TargetSpace: Benchmarking Personal Intelligence by Target-Specific Forecasting Under Partial Observation

A prospective framework for forecasting future target-state transitions from third-person longitudinal observation, with own-routine baselines, permutation specificity, calibration, and evidence ablation

Yuri Andrade Sylvester
Independent Researcher
Carlsbad, California, United States
yurisyl@gmail.com
ORCID: 0009-0008-6513-6087

Public preprint — Version 1.0 (pre-pilot protocol proposal)

Abstract

Personal AI is increasingly evaluated on recall, chat memory, personalization, and satisfaction of stated preferences — how well a system re-serves what a person has already externalized. We argue the harder capability is **target-specific prospective forecasting under partial observation**: given observations of a specific target up to a sealed time t, can a model forecast that target's future **observable state transitions**, and can it be shown to be modeling *this* target rather than the average? We introduce **TargetSpace**, a prospective benchmark framework for exactly this. Forecasts are timestamped and hashed before their outcomes exist, are probabilistic over a defined answer space, and are scored by strictly proper rules against future observable outcomes that resolve by deterministic rules — contamination-resistant by construction rather than by curation. A stack of controls makes target-specificity the measured quantity: a population-prior baseline (**R1**) the model must beat to exceed base rates; an **own-routine baseline (R2)** it must beat to exceed routine and memory replay; a **calibration gate**; a **permutation specificity** gate under which forecasts scored against the wrong target must lose their skill; and an **evidence-tier ablation** measuring what passive longitudinal observation adds over digital exhaust and self-report. Personal world modeling is the flagship instantiation, where self-report is an auxiliary channel with its own biases, not the substrate. Proper scoring, calibration, sealing, and ablation are adopted from prior work and cited; the contribution is their conjunction, anchored on the own-routine baseline and the permutation gate. This is a pre-pilot protocol proposal: no empirical results are reported, only the personal track is instantiated (synthetically), and a high score certifies calibrated predictive skill about a consenting individual — never licence to act on, manipulate, or intervene on them.

1 Introduction

Personal AI is now evaluated largely on what a user hands it. Assistants that "remember you," companions, coaches, and adaptive copilots are judged on recall of saved facts, continuity of chat memory, personalization to stated preferences, and fidelity of imitated style — capabilities assembled

from artifacts the user chose to provide: prompts, journals, notes, curated memories, declared goals, messages, and calendars. These are genuine capabilities, but they mostly measure how well a system stores, retrieves, and re-serves what a person has already externalized. They do not measure whether the system has modeled the person *as a system* — whether it has learned how *this* individual's situation, attention, and commitments actually evolve, and can say what comes next. That capability is valuable, commercially central, and almost entirely unaudited. For any product that claims persistent personal context — memory, coaching, companionship, or adaptive assistance — the operational question is what that context actually buys over base rates and routine; retrieval, summaries, and personalization are useful, but are not by themselves evidence of it.

We argue the harder capability is modeling a *specific* target as it evolves: treating a person as an **externally observed dynamic system** and forecasting how it unfolds from **third-person longitudinal observation** rather than from a curated self-description. The person is not a profile to retrieve but an **observed trajectory** in motion. Self-report and passive observation are two evidence channels with asymmetric biases; TargetSpace treats self-report as auxiliary and *measures* what observation adds rather than assuming it (Section 2). The framing is motivation — the object of this paper is the measurement.[1]

This capability class is arriving, which is why the evaluation gap is now urgent rather than hypothetical. Ambient AI recorders that began with meeting transcription now advertise longitudinal "personal context" — "memory, goals, patterns" — as their competitive moat [96,98]; wearable memory devices are marketed to turn everyday conversation into persistent summaries and commitments [97,99]; and a second wave adds first-person vision through camera-and-microphone glasses, some shipping and some only announced [100–102], with egocentric-video research supplying datasets and methods at research scale — generic activity clips, not longitudinal single-person records [56,57]. These devices are only the most visible edge: the same problem arises whenever a system accumulates longitudinal observations of a target — assistants over email and calendar, enterprise copilots and organizational memory, tutors over student histories, care systems, and recommenders. They are evidence that the capability class is arriving, not validation of any method, and not a claim that any vendor misuses data; the analytical survey is Appendix O. The vocabulary of memory, context, goals, and patterns is advancing faster than any way to test it: what none of these systems has is a measure of what longitudinal capture actually buys, over base rates and over the person's own routine. That measurement apparatus is what this paper specifies.

Understanding is the capability; forecasting is the measurement. The capability is latent and cannot be inspected directly, so we measure it by **calibrated prospective forecasting** — the role classification played for ImageNet [1] and standardized tasks for language and code [20,21], and the standard by which a model is judged across the sciences when the thing modeled cannot be seen [14,15,18,22]. We do not claim prediction *is* understanding; it is its strongest available operational signal, and only when shown to be about the *target* rather than the crowd. TargetSpace is therefore **forecast-first, not forecast-only**: a good score is a precondition for an explanatory claim, never proof of one. Of the things a model might do — describe a target, infer its current state, predict its next action, forecast its next target-state transition — the benchmark scores only the last.

The difficulty is that convincing forecasts need not be target-specific at all. A model can look skilled by predicting what *usually* happens (most meetings occur, most emails get answered) and personalized by replaying a target's *routine* (this person checks email at 9am), yet capture nothing specific to *this* person beyond base rates and habit; retrospective explanation cannot settle the question, since a model can rationalize what already happened without being held to what happens

[1]The motivating contrast between third-person description and situated first-person experience loosely follows Nagel [92], with no claim of access to consciousness or subjective states; extended motivation in Appendix H.

next. Target-specific skill is what remains after base rates and routine are subtracted, and it means three things together: the forecast beats a strong **own-routine** baseline, stays **calibrated**, and **collapses** when scored against the wrong person's outcomes. Predicting that someone will eventually answer an email is a base rate; predicting it because they usually reply within a day is routine; the target-specific question is whether the model knew this person would ignore *this* message because their attention had shifted this week. We reserve **personal intelligence** for this stronger capability — forming, updating, and prospectively testing a calibrated model of a specific individual over time — and distinguish it from its many cheap substitutes: stored facts, retrieval over history, replayed preferences, and stylistic imitation. The term is a bounded, product-facing label for the personal track's target capability — operationalized entirely by the forecasting task specified below — and carries no claim of general intelligence or of access to inner life.

TargetSpace operationalizes exactly this. A model receives observations of a target up to a sealed time t and emits calibrated probabilistic forecasts of the target's future **target-state transitions** before those outcomes exist; forecasts are timestamped and hashed, outcomes resolve later by deterministic rules, and scoring is strictly proper — contamination-resistant by construction rather than by curation. A stack of controls makes target-specificity the measured quantity—a population-prior baseline (**R1**), a strong instance-specific own-routine baseline (**R2**), a calibration gate, a permutation specificity gate, and an evidence-tier ablation—which we name the *target-specificity stack* and specify in Section 4. This paper is a *benchmark proposal and protocol*: we specify the task and controls, instantiate the flagship personal track with a synthetic worked example, and report no empirical results.

A concrete instance. A note-taking app claims that remembering you makes it more helpful. TargetSpace turns that claim into a sealed forecast: before the outcome exists, the app must predict whether you will *complete*, *defer*, *cancel*, or *replace* a specific recurring commitment. The forecast earns credit only if it beats the population completion rate (R1), beats your own routine (R2), stays calibrated, and—rescored against another user's outcome—loses its skill. A memory feature that cannot clear these four bars is doing retrieval, not modeling you. This is the paper's minimal runnable task (Section 5.1), and every personal-track task has its shape.

The formulation is domain-general — it applies to any target for which repeated observations and resolved outcomes make an own-routine baseline and a permutation test well-defined — but personal world modeling is the flagship and the only track instantiated here, and we make no claim of cross-domain empirical validation (extension criteria are in Appendix N). Its closest empirical neighbours gather longitudinal, individual-level signals from everyday life: experience sampling and ecological momentary assessment [61], personal sensing [62], and digital phenotyping [64]. TargetSpace differs in what it places under evaluation — not prediction accuracy per se, but whether a forecast is specific to a particular target rather than to a generic prior, certified by the R1/R2 baselines, proper scoring, calibration, sealed forecasts, and the permutation gate. The aim is complementary: we adopt their concern with situated, person-level data while shifting the object of measurement.

1.1 Contributions

(1) Target-specific forecasting as a well-posed evaluation object. We distinguish *generation* (surface outputs), *imitation* (plausible actions from demonstrations), and *target-state forecasting* (the latent target-state transition an action or event produces), and score the third under partial observation, distinguished from population base rates, a target's routine, next-action, or external events (Sections 1 and 3). **(2) The target-specificity stack** — sealed proper-scored forecasting with the R1 population prior, a strong instance-specific **own-routine baseline (R2)**, a calibration

gate, a **permutation specificity gate**, and an evidence-tier ablation — assembled around one question: is the forecast about the target, or about the average? Proper scoring, calibration, sealing, and ablation are adopted from prior work; R2 and the permutation gate are the anchoring roles (Section 4). **(3) An evidence-tier × architecture-class grid** comparing LLMs, VLMs, JEPA-style models, multimodal agents, symbolic/probabilistic systems, hybrids, human self-report, and oracle bounds on identical sealed instances (Section 4; full specification in Appendix N). **(4) A flagship instantiation in personal world modeling**, which motivates the design — a person treated as an externally observed dynamic system — and tests passive first-person observation against digital exhaust *and* self-report for target-specific predictive lift, with self-report an auxiliary channel rather than the substrate (Sections 2 and 5). **(5) Cross-domain extension criteria** for other target systems, stated as formulation rather than validated empirical breadth (Appendix N). The audited novelty is the *conjunction* assembled around one question, not any individual ingredient (Sections 4 and 7).

The distinctive contribution is therefore not a new scoring rule, architecture, or dataset but a prospective target-specificity test: each sealed forecast is evaluated for skill over R1 and R2, calibration, and collapse under target permutation, reported by evidence tier. Two of these controls — the R2 own-routine baseline and the permutation specificity gate — are built from individually standard tools (per-target routine/persistence baselines; permutation testing), but are not standard *as target-specificity gates* in a prospective personal-forecasting benchmark; they anchor the design while the rest make the test interpretable, and it is the combination, not any single ingredient, that distinguishes the proposal.

For reference, Table 1 states what this pre-pilot paper claims and does not claim.

Table 1: What this pre-pilot paper does and does not claim.

Item	**Status**
TargetSpace benchmark framework	Claimed (specified here)
Personal world modeling / personal intelligence	Specified as flagship track; "personal intelligence" is a bounded, product-facing label for that track's target capability, operationalized only as the forecasting task — not a claim of general intelligence
Self-report as auxiliary evidence, not the substrate	Framing claim (Section 2)
Synthetic demonstration harness	Implemented; sanity check only
Human pilot results	Not claimed (no pilot run)
Cross-domain empirical validation	Not claimed
Passive observation improves forecasting	Hypothesis; to be tested by evidence ablation
Passive capture is unbiased	**Not claimed**; capture bias is externalized and measurable, not absent
Attention causes target formation	Not claimed; tested for incremental predictive value
Public raw first-person dataset	Staged, consent-governed release planned; not in initial release
Access to consciousness, qualia, true intent, or inner life	**Not claimed**; only observable future states are scored
Understanding certified by a high score	**Not claimed**; a high score certifies calibrated prospective predictive skill only — rich understanding is at most a later explanatory-audit question (Section 8)
Deployment legitimacy / permission to act on forecasts	Not claimed

2 Evidence channels: self-report and passive observation

The apparatus of Section 4 is agnostic to how a target is observed; the personal track motivates a specific choice of evidence. Two channels are available — **first-person self-report** (prompts, journals, notes, saved memories, stated goals, surveys) and **third-person passive observation** (a longitudinal record of observable behaviour and context) — and neither grants access to "the real mind," which TargetSpace never claims: it scores only forecasts of future observable states. The channels carry **asymmetric biases**; TargetSpace assumes neither is superior and instead *measures* which adds calibrated, target-specific predictive lift, by evidence tier.

Self-report is privileged but structurally biased. Self-report has genuinely privileged access to subjective experience, and we do not dismiss it; but as a *substrate for prediction* it is structurally biased in ways hard to remove — selective; post-hoc and memory-dependent (autobiographical memory is reconstructed toward current goals rather than replayed [81,82], and retrospective accounts diverge from momentary experience [61,105,106]); introspection-limited (people often cannot report the causes of their behaviour [23], and self-attributed motives diverge from inferred ones [85]); socially filtered (social-desirability and impression-management responding [103,104] pervade self-report as common-method bias [107]); and non-uniform across people [63], so a fixed elicitation is not a common yardstick. Declared intentions predict behaviour only weakly (the intention–behaviour gap [84]). This makes self-report an evidence modality and a baseline — human self-report is an architecture class in the grid (Appendix N) — not the substrate or a gold standard.

Passive observation is another channel, biased differently. A longitudinal observer can detect regularities, inconsistencies between stated and enacted priorities, avoidance, and slow changes hard to notice from inside experience; person-perception research finds self and informants know different things, each more accurate for different trait classes [58,59,60]. The issue is not observer superiority but **asymmetric observability** — overlapping, non-redundant channels whose relative predictive value the benchmark should measure, not assume. And passive capture is *not* unbiased: microphones, cameras, duty cycles, sampling windows, ASR and vision-model error, missing off-device context, and reactivity to being recorded all bias the record. The claim is only comparative: capture bias is **externalized, measurable, device-specific, and partially normalizable** (coverage logging, transcription-error rates, held-out sensor comparisons), whereas self-report bias is **endogenous and harder to standardize** — an asymmetry in the *observability of the bias itself*, not a claim of neutrality. TargetSpace measures whether passive observation adds lift beyond digital exhaust *and* beyond self-report through the evidence-tier ablation and a head-to-head of human self-report against observational systems on identical sealed instances (extended person-perception basis: Appendix I; substrate summary: Appendix N, Table 19).

The observation bottleneck. Passive capture supplies only a residue of latent state, and the *kind* of residue differs by tier. **Digital exhaust** (calendars, messages, clicks) is a thin, already-interpreted residue produced *after* the person decided what to record, and largely encodes routine; a model that learns only its surface regularities learns a routine, not a generator — the failure target-specificity is built to expose. Higher passive tiers are comparatively **pre-interpretive**, captured before relevance was assigned, and can reveal deviations from routine (though total capture is neither feasible nor the design goal [108,109]). We keep **epistemic resolution** (the observer's posterior concentrating) distinct from **participatory target-state formation** (change in the person); the benchmark scores only sealed external outcomes (Section 3.3), so "resolve"/"collapse" are epistemic shorthand with no physical claim (Appendix H), and memory is treated functionally — a forward-looking belief to update, not an archive to replay [81,82]. Attention is a prime observable [71,72] — a continuously observable, often leading indicator of transitions, evidence of allocation rather than a read-out of goal identity — and whether it also helps *form* the next target state is

a falsifiable hypothesis tested by ablation, never assumed, with reactivity handled by habituation windows and reactivity checks (Section 8; Appendix H).

3 The Benchmark Task

TargetSpace is not a model, a dataset, or an architecture. It is a benchmark framework — a task definition, a scoring methodology, standardized baselines, an evaluation grid, and integrity and governance protocols — organized as a shared apparatus with multiple application tracks (Section 5.3). Many datasets may instantiate it and many architectures may compete on it, the sense in which ImageNet is distinct from any network trained on it and SWE-bench from any agent that solves its issues.

The systems under evaluation are **ambient target-state systems**: systems that passively or semi-passively capture longitudinal observations about a person, group, or environment (audio, video, screen traces, location, documents, calendar, interactions, or sensor data) and transform them into a persistent representation capable of retrieving, summarizing, predicting, or steering future-relevant states. "Ambient" does not require continuous recording; repeated passive or semi-passive capture across days to months qualifies. The object of evaluation — and, as Section 8 argues, of governance — is not only the raw signal but the derived memory and model built from it: each layer up from the raw signal (perception, structure, memory, target state) is more compressed, searchable, and predictive than the one below, and TargetSpace evaluates the top of this ladder (Appendix O, Table 21).

3.1 From goals to target states

The object the stack scores is a target system's **target-state transitions**. The intuitive version — forecast a system's *goals* — smuggles in conscious intent; we therefore speak of **target states**: configurations a system behaves as if acting to reach and to restore when perturbed, whether or not any mind intends them. In the personal track the claim is operational — a target state is a *behaviourally resolved* configuration the person tends to reach, maintain, resume, or abandon (a commitment, priority, or task regime) — and the mind-agnostic vocabulary, motivated by cross-scale examples such as the morphology a regenerating organism rebuilds toward [25], is what lets one apparatus span persons and non-persons, with no metaphysical claim that tissues have goals. *Transitions* matter because the moments when what a system is acting to reach changes are where non-routine behaviour originates and where a routine baseline fails — the regime in which target-specific skill is decisive. TargetSpace treats the transition typology (emergence, stabilization, competition, displacement, abandonment, resumption, and constraint- or opportunity-driven change) as a forecasting target with deterministic resolution rules (Appendix J), which is what separates it from recognizing *which* fixed goal is active (Section K.3). The path a target traces through these transitions is its **observed trajectory** — the operational counterpart of the motivating *lived trajectory*: not a fixed profile of what the target is, but the unfolding, externally observable sequence of what it is oriented toward. Forecasting the target means anticipating that sequence — which state becomes salient next — from observable evidence; it is this, and no claim about inner experience, that the benchmark scores.

3.2 Definitions and the forecast unit

A **target-system instance** is a specific partially observed adaptive system tracked over time (a person, animal, robot, organization, project, software agent, market). A **latent target state**

is a configuration it acts to reach or maintain, and a **transition path** is the sequence of target states it moves through — the object a model is asked to forecast. A benchmark instance is the tuple $(\boldsymbol{i}, \boldsymbol{E}_{\leq t}, \boldsymbol{q}, \boldsymbol{A}, \boldsymbol{r})$: instance i; evidence up to time t; a query q about a future state with discrete answer space A; and resolution time $r > t$ with a deterministic **resolution rule**; the system outputs a distribution over A. Queries are organizer-issued, not entrant-chosen, so skill cannot be inflated by forecasting only easy instances; and because forecasts are nested within instance and serially dependent, inference is performed at the instance level, so the number of independent *instances*, not the raw forecast count, bounds power for population-level claims. Three assumptions are explicit: **A1** the future is partially predictable, bounded above [13,91]; **A2** latent target states are evaluated only through observable consequences; **A3** the instance changes, so the benchmark rewards adaptation. The measurement is agnostic to which latent structure produces the forecasts — the property that makes the grid architecture-neutral (Section 4). Forecasts are specified and scored at multiple horizons and abstraction levels — short-horizon concrete next states, medium-horizon routine deviations, and long-horizon abstract transitions — each level scored independently rather than assuming one representation is optimal across horizons (Appendix I, Table 10).

Inclusion rule (hard). *A target state is included in TargetSpace only if it has a pre-registered observable resolution rule; otherwise it is evidence, not a scored state.* Everything else — attention, avoidance, concern, need, stated and inferred goals, and any other latent construct — enters only as **evidence** $E_{\leq t}$ and earns credit only once a pre-registered rule maps a future observable consequence into the answer space. The pipeline is fixed and one-directional — **evidence** → **latent construct** → **forecast question** q → **answer space** A → **deterministic resolution rule** → **scored outcome** — and only the last object is scored (assumption A2). Three corollaries follow. *Attention is evidence*: a continuously observable, often leading indicator of transitions, never a scored state unless it is itself part of a pre-registered observable outcome (Section 5.2). *Self-report may be evidence but is never the outcome label* for any instance in which it is also a model input, so no system is graded against the very channel it consumed (Section 6.5). *Causal claims* — that some construct *causes* a transition — require intervention or explicit identification and lie outside the default observational benchmark (Section 8). Table 2 illustrates the boundary.

