CLOCKBRIDGE: PARTIAL IDENTIFICATION OF CROSS-MODAL DISEASE TIMING ACROSS UNPAIRED COHORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical cohorts often measure different modalities in different individuals and at different disease stages, so stage labels that appear comparable can still admit multiple incompatible cross-modal timelines. We formulate cross-modal disease timing across unpaired cohorts as a partial identification problem. ClockBridge places each cohort on its own disease clock, retains the synchronizations compatible with the observed stage-resolved evidence in a synchronization equivalence set (SES), and reports a temporal direction only when it is invariant over the resulting identified interval. Otherwise the claim remains unresolved. On hidden-pair synchronization, ClockBridge substantially lowers selective false-order risk relative to shared-protocol forced-map references and a same-loss point estimator. FO@40 is 12.3% on ADNI versus 52.2–53.5% for forced maps and 17.7% for Point-$\mathcal{L}^\star$, and 3.3% on PD versus 20.8–49.6% and 7.2%, respectively, while decision coverage at low error remains higher. Under degraded stage evidence, unresolved claims increase and unstable directions are withheld. On fully unpaired molecular and imaging cohorts, ClockBridge yields lower aggregate contradiction under cohort replacement and abstains more often on unstable claims. These results support set-valued identification as a principled alternative to forcing a single cross-modal disease timeline when the available stage evidence does not uniquely determine one.

1 INTRODUCTION

Biomedical studies increasingly observe the same disease through different experimental windows. Molecular atlases resolve cell-type programs and pathology-associated modules (Mathys et al., 2023; 2024; Gabitto et al., 2024; Adams et al., 2024), while neuroimaging cohorts measure regional structure, function, or metabolism in living subjects (Okonkwo et al., 2025). These resources rarely share participants or synchronized schedules, and the resulting question is which stage-resolved evidence supports a cross-modal *timing* claim, such as a molecular program onset preceding an imaging change.

Stage labels make the question look easier than it is. ROSMAP donors labeled Braak III–IV and ADNI participants labeled intermediate CDR are both called mid-stage disease, yet that shared vocabulary leaves room for multiple latent phases. Two synchronizations can each fit the observed stage trends inside their own cohorts while disagreeing on which modality's event occurs earlier. Forced point correspondences (Demetci et al., 2022; Cao & Gao, 2022; Bunne et al., 2023; Ashuach et al., 2023; He et al., 2024; Klein et al., 2025; Tran et al., 2023) and common-clock staging models (Salvadó et al., 2024; Therriault et al., 2024; Wijeratne et al., 2023) collapse that ambiguity to a single map, so a reported direction can rest on one of many clocks compatible with the same stage margins.

Figure 1 summarizes the setting. Compatible clocks can disagree on event order. Forced maps can report a direction even when the identified set crosses zero. Unconstrained acquisition can spend budget without shrinking the unresolved region. Existing alignment and agent pipelines address related subproblems (Cao et al., 2022; Hao et al., 2024; Klein et al., 2025; Tran et al., 2023; Samaran et al., 2024; Rainforth et al., 2024; Gao et al., 2024; Roohani et al., 2025). Those pipelines typically

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

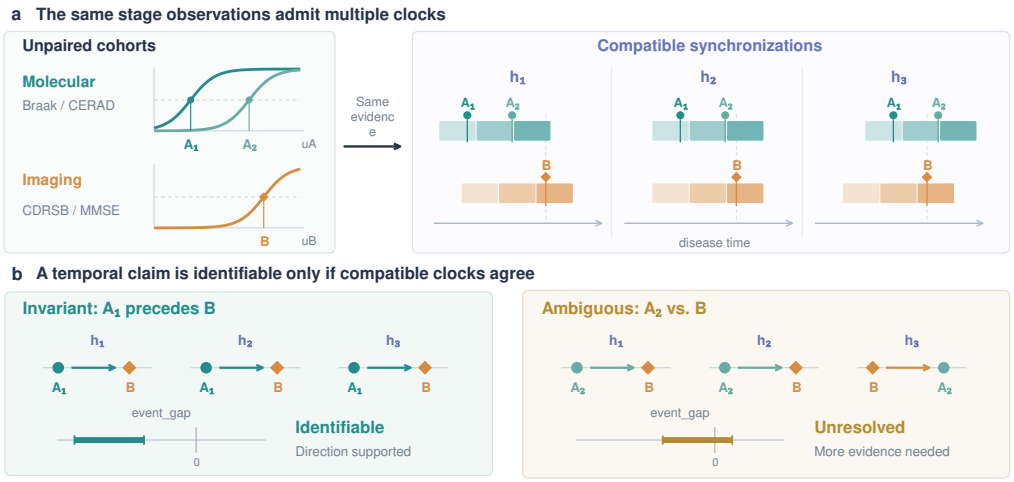


Figure 1: Problem setting for cross-modal timing across unpaired cohorts. (a) Molecular and imaging cohorts can share stage labels without sharing subjects. The same stage-resolved margins admit multiple synchronizations that disagree on which modality's event occurs earlier. (b) A directional claim is identifiable only when every compatible clock agrees on the sign of the temporal gap. Otherwise the claim is unresolved and additional evidence is required.

optimize correspondences or task completion. ClockBridge focuses on which timing claims remain invariant on the identified set (Huang et al., 2024; Chen et al., 2025).

We propose **ClockBridge** (Figure 2) for partial identification of cross-modal disease timing. It places each cohort on its own disease clock, constructs the synchronization equivalence set (SES) of maps compatible with stage-resolved evidence under the stated family, and emits a directional claim only when that claim is invariant on the identified interval. Otherwise the claim stays unresolved. While unresolved and budget remains, a numerical SES utility can select the next witness from a fixed analysis library.

Our contributions are as follows.

- We formulate unpaired cross-modal disease timing as partial identification over synchronization maps. The reporting target is claim invariance over the set of compatible clocks.
- We present ClockBridge, coupling an SES claim layer with optional budgeted resolution of residual ambiguity under a fixed numerical utility over a fixed witness library.
- Empirically, ClockBridge lowers selective false-order risk under hidden pairing, keeps FO@40 low under stage-evidence stress while raising unresolved mass under label corruption and broken stage structure, and improves aggregate claim-level stability under external cohort replacement among bidirectional predictors when evaluated with abstention.

2 Related Work

2.1 Unpaired alignment, staging, and cross-modal bridging

Optimal transport and Gromov–Wasserstein methods align unpaired modalities by matching distributions or geometry (Demetci et al., 2022; Cao et al., 2022; Samaran et al., 2024; Klein et al., 2025), with unbalanced and partially matched variants when cohorts differ in size or contamination (Tran et al., 2023). Propensity-score and spectral unpaired alignment (Xi et al., 2024; Yacobi et al., 2025), snapshot-trajectory OT with dynamics priors (Gu et al., 2025), multimodal OT toolkits (Klein et al., 2025), and single-cell mosaic / neural transport systems (Cao & Gao, 2022; Ashuach et al., 2023; Bunne et al., 2023; Hao et al., 2024; He et al., 2024; Ghazanfar et al., 2024; Wang et al., 2025; Fu et al., 2025; Liu et al., 2025) further expand the forced-map toolkit we adapt as point baselines. Biomarker staging and event-based progression models infer ordered sequences within a shared

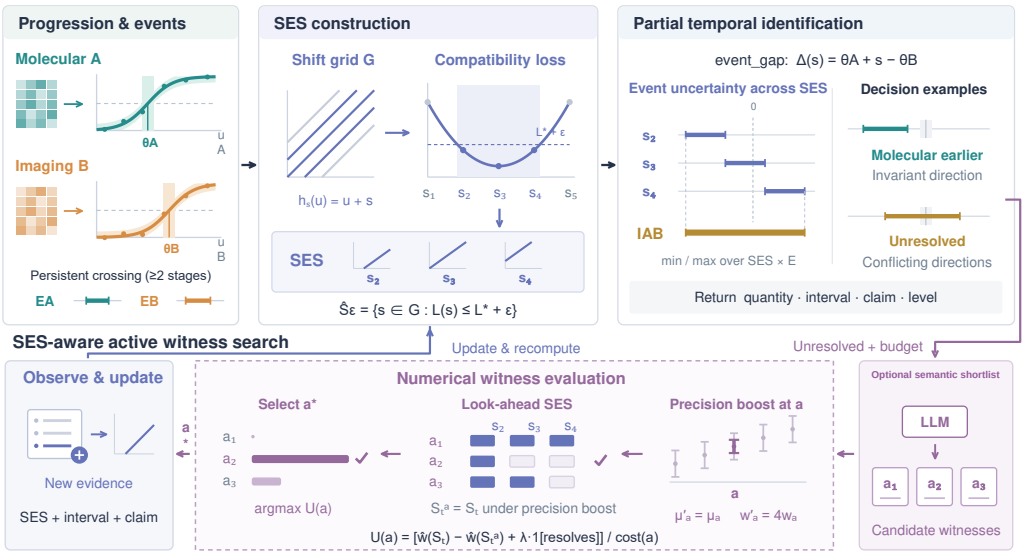


Figure 2: Overview of ClockBridge. Top: stage-guided progression curves with persistent-crossing events, SES construction on a lag grid from the stage-table loss, and partial identification of `event_gap`, which yields an invariant direction or an unresolved interval. Bottom: when the claim is unresolved, SES active witness search ranks fixed-library actions by the numerical utility $U(a)$ (optional semantic shortlist before ranking), observes the chosen witness, and recomputes the SES.

biomarker space (Salvadó et al., 2024; Therriault et al., 2024; Wijeratne et al., 2023), while molecular atlases and imaging cohorts provide complementary disease windows (Mathys et al., 2023; Gabitto et al., 2024; Adams et al., 2024; Okonkwo et al., 2025; Arnatkeviciute et al., 2023). These lines return correspondences, couplings, or common clocks. Stage labels alone still leave an underdetermined set of cross-cohort synchronizations, and that set is the object ClockBridge retains explicitly.

