Vector Summary Pseudo Posteriors for Simulation-Based Inference

STAI-X 2026

with Applications to Cosmology

Pratik Patil[†] · Jonah Rose[‡] · Alex Garcia[§] · **Min Chen**[†] · Paul Torrey[§] · Arya Farahi[†]

[†]Statistics and Data Sciences, UT Austin [‡]Physics, Princeton University [§]Astronomy, University of Virginia

TL;DR *A pseudo posterior is only as identifying as its summaries — VSPP turns summary choice into identification design.*

Problem: weak identification

- Implicit-likelihood simulators are calibrated through low-dimensional summaries (ABC, synthetic likelihood, regression projection).
- A scalar residual discrepancy (Farahi, Rose & Torrey '26) can concentrate on a **ridge or manifold** of parameters — not a point.
- Failures are hard to diagnose; neural SBI can be overconfident under misspecification.

Method: the VSPP update

Reference inputs: relation f_{ref} and target $t = (t_1, \dots, t_K)^\top$ — from observations, published constraints, or a held-out simulation

↓

Residuals $\tilde{r}_m(\theta_j) = \tilde{y}_m - f_{\mathrm{ref}}\{\tilde{x}_m\}$ for each simulated batch under θ_j

↓

Batch summary $\tilde{s}_M(\theta_j) = M^{-1} \sum_m g\{\tilde{r}_m(\theta_j)\}$, $g = (g_1, \dots, g_K)^\top$

↓

Mismatch $d_M(\theta_j) = \tilde{s}_M(\theta_j) - t$, loss $\ell_M(\theta_j) = d_M^\top \Lambda\, d_M$

↓

Weights $w_j \propto \exp\{-\ell_M(\theta_j)/2h^2\}$ → pseudo posterior $\Pi = \sum_j w_j \delta_{\theta_j}$

Generalized Bayes update (Bissiri+ '16). The scalar case $K = 1$, $g_1(r) = r$, $t_1 = 0$ recovers the regression projection baseline of Farahi+ '26. With diagonal Λ the loss splits into named coordinate contributions $\ell_M = \sum_k \lambda_k \{\tilde{s}_k - t_k\}^2$, so ESS and coordinate losses show which relation or moment drives each update.

Two-summary VSPP: moment enrichment

$$g_1(r) = r, \quad g_2(r) = r^2, \quad \text{target } t = (0,\ t_2)^\top$$

The second target separates parameters that match the mean relation but differ in residual dispersion; t_2 comes from a held-out moment or a scientific scatter constraint (DREAMS: $t_2 = 0.09 = 0.3^2$). Adding a moment and adding a relation are the **same operation** — both enlarge the summary vector.

Positioning vs. related approaches

Approach	Interface	Identification
ABC / synthetic likelihood	hand-chosen summaries	implicit, tolerance-dependent
Neural SBI	learned summaries / ratios	high-capacity, opaque under misspec.
GMM / indirect inference	moment maps	point estimates, rank theory
VSPP (ours)	**named vector summaries**	**nested sets, rank, coordinate diagnostics**

Takeaways & limitations

- A lightweight, **transparent calibration layer** — a companion to neural SBI, not a replacement.
- Makes the identification target **explicit and coordinate-wise diagnosable** (GMM / indirect-inference lens).
- Limits: nesting needs exact representability; under misspecification the best matching set may move rather than shrink; h, Λ, and raw moments need care.

Theory: what do the summaries identify?

$$\Theta^\dagger = \arg\min_{\theta \in \Theta} d(\theta)^\top \Lambda\, d(\theta), \qquad d(\theta) = \tilde{s}(\theta) - t$$

T1 · Concentration As $N \to \infty$, then $M \to \infty$ with $h_M \downarrow 0$ and $Mh_M^2 \to \infty$, the pseudo posterior concentrates on the best matching set $\Theta^\dagger$.

