IN-CONTEXT LEARNING OF ANONYMOUS MATRIX OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unpaired distribution examples can determine a new response without identifying the transformation itself. We study this distinction in anonymous matrix operator learning: demonstrations contain independently sampled input and output matrices, rows are unordered, and columns are unnamed but aligned across the task. The query response is unavailable at inference. For shared Markov kernels, we prove that query identification holds uniformly over the mechanism class exactly when the query input law lies in the linear span of the demonstrated inputs. An approximate bound separates span mismatch from amplification of observation error. For affine-Gaussian responses, we characterize compatible mechanisms and establish constructive recovery and local stability conditions. We then introduce ACEMO-ICL, a conditional flow that combines query evidence with unpaired demonstrations while respecting the matrix symmetries. Controlled numerical examples illustrate the identification bounds, and fixed-budget synthetic experiments show gains from correct context and informative background variation. Real-data examples expose the dependence of transfer on representation. Together, the results distinguish information supplied by demonstrations from estimation and modeling error.

1 INTRODUCTION

A rotation is invisible on a centered isotropic Gaussian: it produces the same distribution as the identity map. The two transformations can nevertheless give different responses on a shifted Gaussian query. More samples from the isotropic inputs estimate the demonstrated laws more precisely, but leave this ambiguity intact. An input background that breaks the symmetry can resolve it. This distinction between more observations and more identifying information is central to learning transformations from distribution examples.

We study this distinction in distributional in-context learning. Each task has one shared operator $\mathcal{T}$ and several input backgrounds. Its demonstrations and query satisfy

$$\mathcal{C} = \{(X_i, Y_i)\}_{i=1}^{K}, \quad X_i \sim P_i^{\otimes n_i^x}, \quad Y_i \sim \mathcal{T}(P_i)^{\otimes n_i^y}, \quad X_q \sim P_q^{\otimes n_q}. \tag{1}$$

The matrices are sampled independently conditional on the task. Given $\mathcal{C}, X_q$, the learner generates samples from the unseen response $\mathcal{T}(P_q)$. Rows have no input–output pairing; the learner knows only which input and response sets form each demonstration. Columns have no external names but retain a common correspondence throughout the task. We call these observations anonymous matrices.

Prediction requires agreement on the query response:

$$T_1(P_i) = T_2(P_i) = Q_i \ \ \forall i \quad \Longrightarrow \quad T_1(P_q) = T_2(P_q), \tag{2}$$

for every pair of mechanisms compatible with the visible laws. They may differ on other inputs: determining the requested response requires less information than recovering the entire operator.

We first characterize query identification for unrestricted shared Markov kernels. A guarantee uniform over this class holds exactly when the query input lies in the real linear span of the demonstrated input laws. Outside that span, identical demonstrations can conceal different query responses, imposing a positive prediction-risk floor. An approximate bound separates the effect of span mismatch from the amplification of observation error by the span coefficients.

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.

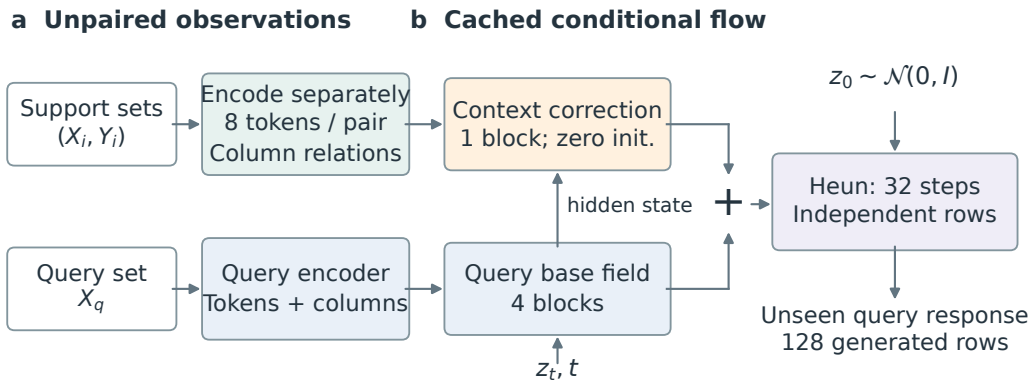

Figure 1: The prediction task and ACEMO-ICL. Independently sampled support inputs and responses supply cached evidence about a shared transformation. A query base field and a context-conditioned correction generate the unseen response. Each output row evolves independently conditional on the cache.

We then study invertible affine-Gaussian responses, where mechanism structure permits a constructive analysis. We characterize every compatible transformation, identify the queries on which they agree, and give a sufficient three-background recovery condition. The accompanying local bounds expose the roles of covariance conditioning, spectral separation, and mean projections.

Finally, ACEMO-ICL combines unpaired distribution evidence in a conditional generator with row and column symmetries. Fixed-budget experiments test its use of demonstrations and informative background variation. The statistical analysis connects conditional flow matching to prediction risk while separating population ambiguity, finite-observation uncertainty, and model error. Population identification and neural recovery remain distinct claims.

2 Task and Related Work

Observation and risk. The visible data are

$$\mathcal{D} = (\mathcal{C}, X_q, \text{masks}, \text{domain}). \tag{3}$$

The predictor has no access to the held-out response $Y_q \sim Q_q^{\otimes m}$, operator parameters, or distribution parameters. Support inputs and responses have independent masks. The population operator is independent of matrix row counts, which control estimation accuracy. Our identification results concern fully observed population laws; partially observed entries require an additional observation model.

A predictor returns a law $\widehat{Q}(\mathcal{D})$. For a bounded continuous characteristic kernel with feature map $\phi(x) = k(x, \cdot)$, write $\Phi(P) = \int \phi \, dP$ and

$$\mathrm{MMD}_k^2(P, Q) = \|\Phi(P) - \Phi(Q)\|_{\mathcal{H}_k}^2. \tag{4}$$

We use a mixture of Gaussian kernels,

$$k_d(x, y) = \tfrac{1}{4} \sum_{h \in \{1/2, 1, 2, 4\}} \exp\{-\|x - y\|^2 / (2dh^2)\}. \tag{5}$$

Neural evaluations scale working coordinates by a divisor computed only from the visible query. The theory either fixes the kernel or conditions on this retained scalar. Positive tasks use log coordinates for both the operator and evaluation, followed by exponentiation of generated samples. Stability bounds refer to these working coordinates.

Anonymous structure. The rows of each X_i, Y_i, X_q and the support pairs are unordered. Pair membership carries information: assigning a response set to a different input changes the demonstrated relationship. A common permutation of input columns must induce the same permutation of the generated output. This requirement concerns column relabeling, rather than arbitrary rotations.

Table 1: Prediction settings. Query output denotes samples of the destination law available at inference. The descriptions concern the cited formulations.

Setting	Query output samples	Across-task variation	External intervention ID
Distribution regression	Unavailable	One fixed distribution map	Not required
In-context OT	Available	Source–target pair	Not required
MapPFN	Unavailable	Interventions	Used
This work	Unavailable	Task-specific operator	Not required

Relation to prior work. Distribution-to-distribution regression learns a fixed relationship between sampled laws (Oliva et al., 2013). In our setting, pretraining ranges over operators and the demonstrations constrain the relationship that applies to the current query. In-context optimal transport receives samples from both the source and destination distributions (Cole et al., 2026); the destination is hidden in our prediction task. MapPFN predicts an unseen response from observational data, labeled intervention-response context, and a query intervention identifier (Sextro et al., 2026). Our demonstrations instead describe a shared operator acting on different input backgrounds, with no external operator identifier. Table 1 summarizes these distinctions.

Set attention (Lee et al., 2019), prior-data fitted inference (Müller et al., 2022), probabilistic symmetries (Bloem-Reddy & Teh, 2020), and equivariant layers (Maron et al., 2019) provide the architectural foundations. Unpaired multi-domain identification (Sturma et al., 2023) and joint-diagonalization uniqueness (Kleinsteuber & Shen, 2012) provide theoretical precedents. We use these established tools to study which query responses unpaired demonstrations determine and to construct a generator for this task.

3 What Can Context Identify?

Fix a mechanism class $\mathfrak{T}$ and feasible population evidence $\mathcal{O} = \{(P_i, Q_i)\}_{i=1}^{K}$. Define

$$\mathfrak{I}(\mathcal{O}) = \{T \in \mathfrak{T} : T(P_i) = Q_i \ \forall i\}, \qquad \mathfrak{R}_q(\mathcal{O}) = \{T(P_q) : T \in \mathfrak{I}(\mathcal{O})\}. \tag{6}$$

Operator identification requires $\mathfrak{I}$ to be a singleton; query identification requires only that $\mathfrak{R}_q$ be a singleton. Repeating a population pair adds no constraint to either set. A distinct background can remove compatible mechanisms and thereby identify additional query responses.

3.1 Shared Markov kernels

A Markov kernel $\mathsf{K}(x, dy)$ induces the linear action

$$(\mathsf{K}P)(B) = \int \mathsf{K}(x, B)\, P(dx). \tag{7}$$

This class includes deterministic maps and stochastic responses acting on each input independently.

Theorem 3.1 (Uniform query identification)**.** *Let $P_1, \ldots, P_K, P_q$ be probability measures on a measurable input space, and suppose the output space admits two distinguishable probability laws. Equality $\mathsf{K}_1 P_i = \mathsf{K}_2 P_i$ for every i implies $\mathsf{K}_1 P_q = \mathsf{K}_2 P_q$ for every pair of Markov kernels if and only if P_q belongs to the real linear span of $P_1, \ldots, P_K$ as finite signed measures. If $P_q = \sum_i a_i P_i$, then $Q_q = \sum_i a_i Q_i$.*

Proof sketch. Linearity gives sufficiency. Outside the span, a bounded measurable function f can have zero integral under every P_i and a nonzero integral under P_q. Mixing two fixed output laws with respective weights $(1 + \epsilon f(x))/2$ and $(1 - \epsilon f(x))/2$ constructs legal kernels that agree on the support and differ at the query. Appendix C.2 supplies the separating-function argument.

The theorem characterizes a guarantee uniform over kernels. A specific observed response can identify queries outside the span: on a finite input space, mapping a full-support input to a point

mass determines the response at every state. The span coefficients sum to one but may be negative, so their empirical combination need not define a sampleable probability law. Normalized population reweighting, $T_w(P)(dx) = w(x)P(dx)/\int w\,dP$, is generally nonlinear in P and falls outside the common-kernel model.

Proposition 3.2 (Approximate identification)**.** *Let k have a measurable feature map in a separable RKHS and $k(y,y) \le \kappa$. Define the variation norm by $\|\nu\|_{\mathrm{var}} = \sup_{\|f\|_\infty \le 1} |\int f\,d\nu|$. For any real coefficients a_i, put $\nu = P_q - \sum_i a_i P_i$. Then*

$$\mathrm{MMD}_k(\mathsf{K}_1 P_q, \mathsf{K}_2 P_q) \le \sum_i |a_i|\,\mathrm{MMD}_k(\mathsf{K}_1 P_i, \mathsf{K}_2 P_i) + 2\sqrt{\kappa}\|\nu\|_{\mathrm{var}}. \tag{8}$$

Proof. Set $\psi_j(x) = \int \phi(y)\mathsf{K}_j(x,dy)$; boundedness gives $\|\psi_j(x)\| \le \sqrt{\kappa}$. The difference of query embeddings equals the weighted sum of support embedding differences plus $\int(\psi_1 - \psi_2)\,d\nu$. The triangle inequality bounds the latter by $2\sqrt{\kappa}\int d|\nu|$, proving the result. □

For probability differences, this variation norm is twice the usual total-variation distance $\sup_B |P(B) - Q(B)|$. Equation (8) identifies two sources of instability: residual error in the query's span representation and large coefficients multiplying support-response discrepancies. Applying it from finite samples also requires control of the input-law approximation. Appendix D develops the finite-observation bound and a four-state example.

3.2 Ambiguity and prediction risk

Theorem 3.3 (Indistinguishable-context lower bound)**.** *Suppose two tasks have identical support laws, query input law, and visible sampling protocol, but query responses Q_0, Q_1. For every prompt-only, possibly randomized predicted law, with $\delta = \mathrm{MMD}_k(Q_0, Q_1)$,*

$$\tfrac{1}{2}\mathbb{E}_0\,\mathrm{MMD}_k^2(\widehat{Q}, Q_0) + \tfrac{1}{2}\mathbb{E}_1\,\mathrm{MMD}_k^2(\widehat{Q}, Q_1) \ge \delta^2/4. \tag{9}$$

The equal-prior optimum is attained by $(Q_0 + Q_1)/2$.

Proof. The prompt and predictor randomization have a common law in both tasks. For each predicted embedding, the parallelogram identity gives its squared distance from the response midpoint plus $\delta^2/4$. Integrating proves the bound, and the midpoint is the embedding of the stated probability mixture. □

Identity and a π rotation on isotropic supports give the indistinguishable demonstrations from the introduction; a shifted query separates their responses. If the identity response is $Q_0 = P_q$ and its expected error is at most r, the alternative task has error at least $(\delta - \sqrt{r})_+^2$ (Appendix C.3). Thus strict identity preservation can conflict with accurate response prediction even at unlimited model capacity.

4 Constructive Identification and Stability

Restrict the operator to

$$Y = AX + b + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, Q), \quad \varepsilon \perp X, \quad A \text{ invertible}, \quad Q \succeq 0. \tag{10}$$

For Gaussian inputs, the response moments are

$$m_i = A\mu_i + b, \qquad C_i = A\Sigma_i A^\top + Q. \tag{11}$$

Taking differences between backgrounds removes b, Q. The resulting constraints use only within-matrix moments, which can be estimated separately from the unpaired input and response sets.

Query-specific ambiguity. Fix a compatible triple (A, b, Q), index a reference background by zero, and write $\Delta\mu_i = \mu_i - \mu_0$, $\Delta\Sigma_i = \Sigma_i - \Sigma_0$, with analogous query differences.

Proposition 4.1 (Compatible affine responses)**.** *The compatible invertible triples are exactly $A_H = AH$, $b_H = b + A(I - H)\mu_0$, and $Q_H = Q + A(\Sigma_0 - H\Sigma_0 H^\top)A^\top$, with invertible H satisfying*

$$H\Delta\mu_i = \Delta\mu_i, \quad H\Delta\Sigma_i H^\top = \Delta\Sigma_i \ \forall i, \quad Q_H \succeq 0. \tag{12}$$

Their query responses are

$$\mathcal{N}(m_0 + AH\Delta\mu_q,\ C_0 + AH\Delta\Sigma_q H^\top A^\top). \tag{13}$$

The query is identified if and only if every feasible H fixes its mean and covariance differences.

Proof sketch. Write an alternative mechanism as $H = A^{-1}A'$. Equality of the support moment differences gives the two invariances, and the reference response fixes the offset and noise. Conversely, these formulas reproduce every support law (Appendix C.4).

Example 4.2 (Nonunique mechanism, unique query)**.** Let $d = 3$, let $A = I, b = 0, Q = 0.2I$ and use supports $\mathcal{N}(0, I)$ and $\mathcal{N}(e_1, 2I)$. The feasible H are precisely the orthogonal transformations fixing e_1. They all agree on the query $\mathcal{N}(2e_1, 3I + e_1 e_1^\top)$, but can disagree on $\mathcal{N}(e_2, I)$. The same nonunique mechanism set therefore identifies one query and leaves another ambiguous.

A sufficient three-background design. Define

$$D = \Sigma_1 - \Sigma_0, \quad F = \Sigma_2 - \Sigma_0, \quad E = C_1 - C_0, \quad G = C_2 - C_0. \tag{14}$$

Theorem 4.3 (Constructive recovery)**.** *For three pairs from* (10)*, let $d \geq 2$. Suppose $D \succ 0$, $B = D^{-1/2}FD^{-1/2}$ has simple spectrum, and every coordinate of $u = U^\top D^{-1/2}(\mu_1 - \mu_0)$ is nonzero, where U is an ordered orthonormal eigenbasis of B. If V is the correspondingly ordered eigenbasis of $E^{-1/2}GE^{-1/2}$ and $v = V^\top E^{-1/2}(m_1 - m_0)$, then the unique parameters are*

$$\begin{aligned} A &= E^{1/2}V\,\mathrm{diag}\{\mathrm{sign}(u_j v_j)\}U^\top D^{-1/2}, \\ b &= m_0 - A\mu_0, \qquad Q = C_0 - A\Sigma_0 A^\top. \end{aligned} \tag{15}$$

Proof sketch. $R = E^{-1/2}AD^{1/2}$ is orthogonal and conjugates the two whitened covariance differences. Their simple spectrum leaves independent eigenvector signs, resolved by the nonzero mean projections. Appendix C.4 proves uniqueness. The argument uses classical whitening and joint diagonalization; three backgrounds are sufficient under these conditions, not universally minimal.