2.2 Partial identification, selective reporting, and evidence acquisition

Partial identification treats parameters as set-identified when data leave a non-singleton feasible set (Manski, 2003; Tamer, 2010; Kline & Tamer, 2023). Selective prediction abstains under low confidence (Wu et al., 2024; Goren et al., 2024), and causal representation / OT study identification under incomplete cross-domain assumptions (Sturma et al., 2023; Dong et al., 2023; Tan et al., 2024). ClockBridge applies set-valued reporting to unpaired disease timing. Claims are synchronization invariants under stage-resolved margins. Unresolved status follows when the identified interval crosses the decision threshold. The feasible set is a synchronization family under a stated stage-table loss. Conformal label sets and causal effect bounds address different estimands. Bayesian experimental design and scientific agents select the next measurement or tool call (Rainforth et al., 2024; Yao et al., 2023; Gao et al., 2024; Roohani et al., 2025; Huang et al., 2024; Chen et al., 2025). Our acquisition step uses a one-step SES utility over a *fixed* analysis library. Reported curves are numerical. Prior partial-identification work typically centers on scalar or causal targets. ClockBridge treats a set of cross-cohort synchronization maps as the identified object and reports only temporal claims that remain invariant over that set.

3 Method

We cast cross-modal disease timing on unpaired cohorts as partial identification over compatible synchronizations, with optional budgeted resolution of residual ambiguity (Figures 1–2). Local clocks place each cohort on its own disease coordinate. Stage-resolved margins induce a synchronization equivalence set (SES). A claim layer emits a directional statement only when that statement is invariant on the identified interval, and otherwise records unresolved or not-evaluable status. When

residual ambiguity remains and budget is available, a numerical SES utility ranks the next witness from a fixed analysis library, updates the SES, and stops on resolution, budget exhaustion, or non-positive utility. Adjudication is numerical throughout; semantic shortlisting of library candidates is optional and unused in the reported acquisition curves.

3.1 TEMPORAL CLAIMS

A cross-modal timing claim is a tuple $(q, I, \mathcal{E})$ where q is a named temporal quantity (`sync_shift` or `event_gap`), $I = [\mathrm{lo}, \mathrm{hi}]$ is the identified interval under the current SES, and $\mathcal{E}$ is the stage-resolved evidence so far. The claim layer maintains a small decision set and records four outcomes.

- **Direction:** every compatible synchronization agrees on the sign of q outside a frozen tolerance δ (molecular earlier or imaging earlier).
- **Near-simultaneous:** I lies inside $[-\delta, \delta]$ with sufficient sharpness.
- **Unresolved:** I crosses $\pm\delta$, so compatible clocks disagree on direction.
- **Not-evaluable:** evidence is insufficient to form a stable I.

Reported directions are restricted to invariants of I. Levels L0–L3 refine an already identified claim by interval width ($\eta_2 > \eta_3$); they summarize sharpness and do not create a new identifying assumption. Outputs always carry the quantity tag, interval, claim, and level, so later tables can separate selective risk from forced point maps.

3.2 PROBLEM SETUP

Let modality/cohort A (e.g., molecular) and B (e.g., imaging) provide stage-resolved features. Stage proxies (Braak, CDR-SB, DX bins, ...) induce *within-cohort* order only: they organize donors inside one study, while a shared absolute clock across studies is additional structure (Bennett et al., 2018; Weiner et al., 2015; Gabitto et al., 2024; Therriault et al., 2024). We therefore fit each cohort independently by soft stage-conditioned phase regression, yielding local coordinates $u_A, u_B \in [0, 1]$ with the locked convention that larger disease time means later.

A synchronization is a monotone map $h : u_A \mapsto u_B$. The main text uses translations

$$h(u) = u + \mathrm{sync_shift}, \tag{1}$$

searched on a frozen lag grid. Richer families (e.g., piecewise-linear) appear in the appendix, where additional flexibility mainly enlarges the unresolved set under the same reporting rule. The scientific object is not a single forced correspondence but the set of maps still compatible with the observed stage-resolved margins,

$$\mathcal{S}(\mathcal{D}) = \big\{h \in \mathcal{H} : \ h \text{ is compatible with } \mathcal{D}\big\}. \tag{2}$$

Partial identification asks which timing statements remain constant for every $h \in \mathcal{S}(\mathcal{D})$.

3.3 PROGRESSION CURVES AND EVENT UNCERTAINTY

Given local clocks, we form stage-guided progression curves μ_{mj}, optionally with donor-bootstrap bands. When an event path is enabled, an onset θ is the *first persistent threshold crossing* of the corresponding curve, with locked persistence of one consecutive stage step on the stage grid (Wijeratne et al., 2023; Salvadó et al., 2024). Persistence is a discrete consecutive-crossing requirement rather than a continuous dwell-time model; Appendix D.2 reports sync-family and onset checks under the same locked persistence. Event uncertainty E collects admissible onsets under bootstrap variability. The identified interval for event-order claims then ranges over both synchronizations and events, so residual onset ambiguity is absorbed into unresolved mass instead of being collapsed to a point.

Curve-alignment benchmarks that score relative lag without named onsets skip this step and use only the synchronization parameter (`sync_shift`). External molecular–imaging analyses use named onsets and score `event_gap`.

3.4 SYNCHRONIZATION EQUIVALENCE SET

Compatibility is scored by a stated stage-table alignment loss. Let $\bar{x}_A(s)$ and $\bar{x}_B(s)$ denote stage-conditional feature means, and let $\tilde{x}_B(\cdot; s)$ denote B's stage table after a candidate lag $s = \text{sync_shift}$. With normalized stage weights $\tilde{w}_k$ (uniform by default),

$$\mathcal{L}(s) = \sum_k \tilde{w}_k \left\| \bar{x}_A(s_k) - \tilde{x}_B(s_k; s) \right\|_2^2. \tag{3}$$

This $\mathcal{L}$ is the compatibility score used throughout; the claim layer asks which directional statements remain constant on the ε-sublevel set of $\mathcal{L}$. Computationally we form a finite ε-SES on a frozen lag grid G (main setting: $[-0.45, 0.45]$, step 0.03),

$$\widehat{\mathcal{S}}_\varepsilon = \left\{ s \in G : \ \mathcal{L}(s) \le \mathcal{L}^\star + \varepsilon \right\}, \tag{4}$$

with $\mathcal{L}^\star = \min_{s \in G} \mathcal{L}(s)$ and locked ε=0.012 chosen on train/dev. $\widehat{\mathcal{S}}_\varepsilon$ is a finite computational object under the stated sync family $\mathcal{H}$. Keeping $\mathcal{H}$ intentionally simple attributes unresolved mass to evidence ambiguity under that family. Under fixed $\mathcal{H}$, $\mathcal{L}$, grid G, and ε, a direction is emitted only when every lag in $\widehat{\mathcal{S}}_\varepsilon$ shares the same sign outside δ, so reported directions are constant on the computational SES by the reporting rule. Finite-sample coverage of $\widehat{\mathcal{S}}_\varepsilon$ for a population set, and consistency under misspecified $\mathcal{H}$ or ε, remain open. Empirical sensitivity to ε and sync family appears in Appendix C–D.2.

3.5 PARTIAL TEMPORAL IDENTIFICATION

From $\widehat{\mathcal{S}}_\varepsilon$ we build a single identified interval $I = [\text{lo}, \text{hi}]$ for a named quantity, reporting invariants of the identified set (Manski, 2003; Kline & Tamer, 2023). The two quantities share the SES machinery and enter separate claim tables.