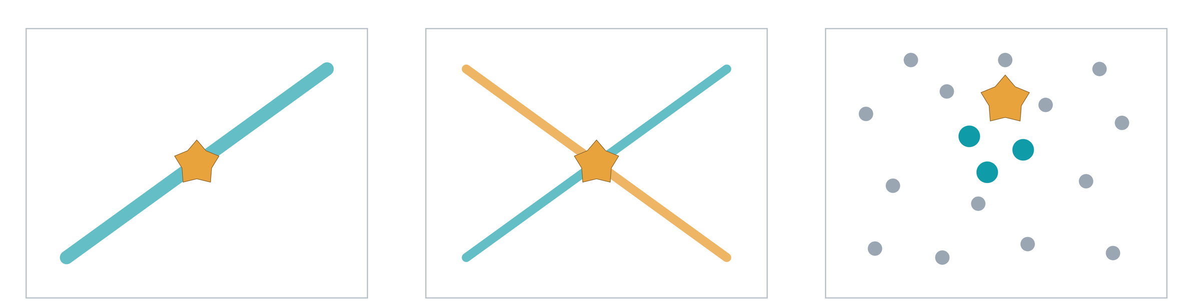
T2 · Nesting Under exact matching, $\Theta^\dagger_{K+1} \subseteq \Theta^\dagger_K$: each added summary imposes a new constraint without changing the old ones.

T3 · Rank geometry The exact match set is locally a manifold of dimension $p - \operatorname{rank} Dd(\theta^\star)$; full rank ⇒ local point identification.

P4 · Local Gaussian limit Under point identification, $(\theta - \theta^\star)/h \Rightarrow N(0, (J^\top \Lambda J)^{-1})$ with $J = Dd(\theta^\star)$, as $h \downarrow 0$.

P5–P6 · Finite banks $\Pi_h^{\mathrm{bank}}(\theta^\star) \geq \{1 + (N-1)\, e^{-\Delta/2h^2}\}^{-1}$; adding an exactly matched summary can only grow the separation gap Δ.

Identification geometry

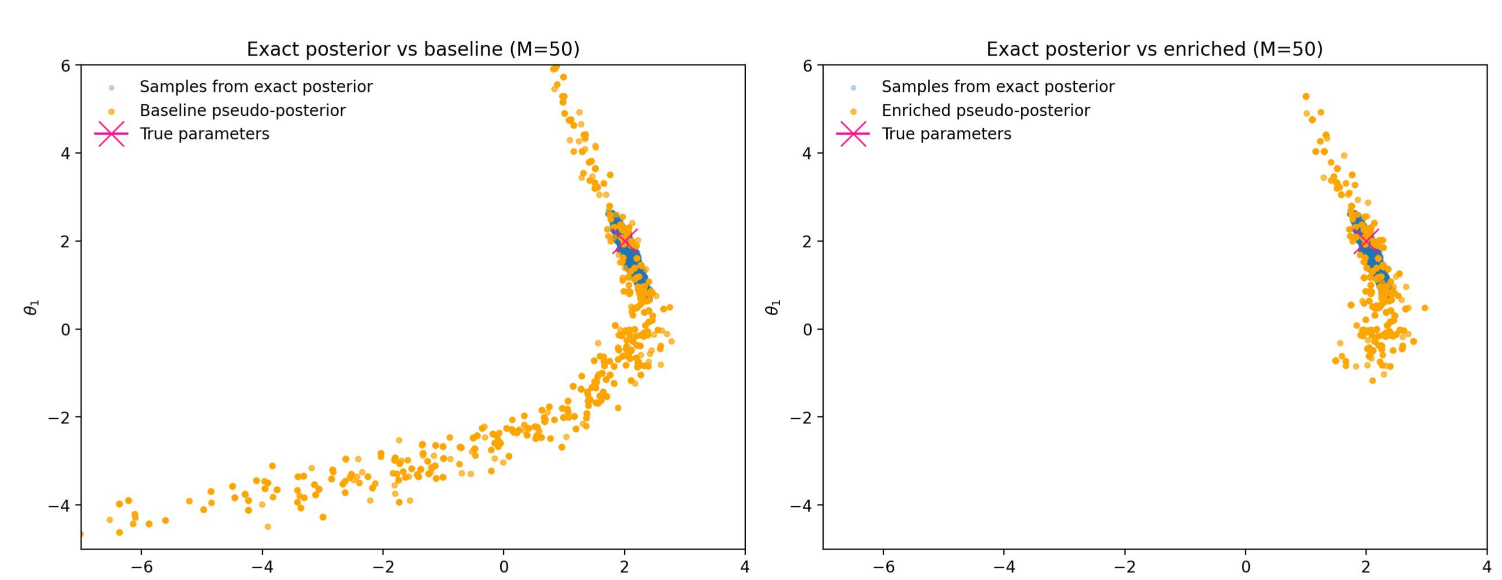


A · one summary: ridge / manifold — B · add a summary: intersection — C · finite bank: separation ↑

What matters is not the number of summaries, but the number of independent summary directions.