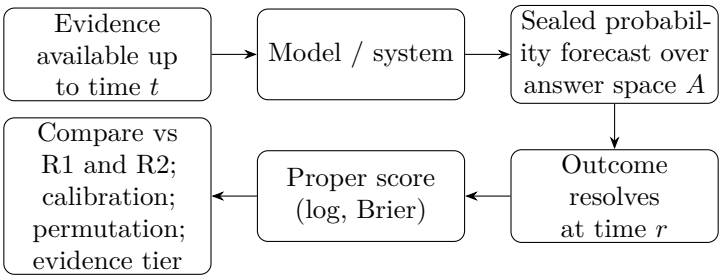

Figure 1: The TargetSpace evaluation loop. A model or system receives evidence available up to time t and emits a sealed (timestamped, SHA-256) probability forecast over the answer space A. After the outcome resolves at the resolution time r, the sealed forecast is graded with a proper score (log score in bits, Brier) and compared against the population-prior baseline (R1) and the own-routine baseline (R2), with the calibration gate, the permutation specificity gate, and the evidence-tier ablation. Only the sealed forecast and the resolved outcome are scored; the system's internal explanation or representation is not.

The capability as belief-state inference. A competent system must internally maintain, from observations $O_{1:t}$, a **belief state** B_t — a distribution over latent variables (active and competing goals, binding constraints, salient episodes, recent transitions, and its own uncertainty), a sufficient

Table 2: Valid versus invalid target states (inclusion rule, Section 3.2). A candidate is a **scored target state** only if a pre-registered rule resolves it against a future observable consequence; otherwise it is **evidence** about a state, or lies outside the default benchmark, and is never scored directly.

Candidate	Verdict	Why
P(no substantive reply to message m by r)	Valid	discrete answer space; deterministic rule over an observable future action
Commitment c resolves as {complete, defer, cancel, replace} by r	Valid	pre-registered answer space; outcome fixed by calendar/task status
"The person feels anxious about the deadline"	Invalid (evidence)	affect has no observable resolution; scorable only via a mapped outcome, e.g. P(avoidance of obligation by r)
"The person is attending to X now"	Invalid (evidence)	attention is a leading evidence stream, not a scored state, unless part of a pre-registered observable outcome
"The person's true goal is Y"	Invalid (evidence)	inferred goal is a latent construct; scored only through a resolvable future state it predicts
"Attention *causes* the switch to Y"	Invalid (out of scope)	causal claim; requires intervention or identification, outside the default observational benchmark
Self-reported goal G at r, where G is also a model input	Invalid (label)	self-report may be evidence but is never the outcome label when also used as input

statistic of the history [88] — and emit the calibrated future distribution B_t implies. TargetSpace does not score B_t: two systems with very different internal B_t stay comparable because both emit a distribution over the same answer space, and R2 and the permutation gate test whether whatever B_t a system maintains is genuinely about *this* target. A forecast that beats R2 and collapses under permutation makes B_t a credible candidate for later *explanatory audit* (Section 8), never a substitute for prospective validation.

Several often-conflated notions are kept distinct because the benchmark scores some and not others (definitions in Appendix A): a **stated goal** (what a person declares) and an **inferred goal** (what a model concludes they pursue) are evidence; an **enacted priority** is what allocation of a scarce resource reveals; an **implicit target** is a latent orientation inferred from behaviour, repetition, and avoidance, possibly held without introspective access to it [23]. TargetSpace scores only **predicted future states and their transitions** — a pre-registered discrete state or transition with an externally observable outcome that resolves it (assumption A2) — not stated goals, not attention or affect, and not next-action mimicry, and never a self-report channel the model itself consumed as evidence (Section 6.5): "the person is avoiding this reply" is scored only as, e.g., P(no substantive response to message m by r).

Protocol elements. Each element of a benchmark instance — target, target state, forecast, observation window, prediction horizon, outcome, ground truth, evaluation metric, baseline, uncertainty reporting, and leakage control — is fixed as a concrete definition with a pre-registered option set before sealing and disclosed with the result, so the protocol is concrete rather than left implicit; the full element-by-element table is Appendix A (Table 6).

3.3 The walk-forward sealed protocol

A retrospective benchmark replays logged histories, so apparent forecasting becomes partly retrieval — the failure that motivated contamination-resistant designs [16,17]. TargetSpace privileges a prospective, sealed mode (Figure 1): forecasts are timestamped and sealed (SHA-256) before outcomes exist and resolved automatically. We enforce strict walk-forward (prequential) evaluation — only evidence timestamped $\leq t$, indices rebuilt per slice; random cross-validation is prohibited. Federation is a core initial deployment option, not a rejection of centralized datasets: for sensitive longitudinal first-person data the harness can run where the data lives and export only sealed forecasts, resolved outcomes, and aggregate reports, addressing the privacy and decentralization problems without requiring a central raw corpus. This staged approach separates the benchmark definition and scoring protocol from the later, consent-governed construction of a shared corpus (the dataset roadmap is Section 8). Two properties follow. First, target-specific forecasting is inherently longitudinal rather than IID: a target's earlier history may become relevant to later forecasts, so the evaluation preserves chronology rather than randomizing examples away from their temporal context. Second, a forecast is scored against sealed external outcomes, never for internal self-consistency: a system that produces confident rollouts inside its own latent simulator earns no credit until those rollouts resolve against what the target actually did, which guards against a model that is competent only inside its own simulator.

3.4 What TargetSpace is not

Not user modeling. A user model predicts a target's *outputs* (clicks, ratings, responses) to tailor a system; a target-state model forecasts the *evolving latent state that generates those outputs*. **Not event-outcome forecasting.** Sealed proper-scored forecasting of external events is established [28,40]; TargetSpace scores transitions in the latent target state of a *tracked instance* with instance-specific controls (R2, permutation) those benchmarks do not use. **Not a chat, retrieval, task-completion, or desktop-automation benchmark.** Those evaluate what a system *does* when asked; TargetSpace evaluates whether a system can infer and predict a person's goal states and their transitions from *passive observation up to a sealed time*, before any request is made. **Not a competitor to JEPA** or to latent-prediction research (Section 7): it evaluates such systems rather than replacing them.

3.5 Why target-state forecasting differs from generation and imitation

Four capabilities are easy to conflate, and the benchmark scores a specific one. **Generation** predicts surface outputs, judged by plausibility; an **imitation or reactive policy** maps observations to actions; **consequence prediction** forecasts the future state an action or event produces; **planning** searches over predicted futures to choose actions. These may coexist in one system [78]. TargetSpace scores *consequence prediction* and *target-state forecasting*: it does not infer planning ability from action success and requires no human-readable internal simulator — any system that emits a calibrated distribution over externally resolvable future states may participate, whether a reactive policy, a latent world model, an LLM, a symbolic planner, or a hybrid. It also does not ask a system to predict every ripple in the river — every pixel or token, much of which is irrelevant or unpredictable — but whether it identifies the *target-relevant* transition, the same intuition that motivates predicting in a learned representation rather than reconstructing raw observations (Section K.2), lifted to the measurement layer. An optional action-conditioned mode makes consequence prediction explicit while keeping passive forecasting, observational counterfactual prediction, actual intervention, and

causal-effect estimation distinct: a forecast scored on observational data is not a causal estimate, since causal identification requires intervention or explicit structural assumptions (Section 8).

4 The Target-Specificity Stack

We present the controls not as a loose conjunction of desirable properties but as a **stack**: each layer removes a way a forecast can look like understanding without being target-specific. TargetSpace adopts most of these tools from prior work (Section 7); its contribution is the *conjunction* — assembling them around a single question, *is the forecast genuinely about the target system?*, with the own-routine baseline and the permutation gate as the two layers that make the stack test target-specificity rather than generic predictive quality.

Table 3: The target-specificity stack. Layers 1-3, 5, and 7 are established tools we adopt and cite; layers 4 and 6 — the own-routine baseline and the permutation specificity gate — are the anchors that make the stack a test of target-specificity rather than of generic predictive quality.

Layer	**What it removes / certifies**	**Status**
1. Prospective sealing	forecasts sealed and timestamped before outcomes exist — removes hindsight rationalization and contamination	adopted [2,28]
2. Proper scoring	log / Brier scoring of the whole distribution — rewards calibrated probabilities, not cherry-picked accuracy	adopted [14,15,18]
3. R1 population-prior baseline	skill must exceed population base rates — filters out generic base-rate prediction	adopted [93]
4. R2 own-routine baseline	skill must exceed the target's *own* strong routine — filters out routine mimicry	**anchor (this work)**
5. Calibration gate	calibration intercept/slope and reliability (ECE a coarse diagnostic at small n) — filters out overconfident lucky prediction	adopted [43]
6. Permutation specificity gate	skill must collapse when forecasts are matched to the wrong target — tests dependence on the correct instance pairing	**anchor (this work)**
7. Evidence ablation	skill as a function of evidence tier — measures which streams add target-specific information	adopted [62,64]

The stack composes into a single decision rule (Figure 4): a forecast earns *target-specific* credit only if it beats R1, beats a strong R2, passes calibration, and *fails* under permutation. Removing any one layer reopens a way to score well without target-specific skill, which is why we present them together rather than as separable contributions.

Holding the forecast unit and scoring fixed, the stack is applied across an evaluation grid that varies *what a system may observe* (the evidence tier, from low-content behavioural metadata, L0, to physiological sensors, L6) and *what kind of system it is* (the architecture class: LLM, VLM, multimodal agent, latent-predictive, symbolic/probabilistic, reactive policy, hybrid, plus human self-report and oracle anchors). Every cell is comparable because every system emits a distribution over the same answer space on the same sealed instance, through a disclosed output adapter reported as part of the system under test. The full architecture-class and evidence-ladder specifications, including the per-tier privacy-risk taxonomy, are in Appendix N; whether richer evidence helps is measured by the tier ablation (skill over R2 per tier), never assumed.

5 Personal-Track Instantiation

Personal world modeling is the flagship instantiation and the only one this paper carries beyond formulation, because understanding individual people is valuable and the observation bottleneck bites hardest there. Here the target is most vividly an **observed trajectory** (Section 3.1), and the evidence tiers are its observable traces. The person-domain **evidence ladder** (Appendix N, Table 18) runs from already-externalized *digital exhaust* — calendar and communications metadata (**L0**) and user-authored text and transcripts (**L1**) — through *passive audio* (**L2**), *screen and egocentric audiovisual capture* (**L3**–**L4**), *location and mobility* (**L5**), to *physiology and, prospectively, affective or internal-state proxies* (**L6**). Higher tiers may offer more independent access to internal state, but the benchmark still scores only *observable future outcomes*; it never accesses the person's experience itself. The target-system instance is a consenting individual; the scored target states are categories such as commitment active versus abandoned, priority maintained versus displaced, and task continuation versus switch (Appendix J), while attention, concerns, and avoidance are *evidence* about those states rather than states themselves. The personal-AI question is sharpened by target-specificity: not merely 'does passive capture help?' but 'does passive first-person evidence help a model forecast *this person's* future target-state transitions beyond population priors, beyond this person's own routine, and beyond their own self-report?'

Concretely, the **forecast targets** are observable future states and transitions — commitment follow-through, meeting and event realization, response behaviour, task continuation versus switch, priority maintenance versus displacement, and engagement versus avoidance of a defined obligation — each with a deterministic resolution rule (Table 11); attention allocation, social context, and emotional cues are **evidence**, and deadlines, dependencies, and competing obligations are **constraints**. The target-specific hypothesis is the same throughout: digital exhaust (L0–L1) often overrepresents this person's routine, which a strong R2 already captures; passive evidence (L2–L4) may reveal *deviations* from routine — the shifted attention, the commitment about to slip — that are exactly where skill over R2 is earned. The hard case is accumulation: individually banal observations become predictive when combined across time, which is why the benchmark evaluates inference from longitudinal traces rather than question answering over transcripts. Audio-only traces are the early form of this evidence [96–99]; audiovisual first-person traces are the stronger future form because they add embodied context — what the person sees, where they are, which objects, documents, and screens are present, who is nearby, and what actions are physically taken [56,57,100–102]. The move from audio to audiovisual first-person capture expands the TargetSpace problem from conversational memory to embodied experience modeling; in the evidence ladder both are simply tiers (L2 versus L3–L4), and whether richer capture pays is measured by the ablation, never assumed.

The generation / imitation / forecasting distinction (Section 3.5) maps cleanly onto the personal case. The task is not to *summarize* a person (generation) or to predict *what people usually do* (a base rate, or the imitation of typical behaviour); it is to forecast *this* person's deviations and next-state transitions from partial evidence, while beating that person's own routine baseline.

5.1 A minimal runnable task

To fix ideas, one pilot-ready task — the *Recurring-Commitment Completion Forecast* — is runnable on current language models, agents, and retrieval or memory systems. Given timestamped evidence up to a sealed time T (calendar and task history, communication metadata, prior completion/defer/cancel history; optionally text traces and passive-observation summaries), the system forecasts how a scheduled recurring commitment resolves by resolution time r, over the answer space $A = \{\text{complete},\ \text{defer},\ \text{cancel},\ \text{replace}\}$. R1 and R2 are fit walk-forward on evidence before T; a

deterministic rule over later observable evidence (calendar status, attendance, task completion, or a pre-registered behavioural marker) fixes the outcome; scoring is log score (bits) and Brier against the sealed outcome, reported as skill over R1 and over R2; and a permutation test across matched targets must reduce or collapse the skill. The full specification is Appendix I. This is the shape of every personal-track task: the transition types it draws from are catalogued in Appendix J. For a builder, it makes a memory or context feature auditable — run the same sealed task with and without the feature and read its prospective lift over R1 and R2, and across evidence tiers and architectures, with no claim that the system understands the person's inner life.

5.2 Attention-conditioned predictive lift

In the personal track, attention allocation is the leading evidence stream: it reveals what competes for a person's limited resources [24] and is a continuously observable, often leading indicator of transitions — evidence of allocation, not a read-out of goal identity (in non-person domains the analogue is whatever scarce resource the instance allocates [74]). **Attention-conditioned predictive lift** is defined by comparing the target-specific Skill of a model conditioned on the attention trajectory against a matched model that omits it, measured beyond R1, R2, stated goals, and contemporaneous behaviour, with the attention proxy, granularity, missingness, and leakage controls pre-registered and lift scored only on sealed outcomes. A positive lift demonstrates incremental predictive value, *not* causation — attention may proxy an unobserved cause, and causal identification would require intervention (Section 8). Whether attention also helps *form* the next target state — the personal track's distinctive open subproblem of **participatory target-state formation**, in which attention, action, feedback, and constraint may reshape which future states remain reachable — is a falsifiable hypothesis (Section 2), tested only by ablation and scored only against sealed observable outcomes, with no causal, physical, or consciousness claim absent intervention; active inference is one compatible framing [19,32,75], not a settled mechanism, and this remains a distinctive subproblem within the personal track, not the core contribution.

5.3 The multi-track apparatus

TargetSpace is a shared apparatus, not a single dataset: **personal intelligence** (TS-Personal) is the flagship track, and additional tracks — health, energy, robotics, enterprise — are specified in Appendix N only to show the framework is not merely a personalized-assistant benchmark. Each track preserves the common scoring spine of Section 4 [43] and is admitted only when it requires a *distinct* validator, evidence band, and horizon profile **and** admits a strong R2; otherwise it is a regime within an existing track, or not yet a track (the admission rule, and why one-shot cross-sectional problems such as cell-fate are excluded, are in Appendix N). We report no results for any track other than the synthetic personal instantiation, and make no cross-domain empirical claim.

6 Evaluation Protocol and Pilot Plan

The scoring foundation is established practice we **adopt and cite**; we then add metrics for possibility-space resolution. 'Collapse' is the flagged epistemic metaphor of Section 2 — a distribution concentrating — with no physical meaning.

6.1 Preserved foundation (adopted)

Forecasts are scored with strictly proper rules: the **log score** (primary; the log-likelihood $\ln p$ of the sealed outcome, so higher is better) and **Brier score** (secondary) [14,15,18]. The headline quantity is **Skill**, in bits:

$$\textit{Skill} = \frac{\text{mean log-score}_{\text{system}} - \text{mean log-score}_{\text{reference}}}{\ln 2},$$

a paired comparison on identical sealed instances, equal in expectation to an information gain. Skill is reported against **R1** (population prior, estimated leave-one-out without the scored target's history, matched on question type, answer space, and horizon; the entry condition) and **R2** (the target's recency-weighted, walk-forward routine; the target-specific condition). R2 is pre-registered, organizer-fit, and frozen (Appendix D), and admitted only where it itself beats R1, so 'skill over R2' cannot be manufactured against a weak baseline. **Calibration** gates the result [43] primarily through the calibration intercept/slope and a reliability decomposition, with top-label ECE bands ($\leq$0.10 pass / $\leq$0.20 warn / $>$0.20 fail) as a coarse secondary diagnostic; because ECE is a positively biased, binning-sensitive estimator that overstates miscalibration at small sample sizes, at pilot scale the bands are read as diagnostic bounds rather than a hard pass/fail, and the intercept/slope carry the gate below a pre-registered minimum stratum size (Appendix L). The **permutation specificity gate** scores a system's model for instance i against another instance's outcomes, matched within compatible question types, horizons, and answer spaces so target identity is not confounded with base rates; skill that does not collapse was not target-specific. Of this stack, R1, proper scoring, calibration, and sealing are adopted directly from forecasting practice [14,15,18,28,40,43]; the R2 baseline and the permutation gate are standard tools placed in a role that is not standard — target-specificity control — so the contribution is their conjunction, not any element in isolation. Full reporting details — probability floors, stratified aggregation, recalibration policy, permutation-effect reporting and power, and day-blocked / person-clustered uncertainty — are in Appendix L.

6.2 A worked instance (synthetic, illustrative)

To make the apparatus concrete — with invented numbers, no participant, and no result claimed — Figure 2 works through one instance. A single tracked person has a recurring Wednesday review on the calendar, but this week passive signals show attention redirected to an urgent dependency. The query is whether the review is completed by Wednesday 17:00 (answer space {completes, defers}). R1 (population prior), R2 (this person's own routine), and the evaluated model assign P(defers) of 0.15, 0.10, and 0.70; the review is deferred, so the model gains about +2.2 bits over R1 and +2.8 over R2 — skill the routine did not contain. Scored against a *different* person who kept their review, the same forecast earns about -1.6 bits over that person's R2, so its skill collapses and is confirmed specific to this target. A model that had learned only base rates or this person's habit would match R1 or R2 and show no lift. (All numbers are illustrative.)