Named quantities. For `sync_shift`, I is the SES lag range (used in Table 1 and Figure 3). For `event_gap`, with onsets $\theta_A \in E_A$ and $\theta_B \in E_B$,

$$\Delta_{AB}(h) = h(\theta_A) - \theta_B. \tag{5}$$

Under translation, the identified range extremizes over SES lags and event bands,

$$\Delta_{\min} = \theta_{A,\text{lo}} + \text{shift}_{\text{lo}} - \theta_{B,\text{hi}}, \quad \Delta_{\max} = \theta_{A,\text{hi}} + \text{shift}_{\text{hi}} - \theta_{B,\text{lo}}, \tag{6}$$

and $I = [\Delta_{\min}, \Delta_{\max}]$. External cohort analyses in Figure 4 and Table 2 use `event_gap`; hidden-pair scoring uses `sync_shift`.

Sign convention and selective reporting. Larger disease time means later, so $\Delta < 0 \Leftrightarrow$ the molecular event is earlier. With sign tolerance δ (frozen on train/dev), the claim tables are opposite for the two quantities: for `sync_shift`, $\text{hi} \le -\delta$ means imaging_earlier and $\text{lo} \ge +\delta$ means molecular_earlier; for `event_gap` the mapping is reversed. If I crosses $\pm\delta$, the claim is unresolved. If support is insufficient to form I, the status is not-evaluable. This selective rule follows the abstention principle of selective prediction under ambiguity (Wu et al., 2024; Goren et al., 2024): a method is not rewarded for guessing a direction whenever the identified set is sign-crossing. Forced-map baselines, by contrast, return a point `sync_shift` and therefore always emit a direction on the same interface.

3.6 ACTIVE TEMPORAL WITNESS SEARCH

When the claim is unresolved and budget remains, ClockBridge selects an action a from a *fixed* witness library by a one-step SES utility (Chaloner & Verdinelli, 1995; Rainforth et al., 2024; Foster et al., 2021),

$$U(a) = \frac{\widehat{w}(\mathcal{S}_t) - \widehat{w}(\mathcal{S}_t^{\text{prec}(a)}) + \lambda \cdot \mathbf{1}[\text{resolves}]}{\text{cost}(a)}. \tag{7}$$

Here $\widehat{w}(\cdot)$ is SES (or identified-interval) width, and resolves indicates that the look-ahead set becomes sign-determined. The look-ahead $\mathcal{S}_t^{\text{prec}(a)}$ uses a **precision boost**: stage means stay fixed,

while the weight of a targeted stage is multiplied by locked $\kappa{=}4$ in $\mathcal{L}$, approximating a reduction in that stage's SE ($w \propto 1/\mathrm{SE}^2$ implies $\kappa{=}4$ halves the SE). The surrogate scores which library analysis would most shrink the unresolved region under the stated loss, using stage means, SES width, and a locked precision-boost look-ahead. Latent disease time and sealed pairing are unused inside $U(a)$.

The witness library is heterogeneous and costed (typical episode: 3–8 actions with integer costs in $\{1, \ldots, 4\}$). The controller proceeds as

$$\text{state} \to \text{candidates} \to a^\star = \arg\max U(a) \to \text{observe} \to \text{update SES}, \tag{8}$$

and stops on resolution, budget exhaustion, or non-positive utility. Ranking and stopping are numerical on the full library; an optional semantic shortlist is not used in the reported curves (Appendix F).

In full, the pipeline fits clocks and events, estimates $\widehat{\mathcal{S}}_\varepsilon$ equation 4, builds I for the named quantity, emits claim and level, and—while unresolved and budget remains—acquires $a^\star$ by equation 7 and recomputes. Local clocks define what may be synchronized, the SES retains every compatible map, the claim layer records which timing statements those maps license, and active search spends budget only when the identified set still crosses the decision threshold.

4 Experiments

We evaluate whether set-valued synchronization with selective reporting lowers false-order risk under hidden pairing, whether unresolved mass rises when stage evidence is thinned or corrupted, whether directional claims remain more stable under external cohort replacement among bidirectional predictors, and whether SES width and abstention—rather than point lag selection alone—drive those gains. Budgeted SES witness selection on a shared bank is reported in Appendix F. Unless noted, hidden-pair summaries use 12 seeds; external replacement uses matched claim pools of size $K{=}63$.

4.1 Evaluation protocol

Datasets and scoring interfaces. Hidden-pair synchronization uses ADNI fMRI–MRI morphology with exact `VISCODE` matches and PD FDG–CFT under subject co-enrollment, scoring `sync_shift` after fit. Evidence-stress experiments reuse the ADNI split. External claim stability uses unpaired molecular–imaging margins (ROSMAP / SEA-AD with ADNI2 / ADNI3), scoring `event_gap` under cohort replacement. Stage proxies define within-cohort clocks and the SES loss $\mathcal{L}$, but never serve as cross-modal claim labels. Hidden pairing is revealed only for scoring Table 1 / Figure 3 and is held out from fit, SES construction, and claim selection. External cohorts are not used to re-fit then rescore the same subjects.

Baselines. Shared-protocol forced-map references return a point `sync_shift` under the same stage-resolved feature interface: PSA-OT adapts propensity-score OT matching (Xi et al., 2024); SUE adapts spectral unpaired embedding alignment (Yacobi et al., 2025); PO-MFL adapts adjacent-snapshot OT with a constant-velocity lag prior (Gu et al., 2025); and FGW-style is a POT fused Gromov–Wasserstein surrogate inspired by multimodal OT (Klein et al., 2025), not the official `moscot` release. Comparisons concern these implementations, not verbatim reproductions of the source papers. Adaptation hyperparameters are selected on train/dev under the shared interface and then frozen.

Metrics and locks. Primary selective metrics are FO@40 (false-order risk at matched 40% decision coverage) and AURC (area under the risk–coverage curve), together with Ord. / FO on ordered pairs whose true gap exceeds δ, and Cov@FO15 in the appendix. External analyses report supported / contradicted / abstained shares at matched K, plus bidirectional capacity. Hyperparameters ($\varepsilon{=}0.012$, $\delta{=}0.04$, $\eta_2{=}0.2$, $\eta_3{=}0.08$, $\kappa{=}4$, $\lambda{=}0.35$) are frozen on train/dev before external replacement; metric denominators and adaptation details appear in Appendix A–C.

4.2 HIDDEN-PAIR SYNCHRONIZATION ACCURACY

Pairing is hidden during fitting and revealed only for scoring. We emphasize selective false-order metrics because accuracy without coverage control rewards methods that guess whenever evidence is thin, consistent with selective prediction evaluations (Wu et al., 2024; Goren et al., 2024). FO@40 reports false-order risk at a matched 40% decision coverage, and AURC integrates risk across the coverage curve, so methods with different unresolved rates remain comparable at a fixed decision mass.

Table 1 summarizes ADNI and PD under this protocol. On ADNI, ClockBridge attains FO@40 12.3% versus 52.2–53.5% for forced-map references, AURC 0.147 versus 0.522–0.537, and order accuracy 72.5% (MAE 0.156 versus 0.203–0.277). Matched risk–coverage summaries further give Cov@FO15 of 55.1% for ClockBridge versus 0.1–0.2% for forced maps on ADNI (38.8% for Point-$\mathcal{L}^\star$), and 99.2% versus 0.3–26.2% on PD (80.5% for Point-$\mathcal{L}^\star$; Appendix C): ClockBridge sustains a low false-order regime at usable coverage, while forced maps only approach that risk after coverage collapses. On PD FDG–CFT, ClockBridge similarly leads (FO@40 3.3% versus 20.8–49.6% for protocol adaptations and 7.2% for Point-$\mathcal{L}^\star$; AURC 0.043; Order 86.3%), providing a second disease setting for scoring `sync_shift`. Seed-level bootstrap intervals on ADNI place ClockBridge FO@40 in $[8.3, 16.5]$ and AURC in $[0.111, 0.183]$, separated from forced-map references whose FO@40 intervals sit entirely above 46% (Appendix C). Point-$\mathcal{L}^\star$ uses the identical stage-table loss $\mathcal{L}$, lag grid, features, and predict map as ClockBridge, but without SES-based interval reporting, confidence scaling, or abstention. Identical MAE with higher FO@40 and AURC shows that the underlying point predictor is unchanged. We therefore interpret Table 1 as evidence for the full SES-based selective-reporting mechanism, rather than as an improvement in the underlying point predictor alone (cf. Table 3).

Table 1: Hidden-pair synchronization on ADNI and PD FDG–CFT (12 seeds). Lower is better except Ord. Ord/FO: ordered pairs with true gap $>\delta$; @40=FO@40 and ARC=AURC at fixed coverage. Forced maps return point `sync_shift`; Point-$\mathcal{L}^\star$ shares $\mathcal{L}$ and the predict map with ClockBridge without SES interval / confidence / abstention.