Synthetic checks

- **Ridge to point**: the mean target identifies $a + b = 3$; the two-summary update identifies the true point $(1, 2)$.
- **Plane to curve** (σ unknown): rank gain without point identification.
- **Nonlinear benchmark**: the enriched pseudo posterior is tighter and closer to the exact MCMC posterior.

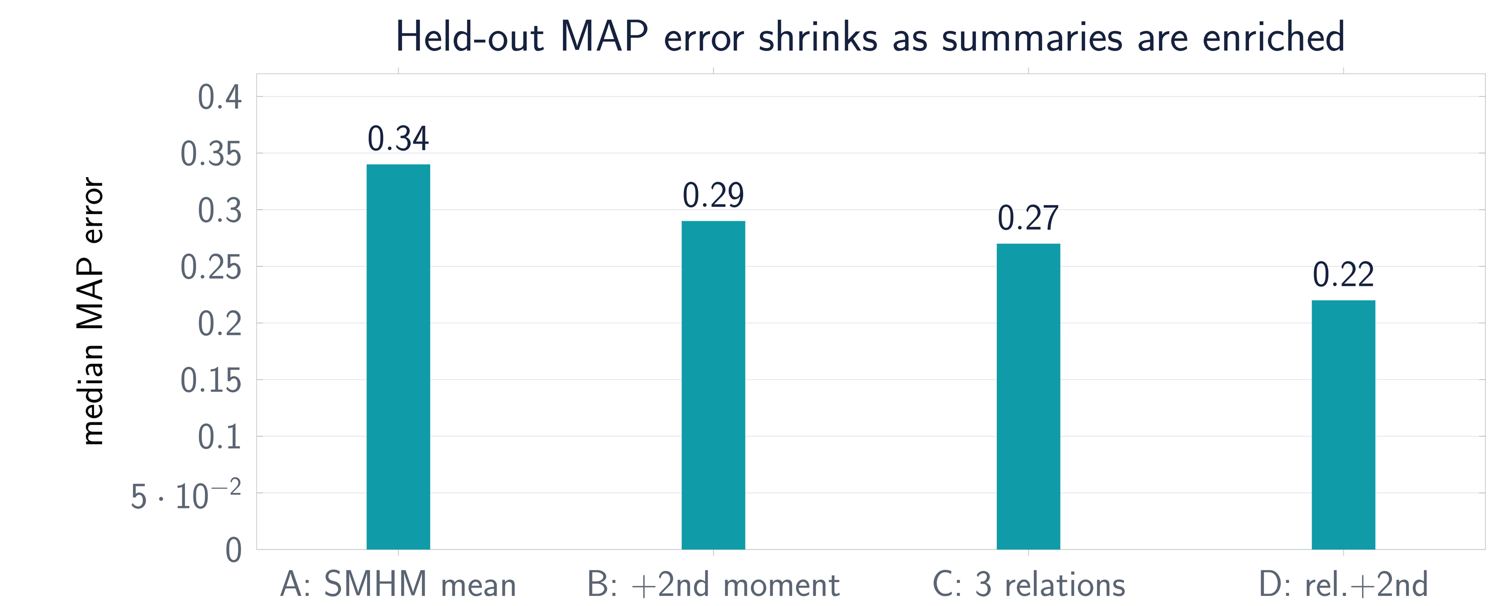


Left: scalar update — **broad manifold** Right: two-summary — **collapsed onto the exact posterior**

Nonlinear benchmark at $M = 50$ with a deliberately misspecified linear f_{ref}; blue points are exact MCMC posterior samples, the cross marks the true $\theta^\star = (2, 2)$.