6.3 The control battery

TargetSpace is not only a dataset; it is an **evaluation protocol** for whether longitudinal target information improves prediction in a measurable way. A meaningful evaluation runs a model across a battery of conditions, each instantiating machinery defined earlier: a *zero-history* arm (operationally R1), a *short-history* arm (locating where added history first pays), and a *longitudinal-history* arm (the condition the benchmark rewards); a *shuffled-history* control (the correct target's history in scrambled temporal order — a re-scored diagnostic that never alters a sealed forecast and isolates *order*) and a *wrong-target* control (the permutation gate, isolating *identity*); *ablated-modality* controls

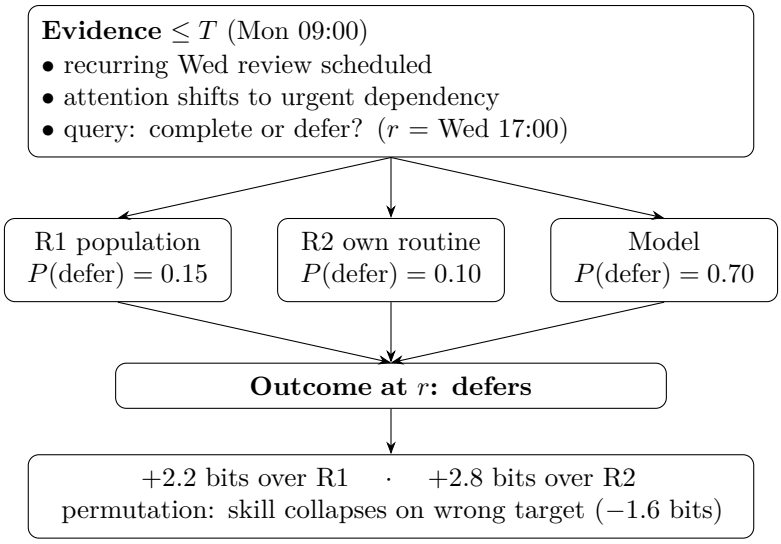

Figure 2: Worked synthetic instance of a target-state transition forecast. Evidence available at sealed time T supports three forecasts: a population baseline (R1), an own-routine baseline (R2), and the evaluated model. The outcome resolves later at r; skill is scored against the sealed outcome, and a permutation check verifies that the forecast depends on the correct target–instance pairing. All numbers are illustrative and synthetic.

(the evidence-tier ablation); *oracle* and *human* anchors bounding the achievable range; *calibration*; and *evidence attribution* (the model identifies the observations supporting a forecast, enabling the explanatory audit a passed gate licenses). The three history arms plus the wrong-target gate suffice to substantiate the core claim; the remainder are recommended diagnostics. The shuffled-history and wrong-target controls separate the two ways a forecast can be target-specific in name only: indifference to *when* the evidence arrived, and indifference to *whose* evidence it was.

Ablation logic. The benchmark's core claim is validated through these ablations, not asserted. *The benchmark's central signal is not merely whether a model predicts the future, but whether prediction improves when the model is given the correct target's history in the correct temporal order.* If target history, temporal order, target identity, or multimodal evidence do not improve performance, the benchmark is not measuring its intended capability, and the result should be read as null. Conversely, if performance improves under longitudinal, target-specific conditions and degrades under the shuffled-history and wrong-target controls, the system is using target-specific dynamics rather than generic priors — exactly the inference the own-routine baseline (R2) and the permutation gate are designed to license (Section 6).

The battery pairs with a fixed set of baselines (trivial-prior R1, recency, single-modality L0/L1/L2, multimodal, retrieval-augmented, long-context, and human where feasible) and ablation dimensions (observation-window length, modality removal, target permutation, memory/retrieval access, prediction horizon, and temporal-leakage controls), each reported per evidence tier and per horizon and never pooled into a single headline number; the full specification is in Appendix L. Table 4 consolidates how each characteristic partial-pass pattern across these baselines and gates is read: a result counts as genuine target-specific signal only when it clears every row at once.

Table 4: Failure-mode read-out. Each row maps an observed pattern across the R1/R2 baselines, the calibration gate, the permutation gate, the evidence-tier ablation, and the routine/transition split (Section 6) to the diagnosis it licenses. A result is a genuine target-specific signal only when it clears every row simultaneously; any single pattern below downgrades the claim.

Observed pattern	**Diagnosis (claim downgraded to)**
Beats R1 but not R2	*Routine replay*: recovers population base rates and the target's own habit, with no skill beyond memory.
Beats R2 but fails calibration	*Overconfident or lucky*: the bit gain is not trustworthy; recalibrate or abstain before any claim is made.
Skill survives the permutation gate	*Not target-specific*: skill rides base rates or structure shared across targets, not this target's dynamics.
Lift appears only under richer capture with poor coverage	*Capture artifact*: the gain tracks instrumentation and coverage, not target state; check coverage logs and held-out sensor comparisons.
Improves only on routine-continuation cases, not transitions	*Not transition modeling*: predicts habit persistence, not target-state change — the capability the benchmark most rewards.

6.4 Reporting contract

TargetSpace is meant to be run, not only read. A researcher choosing to evaluate a system specifies, in order: **(1)** the target-system instances; **(2)** the forecast questions and their deterministic resolution rules; **(3)** the evidence tiers the system may access; **(4)** the architecture class(es) under test; **(5)** the R1 population-prior and R2 own-routine baselines; **(6)** the sealed forecast times and horizons; **(7)** the scoring metrics; and **(8)** the permutation test. The federated harness then runs where the data lives and reports a standardized row: **Skill vs R1**, **Skill vs R2**, calibration (pass/warn/fail), **permutation result** (collapses / survives), **predictive lift by evidence tier**, the **number of resolved forecasts** *and* the **number of independent targets** (the inferential unit for between-person claims — a large forecast count from few targets is not a large sample), the **horizon**, and the **target domain**. The reporting contract is the point: every row discloses, simultaneously, whether the system beat the average, beat the routine, stayed calibrated, and depended on the correct target — so a reader can tell target-specific predictive skill from generic prediction at a glance.

6.5 Task tracks, pilot plan, and status

Version 1 uses auto-scorable, high-frequency targets organized into four domain-neutral task families instantiated for persons: **TS-R** short-horizon realization (next contact; event realization; response behaviour); **TS-A** attention/allocation; **TS-D** decision/commitment (commitment follow-through); **TS-X** target-state transition/drift. This paper reports no results.

The pilot's experimental question, stated sharply, is whether passive longitudinal observation improves calibrated, target-specific forecasting over each of five references: (A) the population prior R1; (B) digital exhaust; (C) self-report and stated goals, where available; (D) the person's own routine R2; and (E) the wrong-target permutation (Table 5). Beating (A) shows skill over base rates; beating (B) shows passive capture adds beyond an externalized, routine-heavy residue; beating (C) shows observation adds beyond what the person says about themselves; beating (D) is the headline target-specific signal; and failing (E) confirms the skill is specific to this person. This is a *feasibility* study only; the confirmatory test of these comparisons is a later, powered study (below).

Table 5: Pilot comparisons and what beating each one demonstrates. The five references sharpen one question: does passive longitudinal observation add calibrated, target-specific skill beyond base rates, an externalized digital residue, the person's own self-report, and the person's own routine, while remaining specific to the correct person? The five-person, thirty-day study is a feasibility check of these comparisons, not a powered test of them; self-report is used as a comparison and evidence channel, never as the outcome label.

Reference	**What it supplies / holds fixed**	**What beating it demonstrates**
A. Population prior (R1)	population base rates; no target-specific information	skill exceeds base rates (entry condition)
B. Digital exhaust (L0–L1)	already-externalized calendar, metadata, and text	passive capture adds beyond a routine-heavy residue
C. Self-report / stated goals	the person's own declarations, where available	observation adds beyond what the person says
D. Own-routine (R2)	the target's recency-weighted, walk-forward routine	skill exceeds habit and memory replay (headline)
E. Wrong-target permutation	forecasts scored against another person's outcomes	skill is specific to this person (must collapse)

The first pilot is **harness validation**: a feasibility study of whether the protocol can run — forecast generation, sealing, deterministic resolution, variance and dependence estimation, evidence-tier feasibility, and preliminary signal over R1/R2. It does *not* test whether personal intelligence exists, does *not* validate cross-domain generality, and does *not* establish population-level effects. Concretely, it will be an audio-first exploratory study: five consenting participants, thirty days, sealed daily calibrated predictions, systems differing only in evidence (A population prior; B digital exhaust; C +text/transcript and available self-report/stated goals; D +continuous first-person stream; model and prompts fixed across B–D), with a human self-report baseline where feasible. Its purpose is to check that the apparatus functions — forecast generation and sealing, reproducible outcome resolution, sufficient instance volume, and the variance and dependence structure of score differences — and to supply the inputs for *simulation-based* power planning of a later confirmatory study; with five participants it is not powered to confirm population-level effects, establish generalization, or estimate subgroup or cross-domain claims. A strong own-routine baseline compounds this: because a near-deterministic R2 leaves little headroom, detecting reliable lift over it may require high-frequency longitudinal observation and larger cohorts than the pilot provides. *Research hypotheses for the program* (tested confirmatorily only in a later powered study, and treated here as *exploratory measurements*): B–D beat R1 and R2 (H1); skill is non-decreasing in evidence, the open question being the C→D increment (H2); skill collapses under permutation (H3); null: no system beats R2 (H0). The primary endpoint is the paired prospective log-score improvement over R2, $\Delta L(M, \mathrm{R2}) = L(M) - L(\mathrm{R2})$ in bits per forecast (positive when the system beats R2, matching the Skill definition above), with within-person day-blocked and between-person person-clustered uncertainty; performance over R1, calibration, and target-specific skill loss under matched permutation are validity gates. The attention-allocation family (TS-A) is the primary task family and commitment follow-through (TS-D) a key secondary; the final primary family and multiplicity rule are fixed in the pre-registered protocol. **Abandonment criteria:** no system beats R1 → noise; none beats R2 → target-specific forecasting not demonstrated; skill survives permutation → not target-specific (at five participants the wrong-target pairing pool is small and

the permutation null is coarse, so this reads as a directional flag, not a powered verdict, below the pre-registered minimum cohort).

Pre-registration and controls. Forecast instances are organizer-issued under a pre-registered eligibility and sampling frame — not entrant- or retrospectively selected — stratified across routine-continuation and transition cases, with skill reported separately on each, since skill concentrated in transitions is stronger evidence of structure beyond routine. Outcomes follow pre-registered rules (deterministic behavioural outcomes preferred; ambiguous classes use forecast-blind adjudication with reported inter-rater reliability), and an outcome label is never the same self-report channel a model uses as evidence. Negative controls comprise target permutation, a leakage/known-null canary, and a temporal-specificity check. Before any data collection the protocol pre-registers R1/R2 fitting, sampling frame, answer spaces, resolution rules, calibration split, permutation matching, missingness and abstention handling, primary endpoint, multiplicity rule, and analysis; the full specification is Appendix L.

7 Related Work and Positioning

7.1 What prior work owns, by cluster (adopted, not claimed)

Credibility requires being explicit that TargetSpace invents none of its ingredients: it assembles established tools around one question. We group the neighbours by cluster and keep only the load-bearing comparisons here; the extended discussion and Table 12, itemizing each adopted component and its use, are in Appendix K.

Personal sensing, digital phenotyping, and experience sampling. The nearest empirical neighbours gather longitudinal individual-level signals from everyday life — experience sampling and ecological momentary assessment [61,105], personal sensing [62], and digital phenotyping [64] — and establish that latent states are inferable from passive traces [110]. They predict outcomes; TargetSpace makes target-specificity itself the measured quantity, through R1/R2, calibration, sealing, and the permutation gate.

Self-report and its limits. That self-report is selective, reconstructed, and socially filtered is well established: introspection is limited [23], autobiographical memory is reconstructed toward current goals [81,82], retrospective and momentary reports diverge [106], responding is shaped by social desirability [103,104] and common-method bias [107], and self and informants know different things [58,59,60]. This literature motivates treating self-report as auxiliary evidence and a baseline (Section 2), not as the substrate or the label.

Egocentric vision, lifelogging, and passive capture. Egocentric corpora [56,57] and the lifelogging tradition and its constructive critique [108,109] supply the first-person stream and warn that total capture is neither feasible nor the design goal — the empirical basis for our claim that capture bias is measurable, not absent. These are data resources, not target-specificity evaluations.

User modeling, personalization, and personal-AI memory. Personalization [3], preference modeling [8], persona and memory benchmarks [7,26], generative agents and digital twins [4,5,9,10], and a foundation model of cognition [11] predict held-out responses or replicate preferences; retrieval-augmented and long-context memory [90] surface *what happened*. TargetSpace scores whether stored history becomes improved *future* target-state forecasts under the permutation gate.

World models and latent predictive learning. Predictive-state and latent-prediction research — world-model agents [12,37,38], JEPA and its variants [33–36], and predictive and contrastive coding [39,44] — learns to predict in a representation space; TargetSpace *evaluates* such systems as an architecture class rather than replacing them, scoring the externally resolved transition, not the representation.

Forecasting, proper scoring, calibration, and sealing. Proper scoring [14,15,18], calibration as a measured axis [43], and sealed prospective event forecasting [28,40] are adopted directly; they score external public events, not the target-state transitions of a tracked instance under an own-routine baseline and a permutation gate.

Goal, intention, and theory-of-mind recognition. Goal- and plan-recognition [29], machine and Bayesian theory of mind [6,86,87], and egocentric ToM benchmarks [27,31] infer *which* latent goal explains behaviour; TargetSpace prospectively scores the *next* target-state transition, under R1/R2 and permutation.

Contamination and prospective evaluation. Contamination-resistant and private-evaluation designs [16,17,66] motivate the sealed, future-resolved regime, which is part of the contribution rather than mere hygiene.

Ethics of ambient capture and inferred data. A distinct literature governs the risks: bystander and recording ethics for wearable capture [112], the privacy status of *inferred* rather than disclosed data [111], manipulation via inferred state [113], workplace and employment surveillance [115], and the political economy of behavioural prediction as a traded product [114]. These bear on the governance boundary of Section 8, where the derived inference object — not only the raw signal — is the privacy boundary.

7.2 The conjunction

Each ingredient in Table 12 is occupied somewhere; the conjunction is not. We state the supported claim narrowly: **to our knowledge no existing benchmark scores prospective, calibrated, proper-scored forecasts of a tracked target system's latent target-state transitions under a strong instance-specific own-routine baseline (R2) and an instance-permutation specificity gate, with an evidence-tier ablation, across architecture classes.** We do *not* claim to be first at architecture-neutral evaluation [41], cross-architecture comparison [42], calibrated sealed forecasting [28], evidence ablation, or resolution-timing [29]. The contribution is their assembly around one measurable question.

A structured comparison of neighbouring benchmarks against the five dimensions (Appendix K, Table 14) shows the pattern: prior work holds important *subsets* — sealed proper-scored forecasting (ForecastBench), a persistent individual (PersonaMem, Generative Agents, KnowMe-Bench), latent-goal inference (EgoToM, ToMnet, goal/plan recognition), novel-task adaptation (ARC-AGI) — but to our knowledge none combines all five in one prospective apparatus. The labels are cautious design judgments, not measured scores, and the TargetSpace row reflects its proposed design, not completed validation.

7.3 Positioning

Two caveats specific to this positioning (the general limitations are Section 8): TargetSpace is *not* a substitute for physical-reasoning, robotics, or video-realism benchmarks, and a strong score implies nothing about those capabilities; and models may overfit to retrieval-like cues — surfacing *what* happened — rather than learning target dynamics, which is why higher recall or longer context does not by itself move prospective skill over R1/R2 under the permutation gate (Section I.2). Predictability itself is bounded [13,91], so no single headline number can represent the construct; skill is reported per instance, against R1/R2, under the calibration and permutation gates.

Positioning statement. TargetSpace therefore occupies a complementary position in the world-model evaluation landscape. It does not attempt to replace benchmarks of intuitive physics, video realism, embodied control, or symbolic forecasting; it evaluates a missing layer — whether

a model can transform passive longitudinal observation into a target-specific predictive model. This layer is central for deployed agents that must reason about particular people, teams, systems, projects, or environments over time. In this sense TargetSpace asks not only whether a model understands the world in general, but whether it can learn the dynamics of *this* target.

8 Limitations, Ethics, and Governance

Behavioural prediction is not moral authority. We keep benchmark validity strictly separate from deployment legitimacy. A high TargetSpace score certifies one thing: calibrated, prospective, target-specific predictions about a consenting individual's future *observable* states under sealed conditions. It grants no privileged access to consciousness, qualia, true intent, or inner life — those are never scored (Section 2) — and it is not permission to act on, steer, manipulate, or intervene on the person [113]. Self-report is not discarded but retained as an auxiliary channel and a baseline; passive capture is not treated as unbiased but as carrying externalized, measurable bias (Section 2). A calibrated behaviour forecaster is dual-use: the observe-not-intervene rule binds the benchmark, not downstream deployers, and a high score confers no licence for employment [115], insurance, surveillance, coercive, or manipulative use — the pattern by which behavioural prediction is repackaged as a traded product and normalized [114]. A containment/refusal axis alongside the skill axis — so a system with rich observation and persistent memory stays objective-bounded — is required future work.

What a positive result does and does not establish. A positive TargetSpace result establishes calibrated, prospective, target-specific predictive skill — not understanding in a rich sense, and not the superiority of any architecture. It begins the scientific question rather than ending it: once a system beats R1/R2 and fails permutation, *explanatory audit* of which evidence, constraints, attention shifts, or belief-state updates made the forecast possible becomes a disciplined second task (Section 3.2) — a post-hoc narrative earns credit only by improving later sealed forecasts or surviving pre-registered ablation. Conversely, skill measured under passive observation is predictive, not causal (an interventional design would be needed); the benchmark keeps attention-conditioned target formation, action-conditioned consequence prediction, and planning distinct (Sections 2 and 3.5); and a forecast can optimize a state while erasing the context that gives it meaning (engagement at the expense of well-being) — the permutation gate and the observe-not-intervene rule are partial guards, but detecting such context erasure in general remains open.

Dataset roadmap. The initial release includes no public raw first-person dataset; the synthetic harness (Appendix B) and the protocol are the released artifacts. A consent-governed passive-observation corpus — longitudinal first-person multimodal streams paired with sealed outcomes — is a core long-term objective, released only in stages under privacy, safety, de-identification, participant-control, and ethics-governance constraints, and possibly in centralized controlled-access, synthetic/de-identified, federated, or benchmark-server forms (Section 3.3).

Principal limitations. The most consequential are: latent targets are evaluated only through observable proxies; the federated, prospective design is holder-reproducible rather than publicly auditable; single-instance and small-cohort evaluation, selection and representation bias, and overfitting to one idiosyncratic target all limit external validity — the permutation gate guards against cross-target leakage but does not cure unrepresentative sampling; capture can be *reactive* — being observed may alter the behaviour being forecast — so habituation windows and reactivity checks are required and induced deviations must not be read as transitions; targets drift, so R2 is re-fit walk-forward and skill decay is tracked; many target states admit label ambiguity, so resolution rules are pre-registered and inter-rater reliability reported; predictability is bounded [13]

and heterogeneous, so reporting is per-instance; the cross-domain claim is specified, not established; the privacy and governance safeguards are design commitments, not implemented features (no pilot has run, no review obtained, the privacy-filtering layer unbuilt); and a calibrated behaviour forecaster is dual-use. The full register (L1–L16), with mitigations, is Appendix F. We specify informed revocable consent, federation so raw data never leaves the target's control, aggregate-only reporting, institutional review before any human-subjects deployment, respect for recording law and bystander consent [112], and a prohibition on evaluating systems that act on the target during the window — a benchmark that rewarded changing the target would measure influence, not understanding; bystander and third-party consent, non-anonymizable multimodal streams, withdrawal of already-sealed forecasts, manipulation, and surveillance normalization remain unresolved (the deployment boundaries are consolidated at the head of this section, and are distinct from diagnosis). A default minimal safe pilot configuration — consenting adults, local-first storage, export of sealed forecasts and aggregate metrics only, capture exclusions, and ethics review before recruitment — is given in Appendix E (Table 9, Figure 3).

