unary relation logits encode UNKNOWN/NONE/HELD, and an MLP $3d \to d \to 2$ applied to $[z_i, z_j, z_i - z_j]$ supplies IN/ON scores for each anchor. Their concatenation yields the 13-way relation distribution. All five anchors are available for prediction. Validity is Bernoulli, while visibility and articulation are categorical. Each geometry head predicts three means and log variances. Effect heads concatenate field latent, scene context, and interval context, then estimate the relation/articulation state at the interval endpoint and a change-support score. E/P use $d = 304$ and R uses $d = 256$. There is no action-conditioned geometry head.

Visibility and validity. Writing v_t^{E} for E's private visibility distribution and $\epsilon_t^V \in [0, 1]$ for its observation support, the readout uses $(1 - \epsilon_t^V) b_{t-1}^V + \epsilon_t^V v_t^{\mathsf{E}}$ when support is positive, with an explicit tensor copy when it is zero. Object and container validity use E's locally updated validity estimates, copied into the public field only with their respective entity observation support. The underlying E update heads remain trainable during E learning. Initial validity is taken from the designated local output.

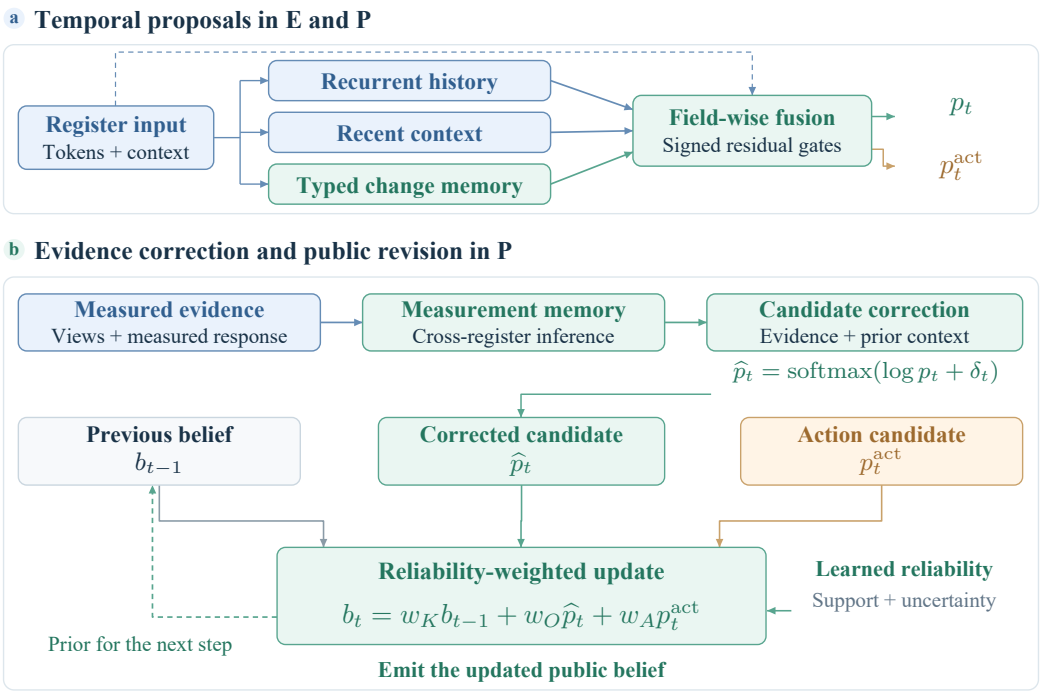


Figure B2: **Temporal inference and physical revision.** (a) E and P use the same proposal structure with separate parameters. Current input, recurrent history, recent context, and typed change memory form temporal and action-conditioned candidates. (b) P corrects the temporal candidate using measured evidence and detached prior/candidate context. Learned reliability weights pool the corrected candidate with the previous public belief and the action-conditioned candidate for relation and articulation. The resulting posterior is emitted and recurs as the next prior. An unavailable action candidate is replaced by the prior. Without either direct field support or a completed interval, the update retains the prior exactly. Register and field indices, and the time indices of the weights, are omitted for readability.

Geometry. Both object-geometry fields copy P's locally updated geometry estimates when direct geometry support is positive or when direct relation support / supported E–P relation agreement authorizes a spatial refresh. In symbols, at $t > 0$,

$$U_t^G = \mathbf{1}[\epsilon_t^G > 0] \vee \mathbf{1}[\epsilon_t^J > 0] \vee \left(\upsilon_t^J \wedge [\arg\max b_t^{J,\mathsf{P}} = \arg\max b_t^{J,\mathsf{E}}]\right)$$

Here U_t^G is the binary permission to refresh geometry, $b_t^{J,\mathsf{P}}$ and $b_t^{J,\mathsf{E}}$ are the two branches' relation distributions, and υ_t^J is P's thresholded, interval-masked transition support. The operators $\vee$ and $\wedge$ denote logical OR and AND. Otherwise the previous public geometry is retained. This spatial rule does not gate P's accepted relation or articulation posterior. Copying geometry after relation agreement also does not synthesize a new action-based spatial prediction.

Role subsets. Queries project from width 768 to R's width 256 and read detached physical-carrier features. Let ζ_i^r score object membership and κ_k^r cardinality. Write $\mathcal{C}^r$ for the candidate registers of role r, and $\mathcal{S}_k^r = \{S \subseteq \mathcal{C}^r : |S| = k\}$ for its admissible k-element subsets. For nonempty $S \in \mathcal{S}_k^r$,

$$\psi_r(S) = \kappa_k^r + \sum_{i \in S} \zeta_i^r - \log \sum_{S' \in \mathcal{S}_k^r} \exp\left(\sum_{i \in S'} \zeta_i^r\right) \tag{21}$$

Here $\psi_r(S)$ is the unnormalized log score of subset S. The log-sum-exp term normalizes membership scores among subsets of the same size, so κ_k^r determines their total cardinality score rather than rewarding sizes with more possible subsets. The empty and UNKNOWN potentials average their membership and cardinality logits and are $(\zeta_\emptyset^r + \kappa_0^r)/2$ and $(\zeta_?^r + \kappa_7^r)/2$, respectively, where $\zeta_\emptyset^r$ and $\zeta_?^r$ are dedicated empty-set and UNKNOWN logits, κ_0^r scores zero cardinality, and κ_7^r scores unknown cardinality. Inadmissible subsets have potential $-\infty$. A softmax over all 2^6 subsets plus UNKNOWN defines the posterior. Summing its probabilities over subsets containing an object gives that object's membership marginal, summing over subsets of each size gives cardinality probabilities, and its maximum-a-posteriori (MAP) outcome gives the predicted subset or UNKNOWN. The main condition admits every object candidate. All role outputs come from R together, preserving a coherent subset distribution. Later query changes affect this readout but neither physical recurrence nor the observer conditioned on the original instruction.

B.5 Common-observer transfer

The earlier four-container observation model is adapted to five containers by copying 587 of 600 name-and-shape-compatible tensors and reinitializing 13 capacity-dependent arrays, including addresses, proposal queries, and relation/binding projections. No additional alignment optimization follows this capacity change. The resulting frame-local measurements are shared by every temporal backend, and their common provenance is separated from Stage-II source holdout in Appendix H.

C Multistage Learning

Observation learning supplies grounded measurements, and temporal learning integrates them into belief. This section specifies the two learning stages, their optimization settings, and the training objectives.

C.1 Learning the Shared Observation Interface

Stage I learns entity proposals, foreground evidence, instance association, typed measurements, and task-role binding. Figure C1 summarizes the learning sequence, and Table 5 gives the data partitions.

Training schedule. Multiview grounding uses three epochs for fusion and entity evidence, followed by one for occupancy and geometric uncertainty. The language proposer is selected at epoch 54 of a 120-epoch run, and concept foreground is trained for two epochs. Structured measurement learning uses successive phases of 700, 210, and 140 updates for measurements/roles, validity/subsets, and relation/articulation. Association, role-readout, and joint physical refinement use 40, 45, and 80 updates, respectively. SigLIP 2 remains frozen. Phase-specific learning rates and batch sizes are provided in `configs/reproduction_schedules.json`.

Language proposals and foreground. The instruction model predicts matched object/container semantic sets and proposal validity. Its loss combines matched cosine error $\mathcal{L}_{\text{cosine}}$, balanced validity cross-entropy $\mathcal{L}_{\text{validity}}$, and a capacity-normalized count penalty $\mathcal{L}_{\text{count}}$,

$$\mathcal{L}_{\text{proposal}} = \mathcal{L}_{\text{cosine}} + \mathcal{L}_{\text{validity}} + 0.25\mathcal{L}_{\text{count}} \tag{22}$$

Concept foreground is trained independently of physical-instance identity using union-mask, per-concept mask, and balanced presence objectives. This ordering lets the later association update redistribute a fixed foreground estimate rather than relearn whether task-relevant evidence is present.

Structured measurements, association, and roles. Measurement supervision covers validity, direct observability, relation–anchor state, articulation, supported position, and uncertainty. After these heads are trained, a 256×256 identity-redistribution matrix is optimized with family/register-balanced conditional instance NLL while the foreground predictor and remaining observer are fixed. Because the association update changes pooled instance features, the role readout is subsequently refined with detached physical features, followed by joint relation/articulation learning.