Local stability. Let $r = \lambda_{\min}(D)$, $s = \lambda_{\min}(E)$, $\gamma = \min_{j \neq l} |\lambda_j(B) - \lambda_l(B)|$, and $\eta = \min_j |u_j|$. Theorem A.3 gives the local bounds

$$\|\widehat{A} - A\|_2 \leq K_A e, \quad \|\widehat{b} - b\| \leq K_b e, \quad \|\widehat{Q} - Q\|_2 \leq K_Q e$$

when support mean and covariance errors are at most $e < e_0$. The constants depend explicitly on inverse powers of r, s, γ, while η controls the neighborhood in which the sign branch remains fixed. Query mean and covariance errors enter separately. With a positive query covariance margin, the Gaussian query MMD^2 is bounded by a constant times the squared query-moment errors. Appendix C.5 gives the constants; the theorem is accompanied by the full query bounds and a Gaussian finite-sample corollary.

The numerical estimator floors covariance-difference eigenvalues at 10^{-4} and projects the recovered noise covariance onto the PSD cone. Flooring is inactive if $\min(r, s) - 2e > 10^{-4}$. PSD projection cannot increase Frobenius error relative to the true covariance and increases the spectral-error bound by at most a factor of two. The local radius remains a condition on the estimation error; we evaluate its applicability alongside prediction accuracy.

5 ACEMO-ICL

Independent evidence and conditional generation. A shared encoder maps each matrix to eight higher-order set tokens, anonymous column nodes, and second-order column relations. Inputs and

responses are encoded independently and compared at aligned columns and learned pooling slots. The decoder retains eight comparison tokens per demonstration; masked aggregation combines column and relation states. Query evidence is encoded separately. Entry, node, and relation widths are 96, 64, and 24. Only visible inputs enter the cache; response targets, source references, and zero-effect labels are reserved for training.

The velocity decomposes as

$$v(z, t \mid \mathcal{D}) = v_{\text{base}}(z, t \mid \mathcal{D}_q) + \Delta v(z, t \mid \mathcal{D}_q, \mathcal{C}). \tag{16}$$

Four base blocks read only query evidence. One correction block reads the base hidden state, demonstration tokens, and aggregated column relations, with its output head initialized to zero. Columns interact within each generated row; generated rows evolve independently given the cache. The correction models the predictive response, which need not correspond to a uniquely identified physical or causal mechanism.

Shared functions, masked reductions, and equivariant relation updates ensure input-row and support-pair invariance and joint column equivariance. Each field and every fixed-step Heun update preserves these properties in exact arithmetic. With iid noise rows, outputs are iid conditional on the cache, and a fixed noise prefix produces the same output prefix at any requested row count. Appendices C.1 and E give the proof and finite-precision checks.

Statistical target and training. Under squared MMD, the Bayes predicted law is the conditional mixture $Q^*_{\mathcal{D}} = \mathbb{E}[Q_\Theta \mid \mathcal{D}]$. Conditional expectation in the RKHS decomposes risk into irreducible prompt uncertainty and distance from this law. Ideal conditional flow matching (Lipman et al., 2023; Tong et al., 2024) regresses $Y - Z$ along $Z_t = (1-t)Z + tY$, where Gaussian noise Z and the response target Y are independent. Under the continuity-equation and flow-uniqueness assumptions in the appendix, it generates $Q^*_{\mathcal{D}}$. This is a predictive marginal law: iid draws from it do not reproduce the joint posterior dependence induced by one shared unknown mechanism.

We train on four-task blocks with loss

$$\ell_h(B) = \tfrac{1}{4}\sum_{j=1}^{4} \ell_j^T + \tfrac{0.5}{4}\sum_{j=1}^{4} \ell_j^S + 0.1\,\frac{\sum_j a_j \ell_j^I}{\max(1, \sum_j a_j)}. \tag{17}$$

Here ℓ^T trains the total response field, ℓ^S trains the base on independent query-source samples, and ℓ^I penalizes correction velocity on true zero-effect tasks ($a_j = 1$). Whole-pair leave-one-out makes the held-out input background the new query and supplies fresh source and response targets from that background. Context dropout acts at the episode level. The pretraining analysis separates representation error, joint-objective bias, optimization error, loss tails, and solver error, making explicit the assumptions needed to connect this objective to distribution risk.

Training protocol. ACEMO-ICL has 2,817,612 parameters. We train all eight neural variants with one initialization seed (20260905), the same task schedule, and 30,000 updates at effective batch size four, giving each final checkpoint 120,000 task exposures. AdamW uses learning rate 3×10^{-4}, weight decay 0.01, 1,000 warmup steps, cosine decay, gradient clipping at 1.0, and CUDA BF16. Training covers fourteen programs, six background families, real and log-positive domains, dimensions 16 and 32, and one to four demonstrations. Each support side independently contains 32–128 rows; query inputs contain 64 rows and targets 128. Nonidentity programs have exact zero, weak, and moderate strengths with probabilities 0.1, 0.2, and 0.7.

The set conditional flow baseline has 2,736,545 parameters, independent support encoders, and joint column-permutation augmentation. Task exposures are matched; ACEMO-ICL additionally receives source targets and trains for 52.4 minutes versus 31.5 minutes. The no-base ablation removes the decomposition and auxiliary losses. Appendix E compares all variants. At inference, `encode_context` caches evidence, `velocity` evaluates the field, and `sample` defaults to 128 rows and 32 Heun steps.

6 Numerical Illustrations and Experiments

Design and uncertainty. The IID test comprises 672 tasks balanced over fourteen mechanisms, six backgrounds, two domains, and two dimensions; validation uses 336 separate tasks. Each task

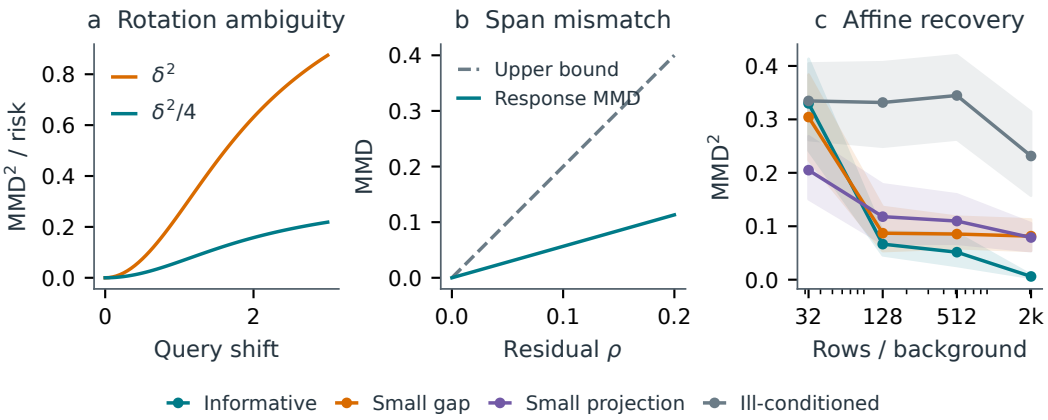

Figure 2: Identification and stability. Panels report analytic ambiguity, approximate identification, and affine recovery at dimension 4. Affine bands are pointwise 95% intervals over 20 mechanism clusters; all regularized estimates are included. Other dimensions and the complete finite-observation study appear in the appendix.

result averages three independent row draws and two generation-noise draws per row draw. The primary metric is multiscale MMD^2 estimated by the V-statistic, with the untruncated U-statistic checking sensitivity to finite-row bias. Two independent response samples provide a sampling reference. We compute confidence intervals from 2,000 mechanism-stratified, task-level paired bootstrap draws and cluster conditions that share an operator. The reported p_H values use the original Holm comparison families and paired sign-flip null assumptions. These intervals quantify task variation at fixed weights; one training seed does not measure retraining variability.

Identification and finite observations. The rotation example exhibits the response separation and risk floor of Theorem 3.3. The four-state example then separates span mismatch from background conditioning: two nearby input laws can identify the query response while requiring large span coefficients. The signed response-embedding estimator has an exact risk decomposition into residual bias and coefficient-amplified variance (Appendix D). We evaluate both the raw estimator and its probability projection over all realizations. Known input laws and span coefficients isolate the effect of observing finite response samples.

The affine study contains 960 records spanning dimension, conditioning regime, and sample size for 60 independent mechanisms. Under informative conditions, error decreases with row count: at $n = 2048$, mean query MMD^2 is 0.00268, 0.00606, and 0.04886 for $d = 2, 4, 8$. Smaller spectral gaps, smaller mean projections, and ill-conditioned covariance differences impair recovery. The local theorem describes a narrower regime than these sample sizes reach: none of the sampled moment errors satisfies its conservative radius, whereas all 240 deliberately local perturbation checks satisfy the explicit parameter bounds. PSD repair is frequent, occurring in every dimension-4 and dimension-8 condition. The reported means include every repaired estimate.

Response transfer. ACEMO-ICL improves IID response prediction, with mean MMD^2 of 0.01524 compared with 0.01715 for set conditional flow and 0.02897 for marginal calibration. The corresponding paired gains are 0.00191 (95% CI 0.00168–0.00215) and 0.01372 (0.01290–0.01460). The U-statistic preserves both rankings. Table 2 compares simple baselines and the separately trained query-only model; the appendix reports all neural variants and auxiliary metrics.

Use of demonstrations. The model uses both the response examples and their assignment to inputs. With queries and targets held fixed, replacing the context with a wrong operator from the same mechanism family increases MMD^2 by 0.02519 on average. Reassigning entire support response sets increases it by 0.00606, even though every sample set remains unchanged. This intervention

Table 2: IID response prediction on 672 tasks. MMD values are multiplied by 10^3; lower is better. The independent-target reference measures finite-row sampling effects. Neural models use the final 30,000-update checkpoint.

Method	MMD V	MMD U	Sliced W_1	Cov. error
ACEMO-ICL	15.24	7.01	0.269	0.488
Set conditional flow	17.15	8.90	0.283	0.544
Query-only model	29.01	20.69	0.359	0.553
Marginal calibration	28.97	20.80	0.365	0.621
Gaussian OT	50.49	42.11	0.520	0.856
Query resampling	35.05	27.06	0.374	0.554
Support Y resampling	133.94	125.49	0.910	0.996
Independent target reference	8.22	0.03	0.195	0.375

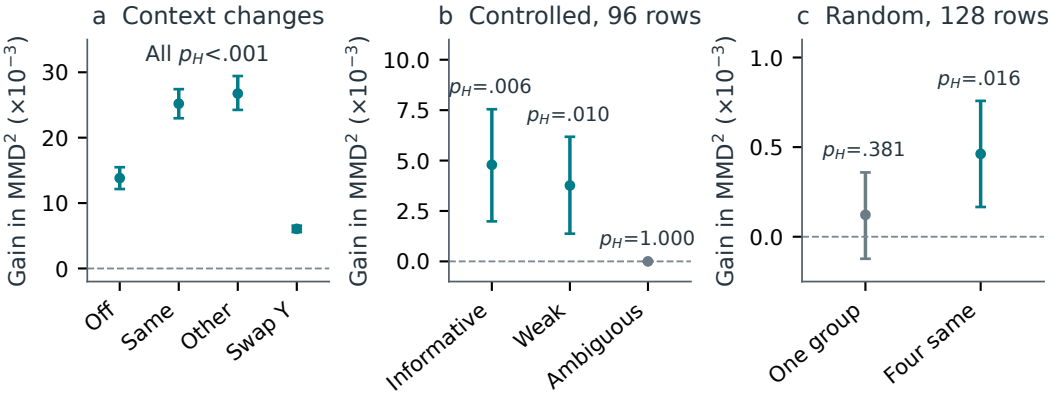

Figure 3: Context and background information. Positive differences favor correct context (a) or varied backgrounds (b,c). Same/other denote wrong operators from the same/another mechanism family; Swap Y reassigns support responses. Panel c compares varied backgrounds against one group or four same-background splits. Intervals are pointwise 95% paired bootstrap intervals; p_H retains the original Holm families. Panels a,b,c use 672 tasks, 32 operator clusters per regime, and 96 tasks, respectively.

isolates the information in input–response membership. Correct context also improves prediction over the separately trained query-only model.

Architectural gains vary in size. Retaining all demonstration tokens improves MMD^2 by 5.19×10^{-5}, about 0.34%, with CI $[3.38 \times 10^{-6}, 9.88 \times 10^{-5}]$ and $p_H = 0.0323$. The other ablations favor the full model at this budget, but mix architectural and resource changes: removing the base also removes auxiliary supervision, and the relation and moment ablations alter capacity. They therefore measure the combined effect of these changes. The appendix gives the complete post hoc breakdown by mechanism and strength.

Information from different backgrounds. At 96 total rows per side, varied backgrounds improve over repeated backgrounds by 0.00479 (CI 0.00198–0.00754) in the informative three-background panel; the ambiguous contrast is exactly zero. This panel uses broader affine rotations than the training prior, combining a controlled information contrast with mechanism extrapolation. In the separate 128-row random-background panel, four varied backgrounds outperform four same-background splits. Their gain over one larger example is unresolved: 0.00012, CI $[-0.00012, 0.00036]$. The one-example and split conditions use identical observed matrices. The benefit of multiple groups therefore depends on the information their backgrounds supply.

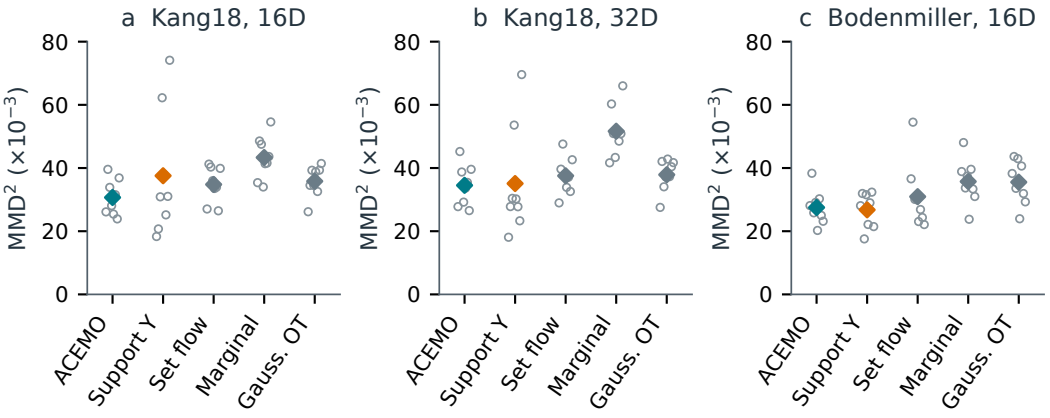

Figure 4: Observed-data examples. Points are the eight split means and diamonds their averages. Kang18 uses a fixed post hoc rotation; Bodenmiller uses original PCA coordinates. Splits share support data, so the comparison is descriptive, without independent-fold significance tests. Full preprocessing and original outcomes are retained in the appendix.

On 117 true zero-effect tasks, the upper confidence bound for model-minus-query error, after subtracting 25% of the target sampling reference, is -0.00497. This meets the prespecified average tolerance. Identity regularization still has a persistent cost in the ambiguous finite-candidate study (appendix): satisfying the average criterion on the test prior does not remove the tradeoff in indistinguishable contexts.

Real-data transfer. We evaluate the pretrained model on Kang18 and Bodenmiller, each with eight held-out backgrounds and four support backgrounds per split. Feature selection and PCA use only visible support inputs. Kang18 uses a fixed post hoc rotation chosen after the original evaluation failed. Mean MMD^2 is 0.03068 and 0.03452 at dimensions 16 and 32, compared with 0.03756 and 0.03509 for Support Y. ACEMO-ICL has lower error on 4/8 and 2/8 splits, so the average gains are concentrated in a subset of backgrounds. On Bodenmiller's original dimension-16 PCA coordinates, ACEMO-ICL gives 0.02751 against 0.02676 for Support Y, with lower error on 3/8 splits. It improves over set flow, marginal calibration, and Gaussian OT on this primary metric. These results show both the influence of representation and the strength of reusing support responses.

7 Discussion and Limitations

Demonstrations are informative when they constrain the requested response. The support span characterizes a uniform guarantee for unrestricted Markov kernels; mechanism restrictions can identify additional queries. Finite-sample prediction also depends on conditioning: nearly redundant backgrounds can amplify observation error despite population uniqueness.