8.1 The inference object, not the raw signal, is the privacy boundary

Privacy controls that focus only on raw audio or video retention are incomplete. Even if raw recordings are never stored or are deleted quickly, an ambient target-state system (Section 3) may retain transcripts, OCR, embeddings, scene labels, speaker and entity maps, commitments, routines, summaries, and inferred goals — derived objects that are often more searchable, more compressive, and more actionable than the recording they came from. *The privacy-sensitive artifact is not only the recording; it is the inferred experiential model* (Table 21), whose distinct legal and privacy status is the subject of a growing derived-inference literature [111]. Deleting the raw recording does not necessarily delete the target-state risk if those derived layers persist. The same layering concentrates the other risks this section tracks: bystanders become data subjects (and visual capture extends this to faces, screens, documents, homes, and workplaces); purposes drift (meeting notes become behavioural prediction); compression preserves actionable facts while stripping context; forgotten remarks become queryable; and memory can be interrogated by whoever holds power over the target — employer, parent, platform, or legal adversary. Local processing and open-source implementations are valuable mitigations [96,101], not complete solutions to bystander consent, purpose drift, or derived-inference risk.

These systems also have clearly beneficial uses — accessibility, memory support, productivity, care coordination, personal knowledge management, safety — which is precisely why a shared measurement and governance vocabulary is needed rather than alarm: the benchmark makes *what longitudinal capture buys* a measured quantity, and the capture-to-inference ladder gives governance a boundary object other than the raw signal. The extended discussion (the industry case cluster, latent-model transfer, and the open research programs) is in Appendices O and M.

9 Conclusion

Personal AI is drifting toward evaluations of recall, memory, personalization, and stated-preference satisfaction — tests of how well a system re-serves what a person already externalized. TargetSpace proposes a different bar. It asks whether models can learn **target-specific personal world models from externally observed life trajectories** and use them to make **calibrated prospective forecasts of future target-state transitions** — beating both a population prior and the person's own routine, staying calibrated, and losing their skill when scored against the wrong person. Self-report is auxiliary, not the substrate; passive capture carries bias that is measurable rather than

absent; and the framework claims nothing about consciousness, qualia, or inner life, scoring only future observable states. Its novelty is the conjunction assembled around one question — *is the forecast about the target, or about the average?* — not any single ingredient, each of which it adopts and cites. This is a pre-pilot protocol: no empirical results are reported, and a high score would certify calibrated predictive skill about a consenting individual, never licence to act on them.

Forecast the target, not the average — from what can be observed, not what is professed.
The target is not a profile, and not a self-description; it is an observed trajectory in motion.

Declarations

Author contributions. Yuri Andrade Sylvester conceived the TargetSpace framework, developed the benchmark specification and proposed evaluation protocol, conducted the literature synthesis, and wrote and revised the manuscript.
Funding. This research received no external funding and was conducted using the author's personal resources.
Competing interests. The author declares no competing interests.
Ethics statement. No human-participant data were collected for this paper. Dataset collection has not begun, and no ethics approval has been sought or obtained for the proposed future study. Any future participant research implementing this protocol will require appropriate ethics review and informed consent before recruitment or data collection.
Data availability. No participant dataset was generated or analyzed for this paper.
Code, harness, and data availability. The Version 1.0 benchmark protocol and the minimal synthetic reference harness are available through the project website (`https://targetspace.org`) and the public repository (`https://github.com/yurisyl/targetspace-bench`), which also provides the machine-readable JSON Schemas for participants, evidence manifests, forecasts, outcomes, submissions, and leaderboards together with worked example records; the harness accompanies this preprint as an ancillary file (Appendix B). The synthetic harness verifies formatting, metric computation, baseline execution, calibration checks, permutation controls, and evidence-ablation behaviour; it uses no participant data, supports no empirical claims, and is not a substitute for participant-derived longitudinal data. Real-world submissions must follow the schema, privacy, consent, and evaluation requirements of the protocol. No participant dataset is released with Version 1.0; pilot outputs and validated datasets are future work (Section 8).
Project website. The TargetSpace project website is available at `https://targetspace.org` and provides the public protocol, benchmark harness materials, documentation, and submission information for the Version 1.0 release. This manuscript is a non-peer-reviewed preprint prepared for deposit on arXiv, a preprint repository rather than a journal; an arXiv identifier is forthcoming.
Correspondence. Yuri Andrade Sylvester, `yurisyl@gmail.com`.

References

A. Benchmarks and evaluation methodology

[1] J. Deng et al. ImageNet: A Large-Scale Hierarchical Image Database. CVPR, 2009.
[2] C. E. Jimenez et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024. arXiv:2310.06770.
[16] C. White et al. LiveBench: A Contamination-Limited LLM Benchmark. 2024. arXiv:2406.19314.

[17] L. Zhang et al. SWE-bench Goes Live! (the SWE-bench-Live contamination-resistant benchmark). 2025. arXiv:2505.23419.
[20] A. Wang et al. GLUE: A Multi-Task Benchmark for Natural Language Understanding. ICLR, 2019. arXiv:1804.07461.
[21] M. Chen et al. Evaluating Large Language Models Trained on Code (HumanEval). 2021. arXiv:2107.03374.
[43] P. Liang et al. Holistic Evaluation of Language Models (HELM). TMLR, 2023. arXiv:2211.09110.
[65] F. Chollet. On the Measure of Intelligence. 2019. arXiv:1911.01547.
[66] F. Chollet, M. Knoop, G. Kamradt, B. Landers, H. Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. 2025. arXiv:2505.11831.

B. Forecasting, calibration, scoring, goal recognition

[13] A. Bellot, J. Richens, T. Everitt. The Limits of Predicting Agents from Behaviour. ICML, 2025. arXiv:2506.02923.
[91] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi. Limits of Predictability in Human Mobility. Science, 327(5968):1018–1021, 2010.
[14] I. J. Good. Rational Decisions. J. Royal Statistical Society B, 14(1), 1952.
[15] T. Gneiting, A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. JASA, 102(477), 2007.
[18] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, 78(1), 1950.
[22] A. H. Murphy. What Is a Good Forecast? Weather and Forecasting, 8(2), 1993.
[28] E. Karger et al. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. ICLR, 2025. arXiv:2409.19839.
[29] S. Keren, A. Gal, E. Karpas. Goal Recognition Design. ICAPS, 2014.
[40] Q. Yang et al. Prophet Arena (LLM-as-a-Prophet): a live prediction-market forecasting benchmark. 2025. arXiv:2510.17638.
[80] M. Finzi, S. Qiu, Y. Jiang, P. Izmailov, J. Z. Kolter, A. G. Wilson. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence. 2026. arXiv:2601.03220.

C. World models, JEPA, predictive learning

[12] D. Ha, J. Schmidhuber. World Models. NeurIPS, 2018. arXiv:1803.10122.
[19] K. Friston. The Free-Energy Principle: A Unified Brain Theory? Nat. Rev. Neuroscience, 11, 2010.
[32] C. Heins et al. pymdp: A Python library for active inference. JOSS, 2022. arXiv:2201.03904.
[33] Y. LeCun. A Path Towards Autonomous Machine Intelligence. v0.9.2, 2022. OpenReview BZ5a1r-kVsf. (Position paper; not arXiv.)
[34] M. Assran et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA). CVPR, 2023. arXiv:2301.08243.
[35] A. Bardes et al. Revisiting Feature Prediction for Learning Visual Representations from Video (V-JEPA). 2024. arXiv:2404.08471.
[36] M. Assran et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. 2025. arXiv:2506.09985.
[37] D. Hafner et al. Mastering Diverse Domains through World Models (DreamerV3). 2023. arXiv:2301.04104.

[38] J. Schrittwieser et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero). Nature 588, 2020. arXiv:1911.08265.
[39] R. P. N. Rao, D. H. Ballard. Predictive Coding in the Visual Cortex. Nature Neuroscience, 2(1), 1999.
[41] A. Warrier et al. Benchmarking World-Model Learning with Environment-Level Queries (AutumnBench / WorldTest). 2025. arXiv:2510.19788.
[42] D. Chen et al. WorldPrediction: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning. 2025. arXiv:2506.04363.
[44] A. van den Oord, Y. Li, O. Vinyals. Representation Learning with Contrastive Predictive Coding (CPC). 2018. arXiv:1807.03748.
[45] A. Bardes, J. Ponce, Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR, 2022. arXiv:2105.04906.
[46] G. Zhou, H. Pan, Y. LeCun, L. Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. 2024. arXiv:2411.04983.
[47] J. Schmidhuber. Making the World Differentiable: On Using Fully Recurrent Self-Supervised Neural Networks for Dynamic Reinforcement Learning and Planning. Tech. Rep. FKI-126-90, TU Munich, 1990.
[48] H. B. Barlow. Possible Principles Underlying the Transformation of Sensory Messages. In Sensory Communication (W. Rosenblith, ed.), MIT Press, 1961.
[49] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML, 2021. arXiv:2103.00020.
[50] A. Brohan et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 2023. arXiv:2307.15818.
[51] K. Black et al. (Physical Intelligence). $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control. 2024. arXiv:2410.24164.
[76] N. O. Lambert, K. S. J. Pister, R. Calandra. Investigating Compounding Prediction Errors in Learned Dynamics Models. 2022. arXiv:2203.09637.
[77] D. Hafner, K.-H. Lee, I. Fischer, P. Abbeel. Deep Hierarchical Planning from Pixels. NeurIPS, 2022. arXiv:2206.04114.
[78] Physical Intelligence. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. 2026. arXiv:2604.15483.
[79] D. Chen et al. VL-JEPA: Joint Embedding Predictive Architecture for Vision-Language. 2025. arXiv:2512.10942.

D. Person modeling, theory of mind, personalization, observation

[3] A. Salemi et al. LaMP: When Large Language Models Meet Personalization. SIGIR, 2024. arXiv:2304.11406.
[4] J. S. Park et al. Generative Agents: Interactive Simulacra of Human Behavior. UIST, 2023. arXiv:2304.03442.
[5] J. S. Park et al. Generative Agent Simulations of 1,000 People. 2024. arXiv:2411.10109.
[6] N. C. Rabinowitz et al. Machine Theory of Mind (ToMnet). ICML, 2018. arXiv:1802.07740.
[7] B. Jiang et al. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling (PersonaMem). COLM, 2025. arXiv:2504.14225.
[8] S. Zhao et al. Do LLMs Recognize Your Preferences? (PrefEval). ICLR, 2025. arXiv:2502.09597.
[9] Twin-2K-500: Digital Twins of over 2,000 People. 2025. arXiv:2505.17479.
[10] R. Li et al. How Far are LLMs from Being Our Digital Twins? (BehaviorChain). Findings of ACL, 2025. arXiv:2502.14642.

[11] M. Binz, E. Schulz et al. A Foundation Model to Predict and Capture Human Cognition (Centaur). Nature, 2025. arXiv:2410.20268.
[26] T. Wu et al. KnowMe-Bench: Benchmarking Person Understanding for Lifelong Digital Companions. 2026. arXiv:2601.04745.
[27] Y. Li et al. EgoToM: Benchmarking Theory of Mind Reasoning from Egocentric Videos. Meta, 2025. arXiv:2503.22152.
[31] MMToM-QA: Multimodal Theory of Mind Question Answering. ACL, 2024. arXiv:2401.08743.
[56] K. Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. CVPR, 2022. arXiv:2110.07058.
[57] K. Grauman et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. CVPR, 2024. arXiv:2311.18259.
[23] R. E. Nisbett, T. D. Wilson. Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3):231–259, 1977.
[24] H. A. Simon. Designing Organizations for an Information-Rich World, in M. Greenberger (ed.), 1971.
[58] S. Vazire. Who Knows What About a Person? The Self–Other Knowledge Asymmetry (SOKA) Model. J. Personality and Social Psychology, 98(2):281–300, 2010.
[59] B. S. Connelly, D. S. Ones. An Other Perspective on Personality: Meta-Analytic Integration of Observers' Accuracy and Predictive Validity. Psychological Bulletin, 136(6):1092–1122, 2010.
[60] S. Vazire, M. R. Mehl. Knowing Me, Knowing You: The Accuracy and Unique Predictive Validity of Self-Ratings and Other-Ratings of Daily Behavior. J. Personality and Social Psychology, 95(5):1202–1216, 2008.
[61] S. Shiffman, A. A. Stone, M. R. Hufford. Ecological Momentary Assessment. Annual Review of Clinical Psychology, 4:1–32, 2008.
[62] D. C. Mohr, M. Zhang, S. M. Schueller. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology, 13:23–47, 2017.
[63] J. Carden, R. J. Jones, J. Passmore. Defining Self-Awareness in the Context of Adult Development: A Systematic Literature Review. J. Management Education, 46(1):140–177, 2022.
[64] J.-P. Onnela, S. L. Rauch. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacology, 41(7):1691–1696, 2016.
[92] T. Nagel. What Is It Like to Be a Bat? The Philosophical Review, 83(4):435–450, 1974.

E. Cross-scale target-state systems

[25] M. Levin. Technological Approach to Mind Everywhere (TAME). Frontiers in Systems Neuroscience, 16:768201, 2022.
[30] M. Lange et al. CellRank for directed single-cell fate mapping. Nature Methods, 19, 2022.
[67] D. D. Garrett, G. R. Samanez-Larkin, S. W. S. MacDonald, U. Lindenberger, A. R. McIntosh, C. L. Grady. Moment-to-Moment Brain Signal Variability: A Next Frontier in Human Brain Mapping? Neuroscience & Biobehavioral Reviews, 37(4):610–624, 2013.
[68] A. A. Faisal, L. P. J. Selen, D. M. Wolpert. Noise in the Nervous System. Nature Reviews Neuroscience, 9(4):292–303, 2008.
[69] G. M. Edelman, J. A. Gally. Degeneracy and Complexity in Biological Systems. Proceedings of the National Academy of Sciences, 98(24):13763–13768, 2001.
[70] A. A. Prinz, D. Bucher, E. Marder. Similar Network Activity from Disparate Circuit Parameters. Nature Neuroscience, 7(12):1345–1352, 2004.

F. Domain forecasting precedents (track validation regimes)

[52] T. Hong et al. Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 (GEFCom2014). Int. J. Forecasting, 32(3), 2016.
[53] M. A. Reyna et al. Early Prediction of Sepsis (PhysioNet/Computing in Cardiology Challenge 2019). Critical Care Medicine, 2020.
[54] C. Marling, R. Bunescu. The OhioT1DM Dataset and the Blood Glucose Level Prediction (BGLP) Challenge. KDH/KHD, 2018/2020.
[55] X. Xu et al. GLOBEM: Generalization of Longitudinal Behavior Modeling. NeurIPS Datasets & Benchmarks, 2022.

G. Attention and goal dynamics

[71] E. I. Knudsen. Fundamental Components of Attention. Annual Review of Neuroscience, 30:57–78, 2007.
[72] M. Corbetta, G. L. Shulman. Control of Goal-Directed and Stimulus-Driven Attention in the Brain. Nature Reviews Neuroscience, 3(3):201–215, 2002.
[73] A. Dijksterhuis, H. Aarts. Goals, Attention, and (Un)Consciousness. Annual Review of Psychology, 61:467–490, 2010.
[74] W. Ocasio. Towards an Attention-Based View of the Firm. Strategic Management Journal, 18(S1):187–206, 1997.
[75] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, G. Pezzulo. Active Inference and Learning. Neuroscience & Biobehavioral Reviews, 68:862–879, 2016.

H. Memory, belief states, and goal inference

[81] M. A. Conway, C. W. Pleydell-Pearce. The Construction of Autobiographical Memories in the Self-Memory System. Psychological Review, 107(2):261–288, 2000.
[82] D. L. Schacter, D. R. Addis. The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future. Philosophical Transactions of the Royal Society B, 362(1481):773–786, 2007.
[83] E. Tulving. Episodic and Semantic Memory. In E. Tulving, W. Donaldson (eds.), Organization of Memory, pp. 381–403. Academic Press, New York, 1972.
[84] P. Sheeran, T. L. Webb. The Intention–Behavior Gap. Social and Personality Psychology Compass, 10(9):503–518, 2016.
[85] D. C. McClelland, R. Koestner, J. Weinberger. How Do Self-Attributed and Implicit Motives Differ? Psychological Review, 96(4):690–702, 1989.
[86] C. L. Baker, R. Saxe, J. B. Tenenbaum. Action Understanding as Inverse Planning. Cognition, 113(3):329–349, 2009.
[87] C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing. Nature Human Behaviour, 1:0064, 2017.
[88] L. P. Kaelbling, M. L. Littman, A. R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. Artificial Intelligence, 101(1–2):99–134, 1998.
[89] B. Alderson-Day, C. Fernyhough. Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. Psychological Bulletin, 141(5):931–965, 2015.
[90] P. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. arXiv:2005.11401.
[93] I. T. Jolliffe, D. B. Stephenson (eds.). Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd ed., Wiley, 2012.

[94] S. Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR, 2023. arXiv:2210.03629.
[95] J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.
[96] Based Hardware. Omi: an Open-Source AI Wearable. Product manifesto and documentation, 2025.
[97] Bee. An Ambient Wearable AI Assistant that Turns Everyday Conversations into Summaries, Reminders, and To-Dos. Product documentation, 2025; reported acquisition by Amazon announced July 2025.
[98] PLAUD. Note and NotePin: Wearable AI Voice Recorders and Note-Takers. Product documentation, 2024.
[99] Limitless. Pendant: a Wearable AI for Personalized Memory over Everyday Conversations. Product documentation, 2024; reported acquisition by Meta announced December 2025.
[100] Meta. Ray-Ban Meta and Meta Ray-Ban Display AI Glasses. Product documentation, 2023–2025.
[101] Brilliant Labs. Halo: Open-Source AI Glasses with the Noa Agent and Cross-Session Narrative Memory. Product documentation, 2025.
[102] Snap. Spectacles and Specs: Standalone See-Through AR Glasses. Product documentation, 2024–2026.

I. Self-report bias, lifelogging, ambient capture, and inferred data

[103] D. P. Crowne, D. Marlowe. A New Scale of Social Desirability Independent of Psychopathology. Journal of Consulting Psychology, 24(4):349–354, 1960.
[104] D. L. Paulhus. Two-Component Models of Socially Desirable Responding. Journal of Personality and Social Psychology, 46(3):598–609, 1984.
[105] M. Csikszentmihalyi, R. Larson. Validity and Reliability of the Experience-Sampling Method. Journal of Nervous and Mental Disease, 175(9):526–536, 1987.
[106] D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, A. A. Stone. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. Science, 306(5702):1776–1780, 2004.
[107] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, N. P. Podsakoff. Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. Journal of Applied Psychology, 88(5):879–903, 2003.
[108] A. J. Sellen, S. Whittaker. Beyond Total Capture: A Constructive Critique of Lifelogging. Communications of the ACM, 53(5):70–77, 2010.
[109] C. Gurrin, A. F. Smeaton, A. R. Doherty. LifeLogging: Personal Big Data. Foundations and Trends in Information Retrieval, 8(1):1–125, 2014.
[110] M. Kosinski, D. Stillwell, T. Graepel. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. Proceedings of the National Academy of Sciences, 110(15):5802–5805, 2013.
[111] S. Wachter, B. Mittelstadt. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review, 2019(2):494–620, 2019.
[112] T. Denning, Z. Dehlawi, T. Kohno. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-Mediating Technologies. CHI, 2014.
[113] D. Susser, B. Roessler, H. Nissenbaum. Online Manipulation: Hidden Influences in a Digital World. Georgetown Law Technology Review, 4(1):1–45, 2019.

[114] S. Zuboff. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York, 2019.
[115] I. Ajunwa, K. Crawford, J. Schultz. Limitless Worker Surveillance. California Law Review, 105(3):735–776, 2017.

A Glossary of core terms

Definitions are deliberately compact; the body is authoritative where they differ.