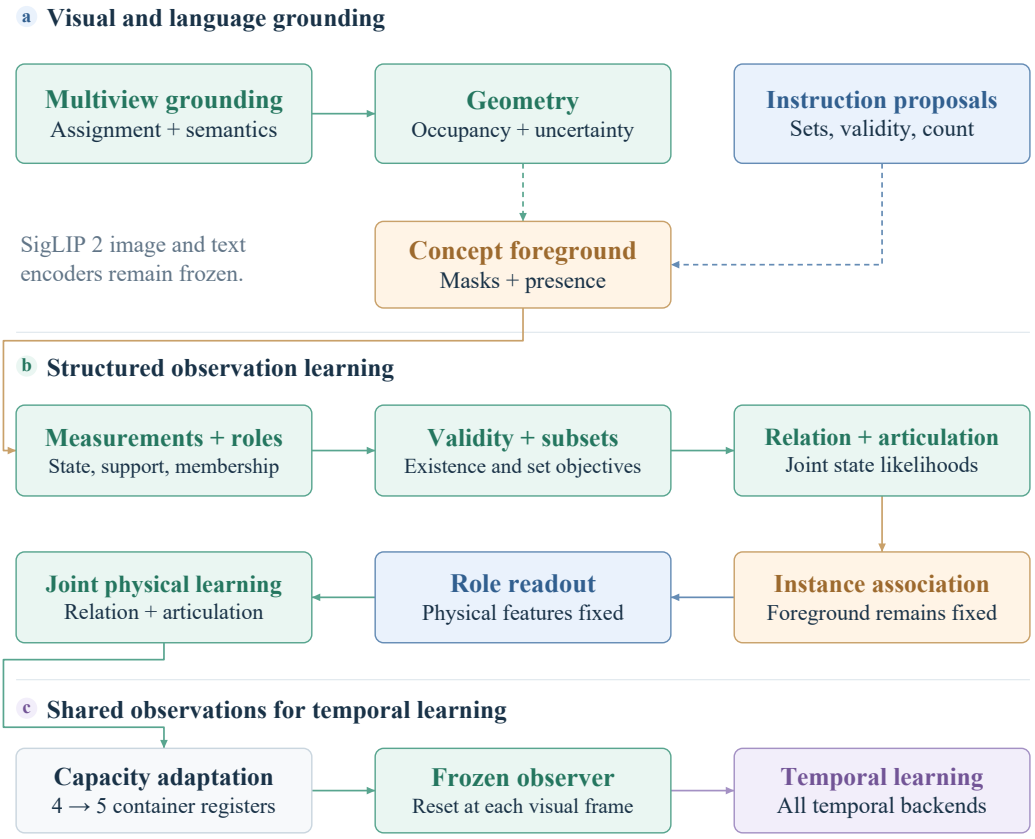


Figure C1: **Training the shared observation model.** Multiview grounding, geometry, instruction proposals, and concept foreground are learned before structured measurement and role objectives. Instance association is refined with foreground fixed, followed by role-readout refinement and joint physical learning. The resulting frame-local model is adapted to five container registers and shared by every temporal system.

Shared frontend for temporal comparison. The observer is adapted to five container registers using the parameter transfer in Appendix B.5, then reset at every frame. Its measurements remain fixed across temporal backends and training repetitions. Stage II uses the temporal-learning partition in Table 5 and does not update the observer.

Observer-stage evidence. On the structured-observer development partition, the identity-redistribution update raises conditional instance accuracy from 69.47% to 82.93% and same-class accuracy from 35.30% to 49.59%. After the association change, refining the detached role readout yields 85.36% exactness. Instance accuracy measures correct register assignment on reference-labeled foreground patches, averaged over source families; same-class accuracy restricts this evaluation to targets whose semantic category is shared by another registered object. Observer-stage role exactness requires correct membership and cardinality on valid nonempty queries whose target members have all been observed. These measurements verify that the shared interface learns association and role binding. Primary35 then holds that interface fixed to compare complete temporal belief backends under common visual measurements.

C.2 Temporal initialization and specialization

E and P share a recurrent initialization followed by typed-proposal learning. The added residual paths start at zero, preserving the recurrent predictor at transfer. The branches then specialize: E learns observation-conditioned corrections, while P learns physical revision from executed actions,

proprioceptive feedback, and measurements. R reuses the independently trained full-output SlotSSM backend without further role-specific tuning. Table 4 gives the temporal training schedule.

P first continues its specialized backbone and fits reliability heads on a parallel initialization path. Joint physical learning restores that backbone, loads the fitted heads, and successively introduces posterior coupling, continuous action response, measurement correction, and command–outcome balancing. Phase-specific settings are provided in `configs/reproduction_schedules.json`.

AdamW uses batch size 16, weight decay 10^{-4}, gradient norm clipping at 1, FP32 master weights, and bfloat16 autocast. Recurrent and typed-proposal initialization use learning rate 3×10^{-4}, and E/P specialization and correction use 10^{-4}, reduced to 5×10^{-5} for final outcome balancing. Optimizer moments reset at stage boundaries. Full histories are length-bucketed, with batches sampled cyclically from action-outcome interventions, natural events, action-outcome interventions, and stable intervals. Reference validity masks select supervised entries without closing the forward candidate domain. All active student parameters are trainable within the relevant branch, and final physical learning updates P while E, R, and the observer remain fixed.

Adapted temporal baselines. All four temporal baselines use the same inputs and full-output objectives. The Gated DeltaNet-2 and Mamba-3 baselines use two width-512 blocks, each consisting of within-frame register attention, per-register causal recurrence, and a residual FFN. Mamba-3 uses the official SISO operator with state width 128, expansion 2, and head width 64, and Gated DeltaNet-2 uses four width-128 heads, value expansion 1, and width-4 short convolution. Both use eight-head register attention. Their FFN widths are 1,088 and 1,024 respectively. The latter two training phases use preservation weights .1/.5, and the final phase reduces learning rate by three. Development-selected initial learning rates are 3×10^{-4} for Mamba-3 and 1.5×10^{-4} for Gated DeltaNet-2, with the development runs specified below.

Table 4: Temporal training schedule. CERB's shared initialization consists of recurrent and typed-proposal learning, inherited by both physical branches. Cumulative updates follow each branch's training path.

Backend / branch	Training path	Cumulative updates
GRU	Full-output training	2,000
SlotSSM	Full-output training	2,000
Gated DeltaNet-2	$1{,}200 + 400 + 400$	2,000
Mamba-3	$1{,}200 + 400 + 400$	2,000
CERB: E	$1{,}600 + 200 + 480$	2,280
CERB: P	$1{,}600 + 200 + 1{,}520$	3,320
CERB: R	Reused SlotSSM role readout	2,000

The shared 1,600-update initialization comprises 1,200 recurrent and 400 typed-proposal updates. A separate 240-update fit initializes P's reliability heads. Gated DeltaNet-2 and Mamba-3 each use two 600-update development runs for learning-rate selection.

Evaluation processes complete causal histories. The observation and action frequencies describe the input streams, rather than measured inference throughput.

C.3 Temporal objective: what each term learns

The temporal objective separates current estimation, event response, hidden-state correctness, role binding, action outcomes, update reliability, and preservation of the initialization predictor. The base terms below train the full-output initialization and remain active during physical learning. For P, the relation/articulation entries are the actually emitted accepted posterior and its actual candidate pool. Additional proper-score and balanced-outcome terms are defined at the end of this section.

The coefficients below are fixed training weights: event response receives unit weight, while the remaining terms supervise complementary state, role, and update behavior. They weight complementary training objectives. For trainable branch parameters θ, each member uses the objective

$$\begin{aligned}\mathcal{L}(\theta) =& .25\mathcal{L}_{\text{frame}} + \mathcal{L}_{\text{event}} + .5\mathcal{L}_{\text{retain}} + .25\mathcal{L}_{\text{set}} \\ &+ .5\mathcal{L}_{\text{effect}} + .5\mathcal{L}_{\text{route}} + \lambda_{\text{pre}}\mathcal{L}_{\text{preserve}}\end{aligned} \tag{23}$$

The typed-proposal initialization uses $\lambda_{\text{pre}} = .1$ and E/P specialization uses .5. Subsequent physical correction uses no teacher-preservation term. Baseline continuation weights follow Appendix C. Physical losses train the appropriate branch, while deterministic field assembly has no separate optimizer.

C.3.1 Masks and episode-balanced reduction

Let j index a sampled episode, a a time–register entry for field f, and ℓ^f_{ja} its scalar loss. Mask m^f_{ja} intersects frame validity, reference-defined register eligibility, and field-label validity. Define $n^f_j = \sum_a m^f_{ja}$ and the supported episodes $\mathcal{J}^+_f = \{j : n^f_j > 0\}$. The reduction operator is

$$\mathsf{M}_f(\ell; m) = \begin{cases} \dfrac{1}{|\mathcal{J}^+_f|} \displaystyle\sum_{j \in \mathcal{J}^+_f} \frac{\sum_a m^f_{ja} \ell^f_{ja}}{n^f_j}, & |\mathcal{J}^+_f| > 0, \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

Here $|\cdot|$ denotes set size. First average valid entries within each episode, then average supported episodes. Longer episodes therefore do not automatically dominate. Most components average a fixed list of field terms: a batch-unsupported field contributes a differentiable zero, rather than changing that denominator. The role-configuration term is added only when supported. Training does not apply the source-family grouping used in evaluation.

C.3.2 Frame, event, and hidden-state estimation

The frame loss averages seven terms: object validity, visibility, joint relation–anchor, current geometry, last-supported geometry, container validity, and articulation. For categorical target Y, its predicted probability is $b(Y)$ and the loss is $-\log b(Y)$, while validity uses binary cross-entropy (BCE). For geometry, let $\mu \in \mathbb{R}^3$ and $v \in \mathbb{R}^3_{>0}$ be the predicted mean and diagonal variance, and $\mu^\star \in \mathbb{R}^3$ and $v^\star \in \mathbb{R}^3_{\geq 0}$ their target counterparts. The implemented Gaussian cross-entropy, omitting an additive constant, is

$$\ell_{\text{geo}}(\mu, v; \mu^\star, v^\star) = \frac{1}{6} \sum_{d=1}^{3} \left[\log v_d + \frac{(\mu_d - \mu^\star_d)^2 + v^\star_d}{v_d} \right] \tag{25}$$

The factor $1/6 = (1/2)(1/3)$ combines the Gaussian negative-log-likelihood coefficient with averaging over three spatial coordinates. Coordinate d contributes squared location error relative to predicted uncertainty, together with a penalty for arbitrarily broad variance. Predicted log variance is clamped to $[-8, 6]$, while current-position target log variance to $[-12, 6]$. These clamps are fixed numerical bounds on uncertainty. Last-supported geometry uses a point target, $v^\star = 0$, and a separate head.

Training onsets require three identical valid source labels and three identical valid destination labels. Let e^{evt}_{jtf} indicate the union of windows from each onset through offset eight for episode j, step t, field f. Writing Y^f for the reference field state and suppressing episode, time, and register indices inside the reduction, the event loss is

$$\mathcal{L}_{\text{event}} = \frac{1}{3} \sum_{f \in \{V,J,A\}} \mathsf{M}_f\big(-\log b^f(Y^f);\, m^f e^{f,\text{evt}}\big) \tag{26}$$

The event-window indicator $e^{f,\text{evt}}$ is distinct from the evidence-availability indicator e^f_t used by the public update. Multiplying the validity mask by $e^{f,\text{evt}}$ emphasizes response around changes. Relation certification/supervision excludes UNKNOWN as a physical state, and articulation onsets likewise require known source and destination. Invalid labels remain masked. Unlike the reported event physical NLL, this training term includes visibility and averages valid entries rather than events.

The retention loss equally averages hidden visibility NLL, hidden relation NLL, and loss on valid last-supported geometry targets. It uses strict-hidden validity and, for relation, the relation-persistence mask. These terms learn *correctness* during missing evidence, but they do not replace the exact-copy rule governing unsupported public fields.