ACEMO-ICL improves synthetic prediction by using demonstration content and pair membership. Its gains from informative backgrounds support the role of background design, while random-background and real-data results delimit that benefit. These findings establish predictive context use, without establishing execution of the analytic recovery procedure. The evidence covers one training seed, low-dimensional continuous observations, and a finite operator prior. Broader mechanisms and discrete observations require further study. The appendix reports all failed criteria, regularization rates, and original real-data outcomes.

AI USE STATEMENT

Generative AI tools were used only for language editing of the manuscript and for assisting with code editing and correction. All research ideas, mathematical derivations, experimental design, implementation, and analysis were carried out by the authors. The authors are responsible for the final claims, results, and code.

REPRODUCIBILITY STATEMENT

The supplement contains the complete proofs, the numerical protocols and scripts, the frozen records used in the tables, and the manuscript and figure sources. It distinguishes new CPU illustrations from analyses of previously fixed neural evaluations. The training configuration and implementation are included; rebuilding figures does not retrain the model or resample the formal tests.

ETHICS STATEMENT

The observed-data illustrations reuse public research datasets in reduced representations. No participants were recruited for this study. We report distribution-level prediction, without claims about individual responses or clinical decisions. The original data sources, preprocessing choices, and post hoc representation changes are documented in Appendix E.5.

REFERENCES

Bioconductor. *muscData: Multi-Sample Multi-Group Single-Cell RNA-Seq Data*, 2026. URL https://bioconductor.org/packages/release/data/experiment/vignettes/muscData/inst/doc/muscData.html. Official dataset vignette; Kang18 EH2259 and Bodenmiller EH2254; accessed September 6, 2026.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. URL https://jmlr.org/papers/v21/19-322.html.

Frank Cole, Dixi Wang, Yineng Chen, Yulong Lu, and Rongjie Lai. In-context operator learning on the space of probability measures. *arXiv:2601.09979*, 2026. URL https://arxiv.org/abs/2601.09979.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Martin Kleinsteuber and Hao Shen. Uniqueness analysis of non-unitary matrix joint diagonalization. *arXiv:1111.7088*, 2012. URL https://arxiv.org/abs/1111.7088. Version 3.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.09902.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2112.10510.

Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *PMLR*, pp. 1049–1057, 2013. URL https://proceedings.mlr.press/v28/oliva13.html.

Marvin Sextro, Weronika Kłos, and Gabriel Dernbach. MapPFN: Learning causal perturbation maps in context. *arXiv:2601.21092*, 2026. URL https://arxiv.org/abs/2601.21092v3. Version 3.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *arXiv:2302.00993*, 2023. URL https://arxiv.org/abs/2302.00993.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

A Supporting Theoretical Statements

A.1 Anonymous generation

Let $C = E(\mathcal{D})$ denote the cached evidence and $S_N(C, Z)$ a deterministic N-row sampler with noise $Z \in \mathbb{R}^{N\times d}$. A permutation of support pairs may reorder the associated context-token blocks. The sampler must be invariant to this block reordering and equivariant to a common column permutation.

Theorem A.1 (Structure of the prototype)**.** *Assume shared entry functions, masked symmetric reductions, permutation-compatible attention, equivariant column-relation updates, and a velocity field with no interaction between generated rows. At deterministic inference with fixed weights and domain, and with masks permuted alongside their entries, the sampler has the following exact-arithmetic properties. Independent permutations of input rows and whole support pairs leave its output invariant. Jointly permuting all input and noise columns yields*

$$S_N(E(\mathcal{D}\Pi), Z\Pi) = S_N(E(\mathcal{D}), Z)\Pi. \tag{18}$$

Both the base and correction fields preserve these properties, as does any fixed number of Heun steps. For iid noise rows independent of the context, generated rows are iid conditional on that context. Moreover, for $M \le N$,

$$S_N(C, Z)_{1:M} = S_M(C, Z_{1:M}). \tag{19}$$

Proof sketch. Attention is invariant to a simultaneous permutation of its keys and values, and equivariant to a permutation of its queries. Apply these identities through the independent encoders, pair comparisons, and relation updates. Each generated row then solves the same deterministic recurrence using only its own state and the fixed context. The Heun recurrence preserves the symmetries by composition and addition. The full argument is in Appendix C.1.

Increasing the output count improves Monte Carlo accuracy while preserving each row's law. Independence holds conditional on the context; averaging over a random shared context can induce marginal dependence. Prefix consistency requires the same noise prefix, which a fixed device-RNG seed need not produce across different array shapes.

A.2 Identity in ambiguous contexts

Corollary A.2 (Identity–response tradeoff)**.** *In Theorem 3.3, suppose $Q_0 = P_q$ and the identity risk satisfies $\mathbb{E}_0 \operatorname{MMD}_k^2(\widehat{Q}, Q_0) \le r$. Then*

$$\mathbb{E}_1 \operatorname{MMD}_k^2(\widehat{Q}, Q_1) \ge (\delta - \sqrt{r})_+^2. \tag{20}$$

In particular, strict identity preservation forces risk δ^2 on the indistinguishable alternative.

A.3 Local stability and finite-sample prediction

The recovery formula is smooth within a region where whitening is well conditioned, eigenvalues are separated, and the sign choices are fixed. Let

$$r = \lambda_{\min}(D), \quad s = \lambda_{\min}(E), \quad \gamma = \min_{j\neq l} |\lambda_j(B) - \lambda_l(B)|, \quad \eta = \min_j |u_j|. \tag{21}$$

These quantities govern the sensitivity of an otherwise unique population solution.

Theorem A.3 (Local parameter and query stability)**.** *Under Theorem 4.3, obtain $(\widehat{A}, \widehat{b}, \widehat{Q})$ by substituting estimated moments into* (15)*, before PSD projection. Explicit constants $e_0, K_A, K_b, K_Q > 0$ are given in Appendix C.5. Errors at most $e < e_0$ in each support mean (Euclidean norm) and covariance (spectral norm) imply*

$$\left\|\widehat{A} - A\right\|_2 \le K_A e, \quad \left\|\widehat{b} - b\right\| \le K_b e, \quad \left\|\widehat{Q} - Q\right\|_2 \le K_Q e. \tag{22}$$

The constants retain inverse powers of r, s and γ; η controls the admissible radius. Let $\alpha = K_A e$, $\beta = K_b e$, and let query moment errors be at most e_q. Using a PSD noise estimate with spectral error at most e_Q, the estimated query moments obey

$$\begin{aligned} \|\widehat{m}_q - m_q\| &\le \alpha \|\mu_q\| + (\|A\|_2 + \alpha)e_q + \beta =: a_q, \\ \left\|\widehat{C}_q - C_q\right\|_2 &\le \alpha \|\Sigma_q\|_2 (2\|A\|_2 + \alpha) + (\|A\|_2 + \alpha)^2 e_q + e_Q =: c_q. \end{aligned} \tag{23}$$

If $C_q \succeq cI$ and $c_q \le c/2$, then, for the kernel (5)*,*

$$\mathrm{MMD}_k^2(\widehat{Q}_q, Q_q) \le L_k^2\big(a_q^2 + dc_q^2/c\big), \qquad L_k^2 = \frac{1}{4d}\sum_h h^{-2}. \tag{24}$$

Proof sketch. Control inverse square roots along a positive-definite segment, then use eigenvector perturbation and a sign-stability radius. Telescope the parameter and query formulas. Couple the two Gaussian laws with the same standard Gaussian and bound the difference of covariance square roots; the kernel embedding is L_k-Lipschitz. Full constants and proofs are in Appendix C.5.

At a fixed covariance-difference scale, $r = \|D\|_2/\kappa(D)$ and $s = \|E\|_2/\kappa(E)$ expose the dependence on condition numbers. Small mean projections shrink the valid neighborhood without adding a $1/\eta$ factor to K_A on a fixed sign branch. Outside that neighborhood, a sign switch can produce a large recovery error even when individual covariance estimates are accurate.

To obtain a finite-sample bound, suppose each of the six support laws and the query input is Gaussian, has covariance norm at most M, and is observed with at least $n \ge 2$ independent rows. Define

$$t = d\log 9 + \log(28/\delta), \qquad e_n = \max\left\{2\sqrt{\frac{2Mt}{n}},\ 4M\left(\sqrt{\frac{t}{n-1}} + \frac{t}{n-1}\right)\right\}. \tag{25}$$

For $0 < \delta < 1$, with probability at least $1-\delta$, every mean and unbiased sample covariance has its respective error bounded by e_n (Appendix C.6). If $e_n < e_0$, Theorem A.3 applies with $e = e_q = e_n$, subject to its query covariance condition. The local-radius condition determines when the finite-sample guarantee becomes informative.

Regularized moment recovery. For the numerical illustration, the regularized estimator floors the eigenvalues of estimated D, E at $\tau = 10^{-4}$ and projects the recovered noise covariance onto the PSD cone. If $\min(r,s) - 2e > \tau$, whitening truncation is inactive. For any symmetric raw estimate $\widetilde{Q}$ and true $Q \succeq 0$, PSD projection satisfies

$$\left\|\widetilde{Q}_+ - Q\right\|_F \le \left\|\widetilde{Q} - Q\right\|_F, \qquad \left\|\widetilde{Q}_+ - Q\right\|_2 \le 2\left\|\widetilde{Q} - Q\right\|_2. \tag{26}$$

Thus $e_Q = 2K_Q e$ is valid in (23) when truncation is inactive. Projection ensures a valid covariance, potentially at the cost of fitting the observed moments. We report this estimator's prediction errors and regularization rates separately from the neural model and the earlier unregularized baseline.

B Conditional Generation and Pretraining Generalization

B.1 The Bayes response law and the cost of compression

Finite observations leave uncertainty about both the mechanism and the input backgrounds. Let Θ be the latent task drawn from a specified prior, with query response Q_Θ. Assume standard Borel task and observation spaces and a measurable response kernel $\Theta \mapsto Q_\Theta$. The Bayes statements below use this prior and condition on dimension; risks across dimensions are averaged with the training weights. Define

$$Q^*_{\mathcal{D}}(B) = \mathbb{E}[Q_\Theta(B) \mid \mathcal{D}] \tag{27}$$

as the conditional mixture response. By averaging compatible response laws rather than their means, it can preserve multiple modes.

Theorem B.1 (Bayes risk and evidence compression)**.** *For a bounded continuous characteristic kernel and any $\mathcal{D}$-measurable predicted law $\widehat{Q}$, the population risk decomposes as*

$$\mathbb{E}\,\mathrm{MMD}_k^2(\widehat{Q}, Q_\Theta) = \underbrace{\mathbb{E}\,\mathrm{MMD}_k^2(Q^*_{\mathcal{D}}, Q_\Theta)}_{\mathcal{R}_{\mathrm{Bayes}}} + \mathbb{E}\,\mathrm{MMD}_k^2(\widehat{Q}, Q^*_{\mathcal{D}}). \tag{28}$$

If $E = E(\mathcal{D})$ is an encoded representation and $Q^*_E = \mathbb{E}[Q_\Theta \mid E]$, then any E-measurable prediction satisfies

$$\begin{aligned}\mathbb{E}\,\mathrm{MMD}_k^2(\widehat{Q}, Q_\Theta) = \mathcal{R}_{\mathrm{Bayes}} &+ \mathbb{E}\,\mathrm{MMD}_k^2(Q^*_{\mathcal{D}}, Q^*_E) \\ &+ \mathbb{E}\,\mathrm{MMD}_k^2(\widehat{Q}, Q^*_E).\end{aligned} \tag{29}$$

If $E_0 = g(E_1)$ deterministically, the Bayes risk available from E_1 is no greater than that from E_0.

Proof sketch. Conditional expectation is an orthogonal projection in the Hilbert space of square-integrable RKHS embeddings. Boundedness allows expectation and kernel integration to commute. Apply the projection identity once for full observation and again for the nested representation. See Appendix C.7.

The risk ordering requires nested representations. Separately trained encoders need not satisfy this condition, so the theorem alone does not order the token ablation. Both representations must also retain any query-derived evaluation scale.

Marginal generation versus a shared-task posterior. An iid sampler from $Q^*_{\mathcal{D}}$ has joint law $(Q^*_{\mathcal{D}})^{\otimes N}$. A posterior over a shared unknown task gives $\mathbb{E}[Q_\Theta^{\otimes N} \mid \mathcal{D}]$. In the first law, each row independently averages over task uncertainty; in the second, one latent task is shared across the batch. The laws generally differ. Our MMD objective and sampler target the marginal response law rather than this joint posterior.

B.2 Conditional flow matching for the Bayes response

In working coordinates, sample $Y \sim Q_\Theta$ independently of the prompt given the task, with $Z \sim \mathcal{N}(0, I_d)$ and $t \sim \mathrm{Unif}[0,1]$ independent of each other, the task, prompt, and target. Set

$$Z_t = (1-t)Z + tY, \qquad U = Y - Z. \tag{30}$$

This training path couples noise Z to a target Y and requires no pairing between support inputs and responses.

Proposition B.2 (Ideal conditional FM)**.** *Assume square integrability. The conditional squared-loss minimizer is*

$$u(z,t,\mathcal{D}) = \mathbb{E}[U \mid Z_t = z, t, \mathcal{D}]. \tag{31}$$

The FM excess of any square-integrable field v is exactly

$$\mathbb{E}\frac{\|v(Z_t,t,\mathcal{D}) - U\|^2}{d} - \mathbb{E}\frac{\|u(Z_t,t,\mathcal{D}) - U\|^2}{d} = \mathbb{E}\frac{\|v(Z_t,t,\mathcal{D}) - u(Z_t,t,\mathcal{D})\|^2}{d}. \tag{32}$$

The conditional path laws satisfy the continuity equation with velocity u. If this continuity equation has a unique probability solution with the Gaussian initial law in a class containing the conditional path laws, and that solution is induced by a characteristic flow through $t = 1$, the flow ends at $Q^*_{\mathcal{D}}$. Replacing $\mathcal{D}$ by an encoded representation gives the corresponding Q^*_E.

Proof sketch. The regression identity is orthogonality of the conditional residual. Differentiate $\mathbb{E}[f(Z_t) \mid \mathcal{D}]$ for smooth test functions and condition on Z_t. The resulting weak equation has the required initial and terminal laws. The stated existence and uniqueness conditions identify this solution with the ODE-generated law. See Appendix C.7.

B.3 The joint training objective

The base field reads only query evidence $\mathcal{D}_q$. The total field is $v_h = v_{\mathrm{base},h} + \Delta v_h$, where h denotes all learnable parameters. Let ℓ_j^T be response FM loss for episode j, ℓ_j^S the base-field FM loss on an independently sampled source reference, and ℓ_j^I the squared correction at response-path states. Each is averaged over valid target rows and coordinates. Let a_j indicate a true zero-effect task. For a block $B = (B_1, \ldots, B_b)$ with $b = 4$, the implemented loss is Equation (17). The last term averages the eligible identity episodes and is zero if the block contains none. Under iid episodes within a dimension and zero-effect probability p_d, its expectation equals $[1 - (1-p_d)^b]\mathbb{E}[\ell^I \mid a = 1, d]$. Thus the effective population weight depends on how often a block contains an identity episode.

Write $J(h)$ for the equal-weight mixture over dimensions of the expected block loss, $R_T(h)$ for its target part, $R_S(h)$ for its unweighted source part, and $I(h)$ for the expected ratio term before weight 0.1. Let R_T^* be the unrestricted full-visible-observation FM Bayes loss, and R_S^* the unrestricted query-only source FM Bayes loss. Take a nonempty joint class $\mathcal{F}$ with finite population and empirical losses, and define

$$\mathcal{A}_T = \inf_{h \in \mathcal{F}} R_T(h) - R_T^*, \tag{33}$$

$$\mathcal{B}_{\mathrm{joint}} = \inf_{h \in \mathcal{F}} J(h) - \inf_{h \in \mathcal{F}} R_T(h) - 0.5 R_S^* \;\geq 0.$$

The approximation term measures restrictions imposed by the representation and field class. The joint-objective term includes source-field approximation, parameter sharing across objectives, and the identity penalty. It can remain positive even when the target field is represented exactly. For every h,

$$R_T(h) - R_T^* \le J(h) - \inf_{g \in \mathcal{F}} J(g) + \mathcal{A}_T + \mathcal{B}_{\text{joint}}. \tag{34}$$

B.4 Dimension-stratified finite pretraining

The training schedule alternates dimensions 16 and 32. Within each dimension, assume independent blocks drawn from a common distribution over masks, row counts, context dropout, leave-one-out selection, and path noise. There are $n_{16} = n_{32} = 15{,}000$ blocks, each containing four task exposures. The bound treats blocks as independent units and allows dependence among paths and rows within an episode.