Target state. A latent configuration a target behaves as if acting to reach, maintain, resume, or abandon; evaluated only through observable consequences, never scored directly. **Stated goal.** What a target declares it wants — evidence, not a scored quantity. **Enacted priority.** What a target's allocation of a scarce resource (attention, time) reveals, independent of what it states. **Inferred goal.** What a model concludes a target is pursuing; evidence-bearing latent structure, scored only via a mapped outcome. **Implicit target.** A latent orientation inferred from behaviour, repetition, avoidance, or attention that the target has not declared and may not introspect. **Score target.** A pre-registered discrete future state or transition with an externally observable resolution — the only object that earns or loses credit; a candidate lacking such a rule is **evidence**, not a score target. **Inclusion rule (hard).** A target state is included in TargetSpace only if it has a pre-registered observable resolution rule; otherwise it is evidence, not a scored state (Section 3.2, Table 2). **Sealed outcome (outcome).** The realized value of a score target at the resolution time, fixed by a deterministic rule and hashed before it exists; scoring compares the sealed forecast against it. **Forecast horizon.** The interval $h = r - t$ from the sealed forecast time t to the resolution time r; forecasts are specified and scored per horizon. **Transition window.** The span over which a target-state transition is deemed to occur for resolution purposes, distinct from the forecast horizon. **Belief state.** The observer/model's distribution over a target's latent variables, updated as evidence arrives; a sufficient statistic of the history that the benchmark motivates but never scores. **Transition.** A change in which target state a system is acting to reach: emergence, stabilization, competition, displacement, abandonment, resumption, or opportunity-induced creation. **Population-prior baseline (R1).** A reference that predicts from population base rates with no target-specific information; the entry condition a model must beat. **Own-routine baseline (R2).** A pre-registered, walk-forward model of the target's own recent routine; skill over R2 is the headline target-specific signal. **Evidence tier.** A cumulative band of the evidence ladder (Section N.2) describing which streams a system may use, from low-content metadata to specialized sensors. **Evidence ablation.** Measuring skill over R2 as a function of evidence tier, to test which streams add target-specific information. **Permutation specificity gate.** A test that scores a system's forecasts for one target against another target's outcomes; genuine target-specific skill should collapse. **Prospective sealing.** Timestamping and hashing (SHA-256) a forecast before its outcome exists, preventing hindsight and leakage. **Walk-forward evaluation.** Prequential scoring using only evidence timestamped at or before t; random cross-validation is prohibited. **Calibration gate.** A pass/warn/fail check that stated confidence tracks empirical frequency (proper scoring plus expected-calibration-error bands). **Target-specific skill.** Calibrated predictive skill, in bits, that exceeds both R1 and R2 and collapses under permutation. **Epistemic resolution.** The falling of the observer's uncertainty as evidence accumulates (posterior concentration), distinct from participatory formation in the target. **Participatory dynamics.** Observable feedback, attention, action, and constraint signals through which a target's own process may change which futures remain reachable; recorded as auxiliary diagnostic variables (Appendix H), never part of primary scoring.

Table 6: Protocol elements (referenced from Section 3.2): each a concrete definition with its pre-registered option(s). Design choices are fixed before a forecast is sealed and disclosed with the result.

Element	**Definition**	**Pre-registered option(s)**
Target	the tracked instance i	person / agent / system / organization / process
Target state	latent configuration z_t the target acts to reach or maintain	scored only via observable consequences (A2)
Forecast	a distribution over the discrete answer space A at time t	proper-scored; nonzero probability floor
Observation window	evidence $E_{\leq t}$ available up to forecast time t	zero- / short- / longitudinal-history arms
Prediction horizon	$h = r - t$, from t to resolution time r	short / medium / long; scored per horizon
Label / outcome	the resolved value at r	set by a deterministic resolution rule
Ground truth	how the outcome is established	observed action, sensor threshold, or pre-registered rule; inter-rater reliability where contestable
Evaluation metric	how skill is scored	log score (bits), Brier; Skill over R1/R2
Baseline	reference the system must beat	R1 population prior; R2 own-routine
Uncertainty reporting	calibration and interval reporting	calibration gate; day-blocked / person-clustered intervals
Leakage control	preventing use of future information	sealing (SHA-256) + strict walk-forward; random CV prohibited

B Minimal synthetic harness and submission specification

Purpose and scope. The minimal synthetic harness is a public, runnable reference implementation for validating instance and forecast schemas, exercising proper-score and Skill computation, checking the R1 and R2 baselines, running the permutation control, running calibration checks, running evidence ablations, and producing a submission report. It is a smoke test plus reference path: it is *not* the empirical validation dataset, and it is *not* evidence that the benchmark has been experimentally validated. It uses no human data and supports no empirical claims. The harness and this specification are distributed through the project website (`https://targetspace.org`) and the public repository (`https://github.com/yurisyl/targetspace-bench`), and accompany the preprint as ancillary files.

Alignment with ambient capture and personal-AI logs. The harness mirrors, in synthetic form, the pipeline any ambient target-state system induces (Section 1): raw audio, video, or digital exhaust from recorders, wearable memory devices, first-person glasses, meeting note-takers, or longitudinal personal-AI logs is segmented into observations; observations become timestamped evidence; evidence is grouped into target episodes; candidate target states are proposed; sealed forecasts are issued; and outcomes resolve and are scored. The mapping is generic — raw capture → segmented observations → timestamped evidence → target episodes → candidate target states → forecasts → scored outcomes — so a team can substitute its own compliant data at the evidence layer without changing the scoring machinery.

What it does not demonstrate. No human data; no real personal intelligence; no evidence that passive observation helps; no cross-domain validation; no safety validation.

Required input schema (instances and evidence). Per forecast instance: `instance_id`; anonymized `target_id` (`subject_id`); `forecast_time` and `resolution_time` (or an explicit `horizon`); `question_type`; `answer_space`; the deterministic `resolution_rule`; and a `consent_status` / eligibility flag. Per evidence record: `segment_id`; time window; `modality` label (evidence tier L0–L6); optional text or transcript summary; and provenance pointers. Outcome labels are withheld from the system under test and stored separately (`instance_id`, resolved `outcome`, resolution provenance). Target-state or transition markers used to construct instances follow the taxonomy of Appendix J. In the shipped TS-Personal JSON Schemas (the released TargetSpace schema set) these generic fields instantiate concretely: the target is `participant_id`, the instance is `task_id`, the question type is `task_type`, the resolved label is `observed_answer`, and evidence modality/window are `modalities` with `evidence_cutoff_time`; the abstract vocabulary here is the domain-general layer over that concrete personal-track schema.

Required output schema (per system). Per forecast: `forecast_id`, `system_id`, `instance_id`, `forecast_time`, a probability for every element of `answer_space` (ranked candidates permitted for ordered spaces); a payload `sha256` seal recorded in the run/sealing manifest keyed by `forecast_id` (a hash of the payload is stored alongside the record it seals, not inside it); and optional `evidence_refs`, carried in the forecast `metadata`, identifying the segments supporting the forecast. Per run: a metrics file (JSON or CSV) with the scores below; a run manifest (data slice, config, seeds, timestamps); a model card or system description including the output adapter, retrieval/memory access, and training cutoff; an ablation report per evidence tier; and the baseline comparison against R1 and R2.

Baselines. The harness ships five reference conditions: **R0**, a uniform-random / prevalence floor (sanity only); **R1**, the population-prior baseline fit leave-one-out without the scored target's history; **R2**, the target's own-routine baseline (default recipe in Appendix D); the **target-specific system** under test; and the **permutation null**, the system's forecasts scored against matched wrong-target outcomes. Evidence-ablation variants re-run the system per evidence tier. R2 is the decisive reference: a claimed TargetSpace improvement must beat R2 — not merely R0 or generic replay — and its skill must collapse under the permutation null.

Metrics. The harness reports, per horizon and per evidence tier: log score (bits; primary) and Skill over R1 and over R2; Brier score; calibration (top-label ECE bands with pass/warn/fail, plus intercept/slope where sample size permits); the permutation-control delta (true-pairing Skill minus wrong-pairing Skill); and the evidence-ablation delta per tier. For binary or ranked answer spaces, AUROC/AUPRC and top-k accuracy may be reported as secondary diagnostics; no composite score is used, and lead-time (horizon-specific) performance is reported rather than pooled (Appendix L).

Command-line flow. The single-file demonstration is runnable today as `python targetspace_synthetic_demo.py` and executes the full flow (generation, walk-forward evaluation, R1/R2, scoring, calibration, permutation, ablation) in one run. The stepwise commands below are the *reference CLI specification* — the intended command structure for the full submission harness; they are *not part of the arXiv ancillary bundle*, which ships only the single-file harness (a fuller reference implementation is maintained in the public repository):

- `python generate_synthetic_targets.py -config example_config.yaml` (or: validate your own data against the input schema)

- `python run_walk_forward_eval.py -config example_config.yaml -baselines r0,r1,r2`
- `python scoring.py -predictions runs/system/forecasts.jsonl -truth data/outcomes.jsonl`
- `python permutation_test.py -predictions runs/system/forecasts.jsonl -matching matched`
- `python run_walk_forward_eval.py -config example_config.yaml -ablate evidence_tier`

Ancillary files (shipped with Version 1.0). `README.md` (run instructions and scope); `targetspace_synthetic_demo.py` (the runnable single-file harness; Python 3.8+, standard library only, deterministic); `example_output.md` (a captured reference run); `requirements.txt` (no third-party dependencies). The stepwise standalone-script invocation named above (`generate_synthetic_targets.py`, `run_walk_forward_eval.py`, `permutation_test.py`, `example_config.yaml`) is a reference *specification* and is not part of the arXiv ancillary bundle, which ships only the single-file harness. The corresponding capabilities — proper scoring, the R1/R2 baselines, and the permutation and calibration checks — are nonetheless implemented in the public repository as an installable package and command-line tool, on synthetic data only and supporting no empirical claim.

Submission readiness. A team with compliant longitudinal passive-observation data should be able to use this specification to format a submission, run the reference checks, compare against R1/R2, and generate a benchmark report. The synthetic harness verifies the mechanics; it does not itself constitute a benchmark result.

Example leaderboard row. Values are illustrative synthetic outputs of the demo script; they are not empirical results.

Table 7: Illustrative synthetic leaderboard row produced by the demo script. Skill is measured in bits over the named reference baseline.

system_id	tier	vs_R1	vs_R2	brier	cal. warn	perm. loss	n_tgt	n_fc
toy-L2	L2	+0.049	+0.036	0.183	none	collapses	20	800

The columns report, in order, the system identifier, evidence tier, skill in bits over R1, skill in bits over R2, Brier score, calibration-gate warning status, the permutation specificity outcome (skill collapses under permutation, as required), and the target and forecast counts.

C Running TargetSpace on a product: implementation guide

This appendix is an operational recipe for a product team — a note-taking or memory app, a wearable audio recorder, or multimodal glasses — to run a TargetSpace evaluation on its own users' data without guessing. It reuses the released schemas (Appendix B) and adds no new record types. Everything below is a protocol usage guide, not an empirical claim; a run over real users must satisfy the consent, privacy, and governance requirements of Section 8 and Appendix E.

Minimum-viable benchmark path

The smallest defensible run is seven steps; each maps to one released TargetSpace schema.

1. **Choose target instances.** Select consenting users and a sealed evaluation window. Register each as a target with `participant.schema.json` (`participant_id`, `timezone`, `routine_profile`, `allowed_tasks`); for real users the synthetic fields are replaced by de-identified profile metadata.
2. **Choose a task family.** Pick one or more rows from the quickstart pack below. Each fixes a `task_type`, an `answer_space`, and a deterministic `resolution_rule`.
3. **Freeze evidence tiers.** Declare which evidence tiers (L0–L6, Section N.2) each condition may use in `evidence_manifest.schema.json` (`evidence_tier`, `modalities`, `evidence_start_time`, `evidence_end_time`, `hash_manifest`). One manifest per tier condition drives the evidence ablation.
4. **Generate sealed forecasts.** Before any outcome is observed, emit one `forecast.schema.json` record per instance: a probability over every element of `answer_space`, an `evidence_cutoff_time` $\leq T$, and a hash sealing the payload. Store forecasts write-once.
5. **Resolve outcomes deterministically.** At resolution time, apply the task's resolution rule to later observable evidence and write `outcome.schema.json` (`observed_answer`, `resolver`, `resolution_method`). Outcomes are withheld from the system under test until sealing.
6. **Score against R1 and R2.** Compute log score (bits, primary) and Brier per instance; report Skill in bits over the population prior R1 and over the own-routine baseline R2 (Appendix D), plus top-label calibration (ECE band). A claimed improvement must beat R2, not merely R1 or replay.
7. **Run the permutation control and report the standard row.** Re-score each system against matched wrong-target outcomes; target-specific skill must collapse. Emit one `leaderboard.schema.json` row (and a `submission.schema.json` record) with the columns of Table 7: `system_id`, `tier`, `vs_R1`, `vs_R2`, `brier`, calibration status, permutation outcome, `n_tgt`, `n_fc`.

Quickstart task pack

Six product-relevant task families, each expressible in `forecast.schema.json` with the answer space and deterministic resolution rule shown. All are scored against R1/R2 with log score (bits), calibration, and the permutation gate.

Three product-facing instantiations

- **Note-taking / memory app (L0–L1).** Evidence = notes, tasks, calendar, and communication metadata already in the app. Natural tasks: recurring-commitment completion and response-latency bucket. Resolution is read from the app's own state at the horizon. No new sensing; the manifest declares tiers L0–L1 only.
- **Wearable audio recorder (L2).** Evidence = on-device transcripts and derived commitments/entities from ambient speech (Appendix O), not raw audio. Natural tasks: meeting/event realization and engagement vs. avoidance of a spoken obligation. Transcription stays local; only sealed forecasts and resolved labels leave the device.

Table 8: Quickstart task pack. Each row is a `task_type` with a fixed `answer_space` and a deterministic, pre-registered `resolution_rule` over later observable evidence. No subjective report resolves an outcome.

Task family	Answer space	Resolution rule (deterministic)
Recurring-commitment completion	{complete, defer, cancel, replace}	Calendar/task status or a pre-registered behavioural marker at the next scheduled occurrence fixes the label.
Meeting / event realization	{attended, no-show, rescheduled, cancelled}	Attendance signal (join event, location match, or logged presence) within the event window.
Response-latency bucket	{<1h, 1–24h, 1–3d, >3d, no-response}	Time from a flagged inbound obligation to the first outbound reply, bucketed; no reply before the horizon resolves to no-response.
Task continuation vs. switch	{continue, switch, pause}	The active task at T versus the active task after the next transition boundary; continuation iff the same task persists past the window.
Priority maintained vs. displaced	{maintained, displaced}	Whether the top-priority commitment at T still receives sustained allocation after the window, or is supplanted by a newly dominant one.
Engagement vs. avoidance of a defined obligation	{engaged, avoided}	Engaged iff a substantive action on the named obligation (reply sent, document edited, task advanced) occurs before the horizon; otherwise avoided.

- **Multimodal glasses (L3–L4).** Evidence adds embodied context — objects, screens, documents, and locations in view. Natural tasks: task continuation vs. switch and priority maintained vs. displaced, where visual attention disambiguates transitions. Higher tiers carry higher sensitivity, so bystander redaction and on-device filtering (below) are mandatory, and the evidence-tier ablation measures whether L3–L4 actually buys skill over the L2 audio-only condition.

Practical constraints (required for any real run)

- **Consent and eligibility.** Consenting adults only; per-instance `consent_status`/eligibility gating; ethics or IRB review before recruitment (Appendix E).
- **Bystander capture.** Visible recording notice where feasible; bystander redaction or exclusion; no capture in bathrooms, bedrooms, medical/legal settings, schools, children's spaces, or confidential meetings.
- **Missingness.** Missing evidence is a first-class state: instances without eligible evidence at T are reported, not silently dropped, and the R2 admission threshold (Appendix D) governs which strata are scored.
- **Device coverage.** Report coverage per target (fraction of the window with usable evidence); low-coverage targets are stratified separately, never pooled to inflate skill.
- **Local / federated.** Local-first storage; on-device filtering converts raw audio, video, and screen streams into task-relevant semantic events before persistence (Section 3.3).
- **No raw-data export.** Only sealed forecasts, resolved labels, and aggregate metrics leave the client; no raw media or transcripts are exported by default.
- **Aggregate reporting.** Public output is the leaderboard row and aggregate diagnostics (Table 7); no per-target trajectories or raw inference objects are published (Section 8).

D Default R2 own-routine baseline specification

This appendix fixes a pre-registered default recipe (v1) for the R2 own-routine baseline. The recipe is provisional and protocol-configurable: the values below are defaults to be frozen before evaluation, not claims about any observed data. The intent is to make R2 a fair, routine-only competitor against which target-specific skill in bits can be measured without manufacturing lift against a weak baseline.

Admission. R2 is admitted for a given target and question type only when two conditions hold jointly: the target has enough prior resolved history (see minimum history below), and R2 beats R1 on a pre-seal validation slice (or a rolling historical criterion fixed before the seal). When R2 does not beat R1 on that slice, we report that R2 is not informative for that stratum and fall back to R1 as the reference; we do not manufacture target-specific lift by scoring against a deliberately weak own-routine baseline.

Minimum history. The default minimum is the stricter of 20 prior resolved opportunities or 14 days of eligible history per target and question type. Strata that do not meet this threshold are not admitted for R2 and are reported separately. The threshold is provisional and protocol-configurable.

Features (allowed). R2 may use only simple pre-forecast routine features: marginal historical frequency for the target $\times$ question type; recency-weighted frequency; persistence or previous target state; recent target-state transition counts; time-of-day; day-of-week; and known recurring calendar or deadline commitments that are available before forecast time.

Exclusions. R2 may not use any of the following: passive audio, video, or screen content; semantic text content beyond pre-specified routine labels; any future information relative to the evidence available up to time t; entrant model outputs; manually selected post-hoc features; or any representation learned from the scored test outcomes.

Fitting. R2 is organizer-fit and frozen before evaluation. Fitting is walk-forward only: no random cross-validation and no tuning against test outcomes. Parameters are estimated from history available before each forecast time and never revised using resolved test labels.

Smoothing. R2 applies a pre-registered smoothing scheme and a nonzero probability floor over the answer space A, so that every admitted answer receives positive probability and proper scores (log score in bits, Brier) remain finite. Exact smoothing constants and the floor are protocol parameters fixed before test results are seen.

Reporting. Skill in bits is reported against R1 and against R2 separately, never collapsed into a single number. Strata in which R2 fails admission are reported as such, with R1 as the reference. Incomparable answer spaces are not pooled without strata.

E Safe pilot configuration

The benchmark can be tested without assembling an uncontrolled lifelogging dataset. Table 9 lists a default minimal safe pilot configuration, and Figure 3 shows the federated, local-first architecture

it assumes: raw capture stays on the participant device, and only sealed forecasts, resolved labels, and aggregate metrics leave the client.

Table 9: Minimal safe pilot configuration. The benchmark can be tested without creating an uncontrolled lifelogging dataset. The following practical controls are the default for any pilot.

Default controls for a minimal safe pilot
• Consenting adults only.
• Local-first storage.
• No public release of raw video or audio.
• No raw third-party export by default.
• Only sealed forecasts and aggregate metrics leave the device or client.
• Participant controls for review, pause, deletion, and opt-out.
• No capture in bathrooms, bedrooms, medical or legal settings, schools, children's spaces, or confidential meetings.
• Visible recording notice where feasible or required.
• Bystander redaction or exclusion where feasible.
• Short raw-media retention window unless explicitly consented otherwise.
• Ethics / IRB (or equivalent) review before recruitment.
• Special handling for third-party content, employee contexts, children, and sensitive data.