C.3.3 ROLE SUBSETS AND ACTION OUTCOMES

The role loss averages supported membership BCE and configuration cross-entropy when the latter has support. A fully resolved role specifies one subset. Known cardinality with partial membership yields a normalized target over compatible subsets, while no membership/cardinality knowledge yields UNKNOWN. Partial membership without known cardinality contributes its marginal supervision but is omitted from the joint target rather than falsely labeled UNKNOWN. The physical-feature input is detached, so these losses train the role readout without letting a role query update physical recurrence.

Effect heads predict the current interval's relation/articulation endpoint and whether it changed. The reference precondition is the most recent label-valid state before that endpoint, not necessarily the immediately preceding frame. Only action-outcome intervention episodes with supported outcome, field, and interval labels contribute. Thus an ineffective attempt is a negative change example, not a success copied from the command.

For one effect field, let $y_{\mathrm{ch}} \in \{0, 1\}$ be the reference change label, m_{eff} its validity mask, and ℓ_{BCE} the support-prediction BCE. Positive/negative-balanced support loss is

$$\mathcal{L}_{\mathrm{support}} = \tfrac{1}{2}\mathsf{M}_f(\ell_{\mathrm{BCE}}; m_{\mathrm{eff}} y_{\mathrm{ch}}) + \tfrac{1}{2}\mathsf{M}_f(\ell_{\mathrm{BCE}}; m_{\mathrm{eff}}(1 - y_{\mathrm{ch}})) \tag{27}$$

The two $1/2$ factors give effective and ineffective attempts equal weight at the loss level. Empty groups follow Equation 24. Each field's effect loss averages this term with postcondition NLL, while $\mathcal{L}_{\mathrm{effect}}$ then averages relation and articulation. These outcomes are targets only, never encoder inputs.

C.3.4 LEARNING UPDATE RELIABILITY FROM CANDIDATE RISKS

For field f, candidate index $m \in \mathcal{O} = \{K, O, A\}$ identifies the retained, observation-based, or action-based distribution $Q^f_{t,m}$ at step t. For P's relation/articulation pool, these are b^f_{t-1}, $\widehat{p}^f_t$, and $p^{f,\mathrm{act}}_t$ in Equation 6. The label Y^f_{t+k} is the reference state at offset k. Let $\mathcal{K} = \{0, 1, 2, 4, 8\}$ be training offsets, and $v^f_{t,k}$ indicate that labels are valid and unchanged throughout $[t, t+k]$. Suppressing episode/register indices, candidate risk and target reliability weights are

$$\mathcal{C}^f_t(m) = \frac{\sum_{k\in\mathcal{K}} .85^k v^f_{t,k}[-\log Q^f_{t,m}(Y^f_{t+k})]}{\sum_{k\in\mathcal{K}} .85^k v^f_{t,k}} \tag{28}$$

$$w^{f,\star}_t(m) = \mathrm{softmax}_{m\in\mathcal{O}}[-\,\mathrm{sg}(\mathcal{C}^f_t(m))/.25] \tag{29}$$

The denominator must be positive. The operator sg stops gradients, and .85 is the fixed temporal discount and .25 the fixed softmax temperature controlling target sharpness. In words, score each *present* candidate on a valid persistence segment, then prefer lower-risk candidates. For a fixed current candidate, unchanged valid targets make this weighted average equal to its current-label NLL, while the certified segment determines eligibility rather than rolling the candidate forward. Candidate distributions are evaluated at the current step.

For relation and articulation, w^f_t is the learned source-weight vector in Equation 6; for visibility, it denotes the auxiliary selector described below. The component notation $w^f_t(m) = w^f_{t,m}$ refers to source m. The vector $w^{f,\star}_t$ is the training target, favoring lower-risk candidates. Cross-entropy $-\sum_{m\in\mathcal{O}} w^{f,\star}_t(m) \log w^f_t(m)$ trains eligible, non-forced-hold updates only when the best–second-best risk gap exceeds 10^{-4}. The implementation constructs these risks separately for visibility, relation, and container articulation, reduces each by Equation 24, and averages the three terms to form $\mathcal{L}_{\mathrm{route}}$. Shared source logits can receive several field losses, while their candidate costs are not first pooled into one multi-field risk. Ordinary proper scores continue to train the candidate distributions.

The auxiliary visibility selector uses the retained distribution b^V_{t-1} and the observation-based proposal $p^{V,\mathrm{obs}}_t$, with the latter occupying both the O and A positions. When the duplicated observation candidate has the lowest risk, the best–second-best margin is zero and that update is excluded from this auxiliary term. A uniquely lower-risk retained candidate remains eligible. Proper-score supervision remains active. CERB's public visibility readout uses the E branch's observation-weighted posterior directly, as specified in Appendix B, rather than this selector's discrete decision.

C.3.5 PRESERVATION WITHOUT FREEZING THE STUDENT

Let θ_0 denote the frozen stage-initialization teacher and θ the trainable student. The field list $\mathcal{F}_{\mathrm{pre}}$ contains the seven physical terms above and role configurations. With field-specific protection mask m^f_{pre},

$$\mathcal{L}_{\mathrm{preserve}} = \frac{1}{|\mathcal{F}_{\mathrm{pre}}|} \sum_{f \in \mathcal{F}_{\mathrm{pre}}} \mathsf{M}_f \Big(D_{\mathrm{KL}} \big(\mathrm{sg}[b^f_{\theta_0}] \| b^f_\theta \big); m^f_{\mathrm{pre}} \Big) \tag{30}$$

Here $b^f_{\theta_0}$ and b^f_θ are the teacher and student field distributions on the same input history. The divergence D_{KL} is teacher-to-student Kullback–Leibler divergence, categorical, Bernoulli, or diagonal Gaussian as appropriate, and Gaussian KL averages coordinates. This discourages unnecessary output drift while allowing every retained student parameter to adapt.

Visibility/relation are protected outside their event windows or on supported hidden entries, while articulation outside its event windows. Object validity uses the union of protected visibility/relation masks, and container validity uses protected articulation. Current geometry additionally requires a valid target and protected visibility, while last-supported geometry uses supported hidden entries. Role KL covers valid frames and reference-valid roles. The coefficient depends on the optimization stage as specified above. Field assembly has neither teacher nor optimizer.

C.3.6 PROPER SCORES FOR CORRECTED AND ACCEPTED STATES

Let $\mathcal{L}_{\mathrm{base}}$ be Equation 23 with $\lambda_{\mathrm{pre}} = 0$, evaluated on P's emitted fields and actual candidate pool. Physical learning adds four auxiliary terms with fixed training weights

$$\mathcal{L}_{\mathsf{P}} = \mathcal{L}_{\mathrm{base}} + .2\mathcal{L}_{\mathrm{current}} + .5\mathcal{L}_{\mathrm{endpoint}} + .2\mathcal{L}_{\mathrm{evidence}} + .5\mathcal{L}_{\mathrm{balanced}} \tag{31}$$

The current term averages relation/articulation NLL for the temporal and action candidates, with action supervision requiring a completed interval. The endpoint term scores private, action, and accepted predictions against realized current states at certified intervention endpoints, including no-effect outcomes. The evidence term scores the measurement-corrected candidate wherever field support or a completed interval is available. The base proper scores already supervise the actual accepted relation/articulation output. Thus the public loss can improve the candidates themselves through the probability pool, while detached routing features prevent a separate indirect feature-gradient path. For relation and articulation, the candidate-risk auxiliary term trains the reliability weights of this same soft pool, without introducing a hard selection at inference.

Balancing commands and realized states. The extra relation term partitions eligible training frames into four groups: closing with/without any HELD object, and opening with/without any HELD object. The sign of the final valid gripper command in the completed interval defines opening/closing. Frames with zero command, no completed interval, or any unknown/invalid supervised object relation are omitted from this added term, while remaining eligible for the base objective. Within each nonempty group, NLL is averaged across the temporal, action, measurement-corrected, and accepted relation distributions, then across supervised objects/times within each episode and across supported episodes. Nonempty groups receive equal weight:

$$\mathcal{L}_{\mathrm{balanced}} = \frac{1}{|\mathcal{D}^+|} \sum_{g \in \mathcal{D}^+} \mathsf{M}_J \left(\frac{1}{4} \sum_{Q \in \{p, p^{\mathrm{act}}, \widehat{p}, b\}} - \log Q(J); m_g \right) \tag{32}$$

The group condition refers to whether any supervised object is held, but the score covers all supervised objects in that frame. Here $\mathcal{D}^+$ contains the nonempty groups, m_g selects valid entries of group g, and J inside $Q(J)$ is the reference primary-relation label. The factor $1/4$ averages the four candidate distributions indexed by Q, and M_J is the relation-field reduction in Equation 24. The added term is zero if no group has support. Group labels are used only to balance training losses. Inference uses the observed inputs and learned field distributions.

D BENCHMARK AND EVALUATION PROTOCOL

D.1 DATA PARTITIONS AND THEIR ROLE IN LEARNING

Table 5 distinguishes observation-stage families from temporal-stage episodes. Both test cohorts are held out from temporal optimization and evaluated under a shared frozen observer. Appendix H examines known observer-source overlap.

Table 5: Data partitions across the two learning stages. Counts use the unit in the first column. The shared observer is fixed before Stage II, and test populations do not supply temporal model selection.

Stage / population (unit)	Optimization	Development	Evaluation
Stage-I visual/language grounding (episodes)	320	40	—
Stage-I structured observer (families)	223	54	—
Stage-II temporal belief (episodes)	1,858	498	—
Primary35 IID (natural episodes)	—	—	397
Primary35 target assets (natural episodes)	—	—	267

The combined temporal-training and development pool contains 2,063 natural histories, 59 field-local articulation repairs, and 234 intervention branches from 117 matched success/no-effect pairs. Its development partition has 406 source families, including 46 intervention-based three-episode groups (24 articulation and 22 relation). Related episodes remain in the same partition. Of 670 prospectively assigned test episodes, six were excluded before model scoring: three had ambiguous equal-priority relation anchors, two failed the source labelability check, and one failed a rendered-view nonblank/dynamic check. Three exclusions occurred in each test partition.

Learned measurements and candidate-set inputs. The observer uses rendered RGB, calibrated replay-depth/camera geometry, instruction tokens/spans, and sensor-availability masks. Entity proposals, image associations, and measurement support are neural predictions. Object validity is the Bernoulli estimate of existence for the object assigned to a register. It is distinct from visibility and from label-validity masks, which select supervised or scored entries. Stage-I training also optimizes a role binder, and the final temporal models train their own role readouts (Appendix A.5).

Every declared object register, container anchor, role query, and role–object candidate is available to the main temporal models during training, validation, and testing. Observation features and support are constructed without episode-specific candidate restrictions. Frame availability, action-sequence padding, and reference-label validity still identify usable observations, executed-action samples, and supervised or scored entries. These indicators do not preselect object or role candidates. The supplementary proposal-component study in Appendix G explicitly uses supplied candidate sets, a separate input condition from the main comparison.