	ADNI					PD				
Method	MAE	Ord	FO	@40	ARC	MAE	Ord	FO	@40	ARC
PSA-OT	0.203	48.1	39.7	53.5	0.524	0.157	71.3	19.5	20.8	0.211
SUE	0.221	49.3	39.4	52.2	0.522	0.173	54.7	34.5	42.0	0.421
PO-MFL	0.277	48.1	29.2	52.8	0.523	0.359	70.1	19.7	25.8	0.261
FGW-style	0.232	47.3	42.6	52.7	0.537	0.187	52.1	38.5	49.6	0.500
Point-$\mathcal{L}^\star$	0.156	69.7	25.0	17.7	0.194	0.107	80.9	15.9	7.2	0.096
ClockBridge	**0.156**	**72.5**	**19.2**	**12.3**	**0.147**	**0.107**	**86.3**	**8.8**	**3.3**	**0.043**

4.3 ROBUSTNESS TO LIMITED AND CORRUPTED STAGE EVIDENCE

We next ask whether selective risk remains controlled when stage evidence is thinned or corrupted on the same ADNI hidden-pair protocol as Table 1. Figure 3(a) decreases stage-stratified training coverage while keeping the full stage vocabulary, and reports FO@40 at fixed 40% decision coverage. Figure 3(b) corrupts stage labels and partitions *all* scored pairs into correct, unresolved, and false outcome shares (decision coverage = correct+false). These all-pair shares use a different denominator from Table 1's Ord., which conditions on ordered pairs with true gap $>\delta$ (Appendix A.3).

On the coverage grid, ClockBridge FO@40 is $12.3/10.3/4.6/7.0/7.5/5.8/15.5\%$ at $100/90/80/70/60/50/40\%$ train coverage, remaining well below the forced-map cluster near $\sim$45–52% at every operating point (best forced-map reference $\approx 46\%$ even at 40% train coverage). Thus thinning the training stages does not erase the selective-risk gap to point maps. Under label corruption, the unresolved share rises to approximately 95%: when stage structure is no longer trustworthy, ClockBridge withholds directional claims instead of continuing to emit forced orders. The zero-stress FO@40 shares the seeds and protocol of the ADNI block of Table 1, so panel (a) continues the same operating point under stress rather than introducing a second evaluation contract.

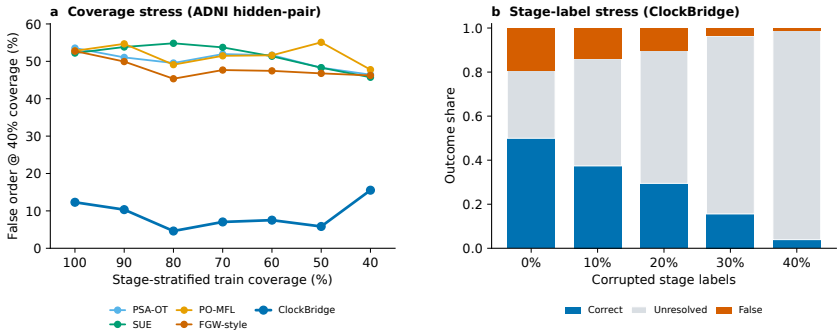


Figure 3: ADNI stage-evidence stress (12 seeds). (a) FO@40 versus stage-stratified train coverage at fixed 40% decision coverage. (b) All-pair correct / unresolved / false shares under stage-label corruption (not Ord. from Table 1).

4.4 Claim stability under external cohort replacement

We then move from sealed pairing to unpaired molecular–imaging claims scored by `event_gap`. This section evaluates *claim-level stability under cohort replacement* on unpaired margins: whether directional claims remain stable when ROSMAP is replaced by SEA-AD, ADNI2 by ADNI3, or both sides are replaced. Figure 4(a) compares bidirectional capacity against external contradiction at matched $K{=}63$. PSA-OT / SUE / FGW-style collapse to unidirectional capacity on this pool, so a low contradiction rate would be vacuous without bidirectional mass; the bidirectional comparison therefore focuses on ClockBridge versus PO-MFL. Figure 4(b) and Table 2 report supported, contradicted, and abstained shares under each replacement. Figure 4(c) destroys stage structure on otherwise identical inputs as a negative control.

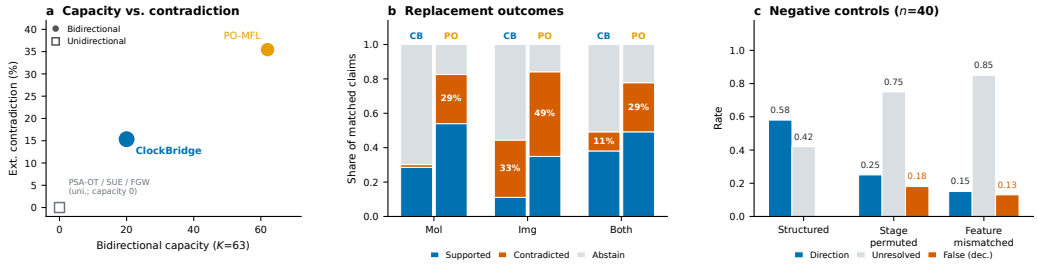


Figure 4: Claim-level stability under cohort replacement (`event_gap`). (a) Bidirectional capacity versus aggregate external contradiction at matched candidate pool $K{=}63$ (read with abstention / conditional rates in Appendix E.2). (b) Supported / contradicted / abstained shares over 63 matched claims per setting. (c) Negative controls ($n{=}40$): breaking stage structure raises unresolved mass relative to structured inputs.

Across the three replacement settings, ClockBridge contradicts $29/189$ matched claims (15.3% aggregate), compared with $67/189$ (35.4%) for PO-MFL. That aggregate gap is inseparable from selectivity: ClockBridge abstains on $111/189$ claims versus $35/189$ for PO-MFL, so fewer directional decisions enter the contradiction count by design. Conditioning on decided claims only, contradiction falls from 43.5% (PO-MFL) to 37.2% (ClockBridge); the remaining gap is smaller than the aggregate contrast, confirming that selective reporting—not only lower conditional error—drives the headline reduction (Appendix E.2). At matched decision counts obtained by ranking within each block by development confidence, contradiction remains lower for ClockBridge ($m{=}20$: 38.4% vs. 81.7%; $m{=}25$: 35.1% vs. 72.0%). Molecular replacement shows the clearest joint pattern (contradicted 1.6% vs. 28.6%; abstention 69.8% vs. 17.5%), and replacing both sides likewise favors ClockBridge (11.1% vs. 28.6%). Imaging-only replacement is the harder shift: ClockBridge still records lower aggregate contradiction (33.3% vs. 49.2%) while abstaining more often when the iden-

tified set is unstable. Matched-K fixes the *candidate* pool size (K=63 per block). It does not fix the number of directional decisions each method emits. External contradiction must therefore be read jointly with decision coverage and abstention, consistent with selective prediction elsewhere in the paper. Under stage permutation and feature mismatch, unresolved rates rise to 0.75–0.85, whereas structured inputs keep a 0.58 direction rate with zero false among decided cases (Figure 4(c); Appendix E). The claim layer therefore tracks stage structure: breaking labels increases unresolved mass rather than preserving a forced directional map.

Table 2: External claim-level stability under cohort replacement (K=63). Entries are % of matched claims.

Replacement	Method	Supported	Contradicted	Abstained
Mol: ROSMAP→SEA-AD	PO-MFL	54.0	28.6	17.5
	ClockBridge	28.6	**1.6**	69.8
Img: ADNI2→ADNI3	PO-MFL	34.9	49.2	15.9
	ClockBridge	11.1	**33.3**	55.6
Both replaced	PO-MFL	49.2	28.6	22.2
	ClockBridge	38.1	**11.1**	50.8

4.5 Ablation of partial-identification components

Finally, we ablate SES width, point estimation from $\mathcal{L}^\star$, progression uncertainty, and unresolved reporting on 120 stress and 120 near-boundary tasks. False rates are fractions of all tasks that emit an incorrect directional claim, so unresolved outcomes count as non-false and must be read with coverage. Near the identification boundary, forcing a direction raises false-direction rate from 3.3% to 65.0%, and reporting only the $\mathcal{L}^\star$ lag yields 55.8% (Table 3). Full ClockBridge keeps near-zero false on the stress suite and low false on near-boundary tasks by resolving only when the identified set agrees. Tight SES collapses calibration ($86.7\% \to 42.5\%$ interval coverage) and raises near-boundary false to 27.5%. Removing progression uncertainty has a milder effect, and a richer piecewise-linear sync family mainly enlarges unresolved mass (Appendix D.2). The ablations isolate the SES claim layer and unresolved reporting as the main drivers of the selective-risk gains in Table 1 and Figure 3.