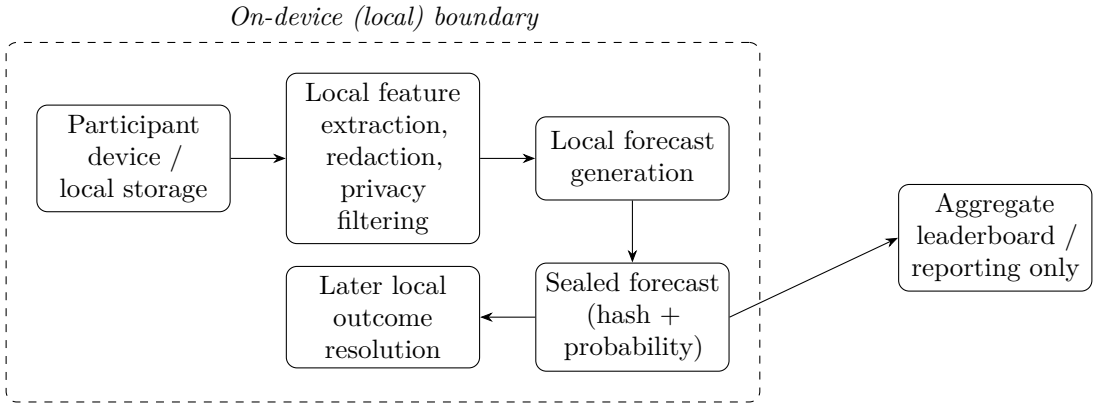

Figure 3: Safe federated pilot architecture. Raw audio, video, and screen data stay on the participant device: feature extraction, redaction, privacy filtering, forecast generation, sealing (hash and probability), and outcome resolution all run locally within the dashed on-device boundary. Only the sealed forecast (and, later, resolved labels and aggregate metrics) leaves the client; the off-device stage performs aggregate leaderboard and reporting computation only. No raw capture is transmitted.

F Limitations register

The full register of limitations summarized in Section 8, with mitigations where they exist. (L1) Latent targets are evaluated only through observable proxies. (L2) The federated, prospective design makes reproduction harder than a downloadable dataset and precludes third-party audit of sealing, labelling, and resolution, so results are holder-reproducible, not publicly auditable; we mitigate with a versioned protocol and shared harness. (L3) Single-instance and small-cohort evaluation

limits external validity. (L4) Predictability is bounded [13] and heterogeneous, so reporting is per-instance. (L5) Capture can be reactive — being observed may alter the behaviour being forecast — so habituation windows and reactivity checks are required, and induced deviations must not be read as target-state transitions. (L6) Domain-generality is a property of the formulation, demonstrated in one domain; the cross-domain claim is specified, not established. (L7) The privacy and governance safeguards are design commitments, not implemented features: no pilot has run, no institutional review has been obtained, and the privacy-filtering layer is unbuilt. (L8) A calibrated behaviour forecaster is dual-use, and the observe-not-intervene rule binds the benchmark, not downstream deployers. (L9) Longitudinal individual-level data carry selection and representation bias; a benchmark validated on a narrow cohort will not transfer, and per-instance reporting does not cure unrepresentative sampling. (L10) Ecological validity is limited: instrumented, consented capture may not reflect unobserved behaviour, and a habituation window does not guarantee it. (L11) Many target states admit annotation and label ambiguity — whether a commitment was 'abandoned' or 'deferred' can be genuinely contested — so resolution rules must be pre-registered and inter-rater reliability reported. (L12) Targets exhibit drift and non-stationarity: the dynamics a model fits can change, so R2 is re-fit walk-forward and skill decay is tracked rather than assumed stable. (L13) For some outcomes ground truth is intrinsically hard to establish — probabilistic, delayed, or only partially observed — which bounds achievable skill and is why scoring is proper rather than accuracy-based. (L14) Multimodal longitudinal capture is costly and burdensome to collect and store, constraining cohort size and biasing toward participants who can be instrumented. (L15) Results risk overfitting to a single person, household, or organization; the permutation gate guards against cross-target leakage but not against a model tuned to one idiosyncratic target, so external validity requires multiple independent targets. (L16) The benchmark measures forecasting of behaviour, not understanding of intent: a high score is predictive alignment with observable future states, and any inference about goals, reasons, or inner life is explanatory audit (Section 3.2) subordinate to the sealed forecast, never established by the score.

G Supplementary figures

These figures restate, in visual form, material presented as tables in the main text: Figure 4 corresponds to the target-specificity stack of Section 4, and Figure 5 to the evidence ladder of Section N.2.

H Participatory dynamics as auxiliary state variables

TargetSpace keeps epistemic scoring separate from target-state formation. Forecast accuracy is scored only against sealed external outcomes (Section 3). Separately, the harness can record observable feedback, attention, action, and constraint signals as *auxiliary* variables, letting researchers test whether adaptive feedback loops change the reachable target-state space and whether those changes improve forecasts. Two processes must not be conflated: *epistemic resolution* is the observer's uncertainty falling as evidence accumulates — posterior concentration over the observer's maintained distribution; *participatory target-state formation* is a change in the target itself — through attention, action, feedback, and constraint — that alters which futures remain salient, reachable, reinforced, or stable. Posterior concentration in the observer is not target-state formation in the observed system; the benchmark continues to score forecasts only against sealed outcomes, and the proxies below are a diagnostic layer, not part of the scoring foundation.

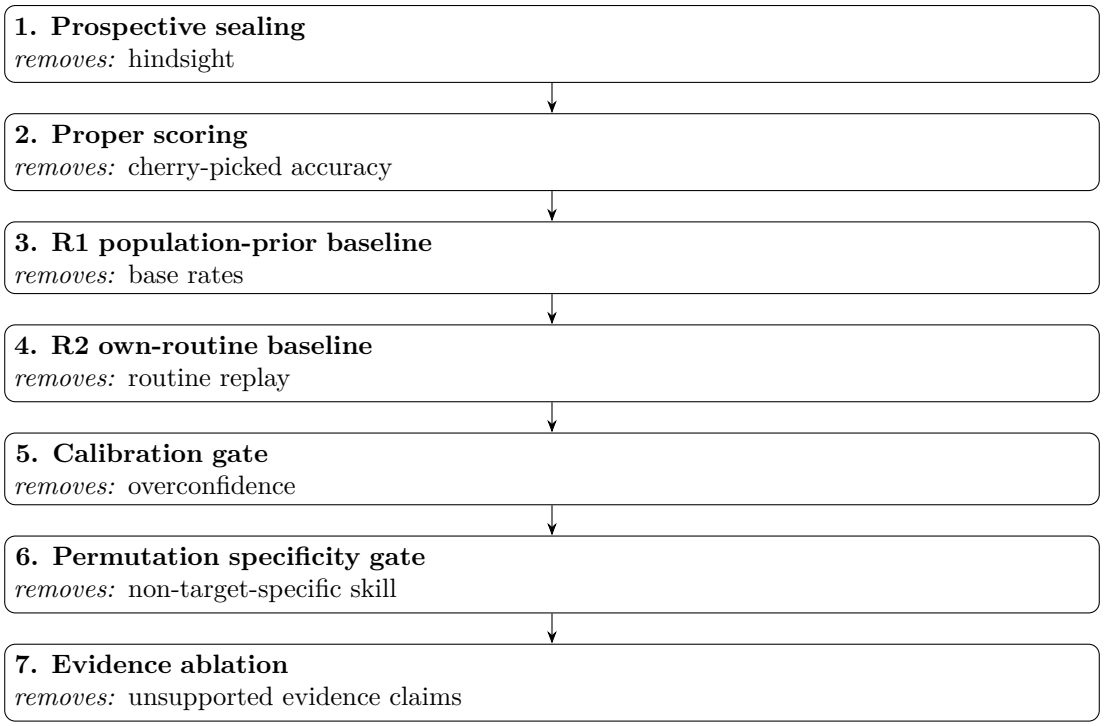

Figure 4: The target-specificity stack, complementing the stack table (Section 4). Each layer removes one way a forecast can appear target-specific without being so: a forecast that survives all seven filters cannot be explained by hindsight, selective reporting, population base rates, replay of the subject's own routine, overconfidence, non-target-specific skill, or unsupported evidence claims. Layers 4 and 6 (the R2 own-routine baseline and the permutation specificity gate) are the anchors contributed here.

- *Feedback exposure index:* count, source, valence, intensity, latency, and channel of feedback the target receives.
- *Attention shift index:* change in time, dwell, mentions, or returns to objects of attention before versus after feedback.
- *Action-coupling index:* whether attention changes are followed by observable actions — emails, calendar changes, task completion, purchases, commitments, code commits, document edits.
- *Constraint / reachability map:* whether resources, obligations, blockers, commitments, permissions, or pathways are added or removed.
- *Forecast sensitivity audit:* compares passive-only, feedback, feedback+attention, and feedback+attention+action/constraint feature sets, testing whether participatory variables improve sealed-outcome forecasts.

A positive finding — forecasts improving when participatory variables are added, and the reachable target-state space measurably changing under feedback — would be prospective evidence that adaptive loops matter, not an assumption. A null finding leaves the primary scoring unaffected.

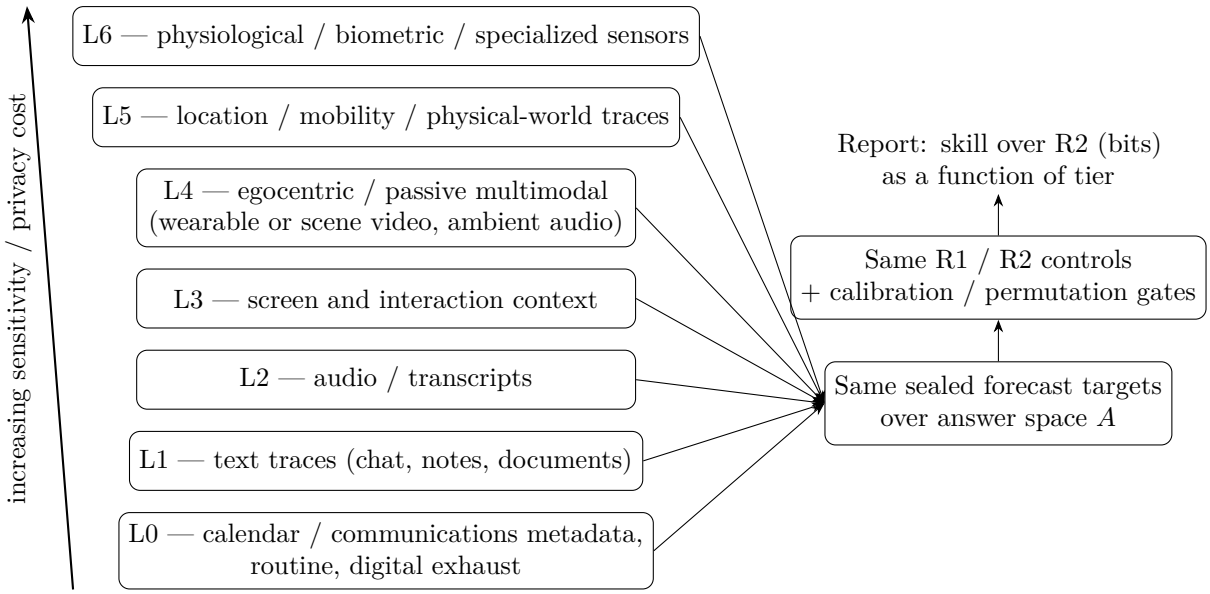

Figure 5: Evidence-tier ablation design (Section N.2). Seven evidence tiers (L0, least sensitive, at bottom; L6, most sensitive, at top) are each evaluated against the *same* sealed forecast targets and the *same* R1 (population-prior) and R2 (own-routine) controls under the calibration and permutation-specificity gates. Whether richer evidence helps is measured as skill over R2 (in bits) as a function of tier, not assumed; sensitivity and privacy cost increase upward.

The attention–target dynamical loop (extended treatment)

The main text (Section 2) treats attention as evidence and as a candidate source of predictive lift. The underlying cognitive-dynamical reading is a loop rather than a one-way law: repeated allocation can reinforce and stabilize a provisional objective; withdrawal can weaken, delay, or displace one; newly encountered evidence or opportunity can make determinate a target that previously was not [72,73]. Schematically,

$$z_t \;\rightarrow\; a_t \;\rightarrow\; (e_t,\, u_t) \;\rightarrow\; z_{t+1},$$

where z_t is the latent target state, a_t attention allocation (or an observable proxy), e_t newly encountered evidence, and u_t action. Attention is thereby both evidence of the current target state and one candidate process through which the next is formed. Prior dispositions, habit, environment, and constraint remain primary; attention does not create targets from nothing; and the loop is exploited by the benchmark only as a source of falsifiable predictive-lift hypotheses, never as an assumed mechanism.

Extended motivation: experience, not knowledge

The bounded ambition behind the personal track is to model not what a person *knows* but — in a precise and deliberately limited sense — the situated stream from which their behaviour follows; we borrow Nagel's "what it is like to be" phrasing [92] only as a motivating contrast between third-person description and situated experience, and explicitly *not* as any claim of access to inner experience, to which a system has none and about which no claim is made, but the lived trajectory from which behaviour emerges — the unfolding, situated stream of attention, context, pressure, memory, constraint, and possibility that shapes what becomes salient and what the person does next.

Knowledge is episodic and declarative; experience is longitudinal, situated, relational, embodied, multimodal, and predictive. TargetSpace evaluates the transition from passive experience capture to target-state inference, and scores it only through externally resolvable forecasts.

I Personal track: task specification and extended discussion

To show that TargetSpace instantiates runnable tasks and not only a framework, we specify one minimal personal-track task, the *Recurring-Commitment Completion Forecast*. It is synthetic and illustrative; no empirical results are reported.

- *Task.* Given timestamped evidence available up to a sealed time T, forecast how a scheduled recurring commitment resolves by a later resolution time r.
- *Inputs (by tier).* Calendar and task history, communication metadata, and prior completion/defer/cancel history (L0); optional user-authored text traces (L1); optional passive-observation summaries (L4). See Section N.2.
- *Answer space.* $A = \{\text{complete}, \text{defer}, \text{cancel}, \text{replace}\}$.
- *Baselines.* R1 (population prior over the four outcomes) and R2 (the target's own recurring-commitment routine), both fit walk-forward on evidence before T.
- *Resolution.* A deterministic rule over later observable evidence — calendar status, attendance, task completion, or a pre-registered behavioural marker — fixes the outcome at r.
- *Scoring.* Log score (bits) and Brier against the sealed outcome; report skill over R1 and over R2 (Section 6).
- *Specificity.* A permutation test across matched targets should reduce or collapse the skill if a model's forecast was genuinely target-specific.

The task is small enough to run on current language models, agents, and retrieval or memory systems, and exercises the full stack: sealed prospective scoring, R1/R2 baselines, calibration, evidence tiers, and the permutation gate.

Horizon–abstraction hierarchy

The four task families are scored across the following horizon–abstraction hierarchy.

I.1 The corpus bottleneck and the capture-to-labels stack

The missing data layer named in Section 2 is concrete in this domain. Digital exhaust (L0–L1) records a person's *outputs* — what they wrote, sent, or scheduled — not the process that produced them. Passive first-person observation (L2–L4) is the stream from which goals, constraints, interruptions, and transitions could be inferred *before* intent is formalized: passive audio, a screen / digital-activity timeline, documents and files touched, calendar and task context, interruptions, spoken intent, environmental and social context, and the timing and sequencing of actions. Egocentric corpora such as Ego4D and Ego-Exo4D [56,57] show passive first-person capture at scale, but as generic clips for activity recognition — not longitudinal records of one tracked individual aligned to that person's goals and outcomes over time.

Table 10: A horizon–abstraction hierarchy for target-state forecasting. Each level is scored independently; no single representation is assumed optimal across horizons, and resolution rules are deterministic or pre-registered. Uncertainty generally grows with horizon.

Level	**Horizon**	**Answer space / resolution rule**	**Baseline; metric**
Immediate action / next state	seconds–minutes	concrete next state; observed action	R2; log/Brier, calibration
Short-horizon task transition	minutes–hours	small discrete set; observed completion or switch	R2; Skill, top-k
Medium-horizon subgoal / commitment	hours–days	commitment set; follow-through within a window	R2; Skill, time-to-resolution
Longer-horizon priority / target-state shift	days–weeks	priority/regime set; sustained reallocation of attention/resources	R1/R2; transition-path likelihood
High-level regime / strategy transition	weeks+	coarse regime labels; observed regime change	R1; Skill, calibration

Raw capture is not, by itself, a benchmark. Turning observation into evaluable state-transition data takes a stack — passive sensing, a longitudinal multimodal corpus, segmentation with goal/task/outcome/transition labels, a temporal state representation, the prediction benchmark, and a privacy-filtering layer (Section 8). **TargetSpace is the evaluation layer**: it fixes the evaluable target and the scoring spine and specifies what the upstream layers must deliver, but does not supply the corpus. Assembling a labelled, privacy-filtered, goal-aligned longitudinal corpus is the open bottleneck, and the pilot (Section 6.5) builds only a thin slice; passive recording alone is insufficient. For high-sensitivity tiers (L2–L6), federation is necessary but insufficient: a viable implementation would likely require on-device filtering that converts raw audio, video, and screen streams into task-relevant semantic events before persistence — mitigating, though not solving, the bystander-consent and recording-law problems of Section 8. **Hardware is the capture layer, not the benchmark**: wearable and passive sensors raise stream fidelity, but TargetSpace takes no position on a device; whether richer capture pays is the evidence-tier ablation (Section N.2). A human-centered world model must be learned from observation, since language and digital exhaust alone are too sparse and post-hoc.

I.2 TargetSpace as a benchmark for human-centered world models

The world-model research direction defines intelligence operationally as the ability to predict the consequences of actions in an abstract representation space and to plan over those predictions [33,36]. TargetSpace applies that idea to human-assistive AI. The 'environment' is not a physical system but the latent state of a person: can a system infer, from passive observation up to a sealed time T and *before the user writes an explicit prompt*, the user's current goal state, hidden constraints, the likely next transition, and the next useful action? It is a world model whose dynamics are a human task process, scored by calibrated skill over R1/R2 rather than by reward or reconstruction.

A behavioral world model, not a physical one. The world-model idea (above) earns its power because predicting the consequences of actions lets an agent test a plan before acting [12], with planning as search over predicted futures [37,38]. TargetSpace generalizes this predictive substrate, not its physical content: it asks for no prediction of torques, contacts, or pixel dynamics and claims no understanding of a physical world it never observes. The classical form is *state + action → next state*; the TargetSpace form is *passive longitudinal observation → belief state over latent targets* (Section 3.2) *→ target-relevant future behaviour*. A competent system must therefore behave as if it had learned a compact behavioral world model of one observed person — equivalently a personal,

or target-state belief, model — capturing recurring contexts, constraints, stable preferences, salient episodes, unresolved decisions, and target transitions, and predicting how that target state evolves as evidence arrives. The analogy is to the *form* of the predictive state, not its physical realization; like any learned dynamics model it can compound error over long horizons [76], and only its sealed forecast is scored — not planning, action, or the model itself (Sections 3.5 and 8).

The shift is clearest by contrast with what existing evaluations ask. A **transcript** or chat benchmark asks *what was said?*; a **memory** benchmark asks *what happened?*; TargetSpace asks a third question — *what goal state is the person in, and what transition is likely next?* (whether any intervention would help or harm is a separate, non-scored governance question bound by the observe-not-intervene rule, Section 8).

Retrieval and long context are not the capability. Two now-standard memory paradigms make the same point concrete. **Retrieval-augmented generation** conditions a model on documents fetched from an external store [90]; **long-context** models simply hold more of the history in the window. Both surface *what happened* more readily — usefully — but neither is scored on whether the system inferred *why* an episode mattered or *what target* it implies: a retrieval system can surface the reread email, whereas TargetSpace asks whether a model reads it as an avoided decision and forecasts the slip. That is why a higher recall@k or a longer context window does not by itself move prospective skill over R1/R2 under the permutation gate; the architecture grid (Section N.1) accordingly places base, long-context, retrieval, episodic-memory, and belief-state–updating systems on identical sealed instances.