Supervision and measurement support. Training distinguishes exact role membership from category-level candidate eligibility. Label-validity indicators select supervised and scored entries, whereas observation support is predicted from current sensory evidence. Missing labels are excluded, and a closed container does not by itself certify that an object is invisible.

Relation-intervention pairs restore a certified simulator snapshot, replay the same tail actions, and disable the gripper in the no-effect branch, while articulation pairs use a high-resistance fixture intervention. Their common starting state and commands can lead to different post-fork RGB/depth histories and proprioception. This divergence occurs in every development intervention pair. Articulation repairs supervise only their supported field. Test streams supply natural physical labels, not the intervention-based action-outcome targets used for effect supervision.

D.2 WHAT CONSTITUTES A GENUINE CHANGE?

The benchmark evaluates whether belief responds to the right event, not just whether a frame is easy to classify. At visual step t, reference label Y_t^f describes field f and b_t^f its predicted distribution, and $\widehat{Y}_t^f$ is the maximum-a-posteriori (MAP) state (threshold .5 for binary validity). Register indices are omitted when unambiguous. The event detector uses labels only, never model predictions.

A dwell-certified transition requires three consecutive valid, identical source labels and three consecutive valid, identical destination labels. Its onset τ_e is the first destination step, not the third step at which certification becomes available offline. The last certified source is retained across invalid gaps. Initial acquisition without a certified preceding state is not automatically a transition. Known relation/articulation events exclude UNKNOWN as source or destination. Dwell certification uses subsequent reference labels only for scoring. The separate grasp-resolved event analysis in Appendix E.3 additionally requires unique adjudication throughout the event dependencies.

For event e in episode j of length T_j, let τ_e^+ be the next onset for the same event stream. Its horizon-h window contains the integer visual indices

$$\mathcal{W}_{e,h} = \{t : \tau_e \le t \le \min(\tau_e + h, \tau_e^+ - 1, T_j - 1)\} \tag{33}$$

Here h is an offset: $h = 8$ permits at most nine observations including onset. The next event censors the window, preventing a later physical change from being credited as recovery of the earlier one. If there is no next event, only the horizon and episode end limit it.

D.3 Complete correctness and actionable recovery

The recovery predicate is a conjunction for one task-addressed object, not simultaneous reconstruction of every object in the scene. Table 6 defines its physical checks. Applicability comes from the reference state. A missing label for an applicable field invalidates that candidate hit, while an inapplicable field imposes no test.

Table 6: Field predicates used in complete-object recovery. Geometry is judged in the declared robot coordinate frame. These recovery checks are distinct from raw five-state visibility classification and probability scoring.

Field	Required prediction at a candidate recovery step
Validity	Correct object-register validity
Visibility	Correct visibility group: VISIBLE/PARTIAL map to observed, OCCLUDED/OUT-OF-VIEW to hidden, and UNAVAILABLE to unavailable
Relation–anchor	Correct relation type and, for IN/ON, correct container anchor
Articulation	For a reference IN(c) relation with an articulation-eligible container c, correct known/unknown status and, when known, open/closed state
Current geometry	Euclidean position error at most .20 m if reference object is observed and not HELD
Last-supported geometry	Euclidean position error at most .20 m if reference object is hidden and not IN a container

The 0.20 m bound treats geometry as coarse task-level spatial evidence rather than a grasp-pose tolerance. Geometry is never credited alone: a recovery hit must simultaneously satisfy relation–anchor, visibility group, validity, applicable articulation, and (for ABR) role subset. The benchmark therefore evaluates state-transition recovery with spatial consistency, not precision localization.

For an object–role event, let Z_t be its reference object subset and $\widehat{Z}_t$ the predicted MAP subset. Let $\mathcal{F}_{e,t}$ contain the physical fields applicable at candidate step t, $\chi_f(b_t^f, Y_t^f) \in \{0, 1\}$ denote Table 6's check for field f, and $v_{e,t}$ indicate valid labels and role eligibility at step t. Joint correctness is

$$C_{e,t} = v_{e,t}\, \mathbf{1}[\widehat{Z}_t = Z_t] \prod_{f \in \mathcal{F}_{e,t}} \chi_f(b_t^f, Y_t^f) \tag{34}$$

A product of binary checks is one only when every applicable requirement holds together. Correct role membership alone, or correct fields at different times, cannot produce a hit. A task-addressed container event instead uses only container validity and articulation in $C_{e,t}$.

Which events enter ABR? Object events are certified known relation–anchor changes or applicable articulation changes. An object–role pair is eligible if the reference role subset is exact, nonempty, supported, contains that object, and every target member has been directly observed somewhere in

the available prefix. Eligibility is checked anywhere in the maximum offset-eight window, then held fixed for all reported horizons. A shorter window without an eligible correct step is a miss, not an omitted event. A second event population covers articulation changes of task-addressed primary containers.

Let $H_{e,h} = \mathbf{1}[\exists t \in \mathcal{W}_{e,h} : C_{e,t} = 1]$ be a recovery hit, as in Equation 3. For episode j, superscripts o, c distinguish object–role and container hit counts $H^o_{j,h}, H^c_{j,h}$ and support counts N^o_j, N^c_j. The episode score is

$$\mathrm{ABR}_j@h = \frac{H^o_{j,h} + H^c_{j,h}}{N^o_j + N^c_j} \tag{35}$$

Object ABR and Container ABR use $H^o_{j,h}/N^o_j$ and $H^c_{j,h}/N^c_j$, respectively, on episodes with support for that event type. For the combined ABR score, counts are pooled within the episode before averaging episodes/families, and the support counts do not shrink at shorter horizons. Zero-support episodes have no score.

A worked example. Suppose an object event begins at step 10 and the next event starts at 15. Its offset-eight window is steps 10–14, not 10–18. Correct relation at 11 and correct role at 12 do not suffice unless all applicable checks hold at one common step. A complete hit at 13 gives $H_{e,4} = H_{e,8} = 1$, but $H_{e,1} = H_{e,2} = 0$. If an episode has three eligible object–role events and one container event, with two and one hits respectively, its ABR is $3/4$, but it is not the unweighted mean of object and container rates.

Why complete-object recovery (Whole) is separate. Whole@8 uses the categorical object-core signature (raw visibility, relation–anchor, applicable articulation) to define events, then checks Table 6 without ABR's actionable-role eligibility. Geometry affects correctness but does not itself create a categorical onset. Whole@8 and ABR therefore differ in event population as well as role requirements, so neither is uniformly stricter.

D.4 Transition timing and event probability

Transition macro-F1. Predicted transitions are changes between consecutive label-valid predictions, including across invalid gaps. A target event matches the earliest unused prediction of its destination in its causal offset-eight window, truncated before the next event. Early predictions cannot match a later event, and each prediction/event is used at most once. This is deliberate: the benchmark estimates the current physical belief rather than forecasting a future transition, so a destination predicted before the certified onset is a reference-stable write at that time. A pre-onset tolerance would reward anticipation and change the target from state estimation to forecasting.

For episode j, field f, and destination class k, write $TP_{jfk}, FP_{jfk}, FN_{jfk}$ for matched transitions, unmatched predicted transitions, and missed target transitions. Let $\mathcal{K}^+_{j,f}$ contain classes with positive target support and $\mathcal{F}^+_j$ fields with such support. Then

$$F_{jfk} = \frac{2TP_{jfk}}{2TP_{jfk} + FP_{jfk} + FN_{jfk}}, \tag{36}$$

$$F_j = \frac{1}{|\mathcal{F}^+_j|} \sum_{f \in \mathcal{F}^+_j} \frac{1}{|\mathcal{K}^+_{j,f}|} \sum_{k \in \mathcal{K}^+_{j,f}} F_{jfk} \tag{37}$$

The three fields are five-state visibility, known relation–anchor, and known articulation. This macro gives supported destination classes equal weight before averaging fields, while prediction-only classes affect micro-F1 and write diagnostics, not the target-supported macro. Matched delay is prediction time minus onset, and missed events reduce recall instead of receiving an arbitrary finite delay.

Event physical NLL. Let $\mathcal{E}_{j,f}$ be supported certified events for field $f \in \{J, A\}$, $\mathcal{W}_{e,f}$ the nonempty label-valid subset of $\mathcal{W}_{e,8}$ for the register associated with event e, and $\mathcal{F}^{JA}_j$ the relation/articulation fields with support. The episode score is

$$L_j^{\text{event}} = \frac{1}{|\mathcal{F}_j^{JA}|} \sum_{f \in \mathcal{F}_j^{JA}} \frac{1}{|\mathcal{E}_{j,f}|} \sum_{e \in \mathcal{E}_{j,f}} \frac{1}{|\mathcal{W}_{e,f}|} \sum_{t \in \mathcal{W}_{e,f}} -\log \max\{b_t^f(Y_t^f), 10^{-12}\} \tag{38}$$

Each event first averages the negative log probability assigned to its reference state over eligible timesteps, and events then receive equal weight within each field. Unsupported fields are omitted. The probability floor prevents undefined logarithms. All logarithms are natural, so NLL is measured in nats. A corresponding Brier diagnostic replaces the final term with $\sum_k (b_t^f(k) - \mathbf{1}[k = Y_t^f])^2$, where k ranges over categorical states. Logarithmic and Brier scores assess probabilistic predictions (Gneiting & Raftery, 2007), while our event-centered aggregation emphasizes post-change intervals. Better proper scores alone do not establish better calibration.

D.5 Persistence, reappearance, and stable writes

For one supported episode, let $\mathcal{T}_f^{\text{stable}}$ contain the register–time pairs (i, t) for which both $t-1$ and t have valid field labels and the reference state of register i is unchanged. The reference-stable write rate (SWR) is

$$\text{SWR}_f = \frac{1}{|\mathcal{T}_f^{\text{stable}}|} \sum_{(i,t) \in \mathcal{T}_f^{\text{stable}}} \mathbf{1}[\widehat{Y}_t^{i,f} \neq \widehat{Y}_{t-1}^{i,f}] \tag{39}$$

The rate pools valid adjacent pairs over registers and time within the episode, then follows the episode–family aggregation below. It counts a prediction change while the reference is stable. For example, revising a wrong ON prediction to the correct IN state can count as a stable-reference write if the reference was already IN at both steps. Thus the rate includes both errors and repairs.

Hidden retention tests relation–anchor and applicable articulation at offset h from the first hidden frame, restricted to objects that remain hidden, label-valid, and physically stable through that endpoint. It does not test visibility or geometry. Its unconditional score can reflect correction of an initially wrong belief, so onset accuracy and endpoint retention must be distinguished.