Table 3: Ablation on stress and near-boundary suites (n=120 each). Point-$\mathcal{L}^\star$ reports the minimizing lag; Force-all emits a direction on every task.

	Stress		Near-boundary	
Variant	False↓	Int.cov↑	False↓	Unres.
Full ClockBridge	**0.0**	**86.7**	**3.3**	92.5
Tight SES ($\varepsilon \downarrow$)	2.5	42.5	27.5	50.0
Point-$\mathcal{L}^\star$ only	14.2	—	55.8	19.2
No progression unc.	**0.0**	80.8	6.7	87.5
Force-all directions	5.0	—	65.0	0.0

5 Conclusion

We study cross-modal disease timing on unpaired cohorts, where shared stage vocabulary leaves multiple compatible synchronizations. ClockBridge places each cohort on its own disease clock, retains a finite SES under a stated sync family, and reports a temporal direction only when it is invariant on the identified interval. Otherwise the claim remains unresolved, with an optional numerical SES utility for residual ambiguity. Empirically, selective SES reporting lowers false-order risk under hidden pairing relative to forced maps and the same-loss Point-$\mathcal{L}^\star$ baseline, raises unresolved mass under degraded stage evidence, and improves claim stability under cohort replacement at aggregate, conditional, and matched decision counts. ClockBridge does not claim a unique biological timeline or causal order from unpaired cohorts. External experiments test claim stability under replacement, while hidden pairing supplies sealed timing supervision.

AI USE STATEMENT

We used generative AI tools during preparation of this manuscript in ways that match the ICLR 2027 submission checkboxes. Specifically, AI assistance was used to aid and polish writing, to assist literature retrieval and discovery of related work, and to support research ideation and technical assistance during research development. All AI-assisted text and research outputs were reviewed and verified by the authors, who take full responsibility for the scientific content and for the final submitted paper. Separately, optional semantic shortlisting of fixed-library witnesses (Section 3.6) is a method component; reported acquisition curves use the numerical controller on the full library.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. Experiments use publicly released neuroimaging and molecular cohorts cited in the text (ADNI, ROSMAP, SEA-AD, and related atlases), plus a co-enrolled PD FDG–CFT cohort for hidden-pair scoring only. We collected no new patient samples. ClockBridge is a research method for selective reporting under unpaired stage evidence and is not intended for clinical decision making.

REPRODUCIBILITY STATEMENT

An anonymous method-core package accompanies this submission (local-clock fitting, SES construction, claim reporting, numerical witness acquisition, selective-risk helpers, and locked hyperparameter contracts). Protocol-adapted baselines and the full evaluation harness will be released as well. Dataset details, metric denominators, locked splits, and baseline adaptations are documented in the appendix. Where third-party cohorts impose access constraints, we provide access pointers so that results can be regenerated from the public releases.

REFERENCES

Levi Adams, Min Kyung Song, Samantha Yuen, Yoshiaki Tanaka, et al. A single-nuclei paired multiomic analysis of the human midbrain reveals age- and parkinson's disease-associated glial changes. *Nature Aging*, 4:364–378, 2024. doi: 10.1038/s43587-024-00583-6.

Aurina Arnatkeviciute, Ross D. Markello, Ben D. Fulcher, Bratislav Misic, and Alex Fornito. Toward best practices for imaging transcriptomics of the human brain. *Biological Psychiatry*, 93(5): 391–404, 2023. doi: 10.1016/j.biopsych.2022.10.016.

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, et al. MultiVI: Deep generative model for the integration of multimodal data. *Nature Methods*, 20:1222–1231, 2023. doi: 10.1038/s41592-023-01909-9.

David A. Bennett et al. Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's Disease*, 2018.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20: 1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Kai Cao, Qiyu Gong, Yiguang Hong, and Lin Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13:7419, 2022. doi: 10.1038/s41467-022-35094-8.

Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40:1458–1466, 2022. doi: 10.1038/s41587-022-01284-4.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022. doi: 10.1089/cmb.2021.0446.

Mingze Dong, Bao Wang, Jessica Wei, et al. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20:1769–1779, 2023. doi: 10.1038/s41592-023-02040-5.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395, 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Shaliu Fu, Shuguang Wang, Duanmiao Si, et al. Benchmarking single-cell multi-modal data integrations. *Nature Methods*, 22:2437–2448, 2025. doi: 10.1038/s41592-025-02737-9.

Mariano I. Gabitto et al. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, 27:2366–2383, 2024. doi: 10.1038/s41593-024-01774-5.

Shanghua Gao et al. Empowering biomedical discovery with AI agents. *Cell*, 187:6125–6151, 2024. doi: 10.1016/j.cell.2024.09.022.

Shila Ghazanfar, Carolina Guibentif, and John C. Marioni. Stabilized mosaic single-cell data integration using unshared features. *Nature Biotechnology*, 42:284–292, 2024. doi: 10.1038/s41587-023-01766-z.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Anming Gu, Edward Chien, and Kristjan Greenewald. Partially observed trajectory inference using optimal transport and a dynamics prior. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/69a076724e7228aba0272305bb98727e-Abstract-Conference.html.

Yuhan Hao, Tim Stuart, Madeline H. Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42:293–304, 2024. doi: 10.1038/s41587-023-01767-y.

Zhen He, Shuofeng Hu, Yaowen Chen, et al. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology*, 42:1594–1605, 2024. doi: 10.1038/s41587-023-02040-y.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309, 2024.

Dominik Klein et al. Mapping cells through time and space with moscot. *Nature*, 638:1065–1075, 2025. doi: 10.1038/s41586-024-08453-2.

Brendan Kline and Elie Tamer. Recent developments in partial identification. *Annual Review of Economics*, 15:125–150, 2023. doi: 10.1146/annurev-economics-051520-021124.

Katarzyna Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5e07476b6bd2497e1fbd11b8f0b2de3c-Abstract-Conference.html.

Chunlei Liu, Sichang Ding, Hani Jieun Kim, et al. Multitask benchmarking of single-cell multimodal omics integration methods. *Nature Methods*, 22:2449–2460, 2025. doi: 10.1038/s41592-025-02856-3.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Hansruedi Mathys, Zhuyu Peng, Carles A. Boix, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell*, 186(20): 4365–4385.e27, 2023. doi: 10.1016/j.cell.2023.08.039.

Hansruedi Mathys et al. Single-cell multiregion dissection of alzheimer's disease. *Nature*, 632: 858–868, 2024. doi: 10.1038/s41586-024-07606-7.

Ozioma C Okonkwo, Monica Rivera-Mindt, Michael W Weiner, and Alzheimer's Disease Neuroimaging Initiative. Alzheimer's disease neuroimaging initiative: Two decades of pioneering alzheimer's disease research and future directions. *Alzheimer's & Dementia*, 21(1):e14186, 2025. doi: 10.1002/alz.14186.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

Yusuf Roohani, Andrew H. Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html.

Gemma Salvadó, Kanta Horie, Nicolas R. Barthélemy, Jacob W. Vogel, Alexa Pichet Binette, Charles D. Chen, Andrew J. Aschenbrenner, Brian A. Gordon, Tammie L. S. Benzinger, David M. Holtzman, John C. Morris, Sebastian Palmqvist, Erik Stomrud, Shorena Janelidze, Rik Ossenkoppele, Suzanne E. Schindler, Randall J. Bateman, and Oskar Hansson. Disease staging of alzheimer's disease using a csf-based biomarker model. *Nature Aging*, 4(5):694–708, 2024. doi: 10.1038/s43587-024-00599-y.

Jules Samaran, Gabriel Peyré, and Laura Cantini. scConfluence: Single-cell diagonal integration with regularized inverse optimal transport on weakly connected features. *Nature Communications*, 15:7762, 2024. doi: 10.1038/s41467-024-51382-x.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, 2023.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Consistency of neural causal partial identification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Joseph Therriault, Suzanne E. Schindler, Gemma Salvadó, Tharick A. Pascoal, Andréa Lessa Benedet, Nicholas J. Ashton, Thomas K. Karikari, et al. Biomarker-based staging of Alzheimer disease: rationale and clinical applications. *Nature Reviews Neurology*, 20:232–244, 2024. doi: 10.1038/s41582-024-00942-2.

Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced CO-optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10006–10016, 2023. doi: 10.1609/aaai.v37i8.26193.

Gefei Wang, Jia Zhao, Yingxin Lin, et al. scMODAL: A general deep learning framework for comprehensive single-cell multi-omics data alignment with feature links. *Nature Communications*, 16:4994, 2025. doi: 10.1038/s41467-025-60333-z.

Michael W. Weiner et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 2015.