Let $\widehat{J}$ be the empirical average of (17). For cap $c > 0$ and $0 < \delta < 1$, define $\ell_h^{(c)} = \min(\ell_h, c)$ and the corresponding $J_c, \widehat{J}_c$. Suppose the clipped loss functions in dimension d admit a uniform ϵ-net of cardinality $N_d(\epsilon)$. Define

$$U_c = \frac{1}{2} \sum_{d \in \{16,32\}} \left[2\epsilon + c\sqrt{\frac{\log(4N_d(\epsilon)/\delta)}{2n_d}} \right], \tag{35}$$

$$\tau_c = \sup_{h \in \mathcal{F}} [J(h) - J_c(h)], \qquad \widehat{\tau}_c = \sup_{h \in \mathcal{F}} [\widehat{J}(h) - \widehat{J}_c(h)], \qquad V_c = U_c + \tau_c + \widehat{\tau}_c.$$

Assume a measurable, separable loss class and measurable supremum envelopes. The bound is finite when both the clipped class and the uniform tails are controlled.

Theorem B.3 (Pretraining bound for the implemented objective)**.** *Under the block-sampling and covering assumptions above, with probability at least* $1 - \delta$*,*

$$\sup_{h \in \mathcal{F}} |J_c(h) - \widehat{J}_c(h)| \le U_c, \qquad \sup_{h \in \mathcal{F}} |J(h) - \widehat{J}(h)| \le V_c. \tag{36}$$

For any data-dependent $\widehat{h} \in \mathcal{F}$*, define its empirical optimization gap* $\varepsilon_{\text{opt}} = \widehat{J}(\widehat{h}) - \inf_{h \in \mathcal{F}} \widehat{J}(h)$*. Then*

$$R_T(\widehat{h}) - R_T^* \le \mathcal{A}_T + \mathcal{B}_{\text{joint}} + 2V_c + \varepsilon_{\text{opt}}. \tag{37}$$

Proof sketch. Apply Hoeffding's inequality to each finite net, take a union bound over its elements and the two strata, and extend the result using two approximation errors. Add the population and empirical clipping tails. Comparing $\widehat{h}$ with an arbitrary member of the class introduces two uniform-deviation terms; (34) then bounds the target excess. See Appendix C.8.

For example, if a parameter ball of radius R in p dimensions induces losses uniformly L_θ-Lipschitz in the parameters, then $\log N_d(\epsilon) \le p \log(1 + 2RL_\theta/\epsilon)$. Uniform parameter Lipschitzness is an additional condition, typically requiring controlled observations and paths; it does not follow from parameter count alone. Alternatively a finite candidate class has $N_d(0)$ equal to its cardinality. For unbounded losses, a uniform $(1 + \nu)$-moment envelope gives $\tau_c \le M_{1+\nu}/c^\nu$; an empirical counterpart bounds $\widehat{\tau}_c$.

The independent units in this bound are the 30,000 blocks containing 120,000 task exposures, since identity loss is normalized within each block. Additional rows and source targets supply supervision within those tasks.

B.5 From FM excess to generated distribution error

We connect field error to distribution error under the following regularity assumptions. The ideal field in Proposition B.2 admits a flow with the prescribed marginals; the learned field is uniformly L-Lipschitz in state and has a unique ODE solution of finite second moment. Its s-step Heun discretization has normalized strong error at most $H^2 s^{-4}$. Appendix C.9 derives this last bound from a sufficient local-defect condition. These are additional assumptions on the field, not consequences of the training objective.

Table 3: Scope of the theoretical guarantees.

Result	Required setting	Scope limit
Structure, Thm. A.1	Shared, masked, row-separable architecture; exact arithmetic	Bitwise equality of GPU reductions
Span, Thm. 3.1	Exact laws; all shared Markov kernels	Necessity for each particular output; reweighting
Affine, Thms. 4.3–A.3	Invertibility and informative, stable moment differences	Recovery for arbitrary programs or small n
Bayes/FM, Thm. B.1, Prop. B.2	Task prior, conditional integrability; unique flow when used	Exact posterior inference by the network
Pretraining, Thm. B.3	Independent stratified blocks; controlled loss class and tails	Global optimizer, tight numerical certificate
Sampling, Cor. B.4	State Lipschitzness and Heun strong-error control	Automatic regularity or raw-positive-space bounds

Corollary B.4 (Conditional distribution-risk bound)**.** *Let $\widehat{Q}_{\widehat{h},s}$ be the Heun-generated law in working coordinates and set $c_k = \frac{1}{4}\sum_h h^{-2}$. Under the preceding regularity assumptions and the event of Theorem B.3,*

$$\mathbb{E}\,\mathrm{MMD}_k^2(\widehat{Q}_{\widehat{h},s}, Q_\Theta) \le \mathcal{R}_{\mathrm{Bayes}} + 2c_k\left[e^{2L}\left(\mathcal{A}_T + \mathcal{B}_{\mathrm{joint}} + 2V_c + \varepsilon_{\mathrm{opt}}\right) + H^2 s^{-4}\right]. \tag{38}$$

For an independent N-row generated sample and m-row target sample, the expected MMD V-statistic adds nonnegative diagonal terms whose sum is at most $1/N + 1/m$. For $N, m \ge 2$, the untruncated U-statistic is conditionally unbiased for population MMD^2 and can be negative.

Proof sketch. Couple ideal and learned ODEs with the same initial Gaussian, evaluate field error along the ideal path, and use Gronwall followed by Cauchy–Schwarz. The kernel embedding is Lipschitz, so a Wasserstein coupling controls MMD. Add the Heun error with the squared triangle inequality, then use (28). Counting diagonal and off-diagonal kernel pairs gives the finite-row correction. See Appendix C.9.

Equation (38) separates errors that additional training can address from those requiring changes to the model, objective, or sampler. More independent training blocks reduce the sampling term when the class and tails are controlled; optimization enters through its own gap. Prompt uncertainty remains even for an ideal learner. Appendix C.10 illustrates these distinctions with a finite candidate class and a known prior.

C Complete Proofs and Technical Details

C.1 Structural properties through discrete generation

Proof of Theorem A.1. For attention $\mathcal{A}(Q,K,V) = \mathrm{softmax}(QK^\top/\sqrt{h} + M)V$, let P permute keys and values and let R permute queries, carrying the mask with the same indices. Permuting softmax summands gives

$$\mathcal{A}(Q, PK, PV) = \mathcal{A}(Q,K,V), \qquad \mathcal{A}(RQ,K,V) = R\mathcal{A}(Q,K,V).$$

In the first identity, the column permutation of the softmax matrix cancels the permutation of V. The second identity follows by permuting its rows. Invalid keys have zero masked weight, so both identities hold under independent padding that preserves valid indices.

Entry embeddings and feed-forward maps share parameters and have no row or column index embeddings. Normalization acts on the feature dimension of an individual entry, not on an unordered observation index. These operations commute with row and column permutations. Within a matrix, masked sums and attention over rows yield row-invariant tokens and column-equivariant nodes. Internal moment contractions sum over that matrix's own rows. Consequently an edge tensor transforms by $E \mapsto \Pi^\top E \Pi$ under a column permutation. None of these contractions pairs a row from one support side with a row from the other.

Each support comparison applies the same function to the independent X_i and Y_i summaries. Matching token slots refers to fixed learned pooling seeds, while matching node indices refers to the columns aligned across the episode. Independent sample-row permutations therefore disappear before comparison. Shared masked pair gates and sums are invariant to reordering whole pairs. Retained comparison tokens may have their eight-token blocks reordered, along with their validity masks. Decoder cross-attention is invariant to that simultaneous key/value permutation.

The second-order linear maps are the known permutation-compatible contractions and broadcasts generated by identity, transpose, row sums, column sums, diagonals, total sums, and trace (Maron et al., 2019). Each commutes with $E \mapsto \Pi^\top E \Pi$. Their shared linear combinations, entrywise nonlinearities, and equivariant column-message contractions do so as well. Query-only summaries obey the same argument. The normalization scale is a symmetric scalar statistic of visible query entries, including its clamp and fallback, so it is invariant under these permutations.

With fixed encoded evidence, a generated-row field has the form $v_j(z, t, C) = f(z_j, t, C)$, with the same function for every j. Its attention involves either columns of that row or fixed prompt tokens. The four query-base blocks and the correction block each have this property. Their column actions commute with the common permutation, so their sum does too. Context caching consists of applying and storing fixed deterministic functions of C; it does not introduce dependence on generated row count or row index.

For step length h, Heun updates one row by

$$z^+ = z + \frac{h}{2}\left[f(z, t, C) + f(z + hf(z, t, C), t + h, C)\right].$$

Composition and addition preserve row separability and column equivariance. Induction over any finite number of steps proves the same properties for the discrete sampler, without an ODE convergence assumption. Restoring an invariant scalar and applying a coordinatewise exponential preserve them. Thus each output row is a deterministic shared function of its own noise row and the fixed context. Noise rows that are iid and independent of the context give conditional iid outputs, and restricting to a fixed noise prefix gives (19). The pair-order and input-row statements and (18) follow from the preceding encoder identities. □

The proof holds in exact arithmetic. Parallel floating-point reductions can change summation order and introduce small discrepancies. Output independence is conditional on the context; averaging over a random shared context can induce marginal dependence.

C.2 Linear-span characterization

Lemma C.1 (Bounded separation of finitely many laws)**.** *If a finite signed measure ν is not in the real span of probability measures $P_1, \dots, P_K$, there exists a bounded measurable real f with $\int f\, dP_i = 0$ for every i and $\int f\, d\nu \neq 0$.*

Proof. Let $\mathcal{V}$ be the real vector space of bounded measurable functions. Define linear functionals $L_i(f) = \int f\, dP_i$ and $L_\nu(f) = \int f\, d\nu$, and the map $L : \mathcal{V} \to \mathbb{R}^K$, $L(f) = (L_i(f))_i$. If L_ν vanished on $\ker L$, the rule $\ell(L(f)) = L_\nu(f)$ would be a well-defined linear functional on the finite-dimensional image of L. Extend it linearly to $\mathbb{R}^K$ to obtain coefficients a_i. Then $L_\nu = \sum_i a_i L_i$ on all bounded measurable functions. Applying this equality to indicators of measurable sets gives $\nu = \sum_i a_i P_i$, a contradiction. Thus some $f \in \ker L$ has $L_\nu(f) \neq 0$. Boundedness is built into $\mathcal{V}$; no topological separation or density regularity is required. □

Proof of Theorem 3.1. If $P_q = \sum_i a_i P_i$, then for any measurable output set B and any kernel K,

$$(\mathsf{K}P_q)(B) = \int \mathsf{K}(x, B)\, dP_q(x) = \sum_i a_i \int \mathsf{K}(x, B)\, dP_i(x).$$

Hence two kernels agreeing on the supports agree on P_q. This also proves the formula for the query response. Evaluating the input-law equality on the whole space shows $\sum_i a_i = 1$.

Conversely, suppose P_q is outside the span. Lemma C.1 supplies f; rescale it to $\|f\|_\infty \le 1$. Choose distinct output laws ν_0, ν_1 and $0 < \epsilon < 1$. Define

$$\mathsf{K}_\pm(x, dy) = \frac{1 \pm \epsilon f(x)}{2}\nu_1(dy) + \frac{1 \mp \epsilon f(x)}{2}\nu_0(dy).$$

The weights are nonnegative, sum to one, and are measurable in x, so these are Markov kernels. Their signed response difference on a law P is

$$\mathsf{K}_+P - \mathsf{K}_-P = \epsilon \left(\int f \, dP \right) (\nu_1 - \nu_0).$$

This vanishes for every P_i. For P_q, both the scalar integral and the signed measure $\nu_1 - \nu_0$ are nonzero, so their product is nonzero. Thus uniform query identification fails, proving necessity. Distinct Gaussians can serve as ν_0, ν_1 when continuous outputs are required. □

The failure of linearity for reweighting is already visible on a two-state space. Take unequal positive weights and two distinct input laws. The denominator of $T_w(aP + (1-a)P')$ mixes their weight integrals; normalizing this mixture generally differs from mixing $T_w(P)$ and $T_w(P')$. The preceding argument therefore does not apply.

C.3 Ambiguity, repetition, and the identity tradeoff

Proof of Theorem 3.3. Both hypotheses induce the same joint visible-data law: the matrix sampling laws and the row-count, mask, and domain protocol agree, with conditional independence as specified in Section 2. Any randomization of the estimator can be coupled with the same additional independent random variable under both hypotheses. On this common space set $a = \Phi(Q_0)$, $b = \Phi(Q_1)$ and $M = \Phi(\widehat{Q})$. The Hilbert identity

$$\frac{\|M - a\|^2 + \|M - b\|^2}{2} = \|M - (a+b)/2\|^2 + \frac{\|a - b\|^2}{4}$$

holds pointwise. Take expectations to prove (9). The maximum of two risks is at least their average. The embedding of $(Q_0 + Q_1)/2$ is $(a+b)/2$, so the first term vanishes for this predictor, giving attainment under the equal prior. □

Proof of Corollary A.2. Use the same common probability space and apply the triangle inequality in $L^2(\Omega; \mathcal{H}_k)$:

$$\delta = \|a - b\| \le \left(\mathbb{E} \|M - a\|^2 \right)^{1/2} + \left(\mathbb{E} \|M - b\|^2 \right)^{1/2}.$$

The first term is at most $\sqrt{r}$. Rearranging and using nonnegativity of the second gives (20). □

Repeated backgrounds. Adding a pair (P_i, Q_i) already present in (6) repeats an existing constraint and leaves both the compatible class and its query responses unchanged. Independent samples can improve estimation of that pair, but have identical laws under any two mechanisms that agree on it.

Analytic Gaussian MMD. Let $X \sim \mathcal{N}(\mu, C)$ and $Y \sim \mathcal{N}(\nu, D)$ be independent, and write $a = \mu - \nu$, $S = C + D$, and $\ell^2 = dh^2$. Completing the square in the Gaussian integral gives

$$\mathbb{E} \exp\left(-\frac{\|X - Y\|^2}{2\ell^2} \right) = \det(I + S/\ell^2)^{-1/2} \exp\left[-\tfrac{1}{2} a^\top (\ell^2 I + S)^{-1} a \right]. \tag{39}$$

For positive definite S, combine the exponents of its Gaussian density and the kernel, integrate the resulting quadratic, and simplify the determinant and mean terms. For semidefinite S, apply the formula to $S + \epsilon I$ and let $\epsilon \downarrow 0$ by bounded convergence. The three independent-pair expectations XX', YY', XY yield MMD^2; averaging over bandwidths gives the numerical curves. The rotation preserves the centered isotropic support law, but changes ae_1 to $-ae_1$. A characteristic kernel makes the separation positive whenever $a \neq 0$.

C.4 AFFINE FEASIBILITY AND EXACT RECOVERY

Proof of Proposition 4.1. Let (A', b', Q') be a compatible triple with A' invertible and set $H = A^{-1}A'$. Subtracting the reference input/output mean equations gives $A'\Delta\mu_i = A\Delta\mu_i$, hence $H\Delta\mu_i = \Delta\mu_i$. Subtracting covariance equations similarly gives $A'\Delta\Sigma_i A'^\top = A\Delta\Sigma_i A^\top$, hence the congruence constraint. The reference equations then force

$$b' = m_0 - AH\mu_0 = b + A(I - H)\mu_0, \quad Q' = C_0 - AH\Sigma_0 H^\top A^\top = Q_H.$$

A valid noise covariance requires $Q_H \succeq 0$. Conversely, any invertible H satisfying these constraints defines a valid affine-Gaussian triple. Substitution into the reference equations, followed by addition of the differences, reproduces every support mean and covariance and hence every Gaussian support law. Substituting the query moments gives (13). Since A is invertible and identity is feasible, all query responses equal that of the identity choice if and only if every feasible transformation fixes both query differences. □

In Example 4.2, the covariance difference is I and the mean difference is e_1. Thus $HH^\top = I$ and $He_1 = e_1$ are necessary and sufficient. Since $\Sigma_0 = I$, $Q_H = 0.2I$ for every such H, so no feasibility restriction removes this ambiguity. The identified query has mean difference $2e_1$ and covariance difference $2I + e_1 e_1^\top$, both fixed by this group. The alternative query e_2 is not fixed by $\mathrm{diag}(1, -1, -1)$.