CAMELS IllustrisTNG: finite-bank recovery

75 held-out reference simulations · relations: SMHM, SFMS, BH · summary sets A (K=1) → D (K=6)



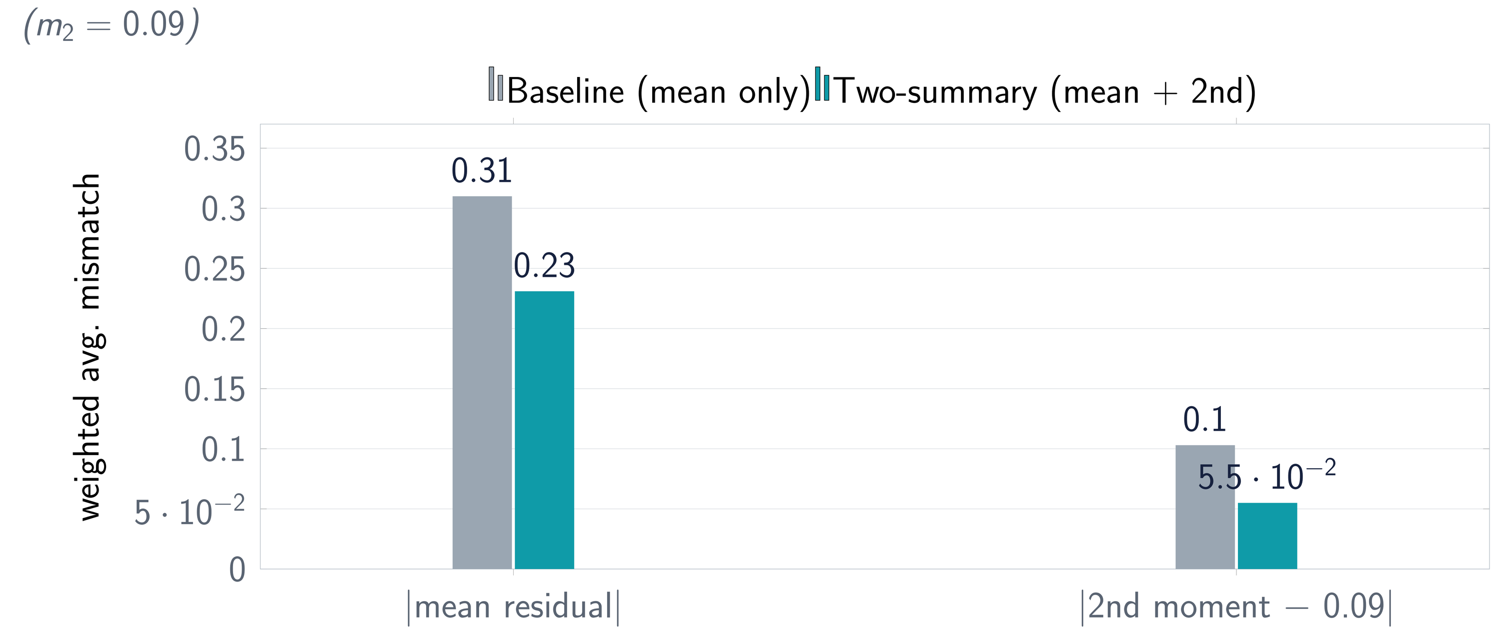
Metric (median)	A (K=1)	D (K=6)
True-run rank	53	**9**
95% high-weight set	795	**410**

−35% MAP error (A → D)

−48% 95% set size (A → D)

DREAMS: generated-sample validation

27,000 astrophysical contexts · 2,500 galaxies each · Milky-Way-mass SMHM · scatter target 0.3 dex ($m_2 = 0.09$)



Both weighted mismatches drop; the 95% high-weight set shrinks **23,441 → 16,251** (ESS 22,450 → 15,181). Enriched weights match the mean relation and the scatter target **jointly**.

Future directions

- Propagate uncertainty in f_{ref} and t; robust moments, quantiles, and distributional summaries; **adaptive summary selection** and calibration of h and Λ.
- Combine relations across redshifts, suites, and fidelities; guide new simulations toward weakly separated regions; **interpretable front end for neural SBI**.

Key references: Farahi, Rose & Torrey '26 (regression projection pseudo posterior) · Bissiri, Holmes & Walker '16 (generalized Bayes) · Hansen '82 (GMM) · Gouriéroux+ '93 (indirect inference) · Kennedy & O'Hagan '01 (Bayesian calibration) · Villaescusa-Navarro+ '21/'22 (CAMELS, Cosmo1gal) · Rose+ '25 (DREAMS)

Acknowledgments: NSF Cooperative Agreement 2421782 and Simons Foundation grant MPS-AI-00010515 (NSF–Simons AI Institute for Cosmic Origins, CosmicAI) · computing by TACC.