Conditional hidden-state persistence. For hidden interval r, the binary indicators $C_{r,0}$ and $C_{r,h}$ record correctness at the first hidden frame and endpoint, and $M_{r,h} \in \{0, 1\}$ records whether the interval remains eligible through offset h. A conditional diagnostic separates retention of an initially correct estimate from later correction:

$$\text{ConditionalHR@}h = \frac{\mathbb{E}_w[C_{r,0} C_{r,h} \mid M_{r,h} = 1]}{\mathbb{E}_w[C_{r,0} \mid M_{r,h} = 1]} \tag{40}$$

Here $\mathbb{E}_w$ follows the episode–family–training-repetition aggregation in Appendix D.6. The ratio is the weighted probability of a correct endpoint conditional on a correct onset among eligible intervals. It is undefined if no eligible onset is correct and does not assert uninterrupted correctness between the endpoints.

Reappearance recovery starts at the first valid direct observation after a certified hidden interval and requires the applicable fields and role subset to be correct together in its eligible window. Role-set exactness separately measures equality of the complete predicted/reference subset on valid queries. Neither register validity nor role exactness is an independent test of physical-instance identity.

D.6 Aggregation and uncertainty

A source family groups related episodes by their shared origin: intervention branches use their common intervention group, and other replays use their source-trajectory identifier. This keeps related branches together when splitting data and prevents them from acting as independent samples in the family-level summary.

For metric m, let $r_{j,m}$ be episode j's score, $\mathcal{J}_{g,m}^+$ the supported episodes of source family g, and $\mathcal{G}_m^+$ the supported families. The point estimate for a fitted model is

$$\widehat{M}_m = \frac{1}{|\mathcal{G}_m^+|} \sum_{g \in \mathcal{G}_m^+} \frac{1}{|\mathcal{J}_{g,m}^+|} \sum_{j \in \mathcal{J}_{g,m}^+} r_{j,m} \tag{41}$$

For ABR, $r_{j,m}$ is Equation 35, and the same hierarchy applies after each metric's episode-level reduction. A family with three related episodes and a singleton family each receive one family weight. No-support episodes/families are omitted rather than scored zero. The main table reports the mean and sample standard deviation of these estimates over three independent training repetitions. The standard deviation describes training variability, not episode variation. All repetitions use the same test episodes, so they are not three independent test sets.

For a fixed metric m and methods C and B, let m_g^C, m_g^B denote their scores for family g, and let $\mathcal{G}_m^+$ contain families supported for both methods. Their paired difference is

$$\widehat{\Delta}_m = \frac{1}{|\mathcal{G}_m^+|} \sum_{g \in \mathcal{G}_m^+} (m_g^C - m_g^B) \tag{42}$$

For the main comparisons, $m_g^C - m_g^B$ in Equation 42 is first averaged over the three training repetitions. We then resample the same families for both methods with replacement 5,000 times and report the 2.5th/97.5th percentiles. Pairing preserves shared task difficulty. These intervals condition on the fitted models rather than estimating uncertainty over future training runs. All intervals are unadjusted for multiplicity. The strongest comparator is defined separately for each endpoint by aggregate point estimate.

Readout and proposal-component analyses are post-hoc comparisons. Each contrast uses identical reference eligibility for both predictions, and the in-distribution and target-asset populations are always evaluated separately.

Evaluation uses trained checkpoints and complete causal histories, reaching 606 visual steps in the in-distribution cohort and 560 in the target-asset cohort. The recovery deadline limits the scoring window, not the input history.

E Primary-Relation Semantics and Reference Reliability

The benchmark represents a primary relation, which requires an explicit rule when several physical relations hold together. This section defines that rule, the grasp-evidence adjudication used throughout the main evaluation, and the resulting event populations. Appendix H reports CERB and baseline results under these reference conditions.

E.1 A single primary relation

For five registered anchors, the relation domain is

$$\mathcal{S}_J = \{\text{UNKNOWN}, \text{NONE}, \text{HELD}\} \cup \{\text{IN}(c), \text{ON}(c) : c = 1, \ldots, 5\}$$

This variable represents the selected *primary* relation used by the benchmark, not a multi-label graph of every simultaneously true edge. A directly grasped object can, for example, remain geometrically inside a bowl while its primary label is HELD. For object o and anchor c, let $d_t^{\text{grasp}}(o)$ indicate direct grasp, $R_t^{\text{scene}}(o, c)$ indicate a registered scene relation, ϖ_c be the stored anchor priority, and $r(c)$ the corresponding IN/ON label. Write $\mathcal{C}_t^{\text{scene}}(o) = \{c : R_t^{\text{scene}}(o, c) = 1\}$ and $\mathcal{C}_t^{\text{top}}(o) = \arg\max_{c \in \mathcal{C}_t^{\text{scene}}(o)} \varpi_c$. On labelable source states, the compiler uses

$$y_t^{\text{raw}}(o) = \begin{cases} \text{HELD}, & d_t^{\text{grasp}}(o) = 1, \\ r(c^*), & d_t^{\text{grasp}}(o) = 0, \quad \mathcal{C}_t^{\text{top}}(o) = \{c^*\}, \\ \text{NONE}, & d_t^{\text{grasp}}(o) = 0, \quad \mathcal{C}_t^{\text{scene}}(o) = \varnothing \end{cases} \tag{43}$$

Here $\mathcal{S}_J$ is the relation-state set, $\mathcal{C}_t^{\text{scene}}(o)$ contains anchors with a registered relation to object o, and $\mathcal{C}_t^{\text{top}}(o)$ contains the highest-priority members of that set. The compiled label $y_t^{\text{raw}}(o)$ is HELD for a direct grasp, the relation of the unique highest-priority anchor c^* otherwise, or NONE when no anchor applies. When no direct grasp is present and multiple anchors share the highest priority, source export rejects the episode as relation-ambiguous rather than breaking the tie by array order or assigning a new state. This source-quality decision is distinct from the later grasp adjudication below,

which can retain unresolved contact evidence within an accepted episode. Stored register-specific priorities take precedence, whereas without them, the exporter orders object receptacles above fixture volumes/refrigerator racks, fixture racks, and fixture surfaces.

E.2 PHYSICS-INFORMED GRASP ADJUDICATION

The original HELD predicate combined gripper contact with a fixed 35 mm closure threshold. The adjudication also uses replayed opposing-finger contacts, object/end-effector motion, jaw aperture, and relative pose on the 20 Hz source timeline. Its central positive rule recognizes bilateral carrying even when a large object remains above the aperture threshold, while incidental contact is left unresolved rather than forced to HELD. The rules were developed on 70 natural development episodes, frozen before predictions were scored against the adjudicated reference, and applied identically to every method. No model output enters an adjudication rule.

Table 7: Frozen adjudicated grasp-evidence rules. Motion thresholds are evaluated on replayed physical states.

Evidence	Decision criterion
No direct contact	No gripper contact with the object at the current source state.
Opposing closed contact	Opposing finger contacts together with the original 35 mm closure criterion.
Bilateral carrying	Bilateral fingerpad contact over 0.2 s, relative displacement at most 2 mm and relative rotation at most 2°, aperture change at most 1 mm, and object and end effector each move at least 10 mm.
Short contact gap	Qualified grasping on both sides of a gap no longer than six source steps, with stable relative pose and carried motion.
Supported incidental contact	Object displacement at most 0.5 mm while end-effector and relative motion are at least 15 mm, with a neighboring scene relation and no opposing grasp.
Other contact	Retain an admissible label set rather than forcing a unique grasp decision.

The adjudication changes 1,928 of 162,463 valid test object–time positions, while 4,019 test positions remain unresolved and are handled by the sensitivity analysis below. All hard corrections add missed HELD labels supported by bilateral carrying. Rendered inputs, executed actions, proprioceptive feedback, queries, prediction-candidate availability, validity, visibility, roles, articulation, and spatial targets are unchanged.

Table 8: adjudicated primary-relation revision. A position is one object slot at one sampled time. Valid and unresolved counts cover the sidecar reference arrays, not a particular metric's event denominator. All episodes are retained. Episodes with a correction may also contain unresolved positions.

Split	Episodes	Corrected episodes	Valid positions	Corrected	Unresolved
IID	397	46	122,036	941	2,782
Asset	267	45	40,427	987	1,237
Development	498	54	127,240	1,138	3,634

E.3 REFERENCE CONDITIONS AND EVENT ELIGIBILITY

Full-population adjudicated reference. The primary tables overlay confirmed corrections on the original point labels and retain every test episode. Event lists and recovery windows are recomputed from the revised relation stream, identically for all methods. This is the adjudicated condition reported in the abstract and Table 1.

Grasp-resolved events. A complementary diagnostic retains only events whose source dwell, destination dwell, and recovery dependencies have a unique grasp adjudication. The requirement is applied equally to hits and misses. Because it concerns object grasp semantics, it retains every container event but only about 40% of Object ABR events, changing the object/container mixture.

Table 9: Coverage of grasp-resolved events within the adjudicated event inventory. The source/destination dwell and complete recovery dependencies must be resolved. All container events pass this grasp-specific filter, whereas only about two fifths of Object ABR events do.

Split	Event population	Resolved / total	Coverage
IID	ABR	524/1,051	49.86%
IID	Whole	1,582/2,304	68.66%
IID	Object ABR	340/867	39.22%
IID	Container ABR	184/184	100.00%
Asset	ABR	365/708	51.55%
Asset	Whole	418/805	51.93%
Asset	Object ABR	233/576	40.45%
Asset	Container ABR	132/132	100.00%

Grasp resolution changes the object/container event mixture, not only annotation certainty. Full-population and resolved-event scores should therefore be read together with their event-type breakdown. The results in Appendix H use the same trained CERB and baseline predictions in both conditions.

Probability sensitivity at unresolved positions. For an unresolved object–time position with admissible relation labels $\mathcal{Y}_t^{\mathrm{adm}} \subseteq \mathcal{S}_J$, let $b_t^J(y)$ be the predicted probability of relation label y. We evaluate the best and worst label-specific negative log probabilities

$$\ell_t^{\mathrm{low}} = \min_{y \in \mathcal{Y}_t^{\mathrm{adm}}} -\log \max(b_t^J(y), 10^{-12}), \qquad \ell_t^{\mathrm{high}} = \max_{y \in \mathcal{Y}_t^{\mathrm{adm}}} -\log \max(b_t^J(y), 10^{-12}) \tag{44}$$

These bounds describe sensitivity to the admissible reference label at an unresolved position. They are distinct from Event NLL in the main tables, which uses the full-population adjudicated point reference and its event windows.