Proof of Theorem 4.3. Subtracting support covariances gives $E = ADA^\top \succ 0$ and $G = AFA^\top$. Set $R = E^{-1/2}AD^{1/2}$. Then $RR^\top = I$ and

$$E^{-1/2}GE^{-1/2} = RBR^\top.$$

Order the simple eigenvalues of B increasingly. Its normalized eigenvectors form U, while V is an ordered eigenbasis of the output matrix. For every j, Ru_j^{eig} is a normalized eigenvector of the output for the same eigenvalue. The eigenspace is one-dimensional, so $R = VSU^\top$ for diagonal S with entries ± 1. The mean difference gives

$$V^\top E^{-1/2}(m_1 - m_0) = SU^\top D^{-1/2}(\mu_1 - \mu_0),$$

or $v = Su$. No u_j is zero, so $S_{jj} = \mathrm{sign}(u_j v_j)$ uniquely. This proves (15). Flipping any eigenvector in either basis flips the corresponding mean coordinate and sign entry, leaving A unchanged. The reference equations uniquely fix b, Q. The recovered Q is PSD because the exact moments were generated by a feasible true model. Any other compatible triple has the same whitened conjugacy and nonzero mean projections, and hence the same recovered parameters. □

C.5 EXPLICIT STABILITY CONSTANTS AND PREDICTION

Use Euclidean vector norm and spectral matrix norm unless a subscript specifies otherwise. Choose $0 < \rho < \min(r, s)/2$. Define

$$\begin{aligned} H_D &= (r - 2\rho)^{-1/2}, & L_D &= (r - 2\rho)^{-3/2}, \\ H_E &= (s - 2\rho)^{-1/2}, & L_E &= (s - 2\rho)^{-3/2}, \\ B_D &= 2H_D L_D(\|F\| + 2\rho) + 2H_D^2, & B_E &= 2H_E L_E(\|G\| + 2\rho) + 2H_E^2, \\ K_U &= 4B_D/\gamma, & K_V &= 4B_E/\gamma. \end{aligned} \tag{40}$$

Put $u_0 = D^{-1/2}(\mu_1 - \mu_0)$, $v_0 = E^{-1/2}(m_1 - m_0)$, and

$$\begin{aligned} a_u &= L_D \|\mu_1 - \mu_0\| + 2H_D, \qquad a_v = L_E \|m_1 - m_0\| + 2H_E, \\ e_0 &= \min\left\{\rho, \frac{\gamma}{4\max(B_D, B_E)}, \frac{\eta}{2\max(K_U \|u_0\| + a_u, K_V \|v_0\| + a_v)}\right\}. \end{aligned} \tag{41}$$

Finally set

$$\begin{aligned} J_E &= \sqrt{\|E\| + 2\rho}, \qquad K_R = \sqrt{d}(K_U + K_V), \\ K_A &= H_E H_D + J_E K_R H_D + J_E L_D, \\ K_b &= 1 + K_A \|\mu_0\| + \|A\| + K_A\rho, \\ K_Q &= 1 + K_A \|\Sigma_0\| (2\|A\| + K_A\rho) + (\|A\| + K_A\rho)^2. \end{aligned} \tag{42}$$

The theorem's strict inequalities ensure that all constants are finite.

Proof of Theorem A.3. Every estimated moment difference has error at most $2e$. Weyl's eigenvalue inequality gives minimum eigenvalues at least $r - 2\rho$ and $s - 2\rho$ for estimated D, E and all intervening line segments. For a positive definite matrix M with minimum eigenvalue a, the derivative norm of its inverse square root is at most $1/(2a^{3/2})$. This follows by differentiating the integral $M^{-1/2} = \pi^{-1} \int_0^\infty t^{-1/2}(M + tI)^{-1} dt$ and bounding the two resolvents. Its square-root derivative solves a Sylvester equation and has norm at most $1/(2\sqrt{a})$. Integrating these derivative bounds along the positive-definite segments avoids any assumption of commuting perturbations. Consequently,

$$\left\|\widehat{D}^{-1/2} - D^{-1/2}\right\| \le L_D e, \quad \left\|\widehat{E}^{-1/2} - E^{-1/2}\right\| \le L_E e, \quad \left\|\widehat{E}^{1/2} - E^{1/2}\right\| \le H_E e.$$

Both true and estimated whiteners are bounded by H_D or H_E. Telescoping $\widehat{D}^{-1/2}\widehat{F}\widehat{D}^{-1/2} - D^{-1/2}FD^{-1/2}$ yields a bound $B_D e$; the output analogue is bounded by $B_E e$.

Consider a normalized perturbed eigenvector $\widehat{u}_j$ and ordered eigenvalue $\widehat{\lambda}_j$. Weyl's inequality moves the eigenvalue by at most $B_D e < \gamma/4$. Projecting $(B - \widehat{\lambda}_j I)\widehat{u}_j = -(\widehat{B} - B)\widehat{u}_j$ onto the complement of its true eigenspace shows that the perpendicular component has norm at most $2B_D e/\gamma$. Align its sign with the true eigenvector. Euclidean distance is at most $\sqrt{2}$ times the sine of the angle, and hence at most $K_U e$. The same reasoning gives $K_V e$ for each output eigenvector. Simplicity and the small perturbation ensure that these ordered alignments are consistent.

Telescoping whitened mean differences yields errors $a_u e$ and $a_v e$. Each aligned projected mean coordinate changes by at most $(K_U \|u_0\| + a_u)e$ or $(K_V \|v_0\| + a_v)e$. Both are less than $\eta/2$. Since the true projected output coordinates equal signed input coordinates, neither side crosses zero. The sign matrix in the recovery formula therefore stays on its true branch. Bounding an eigenbasis difference by its Frobenius norm gives

$$\left\|\widehat{R} - R\right\| \le \sqrt{d}(K_U + K_V)e = K_R e.$$

Telescoping $\widehat{E}^{1/2}\widehat{R}\widehat{D}^{-1/2}$ and using $\left\|\widehat{E}^{1/2}\right\| \le J_E$ proves the K_A bound. In particular, $\left\|\widehat{A}\right\| \le \|A\| + K_A\rho$. In $\widehat{b} = \widehat{m}_0 - \widehat{A}\widehat{\mu}_0$, split the error into the output-mean error, $(\widehat{A} - A)\mu_0$, and $\widehat{A}(\widehat{\mu}_0 - \mu_0)$ to obtain K_b. For $\widehat{Q}$, split the covariance error and use

$$\left\|\widehat{A}\Sigma_0\widehat{A}^\top - A\Sigma_0 A^\top\right\| \le \left\|\widehat{A} - A\right\| \|\Sigma_0\| \left(2\|A\| + \left\|\widehat{A} - A\right\|\right).$$

The estimated-input-covariance contribution is at most $\left\|\widehat{A}\right\|^2 e$. This yields K_Q.

The identical telescoping argument at the query, with its independently estimated moments, gives (23). If $\left\|\widehat{C}_q - C_q\right\| \le c/2$ and $C_q \succeq cI$, then $\widehat{C}_q \succeq (c/2)I$. Let $X = \widehat{C}_q^{1/2} - C_q^{1/2}$. The Sylvester identity

$$\widehat{C}_q^{1/2} X + X C_q^{1/2} = \widehat{C}_q - C_q$$

and separate orthogonal diagonalizations of its left and right multipliers show

$$\|X\|_F \le \frac{\left\|\widehat{C}_q - C_q\right\|_F}{\sqrt{c/2} + \sqrt{c}} \le \frac{\sqrt{d}\, c_q}{\sqrt{c}}.$$

Couple the two Gaussian responses with a shared standard Gaussian. The expected squared distance is $\|\widehat{m}_q - m_q\|^2 + \|X\|_F^2$, so $W_2^2 \le a_q^2 + d c_q^2/c$. For (5),

$$\|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}_k}^2 = 2(1 - k(x,y)) \le \frac{1}{4d} \sum_h h^{-2} \|x - y\|^2.$$

Jensen's inequality on any coupling gives $\mathrm{MMD}_k \le L_k W_2$, proving (24). □

Projection bounds and inactive truncation. Weyl's inequality gives $\lambda_{\min}(\widehat{D}), \lambda_{\min}(\widehat{E}) \ge \min(r, s) - 2e$. When this lower bound exceeds τ, flooring leaves both matrices unchanged.

PSD projection is the metric projection onto a closed convex cone in the Frobenius Hilbert space. Its variational inequality gives, for any $Q \succeq 0$,

$$\left\|\widetilde{Q}_+ - Q\right\|_F^2 \le \left\|\widetilde{Q} - Q\right\|_F^2 - \left\|\widetilde{Q}_+ - \widetilde{Q}\right\|_F^2.$$

This also follows directly in an eigenbasis of $\widetilde{Q}$: on the discarded negative eigenspace, the diagonal entries of Q are nonnegative, giving the required sign of the cross term. For the spectral bound, if $\left\|\widetilde{Q} - Q\right\|_2 \le e_Q$, then $\lambda_{\min}(\widetilde{Q}) \ge -e_Q$ and $\left\|\widetilde{Q}_+ - \widetilde{Q}\right\|_2 \le e_Q$. The triangle inequality gives $2e_Q$. Taking $e_Q = K_Q e$ proves the regularized prediction claim wherever the whitening floor is inactive. □

C.6 A Gaussian finite-sample bound

Proof of (25). For one Gaussian law with covariance norm at most M, the sample-mean error is a centered Gaussian with covariance at most MI/n. A $1/2$-net of the unit sphere has at most 5^d elements and controls vector norm by twice the maximal net projection. The scalar Gaussian tail therefore gives

$$\mathbb{P}\Big(\|\widehat{\mu} - \mu\| > 2\sqrt{2Mt/n}\Big) \le 2\,5^d e^{-t}.$$

For the unbiased sample covariance $\widehat{\Sigma}$, normality and an orthogonal change of coordinates on the n observation indices imply that $(n-1)\widehat{\Sigma}$ is the sum of $n-1$ independent centered Gaussian outer products with covariance Σ. For a fixed unit vector v with $v^\top \Sigma v > 0$, its normalized quadratic form is $\chi^2_{n-1}/(n-1)$. Zero variance directions contribute zero exactly. The moment-generating function $(1-2\lambda)^{-(n-1)/2}$ and Chernoff's inequality give, with probability at least $1 - 2e^{-t}$,

$$\left|v^\top(\widehat{\Sigma} - \Sigma)v\right| \le M\left(2\sqrt{\frac{t}{n-1}} + \frac{2t}{n-1}\right).$$

The negative tail is bounded by the same larger two-sided expression. A $1/4$-net has size at most 9^d. For symmetric A, approximating a maximizing unit eigenvector by a net point and bounding the two bilinear errors gives $\|A\|_2 \le 2\max_{v \text{ in net}} |v^\top A v|$. A union bound therefore gives the covariance term in (25) with failure probability $2\,9^d e^{-t}$.

There are seven sampled Gaussian laws: three support inputs, three support responses, and the query input. A union bound on their mean and covariance events is at most $14(5^d + 9^d)e^{-t} \le 28\,9^d e^{-t} = \delta$. Unequal sample sizes bounded below by n only improve these estimates. Conditional independence is used for each law's sample calculations; the union bound itself does not require independence across laws. Combining this simultaneous event with the deterministic local theorem yields the stated finite-sample consequence. □

C.7 Bayes response and ideal conditional flow

Proof of Theorem B.1. Continuity of the kernel on $\mathbb{R}^d$ makes its RKHS separable; boundedness ensures square-integrability of its probability embeddings and the required Bochner expectations. On the standard Borel spaces in the statement, choose regular conditional laws and form the conditional probability mixture $Q^*_{\mathcal{D}}$. Bounded integration and conditional expectation commute, giving $\Phi(Q^*_{\mathcal{D}}) = \mathbb{E}[\Phi(Q_\Theta) \mid \mathcal{D}]$. Write $M = \Phi(Q_\Theta)$ and $M_D = \mathbb{E}[M \mid \mathcal{D}]$. For any $\mathcal{D}$-measurable $\widehat{M} = \Phi(\widehat{Q})$,

$$\mathbb{E}\left\|\widehat{M} - M\right\|^2 = \mathbb{E}\,\|M_D - M\|^2 + \mathbb{E}\left\|\widehat{M} - M_D\right\|^2,$$

because the cross term has conditional expectation zero. This proves (28) and optimality. A characteristic kernel makes the optimal law unique up to almost-sure equality.

For $E = E(\mathcal{D})$, the tower property yields $M_E = \mathbb{E}[M_D \mid E]$. If $\widehat{M}$ is E-measurable, a second orthogonal decomposition splits $\mathbb{E}\left\|\widehat{M} - M_D\right\|^2$ into $\mathbb{E}\,\|M_E - M_D\|^2 + \mathbb{E}\left\|\widehat{M} - M_E\right\|^2$. This proves (29). For $E_0 = g(E_1)$, the same argument implies that the Bayes risk for E_0 is the risk for

E_1 plus $\mathbb{E}\,\|\mathbb{E}[M \mid E_1] - \mathbb{E}[M \mid E_0]\|^2 \geq 0$. If a visible normalization changes the kernel, condition first on that normalization so the Hilbert space is fixed in each conditional argument, and require it to be retained by both evidence representations. □

Proof of Proposition B.2. Condition on $(Z_t, t, \mathcal{D})$. The residual $U - u$ has conditional mean zero, so expansion of squared norms cancels the cross term and proves (32). Next let f be continuously differentiable with bounded derivative and compact support. Since $dZ_t/dt = U$ and U is integrable, differentiation under expectation gives

$$\frac{d}{dt}\mathbb{E}[f(Z_t) \mid \mathcal{D}] = \mathbb{E}[\nabla f(Z_t) \cdot U \mid \mathcal{D}] = \mathbb{E}[\nabla f(Z_t) \cdot u(Z_t, t, \mathcal{D}) \mid \mathcal{D}]$$

for almost every t. This is the weak continuity equation. The initial conditional law is standard Gaussian, and the terminal conditional law is $\mathrm{Law}(Y \mid \mathcal{D}) = Q^*_{\mathcal{D}}$. By uniqueness in the stipulated class, the conditional path laws coincide with those induced by the characteristic flow, whose terminal law is therefore the mixture. The existence and uniqueness assumptions are needed for this ODE interpretation; the regression identity and weak equation alone do not imply it. The same argument applies with E substituted for $\mathcal{D}$. □

For a target batch conditionally iid given Θ, the posterior cross-moment of a test function on two distinct rows is $\mathbb{E}[(\int f\, dQ_\Theta)^2 \mid \mathcal{D}]$. Two iid rows from the mixture instead give $(\mathbb{E}[\int f\, dQ_\Theta \mid \mathcal{D}])^2$. The difference is the posterior variance of $\int f\, dQ_\Theta$, which quantifies the dependence induced by a shared unknown mechanism.

C.8 Block normalization, uniform convergence, and tails

Expected identity normalization. Fix a dimension and condition on the vector of iid zero-effect indicators. Given an indicator vector with count $k > 0$, the k eligible episode losses all have conditional expectation $\mu_I = \mathbb{E}[\ell^I \mid a = 1, d]$, because episodes are independent and identically distributed within the stratum. Their mean therefore has expectation μ_I; with no eligible episode, the term is zero. A nonempty eligible set has probability $1 - (1 - p_d)^b$, which gives the expectation stated after (17). An episode's indicator and its path loss may be dependent; independence is used across episodes. □

Proof of Theorem B.3. For each dimension take an internal uniform ϵ-net of the clipped block-loss class. For any one net function with values in $[0, c]$, Hoeffding gives

$$\mathbb{P}\{|J_{c,d}(h) - \widehat{J}_{c,d}(h)| > u\} \leq 2\exp(-2n_d u^2/c^2).$$

Choose $u = c\sqrt{\log(4N_d(\epsilon)/\delta)/(2n_d)}$ and union bound over the $N_d(\epsilon)$ net functions. Failure probability is at most $\delta/2$ in that stratum. For an arbitrary function, the uniform approximation changes population and empirical means by at most ϵ each. Union bound over the two strata, then average their bounds with weights $1/2$, to obtain the first part of (36).

For any h, decompose $J(h) - \widehat{J}(h)$ into its clipped difference, population tail, and negative empirical tail. Taking absolute values gives the bound $U_c + \tau_c + \widehat{\tau}_c$ on the same uniform event. The bound for the unbounded loss is informative only if the supremum tails are finite.

For any comparison parameter $g \in \mathcal{F}$, the uniform bound and definition of optimization gap imply

$$J(\widehat{h}) \leq \widehat{J}(\widehat{h}) + V_c \leq \widehat{J}(g) + \varepsilon_{\mathrm{opt}} + V_c \leq J(g) + 2V_c + \varepsilon_{\mathrm{opt}}.$$

Take the infimum over g, or a sequence approaching it. Because $R_S(h) \geq R^*_S$ and $I(h) \geq 0$, $R_T(h) \leq J(h) - 0.5R^*_S$. Inserting (33) proves (37). The optimization gap is defined for the actual unbounded empirical objective, so the result applies without changing the algorithm to minimize a clipped loss. □

Complexity and moment examples. Under the additional uniform parameter-Lipschitz assumption, an ϵ/L_θ Euclidean net for the parameter ball induces a uniform ϵ loss net. Packing disjoint balls of half that radius inside a radius-$R + \epsilon/(2L_\theta)$ ball gives cardinality at most $(1 + 2RL_\theta/\epsilon)^p$.