Peter A. Wijeratne, Arman Eshaghi, William J. Scotton, Maitrei Kohli, Leon Aksman, Neil P. Oxtoby, Dorian Pustina, John H. Warner, Jane S. Paulsen, Rachael I. Scahill, Cristina Sampaio, Sarah J. Tabrizi, and Daniel C. Alexander. The temporal event-based model: Learning event timelines in progressive diseases. *Imaging Neuroscience*, 1, 2023. doi: 10.1162/imag_a_00010.

Yu-Chang Wu, Shen-Huan Lyu, Haopu Shang, Xiangyu Wang, and Chao Qian. Confidence-aware contrastive learning for selective classification. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53706–53729, 2024. URL https://proceedings.mlr.press/v235/wu24s.html.

Johnny Xi, Jana Osea, Zuheng Xu, and Jason Hartford. Propensity score alignment of unpaired multimodal data. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ff56c8dc4c61017eab9771481b7fda0d-Abstract-Conference.html.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/429b76977aa377ed6dff3b91e9163b40-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

APPENDIX

A IMPLEMENTATION DETAILS

This appendix expands the experimental interfaces that the main text only sketches. The goal is operational clarity. A reader should be able to tell what is frozen before evaluation, what a reported claim means, and which quantities unpaired stage evidence leaves open. We therefore separate (i) computational objects (clocks, SES, utilities), (ii) scoring conventions, and (iii) protocol isolation. All design choices below are locked on train/dev before external replacement and before acquisition evaluation.

A.1 LOCAL CLOCKS, SES, AND ACTIVE UTILITY

Why local clocks. Cross-cohort stage labels can look comparable (for example intermediate Braak versus CDR bins) while leaving room for multiple absolute timelines. We therefore fit each modality/cohort independently. Soft stage-conditioned phase regression maps features to a disease coordinate $u \in [0, 1]$ (larger = later). Concretely, (i) z-score features with train-split medians/SDs, (ii) form stage-bin centroids in feature space (bins with fewer than two donors skipped), and (iii) assign each query a soft weight over centroids by a temperature-scaled squared Euclidean kernel (T=0.5), then take the weighted average of centroid phase targets $u_k = (k+1/2)/K$ for K stage bins. No cross-modal pairing enters this fit. Stage proxies induce *within-cohort* order only. This is the modeling premise behind partial identification: many synchronizations can fit the same stage-resolved margins.

Events. When an event path is enabled, onset is the first persistent threshold crossing of a stage-resolved curve (locked persistence =1 stage step on the discrete stage grid). Donor-bootstrap bands around the curve may be retained. The identified interval then ranges over both SES lags and event uncertainty. Appendix D.2 reports sync-family and onset-threshold sensitivity under this locked persistence.

SES as a set. The main synchronization family is translation $h(u) = u + \text{sync_shift}$. We search a frozen lag grid $[-0.45, 0.45]$ with step 0.03 and retain

$$\widehat{\mathcal{S}}_\varepsilon = \big\{\text{sync_shift} : \ \mathcal{L}(\text{sync_shift}) \le \mathcal{L}^\star + \varepsilon\big\},$$

with locked ε=0.012. $\widehat{\mathcal{S}}_\varepsilon$ is a finite computational object under the stated sync family. We keep the family simple in the main experiments so that unresolved mass reflects evidence ambiguity under that class. Piecewise-linear proxies appear as a robustness comparison (Appendix D.2).

Claims and abstention. From $\widehat{\mathcal{S}}_\varepsilon$ we build an identified interval $I = [\text{lo}, \text{hi}]$ for a named quantity (`sync_shift` or `event_gap`). Sign tolerance δ and resolution widths η_2, η_3 are frozen before external evaluation. If I crosses $\pm\delta$, compatible synchronizations disagree on direction and the claim is unresolved. If support is insufficient, the status is not_evaluable. Reporting a direction only when the entire identified set agrees is the operational content of partial identification in this paper.

Active utility. When unresolved and budget remains, action a is ranked by a one-step SES utility that trades expected width reduction against cost. Precision boost keeps stage means fixed and up-weights a stage in the loss (κ=4), approximating a reduction in that stage's SE without inventing new subjects. Latent disease time and sealed pairing are unused inside $U(a)$. A language model may optionally shortlist library actions before numerical ranking. Reported acquisition curves do not use that shortlist. Figure A2 compares acquisition controllers on a fixed witness bank.

A.2 SIGN CONVENTION AND LOCKED QUANTITIES

We keep two named quantities with separate claim tables. `sync_shift` uses $h(u) = u + \text{sync_shift}$: $\text{lo} \ge +\delta \Rightarrow$ molecular_earlier and $\text{hi} \le -\delta \Rightarrow$ imaging_earlier. `event_gap` uses $\Delta = h(\theta_A) - \theta_B$ (A=mol, B=img): $\text{hi} \le -\delta \Rightarrow$ molecular_earlier and $\text{lo} \ge +\delta \Rightarrow$ imaging_earlier. Larger disease time means later, so $\Delta < 0 \Leftrightarrow$ molecular event earlier. Hidden-pair scoring (Table 1,

Figure 3) uses `sync_shift`. External claim-stability analyses (Figure 4, Table 2) use `event_gap`. The same SES machinery underlies both. The scored functional changes with the question (relative lag vs. event order).

A.3 METRICS AND BASELINE SCORING

Let $u_A^{(i)}, u_B^{(i)}$ be true local coordinates on sealed pairs and $\hat{u}_B^{(i)} = h(u_A^{(i)})$ the method's predicted B-clock location. Phase MAE is $\mathbb{E}|\hat{u}_B - u_B|$. Order / FO are computed on subject-collapsed ordered pairs whose true gap exceeds δ: correct / false when the predicted order matches / mismatches, and a predicted near-tie when $|\hat{u}_B^{(j)} - \hat{u}_B^{(i)}| < \delta$. Thus Order + FO can fall below 100% (e.g., ClockBridge $72.5{+}19.2{=}91.7$). The residual is predicted near-ties. Figure 3(b) instead reports *all-pair* correct / unresolved / false shares (denominator = all scored pairs, and decision coverage = correct+false). At zero stress those shares need not match Ord. (ClockBridge correct share ≈ 0.50 vs. Ord. 72.5%), because Ord. conditions on large true gaps and excludes near-ties.

FO@40 is the false-order rate among the top 40% of scoreable pairs (excluding predicted near-ties), ranked by descending pair confidence. AURC is the area under the false-order risk–coverage curve on the same ranking. FO@40 and AURC fix the decision mass being scored, so selective risk is comparable across methods with different unresolved rates.

Pair confidence is margin-based for all methods: $c_{ij} = |\hat{u}_B^{(j)} - \hat{u}_B^{(i)}|$. For ClockBridge only, confidence is further scaled by the identified interval, $c_{ij} \leftarrow c_{ij}/(w{+}10^{-3})$ with level multipliers (L3/L2/L1/L0 $\mapsto$ 1.5/1.2/1.0/0.5), and unresolved/not-evaluable claims are down-weighted by 0.25. Point baselines (PSA-OT, SUE, PO-MFL, FGW-style) use the raw margin. No method uses ground-truth pairing in the confidence score, and thresholds are not calibrated on held-out ADNI test pairs.

Baseline adaptations under shared ADNI features are protocol surrogates under a common feature interface: each returns a forced `sync_shift` point map scored with the metrics above. PSA-OT (Xi et al., 2024) uses propensity logits with stage treated as a treatment indicator, Sinkhorn OT in logit space, and a barycentric push-forward to B phases. SUE (Yacobi et al., 2025) uses per-modality spectral/PCA embedding, stage-mean Procrustes, and a quantile map to B soft phases (without the RID / MMD-net components of the source paper). PO-MFL (Gu et al., 2025) uses adjacent-stage Sinkhorn costs on a lag grid with constant-velocity prediction $\hat{u}_B = u_A + \mathrm{lag}^\star$ (a protocol surrogate for the source paper's latent MFL dynamics). FGW-style (Klein et al., 2025) uses POT fused Gromov–Wasserstein on PCA features with a stage-difference penalty as a shared-interface surrogate. These are intentionally simplified protocol adaptations of external methods. The cleanest within-family contrast is Point-$\mathcal{L}^\star$ in Table 1: it uses identical $\mathcal{L}$, lag grid, features, and prediction map as ClockBridge, while omitting the SES-based interval reporting, confidence, and abstention machinery.

A.4 PROTOCOL ISOLATION

Stage proxies define within-cohort clocks and the SES loss $\mathcal{L}$, but are held out from cross-modal claim labels. Hidden pairing is revealed only after fitting and scores Table 1 / Figure 3. Hyperparameters (ε, δ, η_2, η_3, κ, λ, witness costs) are frozen on train/dev before SEA-AD / ADNI3 replacement. External cohorts evaluate claim-level stability under replacement.