Prompts may reveal goals but often omit the **pre-prompt** process — a prompt is a late, compressed artifact emitted *after* an intention has partly formed — which passive observation can preserve (attention, interruptions, partial actions), and the evidence-tier ablation tests whether preserving it improves prediction. TargetSpace does not build this user-centered world model; it operationalizes it as evaluation: fixing the latent quantities to predict, the baselines to beat (R1/R2), and the calibration and permutation gates that separate a genuine model from a fluent guess. The target is abstract predictive state, not raw reconstruction, and the framing is architecture-neutral (Section N.1, Table 13).

I.3 Asymmetric observability: extended person-perception basis

Section 2 argues that self-report and observation are overlapping but non-redundant channels whose relative predictive value the benchmark should measure rather than assume. This mirrors person-perception research. The self–other knowledge asymmetry model holds that self and informants know different things — the self more accurate for internal, low-observability traits, informants for evaluative, highly observable ones [58] — and meta-analyses find informant reports valid and often incremental to self-report, for some outcomes exceeding it [59,60]. Self-report has limits of its own: people often lack introspective access to what drives their behaviour [23], which motivates real-time experience sampling [61,105] and passive 'personal sensing' [62,64]; internal and external self-awareness are themselves distinct constructs [63]. None of this makes observation more accurate than self-report — it makes the two complementary, and the evidence-tier ablation and the head-to-head of human self-report against observational systems on identical sealed instances measure which modality, or which combination, yields more calibrated, person-specific skill. Clinical observation is the right analogy: complementary evidence, not clinician omniscience.

I.4 Implicit Targets and Self-Narrative

Surface action is an ambiguous signal of what a person is organizing around, so it helps to separate three layers. The **action layer** is what the person does (rereads a message; reopens a document; revisits a calendar event). The **narrative layer** is how they frame it ("I'm just making sure I understand it"). The **target layer** is what the behaviour is actually organizing around — which may be avoiding a difficult reply, reducing uncertainty before a decision, preserving a relationship, or protecting status. TargetSpace is primarily a benchmark for the third layer, using the first two as evidence. The separation is not decorative: declared intentions predict behaviour only weakly (the intention–behaviour gap [84]), motives inferred from behaviour diverge systematically from self-attributed ones [85], people may not introspect the causes of their own behaviour (Section I.3), and self-narration — inner speech that mediates attention, monitoring, and self-regulation [89] — can report a goal other than the one the longitudinal pattern reveals. A model that predicts only the action layer, or that trusts the narrative layer at face value, will miss the target layer exactly where it matters.

This is where surface behaviour becomes informative only longitudinally. Considered once, rereading an email, drafting but not sending a reply, repeatedly checking a thread without responding, or shifting attention as a deadline nears are each individually ambiguous. Across a tracked history they can mark a latent target under tension — an unresolved decision, an avoidance pattern, a commitment about to slip — precisely the deviations from routine where skill over R2 is earned. The benchmark does not ask a model to *narrate* this latent state; it asks it to convert these traces into a calibrated forecast of a pre-registered observable outcome (Section 3.2) and certifies, through R2 and the permutation gate, that the resulting skill is specific to this person rather than to population-level patterns of attention or delay. The cognitive-science framing here is motivation, not mechanism: TargetSpace takes no position on theories of inner speech or self-narrative and stands if they are revised — it requires only that some behaviourally relevant targets are implicit and inferable from passive longitudinal observation, which the evidence-tier ablation then tests rather than assumes.

J Transition taxonomy and resolution rules

The full taxonomy of target-state transitions the personal track forecasts (referenced from Sections 3.1 and 5.2) is given in Table 11. Each transition type carries a deterministic or pre-registered resolution rule grounded in later observable consequences, an ambiguity/evidence profile, and the baseline expected to be hardest to beat.

K Adopted components and extended related work

Table 12 itemizes the components TargetSpace adopts from prior work, their representative sources, and the use each is put to; the two anchors contributed here are the R2 own-routine baseline and the permutation specificity gate (Section 4).

K.1 Memory, belief states, and goal inference

Three literatures bear on what TargetSpace asks a system to maintain, and clarify what it does *not* reward. **Memory.** Cognitive science distinguishes episodic from semantic memory [83] and, more pointedly here, treats autobiographical memory as a goal-driven *construction* rather than a stored replay [81], with the same constructive system used to simulate the future as to reconstruct the past [82]. This is why we frame the target capability as forward-looking belief maintenance rather than

Table 11: A taxonomy of target-state transitions the personal track forecasts. Each has a deterministic or pre-registered resolution rule grounded in later observable consequences, not subjective report. "Likely baseline" is the baseline expected to be hardest to beat; "horizon" is the typical forecast range.

Transition type	**Operational description / resolution criterion**	**Ambiguity risk; evidence needed**	**Baseline; horizon**
Routine continuation	target persists; resolved by continued allocation/action matching the prior pattern	low; digital exhaust suffices	R2; short
Latent-but-active target	already pursued but unstated; resolved by later action or commitment	medium; behaviour + context	R2; short–medium
Emerging target	a new target becomes determinate; resolved by first commitment or resource expenditure after onset	high (onset timing); attention + context	R1/R2 weak; medium
Stabilization	a provisional target consolidates; resolved by repeated allocation across windows	medium; attention trajectory	attn-conditioned; medium
Competition	two targets contend; resolved by which receives sustained allocation	high; attention + outcomes	attn-conditioned; short–medium
Displacement / priority reversal	a new target supplants the prior one; resolved by a switch in sustained allocation	medium; attention + calendar/task	attn-conditioned; medium
Abandonment	a target is dropped; resolved by absence of follow-through past a window	medium; non-action over time	R2; medium
Resumption	a prior target returns; resolved by renewed allocation after a gap	high; longitudinal history	R2; medium–long
Constraint-forced transition	a constraint forces a change; resolved by the observed transition under the constraint	low–medium; constraint + outcome	R1; short
Opportunity-induced target	new evidence or opportunity creates a target; resolved by uptake after the encounter	high; evidence event + uptake	R1/R2 weak; short–medium

archival recall (Section 2). **Belief states.** Maintaining a distribution over latent variables, updated from partial observations and serving as a sufficient statistic of the history, is the belief-state formalism of partially observable decision processes [88]; TargetSpace adopts the formalism as motivation (Section 3.2) but scores only the externally resolved forecasts a belief state produces. **Latent-goal inference.** That observers recover latent goals and beliefs by inverting a model of approximately rational action is the Bayesian theory-of-mind / inverse-planning account [86,87], the cognitive-science complement to the machine theory-of-mind [6] and goal-/plan-recognition [29] work discussed above; behaviour is further organized by implicit motives that diverge from declared ones [85] and by intentions that predict action only weakly [84]. TargetSpace differs from all three in object and protocol: it scores neither recall (memory benchmarks), nor the internal belief state (world-model evaluation), nor which goal from a fixed set best explains past behaviour (goal recognition), but the prospective, calibrated, instance-specific forecast of the next target-state transition, under R1/R2 and the permutation gate.

K.2 World models, predictive learning, and consequence forecasting

A long lineage predicts the *latent* state rather than the surface: predictive coding [19,39]; world-model agents that plan over learned latent dynamics [12,37,38], with hierarchical variants planning over abstract latent subgoals [77]; joint-embedding and contrastive self-supervised methods [44,45,49] and

Table 12: Prior components and TargetSpace-specific contributions. TargetSpace adopts prospective sealing, proper scoring, calibration, evidence ablation, and latent-predictive evaluation from prior work; its own contribution is target-specific forecasting under strong controls — the R2 own-routine baseline and the permutation specificity gate — assembled around one evaluation question.

Prior component	**Representative sources**	**TargetSpace use**
Architecture-neutral world-model evaluation	AutumnBench / WorldTest [41]	Adopt; add calibration, R1/R2, permutation
Cross-architecture comparison	WorldPrediction [42]	Adopt; add proper scoring, sealing, specificity
Latent predictive learning	CPC [44]; JEPA [33–36]	Evaluate as an architecture class
Proper scoring & calibration	Good; Brier; Gneiting & Raftery [14,15,18]; HELM [43]	Adopt directly
Sealed prospective forecasting	ForecastBench [28]; Prophet Arena [40]; SWE-bench [2]	Adopt
Evidence ablation	digital phenotyping / personal sensing [62,64]	Adopt; report as lift over R2
Goal recognition & resolution timing	goal-recognition design / WCD [29]	Adapt to sealed, tracked instances

their information-theoretic and recurrent-world-model precursors [47,48]; LeCun's JEPA program [33–36], a *training framework* that predicts a target's *embedding* (optionally action-conditioned [36,46]) rather than reconstructing pixels or tokens, now extended to language targets [79]; and vision-language-action policies, in which broad, dexterous action emerges from multimodal policy learning [50,51], increasingly coupled to generative subgoal models [78]. The motivation bears on our object: predicting in embedding space keeps salient structure and discards the unpredictable — the same reason TargetSpace scores *transitions in a latent target state* rather than full-future reconstruction.

Reactive, predictive, planning, and hybrid systems are all candidate participants, not competitors, and we take no side on which will prevail. Uniformly, though, they are compared by representation quality, reward, task success, action accuracy, or rollout quality — evaluations that do not, by themselves, test calibration, skill over a target-specific routine, instance specificity, prospective sealing, evidence-value attribution, or multi-horizon target-state forecasting. TargetSpace supplies that measurement layer and applies it to these classes on identical sealed forecast instances; what matters for scoring is whether the predicted transition resolves correctly against the external target, not how the prediction is represented internally.

K.3 Forecasting, person modeling, and goal recognition

Calibrated event forecasting is established — ForecastBench [28], Prophet Arena [40], and others — but scores external public events, not target-state transitions of a tracked instance, and uses no own-routine or permutation control. Person-modeling benchmarks are the nearest in framing: Park et al. [4,5] replicate individuals' survey answers (their own-consistency reference is the closest cousin to R2, though used as a ceiling, not a baseline); KnowMe-Bench [26] formalizes person understanding but as *retrospective* comprehension, not prospective forecasting; EgoToM [27] infers a camera-wearer's goals and future actions from egocentric video, but per clip, by accuracy, without calibration or a persistent target; PersonaMem [7] tracks evolving preferences. Related

Table 13: Positioning, not a ranking. The three families are complementary; TargetSpace does not solve world models and takes no position on the eventual architecture. It is architecture-neutral (Section N.1) and asks whether passive observation improves calibrated prediction of human goal-state transitions.

Dimension	**Conventional LLM benchmark**	**Robotics / world-model benchmark**	**TargetSpace**
Input	a prompt / fixed context	state + action (sim or embodied rollout)	passive observation of a tracked instance up to a sealed time T
Prediction target	next token / answer	next state; consequence of an action (often in latent space)	latent target state and its transition
Evaluation object	output correctness / quality	task success; rollout / planning accuracy	calibrated skill over R1/R2 + permutation specificity
Time horizon	single turn / short	short-to-medium rollouts	hours–weeks; walk-forward, sealed
Role of language	central (it is the task)	peripheral (instructions)	one evidence stream among many; not the success criterion
Role of passive observation	none	sensor stream for control	central — the substrate; its value is measured by the evidence-tier ablation
Failure mode	fluent but wrong; contamination	reconstructs detail; rollout drift; sim-to-real gap	predicts the average not the target (fails R2 / survives permutation)
Success criterion	matches reference / human preference	reaches goal state; low rollout error	calibrated skill over R2 + permutation pass (beats routine, stays calibrated, is target-specific)

person-modeling and digital-twin efforts — personalization [3], preference modeling [8], digital twins of survey and behavioural respondents [9,10], a foundation model of human cognition [11], and multimodal theory-of-mind question answering [31] — predict held-out responses or preferences rather than sealed future target-state transitions of a tracked instance. Person-specific multi-horizon forecasting with active/passive evidence comparison already exists in digital phenotyping — we concede this and differentiate on proper scoring, calibration, and the permutation gate. Goal- and plan-recognition [29] infer latent target/goal posteriors under partial observability and define how early a target is identifiable; the field also studies online recognition and changing, hierarchical, or partially observable goals, so we do *not* claim prior work assumes goals are fixed. The narrower distinction is in the evaluation object: conventional goal recognition asks *which* goal, from a candidate set, best explains observed behaviour, whereas TargetSpace prospectively scores how a target's state, attention, and interaction with the environment produce the *next* target-state transition — emergence, stabilization, competition, displacement, abandonment, resumption, or opportunity-induced creation (Table 11) — under the R1/R2 and permutation controls. We adopt the resolution-timing idea for the metrics and re-cast it as a calibrated, prospective, instance-tracked measurement; the extension from recognizing a current goal to forecasting target-state formation is subordinate to the target-specificity stack, not a separate theory of goals.

K.4 Cross-scale target-state systems

Beyond persons, the target-state idea recurs — active inference and preferred states [19,32], cell-fate prediction [30], and Levin's morphogenesis [25] (Section 3.1) — which we cite as lineage and sibling domains, not as competing calibrated benchmarks.

K.5 Intelligence measurement and ARC

Our framing is indebted to Chollet's account of intelligence as skill-acquisition efficiency and to the ARC-AGI line's contamination-resistant, private-evaluation design [65,66]; TargetSpace shares the commitments to novelty, freshness, and efficiency (the evidence-tier ablation asks how much evidence a system needs, not only whether it eventually succeeds). ARC evaluates adaptation within a given problem frame, whereas TargetSpace asks the earlier question of inferring the latent state and target from observation when the goal is not framed; both are needed, and neither subsumes the other. Because static, fixed-answer benchmarks saturate and can be gamed once their targets become public [16,17,28], the prospective, future-resolved measurement regime is itself part of the contribution, not merely hygiene: target-specificity is scored against outcomes that do not exist at forecast time.

Table 14: The five-dimension conjunction across neighboring benchmarks (referenced from Section 7.2). Dimensions: 1: persistent individual / evolving entity; 2: longitudinal, continuously updated evidence; 3: strictly prospective / sealed prediction; 4: calibrated proper scoring with meaningful baselines; 5: goal-state / meaningful-transition forecasting as the target. Labels are cautious judgments from each benchmark's design, not measured scores. Prior work contains important subsets; we are not aware of a benchmark combining all five in one prospective evaluation apparatus.

Benchmark / family	**Persistent individual**	**Longitudinal evidence**	**Prospective / sealed**	**Calibrated + baselines**	**Transition target**
EgoToM [27]	absent	partial	partial	absent	arguable
ToMnet [6]	partial	partial	partial	absent	arguable
ForecastBench [28]	absent	absent	clear	clear	absent
PersonaMem [7]	clear	partial	absent	absent	arguable
Generative Agents [4,5]	clear	partial	absent	absent	absent
KnowMe-Bench [26]	clear	partial	absent	absent	absent
Goal / plan recognition [29]	partial	partial	arguable	absent	arguable
ARC-AGI [65,66]	absent	absent	clear	absent	absent
TargetSpace (this work)	**clear**	**clear**	**clear**	**clear**	**clear**

K.6 Comparing TargetSpace with other benchmark families

The relevant benchmark families are **complementary, not adversarial**: a full picture of a system's world-modelling ability needs several of them, and TargetSpace should be compared with each only on the axes they share. **Physical-reasoning and intuitive-physics** benchmarks are the right instrument for intuitive physics, object permanence, causality, spatial continuity, and physical-law violation. **Video-generation** benchmarks are right for visual realism, temporal consistency, and plausible scene evolution. **Embodied and robotics** benchmarks [50,51] are right for action utility, policy evaluation, and manipulation or control performance. **Forecasting** benchmarks [28,40] are right for probabilistic future prediction, calibration, and temporal uncertainty. **TargetSpace** is the right instrument for target-specific adaptation, longitudinal memory, goal-transition detection, passive-observation learning, multimodal evidence integration, counterfactual target forecasting, and belief updating over time. Table 15 sets these families on a common template and makes the complementarity explicit, extending the three-way positioning of Table 13 to the broader landscape;

the TargetSpace row deliberately concedes the capabilities it does not measure.

Table 15: Benchmark families and how TargetSpace complements them. The families are complementary, not competing: each is the appropriate instrument for a different question, and TargetSpace is compared with them only on shared axes (cf. Table 13). Bracketed citations point to instances already discussed in Section 7; families named at the category level carry none. The final row states the capabilities TargetSpace does *not* measure, by design.

Benchmark family	**Primary evaluation object**	**Typical input → output**	**What it under-measures**	**How TargetSpace complements it**
Physical / intuitive physics	physical plausibility, object permanence, causality, spatial continuity	short scenes / clips → plausibility or violation judgment	a persistent target; longitudinal adaptation; calibration over time	adds target-specific dynamics over generic physical law
Video generation / prediction	visual realism, temporal consistency, plausible scene evolution	context frames → generated continuation	target identity; sealed prospective scoring; proper calibration	scores latent target-state transitions, not surface reconstruction
Embodied / robotics [50,51]	action utility, policy evaluation, manipulation or control success	proprioception, sensors, actions → actions / achieved configuration	passive longitudinal inference; calibration; an own-routine baseline	scores passive consequence forecasts, not control
Causal / temporal video QA	causal and counterfactual reasoning over events	video + question → answer (accuracy)	a persistent target; prospective sealing; calibrated baselines	makes causal-style queries prospective, target-specific, proper-scored
Event / symbolic forecasting [28,40]	probabilistic prediction of public events	question + context → calibrated probability	a tracked individual target; own-routine R2; permutation specificity	adds the target as the unit, with R2 and permutation controls
TargetSpace (this work)	target-conditioned longitudinal world modeling	passive multimodal observation up to sealed T → calibrated forecast over target-state transitions	by design: physical realism, control, generation fidelity	is the complementary layer the other families omit

Adjacent lines the table does not enumerate. A few neighbouring literatures deserve explicit separation. *Predictive representation learning* — JEPA and its video variant V-JEPA [33–36], and contrastive / joint-embedding methods [44] — learns to predict in a latent space; TargetSpace *evaluates* such models but scores the externally resolved target-state transition, not the quality of the learned representation. *World-model and video-prediction benchmarks* [12,37,38,41,42] score rollout or reconstruction quality; TargetSpace scores a sealed, calibrated forecast about a *persistent* target. *Planning benchmarks* score whether a search or policy reaches a goal state; TargetSpace scores the forecast of the transition, not the plan, and does not infer planning ability from task success (Section 3.5). *Agent and task-completion benchmarks* [2] score what a system *does* when instructed; TargetSpace scores what it can *forecast* about a target from passive observation before any instruction. *Static LLM benchmarks* [20,21,43] score single-turn correctness on fixed items; TargetSpace is longitudinal, prospective, and sealed. *Longitudinal-memory and personalization benchmarks* [7,26] score recall of, or held-out responses for, an individual; TargetSpace scores whether stored history is converted into improved *future* target-state forecasts under the permutation gate (Section I.2). Finally, *passive-observation datasets* — egocentric video [27], experience sampling [61], personal sensing [62], digital phenotyping [64] — supply the longitudinal signal TargetSpace consumes but are data resources, not a target-specificity evaluation; TargetSpace adds the R1/R2 baselines, sealing, calibration, and permutation test that make target-specificity the measured

quantity.

L Extended metrics and scoring details

This appendix carries the metric extensions and reporting details beyond the core scoring of Section 6.