F Functional Decomposition of CERB

Task binding, physical-state recovery, and transition timing place different demands on a belief estimator. The following comparisons identify how CERB's specialized branches contribute to these capabilities.

F.1 Fixed-weight readout comparisons

The comparisons reuse the trained branches without additional optimization. Both E and P produce complete physical predictions and their own role outputs. The E + R and P + R readouts replace only the role predictions, while CERB combines fields as specified in Appendix B. All configurations share the observer, unrestricted prediction candidates, and adjudicated test populations.

Task binding is distinct from physical estimation. The dedicated role branch raises in-distribution role-set exactness by 14.67 percentage points over E's readout and by 8.56 points over P's, with corresponding ABR gains of 2.97 and 1.88 points. Whole@8, transition F1, and Event NLL remain unchanged because the physical predictions are identical. ABR imposes an additional requirement: correct binding must coincide with correct physical fields, and its event population also includes containers. Improved binding therefore contributes to, but does not determine, actionable recovery.

Different physical estimates favor different criteria. Physical revision chiefly improves relation-transition timing. The P readout raises relation F1 over E from .195 to .469 on in-distribution episodes and from .209 to .585 on target assets. CERB retains these relation predictions while improving Whole@8 over P + R by 2.02 points [1.26, 2.79] and 1.46 points [.83, 2.14], respectively. In contrast, E + R achieves higher ABR@8 and Whole@8 and lower Event NLL than CERB. Event NLL measures probability assigned to the reference state, whereas transition F1 requires discrete changes to match within causal windows. These criteria identify different strengths even when the same role predictions are used. Brackets denote paired source-family 95% intervals.

Table 10: Complete-system readout ablations. Entries are mean ± sample standard deviation over three training repetitions. Recovery and role-set exactness are percentages. E and P alone use their own role outputs.

Readout	ABR@8 ↑	Whole@8 ↑	Trans. F1 ↑	Event NLL ↓	Role exact ↑
In-distribution episodes					
E	$57.61 \pm .80$	$46.44 \pm .66$	$.140 \pm .013$	$.321 \pm .038$	74.95 ± 1.98
P	59.45 ± 1.32	43.24 ± 2.95	$.314 \pm .020$	$.416 \pm .005$	81.06 ± 2.33
E + R	$60.58 \pm .26$	$46.44 \pm .66$	$.140 \pm .013$	$.321 \pm .038$	89.62 ± 2.89
P + R	61.33 ± 2.79	43.24 ± 2.95	$.314 \pm .020$	$.416 \pm .005$	89.62 ± 2.89
CERB	$59.60 \pm .79$	45.26 ± 2.16	$.305 \pm .024$	$.416 \pm .005$	89.62 ± 2.89
Held-out target assets					
E	63.55 ± 1.21	53.29 ± 1.08	$.178 \pm .011$	$.229 \pm .052$	$95.32 \pm .46$
P	60.46 ± 3.14	50.24 ± 3.36	$.476 \pm .049$	$.360 \pm .011$	$97.26 \pm .70$
E + R	64.44 ± 1.11	53.29 ± 1.08	$.178 \pm .011$	$.229 \pm .052$	$99.61 \pm .15$
P + R	60.53 ± 3.05	50.24 ± 3.36	$.476 \pm .049$	$.360 \pm .011$	$99.61 \pm .15$
CERB	61.44 ± 2.11	51.70 ± 2.35	$.458 \pm .055$	$.360 \pm .011$	$99.61 \pm .15$

Table 11 separates transition F1 into visibility, relation, and articulation, showing which field predictions each readout preserves.

Table 11: Field-level transition F1 for the same fixed-weight readouts. Each field is averaged over its supported episodes and families, so the three column means need not average to aggregate transition F1.

	In-distribution			Target assets		
Readout	Visibility	Relation	Articulation	Visibility	Relation	Articulation
E + R	.062	.195	.037	.086	.209	.012
P + R	.124	.469	.050	.200	.585	.019
CERB	.062	.469	.050	.086	.585	.019

Appendix H extends this analysis to hidden-state retention, reference-stable writes, and sustained recovery. The next section examines local attention and typed memory within the temporal proposal.

G Temporal-Proposal Component Studies

Local attention and typed change memory refine the recurrent estimate in different ways: one reads recent observations directly, while the other accumulates field-specific changes. Matched training continuations test their individual and combined contributions.

G.1 Matched training design

Each continuation starts from the same repetition-specific 1,200-update GRU checkpoint and receives the same 400 batches, data order, optimizer, objective, and final-checkpoint rule. The common correction architecture S retains the slow GRU, current-observation correction, field router, action-effect pathway, full decoder, and local belief. We then add local attention, typed delta memory, or both. Continuing the unmodified GRU distinguishes these architectural changes from the effect of additional training alone.

These development experiments use supplied object and role candidates, a single-estimator update, and the adjudicated reference. They complement the complete-system readout comparisons in Appendix F, which use unrestricted prediction candidates.

G.2 Correction, short-term context, and typed memory

The correction architecture reaches 38.44% ABR, compared with 28.62% for the continued GRU and 34.49% before continuation. Typed memory adds 5.88 percentage points and reduces Event NLL

Table 12: Proposal-component ablations on adjudicated development streams. Means over three training repetitions, with the same 400 additional batches for each continuation. ABR and Whole are percentages. The initial GRU is shown before continuation. Reported differences use full-precision scores.

Estimator	ABR@8 ↑	Whole@8 ↑	Trans. F1 ↑	Event NLL ↓
Initial GRU	34.49	15.60	.1296	2.5838
Continued GRU	28.62	11.92	.1295	2.8709
Correction architecture S	38.44	23.81	.1509	2.6808
S+local attention	41.87	27.26	.1745	2.2458
S+typed delta memory	44.31	32.75	.1883	1.6636
S+local+delta	44.14	33.02	.1767	1.7271

from 2.681 to 1.664, while local attention adds 3.44 points and reduces NLL to 2.246. Both improve the recurrent estimate when added separately.

Their gains overlap. Adding local attention to typed memory changes ABR by $-.17$ points, with a paired interval spanning zero, and Whole by $+.27$ points. The negative ABR interaction (-3.61 points) is consistent with partially overlapping corrective roles.

On grasp-resolved events, the correction architecture improves ABR by 3.69 points [1.87, 5.58] and Whole by 7.37 points [5.29, 9.61] over continued GRU. The additional attention and typed-memory effects remain uncertain on this smaller event population.

H Recovery, Reference Sensitivity, and Task Composition

These analyses connect aggregate recovery to field-level timing, reference sensitivity, and task composition. They use the same trained models and adjudicated labels as the main comparison.

H.1 Paired comparisons and field-level behavior

Table 13 reports paired CERB–baseline differences, averaging each family's difference over the three training repetitions before resampling families. All primary point estimates favor CERB. The target-asset Whole contrast with Gated DeltaNet-2 is $+1.66$ percentage points, with a 95% interval of $[-.52, 3.94]$, while the other primary intervals exclude zero.

Table 13: CERB minus each adapted temporal baseline, with paired source-family 95% bootstrap intervals. Within-family differences are averaged across the three fitted repetitions before 5,000 resampling draws, and intervals are unadjusted for multiplicity. Recovery differences are percentage points, and Event NLL favors negative differences. GDN-2 denotes Gated DeltaNet-2.

Comparator	ABR@8	Whole@8	Trans. F1	Event NLL
IID				
GRU	29.12 [26.55, 31.79]	23.96 [21.74, 26.21]	0.167 [0.141, 0.193]	−2.113 [−2.291, −1.939]
SlotSSM	21.69 [18.92, 24.61]	17.65 [15.37, 20.05]	0.135 [0.110, 0.159]	−1.717 [−1.901, −1.538]
GDN-2	20.70 [17.09, 24.57]	2.25 [0.39, 4.11]	0.151 [0.123, 0.180]	−1.116 [−1.289, −0.950]
Mamba-3	24.41 [21.21, 27.84]	4.65 [2.85, 6.44]	0.164 [0.138, 0.192]	−1.492 [−1.669, −1.319]
Target assets				
GRU	35.34 [32.47, 38.16]	29.51 [26.85, 32.13]	0.242 [0.211, 0.273]	−3.203 [−3.468, −2.942]
SlotSSM	28.46 [24.57, 32.32]	19.31 [16.49, 22.23]	0.182 [0.148, 0.217]	−2.409 [−2.651, −2.167]
GDN-2	12.34 [8.63, 16.14]	1.66 [−0.52, 3.94]	0.265 [0.229, 0.301]	−0.782 [−0.957, −0.611]
Mamba-3	14.75 [11.01, 18.49]	2.58 [0.33, 4.86]	0.260 [0.225, 0.295]	−1.161 [−1.381, −0.940]

CERB and SlotSSM share the same R role readout, so their difference in joint object recovery comes from physical estimation. Container ABR measures whether a correct state occurs within the recovery window, whereas articulation F1 measures when that state changes. The two can therefore differ substantially. Visibility write activity also rises, motivating the temporal and correctness-based decomposition below.

Table 14: Field-level evaluation under the main evaluation protocol: means over three training repetitions (percentages). Recovery and retention use offset eight. Hidden rel. checks relation–anchor retention, and Hidden joint additionally checks applicable articulation. Role exact is exact task-role subset agreement.

Backend	Object ABR	Container ABR	Role exact	Hidden rel.	Hidden joint
IID					
GRU	16.56	62.47	72.97	62.86	43.09
SlotSSM	26.09	43.01	89.62	59.19	42.69
Gated DeltaNet-2	39.06	25.59	81.14	62.76	43.39
Mamba-3	31.97	29.78	56.16	57.68	37.38
CERB	43.86	100.00	89.62	73.47	59.28
Target assets					
GRU	22.92	52.29	95.72	76.30	42.70
SlotSSM	35.12	6.25	99.61	58.92	15.73
Gated DeltaNet-2	53.90	7.71	99.44	85.12	50.52
Mamba-3	50.63	8.33	84.20	82.80	42.20
CERB	56.39	98.54	99.61	93.37	64.44

Table 15: Reference-stable write rates by physical field, averaged over three training repetitions (percentages). A write changes a prediction while the reference remains unchanged, including corrections of earlier errors.

Backend	Visibility writes	Relation writes	Articulation writes
IID			
GRU	1.42	4.20	1.57
SlotSSM	1.99	4.38	1.92
Gated DeltaNet-2	1.20	4.85	1.57
Mamba-3	1.31	4.39	1.66
CERB	3.66	3.33	1.50
Target assets			
GRU	1.29	4.18	0.59
SlotSSM	1.99	4.84	0.26
Gated DeltaNet-2	0.86	5.01	0.57
Mamba-3	1.23	4.57	0.58
CERB	3.60	2.91	1.32

H.2 When recovery occurs and whether it persists

For each eligible transition, we measure how often the prediction already equals the destination state during the certified three-observation source dwell preceding onset. The early-prediction rate divides destination-state predictions by valid source-dwell observations, not by the number of events. These pre-onset predictions distinguish early destination estimates from recovery after a reference change. Relation events use grasp-resolved dependencies, while visibility and articulation use their existing references. CERB predicts the future relation destination less often than the other backends during the source dwell: 12.87% on IID and 9.90% on target assets. The pattern reverses for visibility and articulation, where early destination predictions are more frequent.