Locked SES settings: lag grid $[-0.45, 0.45]$ (step 0.03), $\varepsilon{=}0.012$, $\delta{=}0.04$, $\eta_2{=}0.2$, $\eta_3{=}0.08$, $\kappa{=}4$, $\lambda{=}0.35$, event persistence $=1$, frozen on train/dev before SEA-AD / ADNI3 replacement. Appendix D reports ε sensitivity and sync-family / onset checks under those locks. Ablation tasks use 120 stress $+ 120$ near-boundary instances. The acquisition bank uses $N{=}60$ unresolved claims with budgets $\{1, \ldots, 8\}$. Primary hidden-pair tables use 12 evaluation seeds.

B DATASETS

B.1 HIDDEN-PAIR EVALUATION SPLITS

Hidden-pair evaluation is the only setting in which we can score synchronization against a sealed truth without claiming biological causality. ADNI seals exact `VISCODE` visit matches at fit time (130 subjects / 160 pairs; mean 6849 order-pairs over 12 seeds), so the held-out pairs are true same-visit modality links that the model never sees during fitting. PD FDG–CFT uses subject co-enrollment (56 subjects / 56 pairs, mean 1273 order-pairs). This is enough to score `sync_shift` under hidden pairing when visit or scan dates are unmatched. We therefore read PD as a second disease setting for selective lag scoring. Order-pair counts explain why selective ranking (FO@40 / AURC) is statistically meaningful even though subject counts are moderate.

B.2 EXTERNAL UNPAIRED POOLS

External evaluation asks a different question. When molecular and imaging cohorts never share subjects, which program and ROI temporal claims remain stable under cohort replacement. Table A1 summarizes native pools before matched-K selection. Molecular stages use Braak/CERAD-style bins on ROSMAP/SEA-AD. Imaging stages use CDR-SB/MMSE/DX-derived bins on ADNI2/ADNI3. Pool composition is intentionally unbalanced (many imaging-earlier native claims on some blocks), reflecting the stage-resolved margins before matched-K selection. Matched evaluation then uses $K{=}63$ claims per replacement block after capacity matching so that methods are compared at a common claim budget (Figure 4, Table 2).

Table A1: Native external claim pools before matched-K selection.

Mol.	Img.	Pool n	Mol. earlier	Img. earlier	Near-sim.	Unres.
ROSMAP	ADNI2	80	10	53	0	17
ROSMAP	ADNI3	80	19	22	2	37
SEA-AD	ADNI2	80	0	29	20	31
SEA-AD	ADNI3	80	0	41	13	26

C HIDDEN-PAIR ANALYSES

C.1 SEED-LEVEL UNCERTAINTY

Point estimates in Table 1 can look decisive even when seed variability is large. We therefore resample the $n_{\text{seeds}}{=}12$ ADNI evaluation seeds (2000 bootstrap replicates), not individual subject-pairs (Table A3). Seed-level resampling matches the experimental unit that actually changes when we redraw train/eval partitions. ClockBridge FO@40 12.3% $[8.3, 16.5]$ and AURC 0.147 $[0.111, 0.183]$ remain well separated from forced-map references near $\sim 52\%$ FO@40, whose intervals sit entirely above 46%. Matched risk–coverage summaries further show that ClockBridge reaches FO$\leq 15\%$ at 55.1% coverage on ADNI (PD: 99.2%), while forced-map coverage at that risk target stays near 0–26% and same-loss Point-$\mathcal{L}^\star$ reaches 38.8% / 80.5% (Table A2). The selective gap is stable across the 12 evaluation seeds.

D ROBUSTNESS ANALYSES

D.1 SES TOLERANCE

The locked ε controls how large the computational SES is allowed to be. Figure A1 and Table A4 multiply ε around the default ($\times 1$). This sweep uses an independently generated stress-style suite with the same generative construction as Table 3, on an independent draw of instances. Each multiplier aggregates 72 tasks (3 noise levels $\times$ 24 seeds). Consequently the $\times 1$ operating point here (Dec. cov. 22.2%, Int. cov. 88.9%) differs slightly from Full CB on the main ablation (25.8% / 86.7%). Larger ε raises unresolved mass and lowers decision coverage while keeping false claims

Table A2: Matched selective-risk summary (locked Table 1 protocol). Cov@FO15: decision coverage at false-order $\leq 15\%$. Point-$\mathcal{L}^\star$: same $\mathcal{L}$ / predict map as ClockBridge without SES interval, confidence, or abstention.

Dataset	Method	Cov@FO15↑	FO@40↓	Dec.cov.
ADNI	PSA-OT	0.2	53.5	76.3
ADNI	SUE	0.1	52.2	77.8
ADNI	PO-MFL	0.1	52.8	54.9
ADNI	FGW-style	0.1	52.7	80.9
ADNI	Point-$\mathcal{L}^\star$	38.8	17.7	81.9
ADNI	**ClockBridge**	**55.1**	**12.3**	69.1
PD	PSA-OT	26.2	20.8	66.3
PD	SUE	9.7	42.0	76.9
PD	PO-MFL	19.2	25.8	63.0
PD	FGW-style	0.3	49.6	80.5
PD	Point-$\mathcal{L}^\star$	80.5	7.2	83.5
PD	**ClockBridge**	**99.2**	**3.3**	63.6

Table A3: Seed-level bootstrap 95% intervals on ADNI (n_{seeds}=12; 2000 replicates).

Method	FO@40 [lo, hi]	AURC [lo, hi]
PSA-OT	53.5 [50.6, 56.2]	0.524 [0.494, 0.548]
SUE	52.2 [48.8, 55.2]	0.522 [0.490, 0.551]
PO-MFL	52.8 [50.0, 55.6]	0.523 [0.496, 0.550]
FGW-style (POT)	52.7 [46.7, 58.2]	0.537 [0.484, 0.585]
ClockBridge	**12.3 [8.3, 16.5]**	**0.147 [0.111, 0.183]**

near zero. The left panel of Figure A1 is identically zero for every multiplier ≥ 0.50: once the SES is wide enough to cover the truth under this generative noise, further widening leaves more claims unresolved. The only setting that admits false claims is the overly tight $\times 0.25$ multiplier (1.4%). Interval coverage is mostly increasing with a small dip at $\times 2.00$. We retain that non-monotonicity in the reported curve.

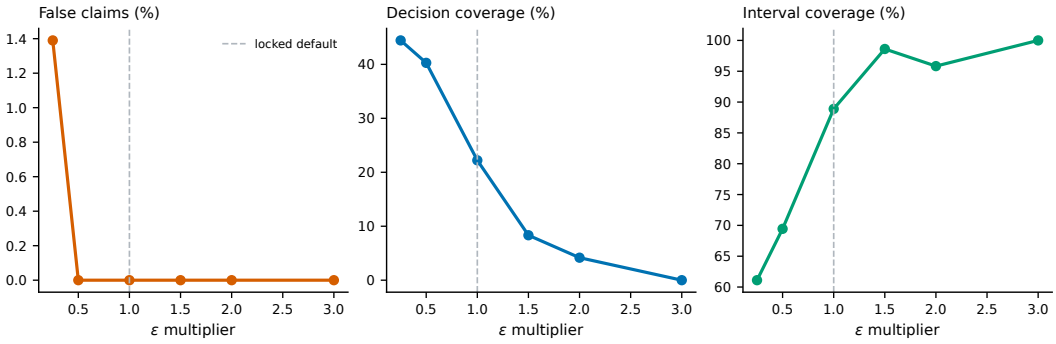


Figure A1: SES tolerance (ε) sensitivity on an independent stress-style suite (72 tasks per multiplier, dashed line at locked default $\times 1$). False claims remain at 0% for all $\varepsilon\times \geq 0.50$.

D.2 Synchronization family and event onset

Translation SES may appear restrictive. On n=60 tasks, a more flexible piecewise-linear family increases unresolved mass (decide rate falls from 0.117 to 0.017) without creating false directions. When both families decide, directional agreement is perfect. We therefore keep translation as the

Table A4: ε multiplier sensitivity (72 tasks per row, locked default $\times 1$).

$\varepsilon\times$	False%	Dec.%	Abst.%	Int.%	Width
0.25	1.4	44.4	55.6	61.1	0.18
0.50	0.0	40.3	59.7	69.4	0.22
1.00	0.0	22.2	77.8	88.9	0.35
1.50	0.0	8.3	91.7	98.6	0.43
2.00	0.0	4.2	95.8	95.8	0.48
3.00	0.0	0.0	100	100	0.58

main family. It already abstains heavily under weak evidence, and a richer family mainly enlarges the unresolved set.

Under the locked persistence ($=1$), sweeping the effect-fraction threshold used to declare event onset (0.10–0.35; $n=24$ per setting) keeps directional rates in 0.75–0.88, with unresolved mass as the primary failure mode. This supports treating event definition as a sensitivity axis under the locked persistence.