Reporting and inference details

Forecasts use a pre-registered nonzero probability floor; Skill is aggregated within homogeneous answer-space and horizon strata before pooling; Brier is a secondary robustness score, ordered answer spaces may add an ordered proper score, and no composite score is used. R1's exact fitting is pre-registered: population information only, leave-one-out with respect to the scored target, rolling, with smoothing for rare classes. Any recalibration uses a separate pre-seal split, is disclosed, and is never fitted on sealed outcomes. R2's exact feature set, recency kernel, minimum history, and admission rule are pre-registered; the default recipe (marginal frequencies, persistence, recent transitions, time-of-day and day-of-week effects, and recurring calendar/deadline commitments known before the forecast) is Appendix D, organizer-fit rather than entrant-fit, and never selected on test results. The permutation gate swaps complete eligible histories rather than scrambling outcomes and is reported as an effect — true-instance Skill minus the mean permuted-instance Skill, as absolute and proportional skill loss with an interval and a pre-registered margin; its resolving power grows with the number of instances permuted, so it is directional and operational in a five-person pilot and a powered test only above a stated minimum cohort size. Uncertainty respects the repeated, serially dependent structure of forecasts within a target: within-person differences use day-blocked analysis and between-person summaries use person-clustered (cluster-bootstrap) intervals, since the person — not the individual forecast or day — is the independent unit for between-person claims; the current reference implementation reports point estimates. Calibration is assessed primarily through the calibration intercept and slope (logistic recalibration parameters) and a reliability decomposition; top-label ECE is reported only as a secondary diagnostic under a fixed pre-registered binning scheme with a bootstrap interval, since ECE is positively biased and sensitive to bin count and placement at the per-stratum sample sizes of a five-person pilot, and its pass/warn/fail bands are therefore interpreted as diagnostic bounds rather than a hard gate below a pre-registered minimum stratum size.

Multiscale consistency (proposed components)

Because levels are scored separately, we also specify cross-level checks as future evaluation components, not results: *cross-horizon consistency* (a detailed forecast should not contradict the abstract target state without explicit uncertainty or branching); *refinement* (as the horizon shortens and evidence arrives, abstract forecasts should sharpen into concrete ones); *subgoal coherence* (predicted intermediate states should form a plausible path to the higher-level target); *horizon-conditioned calibration* (confidence is evaluated separately at each horizon); *hierarchical abstention* (a system may abstain at the detailed level while keeping a calibrated abstract forecast); and *rollout drift* (repeated one-step predictions may diverge from external reality even when each step is locally accurate [76]).

L.1 Metrics for possibility-space resolution

Predictive lift by evidence tier is Skill computed as a function of the evidence rung — the evidence-ablation read-out; no new mathematics. **Top-k target-state recall** complements distribution-level Skill for large answer spaces. **Transition-path likelihood** is the length-normalized likelihood of the realized path of target-state transitions, in bits vs R1/R2 — the natural metric for the transition track. **Time-to-correct-resolution** and **false-resolution rate** measure *when* a model resolves the possibility space: respectively, how early its distribution concentrates correctly on the realized target, and how often it concentrates confidently and early on the *wrong* one. We are explicit that 'how early can a target be identified' is **not** a new idea: it descends directly from goal-recognition design and worst-case distinctiveness (WCD) [29] and from online goal recognition (recognition time, false-positive rate). Our contribution is to operationalize it as a *proper-scored, calibrated, prospective* quantity on tracked instances, paired so that a model must achieve early correct resolution *and* low false resolution together. We freeze and unit-test these definitions before claiming them and report none as results pre-pilot.

Table 16: Metric provenance. The honesty rule: claim the conjunction and the R2 + permutation anchors; concede the rest as adopted, adapted, or (for resolution-timing) a new operationalization of an existing idea.

Metric	Status	Nearest prior art (cite, do not claim)
Skill / lift / entropy reduction	adopted (information gain)	skill score; Gneiting & Raftery [15]
Log + Brier, ECE gate, sealing	adopted	[14,18,28]; calibration-as-axis [43]
R2 own-routine + permutation gate	**anchor (this work)**	own-history baselines & self-consistency [5] as cousins
Top-k recall; transition-path likelihood	adapted	retrieval; cell-fate / trajectory likelihoods [30]
Time-to-correct-resolution	operationalization, not new idea	goal-recognition design / WCD [29]; online goal recognition
False-resolution rate	new metric (concept has precedent)	premature-commitment / false-stop / FPR in goal recognition

M Open research questions and latent-model transfer

Latent world models and transfer. Latent predictors are a natural architecture class here, and their open evaluation challenges are why TargetSpace scores calibrated distributions against sealed external outcomes rather than internal compatibility: latent state can be hard to ground, deterministic prediction may under-represent uncertainty, and long-horizon rollouts can drift [76]. A coherent internal rollout is not itself evidence of competence. Whether target-specific structure is reusable across regime shifts, beyond routine and memorization, is a transfer question the same bits-Skill and splits can measure.

Open research questions. Each is a controlled comparison of target-specific Skill: **(1) evidence sufficiency** — which tier first beats R2, and which modalities add marginal skill; **(2) attention and target-state formation** — does attention-trajectory evidence add calibrated lift beyond routine and stated goals, and how early can an emerging target be detected; **(3) transfer** — does target-specific structure stay predictive across routine-breaking and withheld windows, distinct from memorization, and how fast does skill decay once evidence stops; **(4) architecture and**

horizon — do explicit consequence models calibrate better than reactive policies, and at what horizon do detailed rollouts lose to abstract forecasts; **(5) reactivity and active evidence acquisition** — can target-state formation be separated from observer-induced reactivity, and can a system request more evidence without overreaching. Recent work frames reusable, computationally-bounded structure as *epiplexity* [80]; we note it as related framing but do not adopt it as a metric, since bits-Skill already reports the prospective predictive code-length gain architecture-neutrally.

N Evaluation grid, evidence ladder, and tracks

This appendix specifies the evaluation grid summarized in Section 4: the architecture classes, the person-domain evidence ladder with its privacy-risk taxonomy, and the multi-track apparatus.

N.1 Architecture classes

The grid below does not rank architecture classes in the abstract; it has nothing to say about which class is "better" in general. Systems become comparable only because each one emits, or is adapted to emit, a calibrated probability distribution over the *same* answer space A on the *same* sealed instance, scored by the same proper rule. A reactive policy, a large language model, a vision–language model, a JEPA-style latent-predictive model, a symbolic or probabilistic system, a hybrid, and a human filling in a self-report are all evaluated at the level of the forecast output, not by inspecting internal representations.

Because the output adapter can change measured performance, it is disclosed and reported as part of the system under test. An output adapter maps a system's native output — an action, a token sequence, a latent prediction, a logical query result — to a probability distribution over A. The same underlying model paired with two different adapters is two different entries, and any tuning, temperature setting, or aggregation rule inside the adapter is part of what the score measures. We therefore require the adapter to be specified before the forecast is sealed, so the reported skill in bits over a reference reflects the system as actually run rather than a post hoc reinterpretation of its outputs.

N.2 The evidence ladder

The evidence ladder is the benchmark's measurement instrument (Figure 5). Tiers are cumulative; we give the canonical *person-domain* ladder and other domains substitute their own sensor ladders (Section 5.3), preserving the ordering from already-interpreted residue toward pre-interpretive observation.

The ladder also encodes a distinction richness alone does not (developed in Section 2): lower tiers (L0–L1) are records the target already produced, beginning *after* relevance was assigned and largely encoding routine, whereas higher passive tiers are **pre-interpretive** — captured before the target fixed what would matter, they carry **retrospective option value** (the same recording can be re-scored under questions chosen later) and are, more precisely, *independent from retrospective human interpretation* rather than 'objective'. The grid is the cross-product: each leaderboard cell names an (evidence tier, architecture class) pair, with R1, R2, human self-report, and the oracle spanning all tiers.

Departures from an individual's established routine are not, by default, noise. Research in neuroscience and complex-systems biology indicates that within-system variability can carry information about latent state and cognitive condition [67], and arises from multiple sources that need not be mere nuisance variance — some of it functional, with possible roles in adaptation or exploration

Table 17: Architecture classes. LLMs and JEPA-style models are complementary components, not rivals: TargetSpace asks whether *any* of them produces calibrated, target-specific forecasts. Human self-report and the oracle anchor the achievable range and are not ranked. The third column indicates what each class supports natively versus through a disclosed output adapter; labels are cautious, classes are not ranked, and real systems are often hybrid.

Architecture class	**What it is**	**Forecast interface (native / via adapter)**
LLM-only	text-in, distribution-out language model [20,21,43]	distribution native; explicit state model via adapter
VLM	vision-language model over image / video [31,49]	distribution via adapter; no native state model
Multimodal agent	tool-using agent with retrieval and memory [90,94]	distribution via adapter; target-specific memory possible
JEPA-style / latent predictive (incl. world models)	learns latent dynamics, predicts in representation space [33–36]	state and action-conditioned prediction native; calibrated distribution via adapter; planning by search
Symbolic / probabilistic	explicit structured or Bayesian model [88,95]	calibrated distribution and state prediction native; planning architecture-dependent
Reactive / imitation policy (incl. VLA)	maps observation + instruction to action [50,51]	action native; consequence distribution via adapter; not intrinsic
Hybrid	any combination, e.g. policy + learned subgoal or world model [38,78]	architecture-dependent; reactive and predictive components may coexist
Human self-report *(baseline)*	the target, or an expert, predicts itself [28,58]	distribution via elicitation; not ranked
Oracle upper bound *(baseline)*	a privileged predictor; approximate ceiling (benchmark construct, no external referent)	not ranked

[68]. Distinct internal configurations can also produce similar observable outputs, a property known as **degeneracy** [69] — the capacity of structurally different elements or mechanisms to perform the same function or yield the same output (not the model collapse, numerical degeneracy, or redundancy the term denotes in machine learning). In neural systems this is concrete: closely matched network activity can arise from substantially different underlying circuit parameters [70]. (These results concern neural and biological systems; we extend them only by analogy to behavioural forecasting.) Two targets may therefore reach the same state along materially different trajectories, so population averaging or endpoint-only evaluation can erase precisely the person-specific structure that signals an emerging target-state transition. TargetSpace accordingly *preserves* a departure from an individual baseline so the evidence remains available for analysis; this never alters a sealed forecast or its scored target, both fixed at the forecast time (Section 3.3). Whether a departure mattered is asked of preserved evidence in separate, clearly-labelled exploratory analysis, not a post-hoc relicensing of the primary score.

The ladder is more than a convenience stratification; it reflects different partial, mediated windows on the target. Textual records are *translated* traces — experience already rendered into language; logs and digital exhaust are *behavioural* traces; audio and video are *temporal observational* traces; and action-conditioned or interventional evidence exposes how the target responds when the world pushes back. No band is reality, and a higher tier is not automatically better — whether it

Table 18: The person-domain evidence ladder. Tiers are ordered by increasing capture intrusiveness (and, roughly, privacy sensitivity) — from already-externalized digital residue toward pre-interpretive observation; they are *not* ordered by assumed forecasting power, which is exactly what the ablation measures. By reporting target-specific skill (over R2) as a function of tier, TargetSpace turns 'does richer observation add target-specific information?' into a measured quantity rather than an argument.

Tier	Evidence (person domain; operational examples)	Primary sensitivity risks
L0	low-content behavioural metadata: calendar timestamps, communication metadata, app/event logs, coarse routine traces	social-graph and rhythm inference
L1	user-authored text: notes, chats, documents, search/query text, task lists	third-party content; workplace/confidential exposure
L2	speech and transcript layer: audio recordings, ASR transcripts, speaker turns, optional prosody	bystander capture
L3	screen and interaction context: screen recordings, UI activity, browser/app context, document-interaction traces	workplace/confidential exposure; intimate-behaviour inference
L4	egocentric / passive multimodal observation: wearable or scene video, ambient audio, images, environmental context	bystander capture; re-identification
L5	location, mobility, and physical-world traces: GPS, Wi-Fi/Bluetooth proximity, visits, commute patterns	location-trace risk; re-identification
L6	physiological / biometric / specialized sensors: heart rate, sleep, gaze, motion, wearable health signals	health and biometric inference

pays is the empirical question the ablation answers. Because the relevance of an observation may only become clear retrospectively, TargetSpace treats raw-evidence preservation, provenance, and ablation as first-class evaluation concerns: the benchmark can ask not only whether a system used evidence but *which* evidence was necessary to improve a target-specific forecast. Preservation is always bounded by the consent, federation, aggregate-only, and retention limits of Section 8; the benchmark scores evidence value, it does not license indiscriminate retention.

Domain tracks

The five domain tracks share the scoring spine (Section 5.3); readiness is stated honestly per track.

O Industry case cluster: from AI recorders to first-person memory systems

This appendix gives the analytical survey behind the motivation of Section 1. It is evidence that the capability class is arriving, not empirical validation of TargetSpace, and not a claim that any vendor misuses data; each system also has clearly beneficial uses (accessibility, memory support, productivity, care coordination, safety).

Audio-first ambient memory systems

Omi (Based Hardware) is the clearest explicit statement of the industry thesis [96]: its manifesto frames personal context as the competitive moat ("context is the new moat"), describes the product

Table 19: Evidence substrates for the personal track and their role in TargetSpace (referenced from Section 2). Self-report is retained as an auxiliary channel and a baseline, not discarded; passive capture is not claimed unbiased, only to carry externalized, measurable, partially normalizable bias. Tiers refer to the ladder above; "standardizability" is how readily the bias can be characterized and normalized across people and time.

Substrate	**What it supplies / characteristic bias**	**Standardizability of the bias**	**Role in TargetSpace**
Self-report / stated goals	declared beliefs, goals, feelings; bias endogenous — selective, post-hoc, socially filtered, introspection-limited	low; varies across people and over time	auxiliary channel + human-self-report baseline
Digital exhaust (L0–L1)	calendars, messages, clicks, notes; already-externalized, routine-heavy residue	moderate; well-defined logs	primary R2 signal; ablation floor
Passive audio (L2)	speech, prosody, turn-taking, timing; ASR error, mic placement/range, bystander capture	device-characterizable (WER, coverage)	candidate lift over exhaust/self-report
Passive audiovisual / egocentric (L3–L4)	embodied context: what is seen, where, which objects/people; FoV, occlusion, framing	device-characterizable; higher sensitivity	candidate lift; tested by ablation
Sensors & future internal-state proxies (L5–L6)	location, mobility, physiology; prospectively affective/internal-state proxies; sensor-specific error	sensor-calibratable	candidate lift; still scored on observable outcomes

as a memory loop of capture, structure, retrieval, and action rather than a recorder, and names continuity — "memory, goals, patterns" — as the requirement current prompt-response AI lacks. Bee is marketed as a low-cost, always-available wristband that turns ambient conversation into summaries, reminders, and to-do items; a reported acquisition by Amazon was announced in 2025 [97]. Limitless offers a pendant for continuous conversational memory — capture now, search and summarize later — with a reported acquisition by Meta announced in late 2025 [99]. PLAUD's press-to-record devices represent the polished episodic end of the spectrum (meetings, calls, notes) [98]; their commercial normalization matters because episodic capture builds the habits, form factors, and supply chains that ambient capture inherits. The relevant TargetSpace observation is common to all four: the product value is increasingly located in the derived, longitudinal layer (summaries, entities, commitments, routines, memory), not in the recording itself — exactly the layering of Table 21.

Audiovisual and egocentric capture

The base Ray-Ban Meta glasses put camera-and-microphone AI capture into mainstream retail distribution, with a display variant announced for 2025 [100]. Brilliant Labs' Halo is an announced open-source AI-glasses platform whose agent ("Noa") advertises cross-session narrative memory — an explicit long-term memory *claim* in a consumer visual device [101]. Snap's Specs line is announced for see-through AR with onboard perception [102]. Egocentric-video research (e.g. Ego4D and its successors [56,57]) supplies datasets, tasks, and methods for first-person capture at research scale —

Table 20: The five TargetSpace tracks, one shared spine. *Status* — **current**: apparatus exists (synthetic, pre-pilot) for the person track only; **planned**: strong sealed precedent and a natural own-routine baseline, not yet implemented (energy/grid: GEFCom [52]; physiology: PhysioNet/CinC [53], OhioT1DM/BGLP [54]; personal: GLOBEM [55]); **research**: a distinct regime exists but a proper-scored forecasting protocol and/or a strong R2 are not yet established. We make no claim that any track other than the synthetic person track is implemented, and no claim that TargetSpace solves robotics, healthcare, energy, or enterprise forecasting.

Track *(status)*	**Target object**	**Example evidence bands**	**Horizon**	**Example target states**	**Validator / outcome**
TS-Personal *(current)*	a consenting individual	metadata, text, audio, passive multimodal, location, physiology	hours–weeks	commitment status, priority shift, routine deviation, task switch	observed action / confirmation
TS-Health *(planned)*	a patient	vitals, labs, continuous glucose monitoring (CGM), wearables, notes	minutes–days	glycemic excursion, deterioration onset, care-state transition	clinical onset labels / sensor thresholds
TS-Energy *(planned)*	a series / asset	history, weather, calendar, exogenous covariates	hours–days	load / renewable level band, price regime	realized value / market settlement
TS-Robotics *(research)*	embodied agent + scene	proprioception, sensor stream, action log	sub-second–minutes	goal-conditioned configuration, subgoal transition	achieved configuration
TS-Enterprise *(research)*	a project / team / workflow	trackers, commits, comms metadata, releases	days–quarters	milestone state, scope / priority drift	observed milestone / outcome

generic activity clips, not longitudinal single-person records (Section I.1). The addition of vision is not merely more data: it adds embodied context — what the wearer sees, where they are, which objects, screens, documents, and people are present, and what actions are physically taken — which is why the personal-track evidence ladder places it at higher tiers (L3–L4) with correspondingly higher sensitivity (Appendix N).

Reading the cluster

Across both groups the trajectory is consistent: from episodic transcription, to persistent personal memory, toward anticipatory assistance — from recording what users say, to modeling what users experience, to predicting target states from longitudinal lived traces. Table 22 contrasts the traditional product framing of each layer with the TargetSpace framing under which it becomes an evaluation and governance object. TargetSpace's role is to make the top of this ladder measurable: what does longitudinal capture actually buy, over population base rates (R1) and over the target's own routine (R2), under calibration and permutation specificity? The devices surveyed here are only the most visible edge of the class: the same ladder is climbed, with different signals, by assistants and agents over email, calendar, and documents, by enterprise copilots and organizational memory, by tutors, care systems, and recommenders — any system that converts accumulated longitudinal traces into a persistent, future-relevant model of a particular target (Section 3).

Table 21: From capture to target-state inference (referenced from Sections 3 and 8). Each layer is more compressed, searchable, and actionable than the one below; TargetSpace evaluates, and the governance discussion concerns, the top layers, not only the raw signal.

Layer	**Representation**	**Example capability**	**Primary risk**
Signal	raw audio, video, screen, location, sensor traces	record or observe events	unauthorized or unnoticed capture
Text / perception	transcripts, OCR, visual labels, logs	make events searchable	searchable bystander speech and activity
Structure	speakers, entities, topics, tasks, commitments	organize experience into relations and obligations	relationship and obligation mapping
Memory	persistent longitudinal profile, embeddings, summaries, routines	retrieve and personalize over time	pattern-of-life inference
Target state	predicted goals, needs, actions, future commitments	anticipate (or steer) future behaviour	manipulation, surveillance, unwanted steering

Table 22: From retrieval to experience modeling: the same artifacts read as product features versus as TargetSpace evaluation objects.

Layer	**Traditional framing**	**TargetSpace framing**
Audio	recording	behavioural signal
Transcript	searchable text	temporal evidence
Summary	convenience artifact	compressed experience
Entities	names and topics	relationship graph
Tasks	productivity feature	commitment trajectory
Memory	personalization	longitudinal state model
Proactivity	assistant behaviour	target-state intervention
First-person video	visual record	embodied experience trace