A finite class instead admits itself as a zero-radius net. If a nonnegative envelope $W(B)$ dominates every unbounded loss and $\mathbb{E}W^{1+\nu} \le M_{1+\nu}$ under the stratified mixture, then

$$(\ell_h - c)_+ \le \ell_h \mathbf{1}_{\ell_h > c} \le W^{1+\nu}/c^\nu.$$

Taking expectation gives $\tau_c \le M_{1+\nu}/c^\nu$. Replacing it by the empirical stratified average gives the corresponding bound on $\widehat{\tau}_c$. A tail bound thus requires an envelope estimate; clipping alone also leaves the uniform parameter-Lipschitz condition to be checked on unbounded inputs.

C.9 Velocity, Heun, and finite-row distribution errors

Proof of Corollary B.4. Condition on a visible observation and a dimension. Let X_t follow the ideal field u with its prescribed path marginals, and W_t follow learned field v, sharing the same initial Gaussian. State Lipschitzness of v gives

$$\|W_t - X_t\| \le \int_0^t L\,\|W_r - X_r\|\,dr + \int_0^t \|v(X_r, r) - u(X_r, r)\|\,dr.$$

Gronwall and Cauchy–Schwarz over $[0, 1]$ imply

$$\mathbb{E}\,\|W_1 - X_1\|^2/d \le e^{2L} \int_0^1 \mathbb{E}\,\|v(X_t, t) - u(X_t, t)\|^2/d\,dt.$$

The right side is the conditional target FM excess times e^{2L}; no learned-path density ratio is needed because the error is evaluated along the ideal path. The shared-noise coupling bounds W_2^2 between the exact generated law and $Q^*_{\mathcal{D}}$.

Let $W_1^{(s)}$ be the Heun output, coupled to the same learned ODE initial condition. The assumed normalized strong error is $\mathbb{E}\left\|W_1^{(s)} - W_1\right\|^2/d \le H^2 s^{-4}$. Applying the squared triangle inequality and the kernel bound $L_k^2 = c_k/d$ gives

$$\mathbb{E}\,\mathrm{MMD}_k^2(\mathrm{Law}(W_1^{(s)} \mid \mathcal{D}), Q^*_{\mathcal{D}}) \le 2c_k\left[e^{2L}(R_T(v) - R_T^*) + H^2 s^{-4}\right]$$

after averaging over observations and dimensions. Apply the Bayes decomposition and Theorem B.3. This yields (38). For random trained weights, condition on their independent pretraining sample on the uniform event before taking the fresh-task expectation. □

A sufficient Heun error assumption and proof. For the learned ODE, suppose the defect from inserting an exact trajectory into one Heun step is at most $a(Z)h^3$ at every step, with $\mathbb{E}[a(Z)^2]/d \le H_0^2$, uniformly over contexts and dimensions. Such a condition follows, for example, from bounded Taylor remainder terms of the nonautonomous field along the trajectory and trial-step segments, with a square-integrable bound depending on the initial state. It is stronger than state Lipschitzness alone. A Heun update with an L-Lipschitz field has state Lipschitz constant at most $1 + Lh + L^2h^2/2 \le e^{Lh}$: bound the first evaluation by L and the predicted-stage displacement by $(1 + Lh)$.

Let e_j be the error after j steps of size $h = 1/s$. Then $e_{j+1} \le e^{Lh}e_j + a(Z)h^3$ and $e_0 = 0$. Iteration gives $e_s \le a(Z)h^3 s e^L = a(Z)e^L s^{-2}$. Squaring and averaging proves the strong error bound with $H = e^L H_0$. This argument converts the assumed local defect into a global error estimate. The symmetry proof in Appendix C.1 holds without either error assumption.

Finite-sample MMD formulas. Given a task and prompt, let $\widehat{Y}_1, \ldots, \widehat{Y}_N$ be iid from the predicted law P, and $Y_1, \ldots, Y_m$ independently iid from its true response Q. For the ordinary V-statistic

$$V_{N,m} = \frac{1}{N^2}\sum_{a,b} k(\widehat{Y}_a, \widehat{Y}_b) + \frac{1}{m^2}\sum_{a,b} k(Y_a, Y_b) - \frac{2}{Nm}\sum_{a,b} k(\widehat{Y}_a, Y_b),$$

separate diagonal and off-diagonal terms to obtain

$$\mathbb{E}[V_{N,m} \mid P, Q] = \mathrm{MMD}_k^2(P, Q) + \frac{\mathbb{E}_P k(Y, Y) - \|\Phi(P)\|^2}{N} + \frac{\mathbb{E}_Q k(Y, Y) - \|\Phi(Q)\|^2}{m}. \tag{43}$$

Each numerator is an RKHS variance between zero and one for our normalized kernel, giving the $1/N + 1/m$ upper bound. For $N, m \geq 2$, removing the within-sample diagonals and dividing by $N(N-1)$ and $m(m-1)$ gives the unbiased U-statistic (Gretton et al., 2012), whose realized value may be negative. For a query-derived scale, condition on both the query and task: the kernel is then fixed and the target and generation streams remain independent. Floating-point error is separate from these expectation identities.

C.10 Explicit finite-candidate illustration

The two hypotheses are identity and the rotation $-I_2$. All visible supports are centered isotropic Gaussians and have the same responses under either hypothesis. Query input is the known law $\mathcal{N}(e_1, I_2)$. Let $\pi = 0.3$ be the probability of identity, independent of visible data. The conditional Bayes response is $Q_\pi = \pi\mathcal{N}(e_1, I_2) + (1-\pi)\mathcal{N}(-e_1, I_2)$.

For any candidate α, define $\sigma_t^2 = (1-t)^2 + t^2$, $c_t = (2t-1)/\sigma_t^2$, and component fields

$$u_\pm(z,t) = \pm e_1 + c_t(z \mp t e_1), \qquad w_\alpha(z,t) = \text{sigmoid}\left(\log\frac{\alpha}{1-\alpha} + \frac{2tz_1}{\sigma_t^2}\right).$$

Then $v_\alpha = w_\alpha u_+ + (1-w_\alpha)u_-$ is the ideal conditional mixture field for candidate prior α. To verify this, the component path law is $\mathcal{N}(\pm te_1, \sigma_t^2 I)$, while $\text{Cov}(Y-Z, Z_t) = (2t-1)I$. Gaussian conditioning gives $u_\pm$ and the component likelihood ratio gives w_α. The analytic query-base field is u_+, so the identity penalty acts on $v_\alpha - u_+$ at response-path states. An independent source path supplies the source loss, identical across candidates.

Set $\delta^2 = \text{MMD}_k^2(\mathcal{N}(e_1, I), \mathcal{N}(-e_1, I))$. By linearity of the embedding, the exact task-averaged target distribution risk is

$$\pi\,\text{MMD}_k^2(Q_\alpha, Q_+) + (1-\pi)\,\text{MMD}_k^2(Q_\alpha, Q_-) = \delta^2\left[(\alpha-\pi)^2 + \pi(1-\pi)\right].$$

The first term is the prediction excess within the candidate class; the second is the Bayes floor due to ambiguity. We select candidates by enumerating the empirical block objective and estimate the population objective by independent reference integration. Distribution risks use the exact formula above, so neither ODE simulation nor a sample-based target MMD estimate is needed.

D Approximate Identification and Finite Observations

Proposition 3.2 uses the full variation norm of a signed measure: for $\nu = P - Q$, $\|\nu\|_{\text{var}} = 2\sup_B |P(B) - Q(B)|$. Under its measurability assumptions, the bound also holds for non-characteristic kernels. The proof permits arbitrary real coefficients and any two common Markov kernels on a general measurable input space.

D.1 Finite-observation counterpart

Suppose the input laws and coefficients are known, and $\widehat{Q}_i$ is an empirical estimate of $\mathsf{K}P_i$. The signed embedding $\widehat{\mu}_a = \sum_i a_i \Phi(\widehat{Q}_i)$ obeys

$$\|\widehat{\mu}_a - \Phi(\mathsf{K}P_q)\|_{\mathcal{H}} \leq \sum_i |a_i|\,\text{MMD}_k(\widehat{Q}_i, \mathsf{K}P_i) + \sqrt{\kappa}\left\|P_q - \sum_i a_i P_i\right\|_{\text{var}}. \tag{44}$$

Proof. Let $\psi(x) = \int \phi(y)\mathsf{K}(x, dy)$, whose norm is at most $\sqrt{\kappa}$. Subtract and add $\sum_i a_i \Phi(\mathsf{K}P_i)$. The remaining population term is $-\int \psi\, d\nu$, with $\nu = P_q - \sum_i a_i P_i$. Bounding the empirical sum and this integral separately proves the claim. □

The empirical combination may fail to represent a probability law, so (44) controls embedding error without itself supplying a sampler. The following example takes the input laws and coefficients as known; estimating them would introduce additional error.

D.2 A FOUR-STATE CONSTRUCTION WITH EXACT RISK

Use four input states, two output states, and $k(y, y') = \mathbf{1}\{y = y'\}$. The embedding of $\mathrm{Bernoulli}(q)$ is $(1-q, q)$ and its squared distance from $\mathrm{Bernoulli}(p)$ is $2(p-q)^2$. Let

$$u = (1/4, 1/4, 1/4, 1/4), \quad h = (1/2, -1/2, 0, 0), \quad r = (0, 0, 1/2, -1/2),$$

$$P_1 = u + \delta h, \quad P_2 = u - \delta h, \quad P_q = u + \beta h + \rho r, \qquad \beta = 0.1.$$

Take $\delta \in \{0.02, 0.05, 0.1, 0.2\}$ and $\rho \in \{0, 0.02, 0.1, 0.2\}$. All entries are nonnegative and each law sums to one. The coefficients

$$a_1 = (1 + \beta/\delta)/2, \qquad a_2 = (1 - \beta/\delta)/2$$

give residual $\nu = \rho r$, variation $\|\nu\|_{\mathrm{var}} = \rho$, and coefficient norm $\sum_i |a_i| = \max(1, \beta/\delta)$. At $\rho = 0$ the query is in the support span; smaller δ makes its signed representation less stable.

For the two kernels with success probabilities $(0.5, 0.5, 0.7, 0.3)$ and $(0.5, 0.5, 0.3, 0.7)$, both support responses have probability 0.5. Their query success probabilities differ by 0.4ρ. Thus their MMD is $0.4\sqrt{2}\rho$, below the 2ρ bound, and the equal-prior risk floor is $0.08\rho^2$. This comparison is analytic and has no sampling uncertainty.

For response estimation, use the single common kernel with success probabilities $(0.7, 0.3, 0.7, 0.3)$. It gives

$$q_1 = 0.5 + 0.2\delta, \qquad q_2 = 0.5 - 0.2\delta, \qquad q_q = 0.52 + 0.2\rho.$$

Let $C_i \sim \mathrm{Binomial}(n, q_i)$ independently and define $\widehat{q} = a_1 C_1/n + a_2 C_2/n$. Its expectation is 0.52, so the signed embedding estimate has exact expected error

$$\mathbb{E}\|\widehat{\mu}_a - \Phi(Q_q)\|^2 = 0.08\rho^2 + \frac{[1 + (0.1/\delta)^2](0.25 - 0.04\delta^2)}{n}. \tag{45}$$

Proof. The squared bias is $2(0.52 - q_q)^2 = 0.08\rho^2$. Independence of the two counts gives variance $2\{a_1^2 q_1(1-q_1) + a_2^2 q_2(1-q_2)\}/n$. Since $q_1(1-q_1) = q_2(1-q_2) = 0.25 - 0.04\delta^2$ and $a_1^2 + a_2^2 = [1 + (0.1/\delta)^2]/2$, this is the second term in (45). □

Projection to $\widetilde{q} = \min(1, \max(0, \widehat{q}))$ gives a valid Bernoulli law and cannot increase error: $q_q \in [0, 1]$ implies $|\widetilde{q} - q_q| \le |\widehat{q} - q_q|$. Equation (45) describes the raw embedding; the numerical results include both raw and projected errors.

D.3 FIXED PROTOCOL AND COMPLETE RESULTS

The protocol was fixed before execution. Seed 20260907 specifies separate response and bootstrap streams. Each of the 16 (δ, ρ) cells uses $n \in \{32, 128, 512, 2048\}$ and 20 independent repetitions, yielding 1,280 records. Of these, 11 estimates (0.86%) require projection; all realizations enter the analysis. The analytic error decomposition and pointwise projection inequality pass every numerical check. Intervals use 2,000 bootstrap draws over the 20 repetitions within each cell, quantifying sampling variation for fixed laws rather than variation across tasks or mechanisms.

The accompanying affine and neural analyses use all existing evaluation records, with model weights and formal-test samples held fixed. Post hoc neural groups follow the generator intervals, selected before computing grouped effects: exact zero, weak $[0.01, 0.15]$, and moderate $[0.35, 0.8]$. Appendix F reports every group and its uncertainty.

E EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

E.1 NUMERICAL PROTOCOLS

The theoretical illustrations use the independent numerical seed 20260906. They evaluate analytic estimators and a finite class of flows.

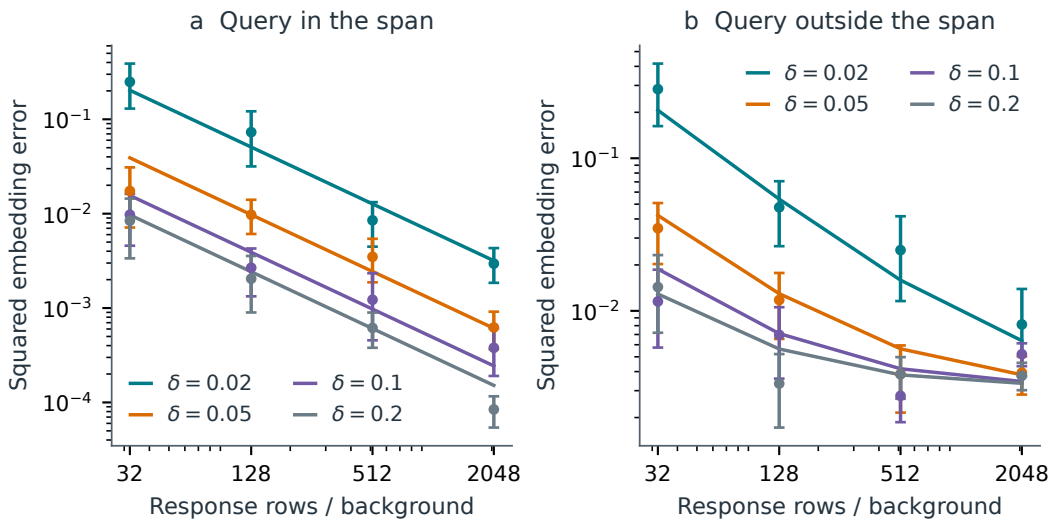

Figure 5: Finite observations in the four-state model. Solid curves give exact expected signed-embedding error; points and intervals summarize 20 independent repetitions. Nearby supports amplify response-sampling error even for an exactly spanned query. All 64 grid cells and both raw and projected errors are supplied as CSV records.

Affine recovery. For each dimension, we draw 20 independent mechanisms and hold the affine matrix, offset, noise covariance, and variable structure fixed across four background-conditioning regimes. Informative settings use eigenvalue spacing 0.4, mean projection 1, and minimum covariance difference 0.7. The three difficult settings respectively change only one of these quantities to 0.02, 0.03, or 0.03. Rows are sampled independently for the three input laws, three response laws, and query input. Row counts are 32, 128, 512, and 2048. The estimator floors both difference spectra at 10^{-4}, uses ordered eigenbases with deterministic sign selection, and projects raw Q onto the PSD cone. Estimates requiring either correction remain in the analysis.

Finite-candidate pretraining. There are 19 fixed candidate priors from 0.05 through 0.95 in increments of 0.05. Each of 20 numerical repetitions uses nested block counts 32, 128, 512, 2048, and 8192, with four independent tasks per block. The two objectives use the same response/source draws and differ only in identity weight. We evaluate the unbounded empirical objective for every candidate and break ties at the smallest candidate index. Each task contributes one path. An independent 262,144-block integration estimates the population objective. The repetitions measure finite-class candidate selection, with no neural training.

At 8,192 blocks, all 20 repetitions select probability 0.3 without the identity penalty; weight 0.1 selects 0.35. Its persistent excess MMD^2 is 6.32×10^{-4}, above the shared Bayes floor of 0.05309. These values use exact Gaussian-mixture response errors.