E External claim stability

External experiments report claim-level stability under cohort replacement: a method may keep emitting the same directional claims, flip, or leave the claim unresolved. The next subsections separate stage-structure controls, exact replacement counts, and capacity matching.

E.1 Negative controls

Figure 4(c) destroys stage structure on otherwise identical inputs ($n=40$). Structured inputs yield direction rate 0.58 with false rate 0 among decided cases (unresolved 0.42). Stage-permuted and feature-mismatched inputs raise unresolved rates to 0.75–0.85 (direction 0.25 / 0.15), which confirms that the claim layer tracks stage structure. Displayed rates are rounded to two decimals so direction+unresolved sums to 1.00.

E.2 Replacement accounting

Percentages in Table 2 can hide absolute mass and selectivity. Table A5 therefore lists per-block supported / contradicted / abstained counts ($n=63$ per block) together with the aggregate over all 189 matched claims. ClockBridge reduces *aggregate* contradiction to $29/189$ (15.3%) versus $67/189$ (35.4%) for PO-MFL, while abstaining on $111/189$ versus $35/189$. Among decided claims only, conditional contradiction is 37.2% (ClockBridge) versus 43.5% (PO-MFL). Molecular replacement contributes the largest absolute gap (1 vs. 18 contradicted). Imaging replacement yields fewer aggregate contradictions for ClockBridge than PO-MFL (21 vs. 31), with higher abstention when the identified set is unstable. Matched-K in Figure 4(a) equalizes the candidate pool ($K=63$). Table A6 further equalizes decision counts by ranking claims within each block by development confidence and retaining the top m decided outcomes.

E.3 Bidirectional capacity at matched K

Not every baseline emits bidirectional temporal claims. At matched candidate pool $K=63$, PSA-OT / SUE / FGW-style adaptations have bidirectional capacity 0 (native pools ≈ 69), so their contradiction cell is unmarked: there is no bidirectional recovery to score. Among methods with nonzero capacity, ClockBridge is more selective (capacity 20 from a native pool of 63) than PO-MFL (capacity 62 from 76) and shows lower external contradiction (15.3% vs. 35.4%). Capacity matching is what makes Figure 4(a) readable. Without it, a method could look stable simply by emitting almost no comparable claims.

Table A5: Replacement outcomes by block (n=63) and aggregate over 189 matched claims. Ext. % = contradicted / 189. Cond. % = contradicted / (supported+contradicted).

Replacement	Method	Supp.	Contr.	Abst.	Ext. %	Cond. %
Mol. ROSMAP→SEA-AD	PO-MFL	34	18	11	—	—
	CB	18	1	44	—	—
Img. ADNI2→ADNI3	PO-MFL	22	31	10	—	—
	CB	7	21	35	—	—
Both replaced	PO-MFL	31	18	14	—	—
	CB	24	7	32	—	—
Aggregate (189)	PO-MFL	87	67	35	35.4	43.5
	ClockBridge	49	29	111	15.3	37.2

Table A6: Contradiction among the top-m decided claims per replacement block (mean over three settings), ranked by development confidence.

m	ClockBridge	PO-MFL
10	43.3	76.7
15	40.0	80.0
20	38.4	81.7
25	35.1	72.0
30	34.5	62.2

F Active witness acquisition

When a claim remains unresolved, ClockBridge can rank the next witness from a fixed library by the SES utility in Eq. (7). The utility depends on SES width, precision-boost look-ahead, and action cost. Latent disease time and sealed pairing are unused inside $U(a)$. This appendix reports budgeted selection on a shared bank against Random, Bayesian/EIG (Rainforth et al., 2024), an Active Disambiguation heuristic, and a fixed BioDiscovery protocol order (Roohani et al., 2025).

F.1 Witness library and protocol

The shared bank contains N=60 unresolved near-boundary claims. Each claim exposes a heterogeneous witness library of 3–8 actions (mean ≈6.5) with integer costs in $\{1, 2, 3, 4\}$ (high-value cost 1, ultra cost 4, medium 1–3). All policies see the same bank and budgets $\{1, \ldots, 8\}$, so differences come from selection. The BioDiscovery protocol uses a fixed order (no live LLM call) and serves as a non-adaptive reference. Active Disambiguation here is an SES-disagreement heuristic inspired by viable-set reduction (Kobalczyk et al., 2025). Optional language shortlisting may restrict candidates before $U(a)$ is evaluated. The reported curves use the numerical controller on the full library.

F.2 Budget efficiency

B_{50} is the linearly interpolated budget where the mean correct-rate curve crosses 0.5. Uncertainty uses claim-level bootstrap (2000 replicates). We resample the 60 claims, recompute the mean curve and crossing, then report bootstrap SE (± 1 SE in Figure A2(b)) and the 2.5–97.5 percentile interval (Table A7).

On this bank, ClockBridge reaches B_{50}=1.89 and Correct@8=0.82, ahead of Bayesian/EIG (3.71 / 0.75) and Random (7.25 / 0.55). Figure A2(a) shows the full acquisition curves under the shared action set. The N=60 bank is a stress suite of unresolved near-boundary claims. Claim-level bootstrap intervals in Table A7 quantify variability under resampling of that bank.

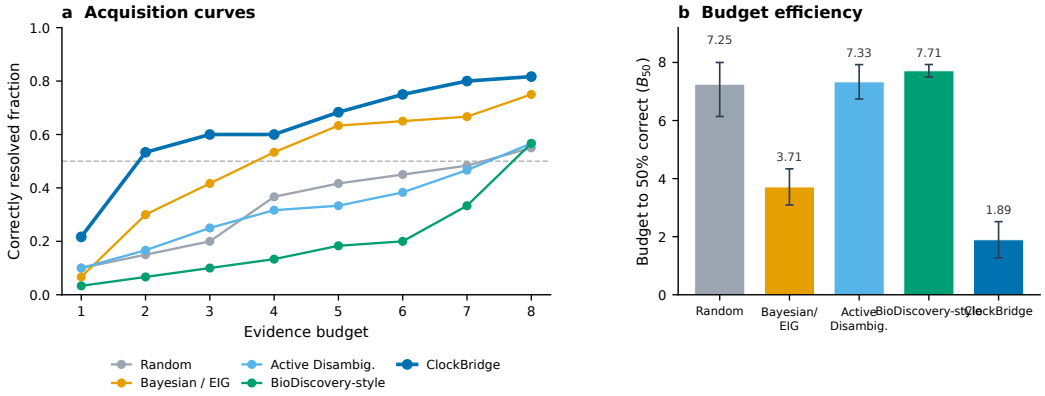


Figure A2: Active witness selection on a shared bank (N=60). (a) Correct fraction versus budget. (b) Budget to 50% correct (B_{50}) with claim-level bootstrap SE.

Table A7: Budget to 50% correct (N=60, budgets $\{1, \ldots, 8\}$, claim-level bootstrap of B_{50}).

Policy	$B_{50} \downarrow$	SE	95% CI	Correct@8↑
Random	7.25	1.11	[4.00, 8.00]	0.55
Bayesian / EIG	3.71	0.62	[2.69, 4.88]	0.75
Active Disambig.	7.33	0.59	[5.81, 8.00]	0.57
BioDiscovery protocol	7.71	0.22	[7.18, 8.00]	0.57
ClockBridge	**1.89**	0.62	[1.57, 4.25]	**0.82**

G METHOD-COMPONENT ABLATIONS

Table 3 in the main text isolates SES width, point estimation from $\mathcal{L}^\star$, progression uncertainty, and the unresolved reporting rule on 120 stress + 120 near-boundary tasks. False rates are fractions of *all* tasks that emit an incorrect directional claim (unresolved outcomes count as non-false), so coverage and false rate are read jointly.

Near the identification boundary, forcing a direction raises false-direction rate from 3.3% to 65.0% ([57.5, 73.3]); reporting only the $\mathcal{L}^\star$ lag yields 55.8% ([47.5, 64.2]). Full ClockBridge keeps stress false at 0.0% and near-boundary false at 3.3% ([0.8, 6.7]) by resolving when the identified set agrees (decision coverage on stress 25.8% [18.3, 33.3], interval coverage 86.7% [80.0, 92.5], mean width 0.335). Tight SES buys coverage (72.5% [64.2, 80.0]) by shrinking width (0.109) but collapses interval coverage to 42.5% ([34.2, 50.8]) and raises near-boundary false to 27.5% ([20.0, 35.0]). Removing progression uncertainty is milder (Dec. 41.7% [33.3, 50.8], Int. 80.8% [73.3, 87.5], boundary false 6.7% [2.5, 10.8]). Point-$\mathcal{L}^\star$ and Force-all variants decide on every task and push boundary false above 55%.