The Retained@8 diagnostic requires the first jointly correct prediction to remain correct until the end of a fully valid Whole-event window through offset eight. Its denominator includes every such window, including those with no recovery hit; it is therefore distinct from ConditionalHR, which conditions on a correct hidden-state onset. CERB scores 49.74%/64.66% on the in-distribution/target-asset cohorts, compared with 47.83%/63.62% for Gated DeltaNet-2. This diagnostic uses a narrower event population than Whole@8 and requires no additional persistence after a first hit at offset eight.

Stable-reference writes can either harm or repair belief. On IID, CERB changes a correct visibility estimate to an incorrect one on 1.73% of eligible stable pairs and repairs an incorrect estimate on 1.59%, and the corresponding Gated DeltaNet-2 rates are .45% and .54%. For relation, CERB's harmful changes are lower (1.17% versus 2.05%), while repairs are similar (1.71% versus 1.62%).

Table 16: Temporal diagnostics on frozen predictions, averaged over three training repetitions (percentages). Early vis./rel./art. are the fractions of certified source-dwell observations predicting the subsequent visibility/relation/articulation destination. Hit@8 and Retained@8 use all complete, valid nine-observation Whole windows: any jointly correct estimate, and correctness from the first hit through offset eight, respectively. Both include unsuccessful windows in the denominator, which differs from main-table Whole@8.

Backend	Early vis.	Early rel.	Early art.	Hit@8	Retained@8
IID					
GRU	47.98	37.42	39.02	23.82	17.74
SlotSSM	33.78	33.53	32.83	35.43	28.71
Gated DeltaNet-2	35.66	28.35	17.20	52.47	47.83
Mamba-3	40.68	26.64	19.63	49.81	46.24
CERB	58.30	12.87	83.67	53.57	49.74
Target assets					
GRU	47.23	28.84	24.79	28.03	19.69
SlotSSM	28.85	30.43	4.86	46.60	36.90
Gated DeltaNet-2	38.86	32.30	10.63	68.45	63.62
Mamba-3	46.76	29.73	8.92	66.47	62.15
CERB	63.37	9.90	90.60	67.93	64.66

These decompositions distinguish harmful changes from repairs, which the aggregate reference-stable write rate counts together.

H.3 Observer-source exposure

We exclude known observer-training sources and, separately, known observer-training and development sources using provenance-defined subsets. Table 17 reports the remaining 589 and 577 episodes. Earlier provenance is incomplete, so these subsets measure sensitivity to known overlap rather than independence from every learning stage. On the strictest source-exclusion subset, the IID Whole difference from Gated DeltaNet-2 is $+1.57$ points with a 95% interval of $[-.62, 3.72]$. Asset Whole differs by $+.40$ points from Gated DeltaNet-2 ($[-2.07, 2.83]$) and $+.78$ from Mamba-3 ($[-1.48, 3.15]$).

H.4 Grasp-resolved event coverage

The grasp-resolved subset requires unique adjudication throughout each event's source dwell, destination dwell, and recovery dependencies. Table 9 gives its coverage. All container events remain eligible, changing the object/container mixture relative to the full cohort. CERB's resolved-event ABR is lower on target assets than on in-distribution episodes, reversing the full-cohort ordering. Event composition thus contributes to the observed difference between cohorts.

H.5 Common-task standardization

IID covers 35 tasks and the asset cohort covers 16. Median history lengths are 107 and 80 visual steps, with ranges 21–606 and 31–560. To examine composition, we average each metric within each common task and weight the IID task means by the asset cohort's eligible-family distribution. All quantities below are metric-specific, with that index omitted:

$$\widetilde{M}_s = \sum_{u \in \mathcal{T}_\cap} w_u^{\text{asset}} \overline{M}_{s,u}, \qquad w_u^{\text{asset}} = \frac{n_{\text{asset},u}^{\text{eligible}}}{\sum_{v \in \mathcal{T}_\cap} n_{\text{asset},v}^{\text{eligible}}} \tag{45}$$

Here s indexes the IID or asset cohort, $\mathcal{T}_\cap$ contains tasks with metric support in both, and $\overline{M}_{s,u}$ is the family-averaged score for task u in cohort s, averaged over the three training repetitions. The count $n_{\text{asset},u}^{\text{eligible}}$ is the number of eligible asset-cohort families for that task, and w_u^{asset} is its normalized weight. ABR uses 15 common tasks, and the other primary metrics use 16. No-event tasks are not assigned zero. Equal-task summaries provide a second descriptive view. The recovery and F1 gaps largely disappear after standardizing task composition. The remaining differences include scene

Table 17: Source-overlap sensitivity under unchanged predictions and references, averaged over three training repetitions. Exclusions follow source provenance, not model performance. Recovery values are percentages.

Backend	ABR@8 ↑	Whole@8 ↑	Trans. F1 ↑	Event NLL ↓
IID: Exclude observer train (352 episodes)				
GRU	33.45	21.81	0.1452	2.4821
SlotSSM	41.12	28.94	0.1807	2.0730
Gated DeltaNet-2	40.77	44.56	0.1587	1.5430
Mamba-3	37.78	42.21	0.1462	1.8565
CERB	61.92	46.20	0.3229	0.3928
IID: Exclude observer train/dev (344 episodes)				
GRU	33.92	21.85	0.1440	2.4658
SlotSSM	41.88	29.24	0.1813	2.0582
Gated DeltaNet-2	41.04	44.53	0.1583	1.5424
Mamba-3	38.41	42.41	0.1465	1.8506
CERB	62.15	46.10	0.3256	0.3853
Target assets: Exclude observer train (237 episodes)				
GRU	26.87	24.03	0.2295	3.4542
SlotSSM	35.83	35.23	0.2915	2.6131
Gated DeltaNet-2	51.60	52.18	0.2014	1.0343
Mamba-3	49.58	51.79	0.2105	1.3528
CERB	59.85	52.84	0.4984	0.3848
Target assets: Exclude observer train/dev (233 episodes)				
GRU	26.96	24.30	0.2318	3.4528
SlotSSM	36.28	35.62	0.2939	2.5770
Gated DeltaNet-2	51.96	52.55	0.2016	1.0215
Mamba-3	50.15	52.17	0.2122	1.3143
CERB	59.49	52.94	0.5034	0.3878

Table 18: Recovery on grasp-resolved events: means over three training repetitions (percentages). The same label-defined dependency filter applies to every model. This nonrandom event subset complements, rather than replaces, full-cohort evaluation.

Backend	ABR@8	Whole@8	Object ABR	Container ABR
IID				
GRU	33.06	17.81	15.63	62.47
SlotSSM	32.37	15.73	14.07	43.01
Gated DeltaNet-2	28.12	25.14	21.84	25.59
Mamba-3	27.68	22.42	18.42	29.78
CERB	55.98	31.60	31.46	100.00
Target assets				
GRU	26.96	18.92	23.47	52.29
SlotSSM	15.53	14.01	16.36	6.25
Gated DeltaNet-2	25.18	25.84	27.44	7.71
Mamba-3	24.35	25.17	26.30	8.33
CERB	53.42	36.28	42.46	98.54

Table 19: Common-task CERB scores, averaged over three fitted repetitions. The first pair uses target-asset eligible-family task weights, and the second weights common tasks equally. These descriptive summaries do not replace the primary cohort scores.

Metric	Tasks	IID (asset wt.)	Asset (asset wt.)	IID (equal)	Asset (equal)
ABR@8	15	61.25	61.44	63.50	63.45
Whole@8	16	51.94	51.70	51.42	51.61
Trans. F1	16	0.4739	0.4584	0.4663	0.4503
Event NLL	16	0.3725	0.3598	0.3808	0.3561

conditions, object identities, history lengths, and reference coverage, so task weighting is descriptive rather than a causal estimate of asset shift.

I BELIEF-GUIDED RECOVERY WITH A SHARED PLANNER

Primary35 evaluates state estimation on fixed streams. This development study tests how those estimates guide action when the robot's decisions determine its subsequent observations. Each trained backend drives the same belief-aware planner and skill executor.

I.1 TASK AND INFORMATION BOUNDARY

The robot must place a target object on a refrigerator shelf despite occlusion or a failed opening or grasping attempt. Seven scenes provide 18 starting states used during development: seven with a hidden target, seven after a failed opening attempt, and four after a failed grasp. Every trained backend repeats these same starts, yielding 54 rollouts per backend across three training repetitions. The scene, rather than each rollout, is therefore the clustering unit for paired uncertainty estimates. Each run allows at most eight high-level decisions and 1,600 robot action steps at 20 Hz.

The planner consumes structured field probabilities and selects among OPEN, SEARCH, GRASP, PLACE, VERIFY, and DONE. A shared skill executor converts those choices into simulator actions. The execution service has known-fixture geometry and local visible-target geometry, shared across all estimators. Each backend supplies belief to this common planning system. Simulator state used by the independent success evaluator is not substituted for the planner's belief.

I.2 A FIXED, BELIEF-AWARE PLANNING RULE

The planner first selects a target using role-membership probability multiplied by object validity. For that target, it summarizes belief as a distribution over five states: in a closed source, in an open source, held, at the goal, or elsewhere. Source membership combines IN and ON probability at the source anchor, and container validity and OPEN probability distinguish the first two states. Visibility and binding confidence remain separate inputs.

At each decision, the planner enumerates admissible sequences of up to three OPEN, GRASP, and PLACE skills, predicts their effect on this five-state distribution, and executes only the first skill of the highest-scoring sequence. It then replans from newly observed belief. If p is the current distribution, p^σ the forecast after sequence σ, and β target-binding confidence, its score is

$$Q(\sigma; p) = \beta\big[U(p^\sigma) - U(p)\big] - \sum_{a \in \sigma} \big[c(a) + .8(1 - \rho_a)\big] - .8\, n_{\text{repeat}}, \tag{46}$$

where $U(p) = 10p_{\text{goal}} + p_{\text{held}} + .2p_{\text{source-open}}$, p_{goal}, p_{held}, and $p_{\text{source-open}}$ are the corresponding state probabilities, a indexes a skill in σ, $c(a)$ is its cost, and ρ_a its forecast success probability. The count n_{repeat} penalizes repeating the same first skill and target from the two most recent decisions. The transitions are hand-specified and identical across estimators. OPEN transfers closed-source mass to open-source mass with probability .9, while GRASP transfers visible open-source mass to HELD with probability .85, with an additional .25 factor for a visible object elsewhere, and PLACE transfers HELD mass to the goal with probability .85. OPEN, GRASP, and PLACE cost .45, .5, and .45, respectively. These are planning assumptions, not measured executor success rates.