E.2 Affine error and regularization at every condition

Table 4 reports all 48 dimension/regime/row-count aggregates, each over 20 mechanisms. "Floor X/Y" denotes the fraction of covariance differences requiring spectral flooring; "PSD" denotes the fraction requiring a nonnegligible PSD repair. The final column gives the mean Frobenius magnitude of the repair. Every estimate enters the MMD average. Per-mechanism records also contain parameter and query-moment errors, eigengap, conditioning, and local radius. All sampled moment errors exceed the sufficient local radius; the local checks instead use perturbations of size one tenth of that radius.

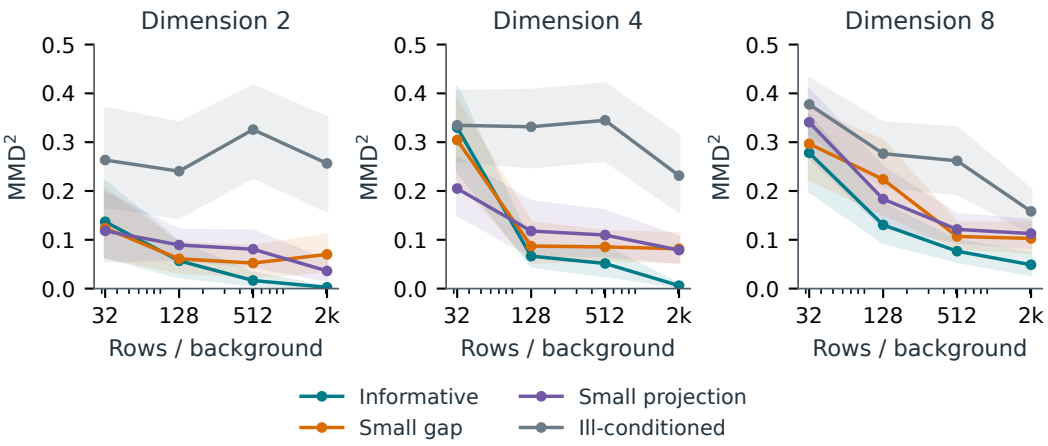

Figure 6: Regularized affine recovery in all dimensions. Pointwise intervals use 20 mechanism clusters; all estimates are included. No sampled condition falls within the sufficient local radius.

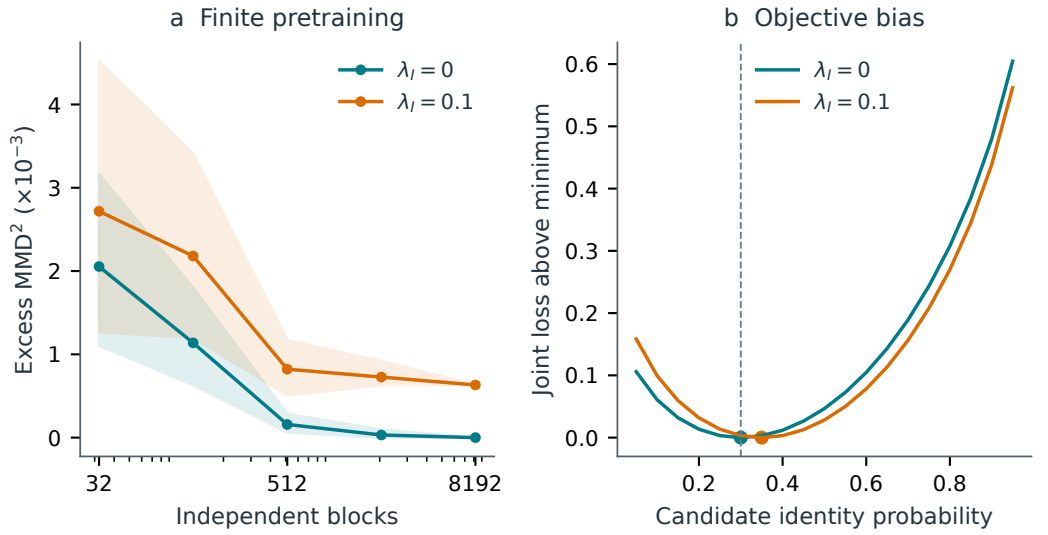

Figure 7: The finite-candidate illustration. The known prior has identity probability 0.3. Increasing the number of blocks reduces selection error, while the identity penalty favors a different candidate. Bands summarize 20 numerical repetitions and do not measure neural retraining variation.

Table 4: Complete regularized affine results. Rates are percentages. Population MMD^2 uses analytic Gaussian distributions.

d	Regime	n	MMD^2	Floor X	Floor Y	PSD	Repair $\|\cdot\|_F$
2	Informative	32	0.1370	15	20	70	2664.37
2	Informative	128	0.0570	0	0	95	1.69
2	Informative	512	0.0166	0	0	70	0.44
2	Informative	2048	0.0027	0	0	75	0.08
2	Small gap	32	0.1239	10	15	95	3551.40
2	Small gap	128	0.0610	0	0	90	1.17
2	Small gap	512	0.0523	0	0	90	0.84
2	Small gap	2048	0.0701	0	0	95	1.01
2	Small projection	32	0.1182	10	15	75	363.28

Table 4 continued

d	Regime	n	MMD^2	Floor X	Floor Y	PSD	Repair $\|\cdot\|_F$
2	Small projection	128	0.0893	0	0	95	1.89
2	Small projection	512	0.0807	0	0	80	0.92
2	Small projection	2048	0.0362	0	0	75	0.53
2	Ill-conditioned	32	0.2634	65	90	95	3256.97
2	Ill-conditioned	128	0.2405	40	55	95	6756.13
2	Ill-conditioned	512	0.3258	55	55	100	9229.30
2	Ill-conditioned	2048	0.2565	35	20	95	6148.27
4	Informative	32	0.3299	70	45	100	5527.10
4	Informative	128	0.0665	0	10	100	2.33
4	Informative	512	0.0515	0	0	100	1.02
4	Informative	2048	0.0061	0	0	100	0.27
4	Small gap	32	0.3043	60	75	100	7853.59
4	Small gap	128	0.0870	5	10	100	568.34
4	Small gap	512	0.0854	0	0	100	1.16
4	Small gap	2048	0.0815	0	0	100	0.89
4	Small projection	32	0.2050	45	65	100	3028.94
4	Small projection	128	0.1180	5	0	100	248.62
4	Small projection	512	0.1097	0	0	100	0.98
4	Small projection	2048	0.0790	0	0	100	0.68
4	Ill-conditioned	32	0.3345	80	90	100	4186.54
4	Ill-conditioned	128	0.3316	65	65	100	5023.86
4	Ill-conditioned	512	0.3447	70	35	100	4304.57
4	Ill-conditioned	2048	0.2315	20	20	100	1116.20
8	Informative	32	0.2782	100	100	100	7930.71
8	Informative	128	0.1305	15	30	100	400.11
8	Informative	512	0.0766	0	0	100	2.09
8	Informative	2048	0.0489	0	0	100	0.94
8	Small gap	32	0.2967	100	100	100	4140.07
8	Small gap	128	0.2237	30	30	100	2235.96
8	Small gap	512	0.1068	0	0	100	2.11
8	Small gap	2048	0.1028	0	0	100	1.28
8	Small projection	32	0.3408	100	100	100	10146.94
8	Small projection	128	0.1834	30	40	100	1092.76
8	Small projection	512	0.1213	0	0	100	2.06
8	Small projection	2048	0.1127	0	0	100	1.27
8	Ill-conditioned	32	0.3774	100	100	100	11974.69
8	Ill-conditioned	128	0.2765	95	100	100	3185.62
8	Ill-conditioned	512	0.2618	70	85	100	2367.06
8	Ill-conditioned	2048	0.1582	25	55	100	321.18

E.3 Training and model counts

Each neural model received 120,000 task exposures over 30,000 updates. The shared schedule contains 22,415 whole-pair leave-one-out exposures and 19,585 true zero-effect eligible exposures. Each variant received 15,360,000 response target vectors. Models using the base objective received the same number of independent source target vectors. These vectors supply supervision within each task and leave the independent task count unchanged. Table 5 reports training time, excluding evaluation.

The set-flow baseline is adapted to this task and has 2.9% fewer parameters than the full model. Relation and moment ablations also reduce capacity, while the no-base comparison removes auxiliary objectives. Only the token-compression comparison retains both the parameter count and the objectives among these ablations.

E.4 Synthetic panels and comparison rules

The six background families are elliptical, factor mixture, graph covariance, nonlinear, clustered, and low-rank anomaly. The fourteen programs are identity, shift, scale, sparse mixing, low-rank mixing, smooth nonlinearity, additive noise, multiplicative noise, heteroskedastic response, corre-

Table 5: Model size and training cost. All models have 120k task exposures and 30k updates. Source column is auxiliary source-objective exposures in thousands. Memory is peak recorded GiB.

Variant	Params (M)	Source (k)	Minutes	GiB
ACEMO-ICL	2.818	120	52.4	5.67
Pooled-context single flow	2.637	0	38.4	5.35
No base decomposition	2.637	0	38.4	5.35
Compressed demonstrations	2.818	120	52.3	5.64
Moments only	2.015	120	36.5	1.47
No explicit relations	1.809	120	37.2	5.43
Set conditional flow	2.737	0	31.5	4.85
Query-only model	2.818	120	48.5	1.94

lated noise, heterogeneous branching, population reweighting, composition, and affine-Gaussian response. The panels include every category. For positive-domain tasks, both the program and evaluation operate in log coordinates.

The default weak strength interval is $[0.01, 0.15]$ and the moderate interval is $[0.35, 0.8]$, with exact zero sampled separately for nonidentity programs. OOD strength uses $[1.0, 1.3]$. The other OOD panels change the query background, use held-out ordered two-step or three-step programs, or increase the dimension to 64. Each panel has 96 new tasks, with the response and generation sample sizes and 32-step Heun setting fixed. Table 6 gives primary errors; Figure 11 reports all declared OOD contrasts with their multiplicity correction.

Table 6: OOD mean MMD^2 ($\times 10^3$), with 96 tasks per panel.

Shift	ACEMO-ICL	Set flow	Marginal
Strength	24.12	26.63	54.26
Query background	15.40	16.70	28.25
Composition	16.14	17.36	30.70
Dimension 64	15.15	16.34	27.23

Support-only OT-CFM adaptation. On the fixed balanced subset of 112 tasks, we fit one shared flow per task from its support pairs, using 500 steps of the OT-CFM approach of Tong et al. (2024). Query responses are excluded from adaptation. Mean MMD^2 is 0.05183, versus 0.01506 for the full pretrained model on the same tasks; the paired gap is 0.03677 with CI 0.03325–0.04076. This comparison measures performance at the fixed adaptation budget; longer fitting may change the result. All primary pretrained models and simple baselines use the same subset.

Uncertainty and multiplicity. Primary intervals are pointwise 95% bootstrap intervals, stratified by mechanism or dimension as appropriate. Paired tests use 99,999 random sign flips with a plus-one Monte Carlo correction when exact enumeration is not used. Repeated conditions for one operator share a sign. The nine comparison families comprise 11 IID contrasts, four context contrasts, three controlled-information contrasts, three balanced-information contrasts, fourteen mechanism contrasts, five strength-bin contrasts, three solver contrasts, eight OOD contrasts, and one OT-CFM contrast. Holm adjustment (Holm, 1979) applies within each declared family, provided its constituent tests are valid. Each correction covers its full family, including comparisons omitted from an individual figure, but does not jointly cover all figures and six metrics. Pointwise intervals and adjusted tests can differ because their constructions and null assumptions differ.

E.5 Observed-data examples: sources, preprocessing, and results

Kang18 is obtained through `muscData::Kang18_8vs8()`, ExperimentHub EH2259 (Bioconductor, 2026). The eight donors have control and IFN-β observations. For every held-out query donor, the next four donors in a fixed order form support. Each support condition contains 128 sampled rows; query input contains 64, while target and generated sets contain 128. Visible support control data determine filtering, up to 2,000 high-variance genes, and 16- or 32-dimensional

unwhitened PCA. Library-size normalization targets 10,000 before log1p. Query output samples do not determine the representation. Each split result averages three independent row draws and two generation noises per draw.

For processed Kang18, we obtain one orthogonal matrix per dimension by QR factorization of a standard Gaussian matrix with seed $20260905 + d$. We rotate the PCA coordinates before prediction and invert the rotation for evaluation. This preprocessing was investigated after the original model failure and is therefore exploratory. Rotation can also change the coordinatewise median scale. The resulting comparison therefore combines a change in orientation with a possible change in normalization.

Bodenmiller CyTOF uses the official muscData record EH2254, eight donors, reference and BCR-XL conditions, 24 protein variables transformed by $\operatorname{arcsinh}(x/5)$, and PCA fitted on visible support controls. We report its 16-dimensional PCA representation; a 32-dimensional independent PCA representation is unavailable from 24 variables. The source and response row counts and cyclic donor policy match those used for Kang18.

Table 7: Real-data mean MMD^2 ($\times 10^3$), including the original Kang18 results. Each row averages eight donor folds with overlapping support and is descriptive.

Representation, dimension	ACEMO	Support Y	Set flow	Marginal	G. OT
Kang18 rotated, 16	30.68	37.56	34.80	43.33	35.82
Kang18 rotated, 32	34.52	35.09	37.54	51.66	37.89
CyTOF raw PCA, 16	27.51	26.76	30.94	35.72	35.53
Kang18 original, 16	78.23	37.56	76.30	50.34	35.82
Kang18 original, 32	97.75	35.09	103.94	52.72	37.89

Donor folds share support data, so we report fold values and their means descriptively, without independent-fold significance tests. Row and noise repetitions refine each fold's estimate while leaving the number of evaluation units unchanged. Some auxiliary metrics favor Gaussian OT where MMD favors the neural model: the methods capture different aspects of the response in these low-dimensional representations.

E.6 Numerical and empirical limitations

Table 8 collects the numerical precision checks, finite-sample recovery failures, and limits of the neural comparisons.

E.7 Sampling, learning curves, and computation

We evaluate the sampler at 8, 16, 32, and 64 Heun steps with cached context. This measures sensitivity to step count, rather than the Lipschitz or local-defect constants in Corollary B.4. Validation error falls from 0.01913 at 5,000 updates to 0.01493 at 30,000 updates, with much smaller gains over the final 5,000 updates. All formal comparisons use the final checkpoint; the learning curve covers only this schedule and task prior.

Table 8: Numerical checks and empirical limitations.

Check	Recorded outcome	Interpretation
FP32 permutation threshold	Positive dimension-16 query-row reversal: 1.2947×10^{-5} versus 10^{-5} threshold	Fails the registered tolerance. The symmetry theorem assumes exact arithmetic.
Higher-precision attribution	CPU FP64 discrepancy about 2.81×10^{-14}	Consistent with rounding error in the failed FP32 check.
BF16 drift	Maximum reported drift/reference ratio about 0.001837, below 0.01	Passes the distribution-drift criterion in the evaluated cases.
Original moment estimator	480 conditions, 24 operator seeds; 109 algebraic returns, zero PSD-valid noise estimates	Population identifiability did not ensure valid finite-sample parameter recovery.
Regularized moment estimator	All 960 conditions retained; frequent PSD projection; zero entries in sufficient local neighborhood	Regularization yields valid covariances, but the sampled errors exceed the sufficient local radius.
Original Kang18	Model worse than marginal calibration for all eight donors at both dimensions	The later preprocessing was chosen after this failure.
Bodenmiller, original PCA	Model mean 0.02751 versus Support Y 0.02676; 3/8 donor wins	No overall advantage over this strong real-data comparator.
Random background count	Varied versus one-large-example CI contains zero	An advantage over one larger example is unresolved at fixed total rows.
Training replication	One neural training seed for each variant	Retraining variability is unmeasured.