A sequence must improve the score by more than .05, and otherwise the planner searches a source or goal region using the remaining uncertain mass and prior search attempts. Goal belief above .55 triggers VERIFY. DONE additionally requires a recent executed VERIFY, goal probability at least .8, visibility at least .6, and binding confidence at least .6. Execution feedback reports skill status and public motion information, not the simulator's true task-success flag. This small planner makes the role of each belief field inspectable without introducing a separately learned policy.

I.3 SUCCESS CRITERIA

Index the recorded post-action success checks by $k = 1, \ldots, K$, where K is the number of executed robot action steps. Define

$$\begin{aligned} C_k = \mathbf{1}&[\text{target lies in the shelf region}] \\ &\cdot \mathbf{1}[\text{target contacts the refrigerator fixture}] \\ &\cdot \mathbf{1}[\text{gripper–target distance} > 0.15\,\text{m}] \end{aligned} \tag{47}$$

The shelf test is geometric, whereas actual fixture contact is checked separately. It does not localize that contact to a particular shelf surface. A rollout succeeds if

$$S = \mathbf{1}[K \geq 20] \prod_{k=K-19}^{K} C_k, \qquad N_{\text{success}} = \sum_{j=1}^{N} S_j \tag{48}$$

The binary check C_k is one only when all three placement conditions hold. Their product over the last 20 checks requires one second of terminal persistence at 20 Hz, and $\mathbf{1}[K \geq 20]$ excludes shorter runs. Here $S_j \in \{0, 1\}$ is the success indicator for rollout j, and N is the number of evaluated rollouts. Task Success counts persistent physical placements, N_{success}/N. The placement must satisfy the geometric and contact conditions throughout the final second, independently of the planner's DONE decision.

I.4 RESULTS AND INTERPRETATION

Table 20: Downstream Task Success by training repetition (successful / total rollouts). DONE counts planner declarations, while False DONE counts declarations without terminal physical success. All repetitions use the same 18 starting states.

Backend	Repeat 1	Repeat 2	Repeat 3	Total	DONE	False DONE
GRU	3/18	2/18	2/18	7/54	0	0
SlotSSM	6/18	9/18	7/18	22/54	4	4
Gated DeltaNet-2	4/18	5/18	8/18	17/54	4	2
Mamba-3	4/18	9/18	0/18	13/54	5	5
CERB	8/18	10/18	9/18	27/54	0	0

Table 21: Paired downstream contrasts. The middle column is the difference in successful rollouts / 54. Confidence intervals are on the corresponding probability-difference scale, from 10,000 paired scene-cluster draws after averaging repetitions within each start. They condition on the seven scenes and three fitted models. Δp is CERB's Task Success proportion minus that of the comparator.

Comparator	Additional successes / trials	95% interval for Δp
GRU	+20/54	$[0.125, 0.550]$
SlotSSM	+5/54	$[-0.042, 0.241]$
Gated DeltaNet-2	+10/54	$[0.000, 0.389]$
Mamba-3	+14/54	$[0.089, 0.400]$

CERB achieves 27/54 physical successes, compared with 7/54 for GRU, 22/54 for SlotSSM, 17/54 for Gated DeltaNet-2, and 13/54 for Mamba-3. Its advantage over SlotSSM comes from hidden-target and failed-open starts, while both complete 10/12 failed-grasp rollouts. Tables 20 and 21 report repetition-level counts and paired scene-cluster intervals. The CERB–SlotSSM interval includes zero, and the CERB–Gated DeltaNet-2 interval reaches zero.

Success is reconstructed from recorded simulator states using Equation 48. The planner never selects DONE in the CERB rollouts, separating physical task completion from recognition of completion in this experiment.

Why offline recovery does not fully determine behavior. For one fitted pair of models, two failed-grasp cases share an execution prefix until a decision at which the container is already open and

the object is not held. CERB assigns source-open probabilities of approximately .521/.745, versus .891/.950 for Mamba-3. The shared planner chooses OPEN with the former beliefs and GRASP with the latter. Swapping only the saved articulation probabilities reverses the decisions, while swapping visibility alone does not. This saved-decision intervention isolates sensitivity to articulation at that decision boundary, without executing a new counterfactual rollout. It illustrates why decision-relevant state errors can matter more than average recovery scores.

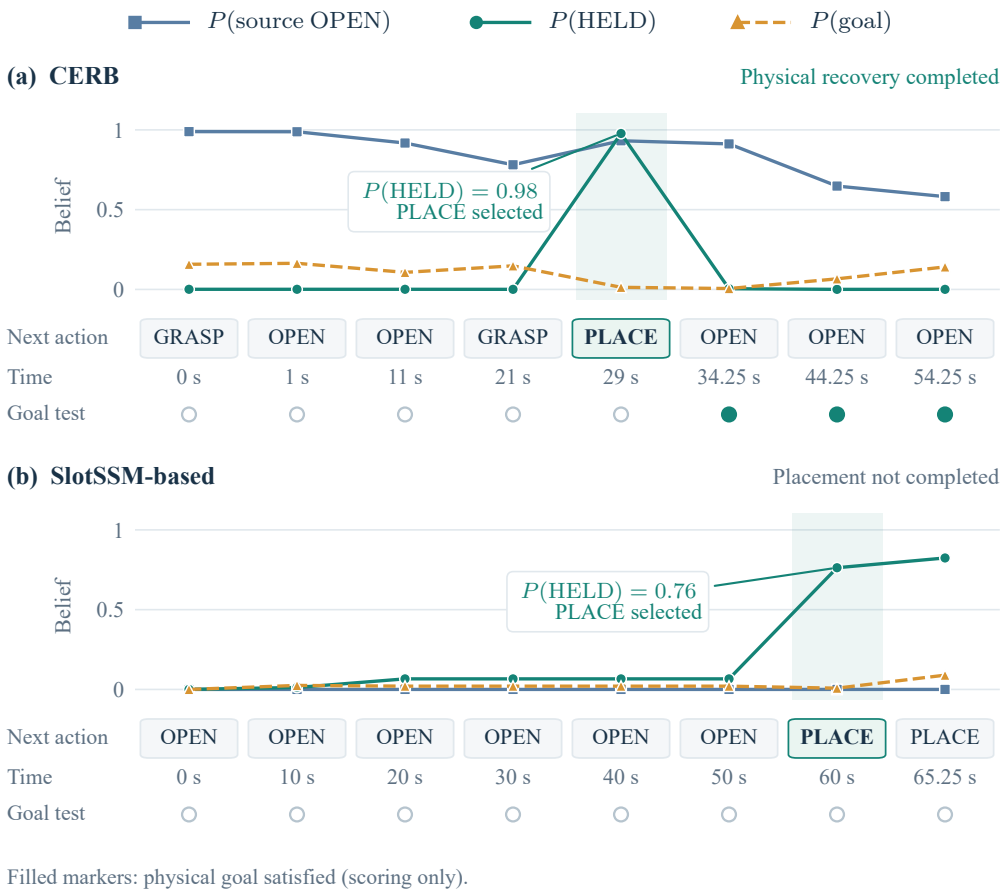


Figure I1: **From belief to decision to physical outcome.** A selected hidden-target development case, with all saved decision points shown. Lines connect decision-time estimates, not continuous frame-level predictions. CERB's HELD probability rises after grasping, leading the planner to select PLACE, and the independent goal test is satisfied at the next saved decision. Its goal belief nevertheless remains low, and it continues acting. The SlotSSM-driven rollout selects PLACE later but does not complete the task. Elapsed times identify each rollout's own action-dependent history, not matched video frames. The goal-test markers are scoring-only observations, and final Task Success additionally requires terminal persistence. Aggregate counts appear in Table 3.

Figure I1 shows how a belief can support successful grasping and placement while remaining slow to recognize the resulting goal state. The shared-planner study connects state estimation to physical recovery and identifies completion recognition as a distinct requirement.

J Properties of Evidence-Conditioned Belief Revision

The update separates whether evidence is available from which state it supports. The following properties characterize the update operator, while prediction accuracy is evaluated empirically.

Normalized legal state. Each relation/articulation candidate is normalized on the same typed support $\mathcal{S}_f$. The bounded logit correction preserves that support, and the reliability weights form a simplex. Equation 6 therefore maps valid categorical distributions to valid categorical distributions. Induction from the normalized initial state gives normalization and legal support at every step. The candidate vocabulary includes every declared anchor without asserting that it exists in the scene.

Exact persistence without new field evidence. If neither direct field support nor an elapsed action–proprioception interval is available over visual steps t_0 through t_1, the explicit copy branch gives

$$b_t^f = b_{t_0-1}^f \quad \text{for all } t \in [t_0, t_1], \qquad f \in \{J, A\} \tag{49}$$

Visibility analogously holds when its direct support is zero, and validity and geometry follow their declared copy permissions. This is equality of the public distribution-parameter tensor, rather than a small-drift penalty. Available action feedback permits reconsideration, including retention, but it does not force a state transition.

Probability-pooling bound. Let $m \in \mathcal{O} = \{K, O, A\}$ index the retained, measurement-corrected, and action candidates $Q_{t,m}^f$, with normalized nonnegative weights $w_{t,m}^f$, as in Equation 6. For reference state y, convexity of $-\log$ gives

$$-\log\left(\sum_{m\in\mathcal{O}} w_{t,m}^f Q_{t,m}^f(y)\right) \le \sum_{m\in\mathcal{O}} w_{t,m}^f [-\log Q_{t,m}^f(y)] \tag{50}$$

The bound compares pooled NLL with the weighted average of candidate NLLs, not with the best candidate. CERB scores the emitted posterior directly and combines correlated candidates discriminatively, without assuming independent evidence sources.

Physical invariance to later queries. With identical original instruction, sensory and action history, initialization, and physical parameters, two different later role queries traverse the same physical computation. The queries enter only R's detached subset readout, so the physical-belief trajectories coincide while the role answers may differ. Changing the original instruction is outside this invariance statement because it conditions visual grounding.

Why physical change is not an update permission. A failed grasp may leave the world unchanged while providing evidence against an earlier HELD estimate. Requiring a predicted physical transition before every revision would prevent that error from being corrected. CERB instead uses field evidence to permit revision and learned reliability to determine how much of the previous belief to retain. Joint recovery, transition timing, and reference-stable writes measure the resulting behavior.