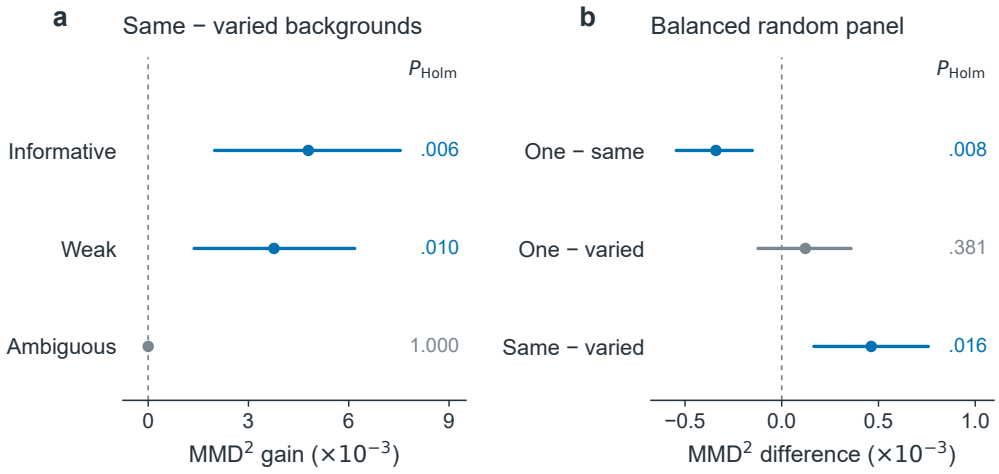

Figure 8: Background-information contrasts. The controlled panel has 96 total rows per side; the balanced random panel has 128. The latter compares one large example, four splits of the same observations, and four varied backgrounds. Intervals are pointwise task/operator bootstrap intervals; p-values use the declared Holm families.

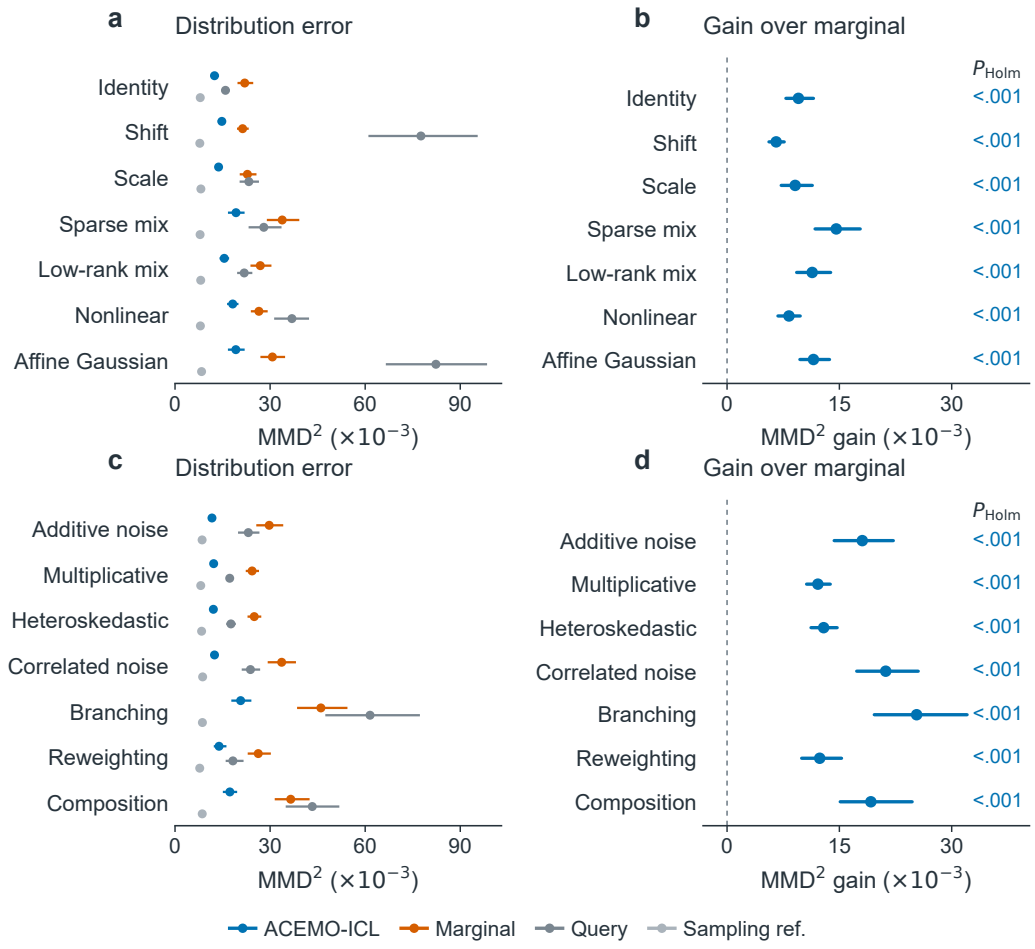

Figure 9: Results for the fourteen mechanisms, with 48 IID task means per mechanism. Left: primary errors. Right: paired marginal-minus-model differences, with one Holm family across all fourteen mechanisms. Intervals are pointwise task-level bootstrap intervals. Mechanisms are categorical.

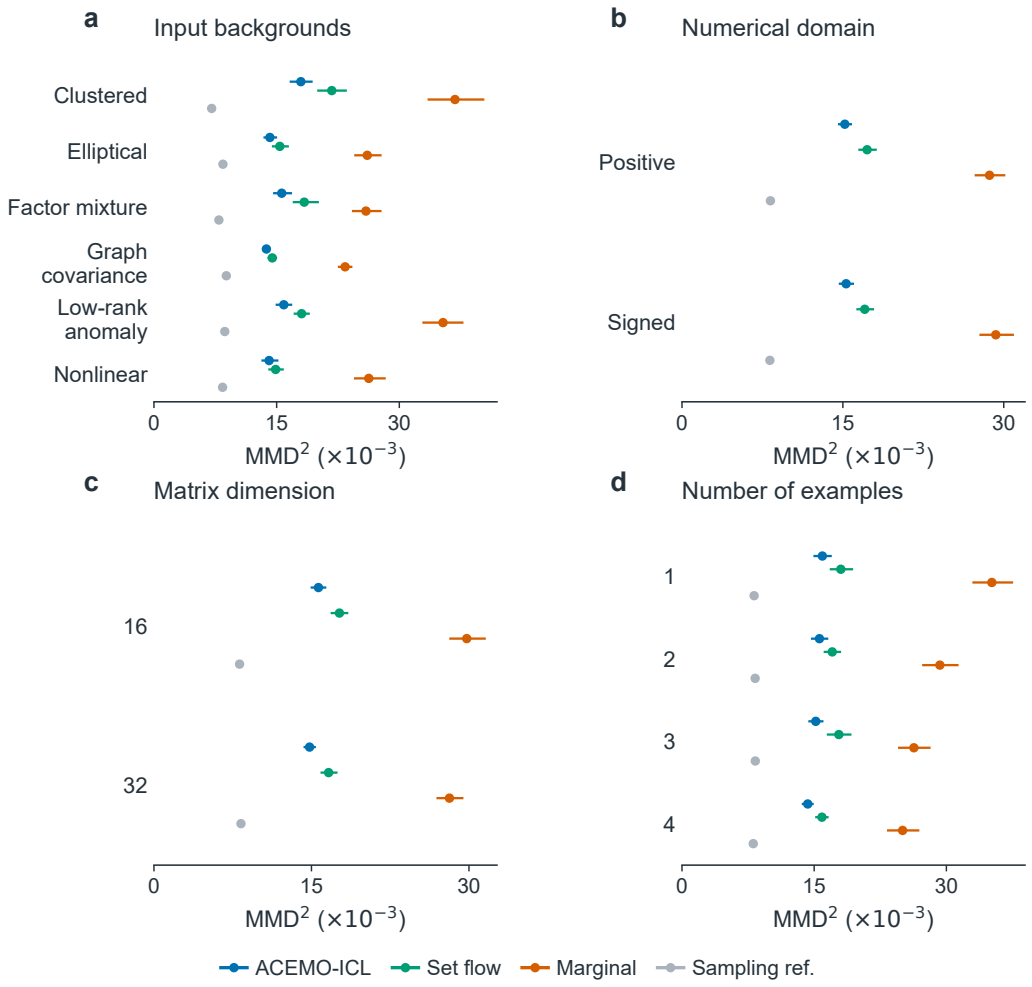

Figure 10: Results by background, domain, dimension, and demonstration count. Strata generally contain different tasks, so comparisons between them are descriptive rather than paired estimates of changing one factor.

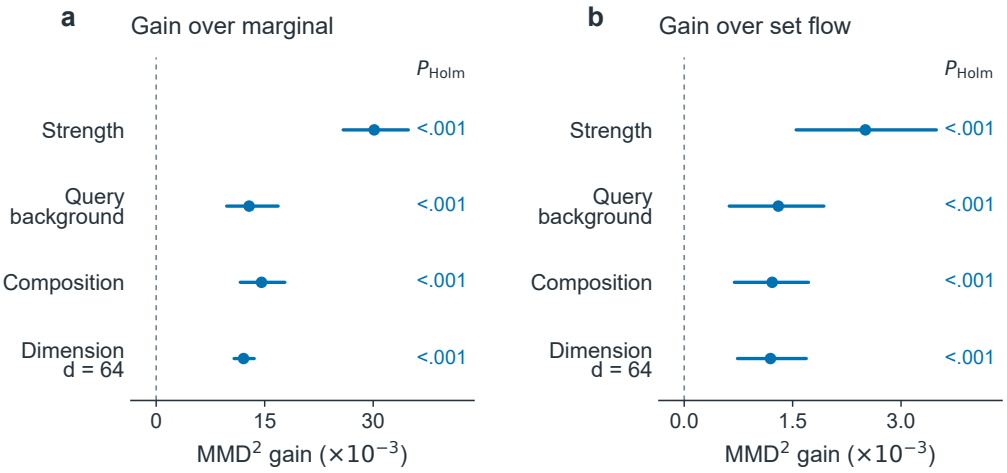

Figure 11: The four OOD panels. Primary paired comparisons against marginal calibration and set flow form one family of eight Holm-corrected tests. Each panel has 96 new task means.

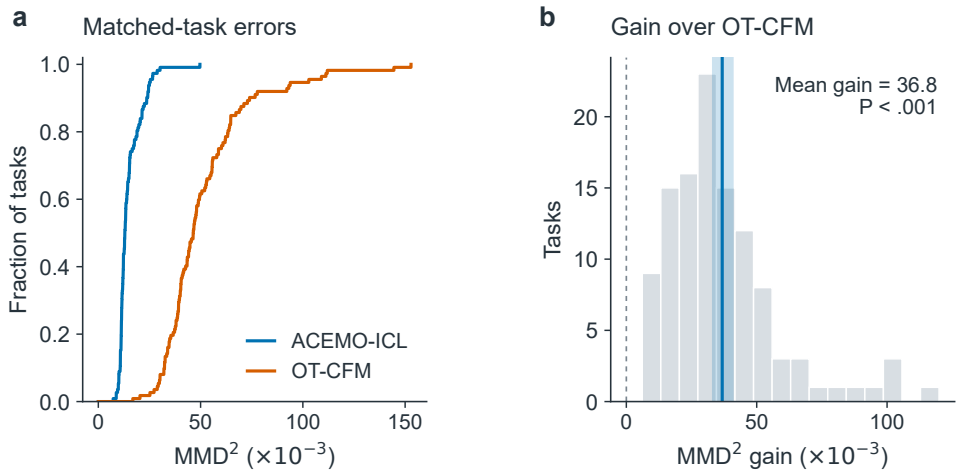

Figure 12: Support-only OT-CFM and pretrained models on the same 112 tasks. The adapted OT-CFM baseline is fitted separately for 500 steps per task, using support observations only.

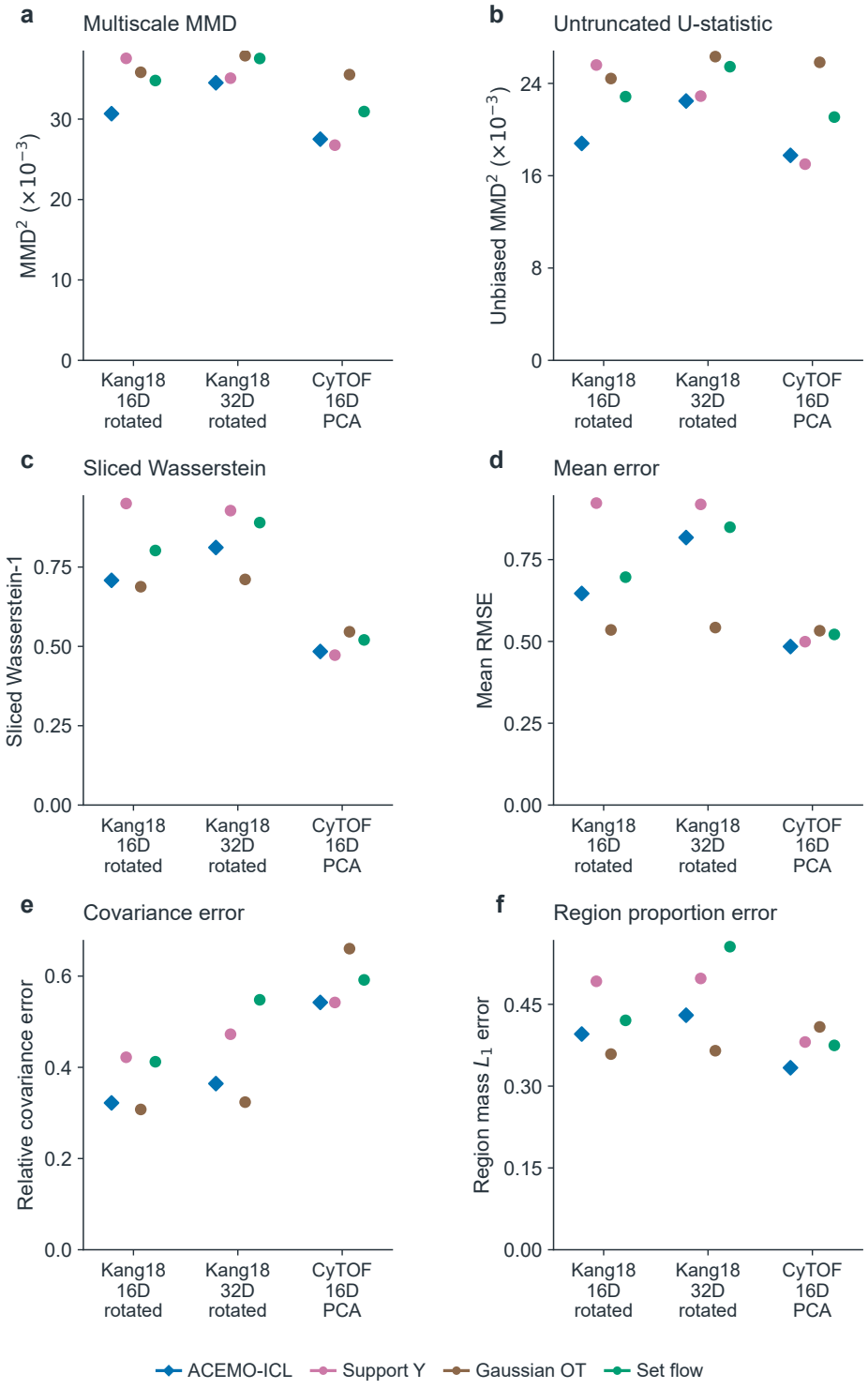

Figure 13: Auxiliary metrics for rotated Kang18 and Bodenmiller in original PCA coordinates. Values are descriptive donor averages; method rankings vary across metrics and folds share support donors.

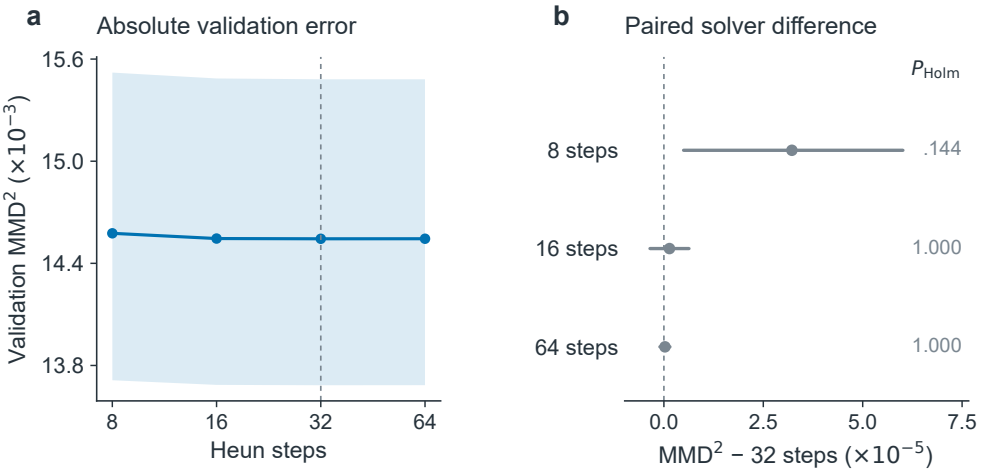

Figure 14: Sampling-step accuracy. The three paired comparisons against the 32-step default form one Holm family. This curve measures step-count sensitivity for the trained field; the theoretical convergence order also requires the stated regularity assumptions.

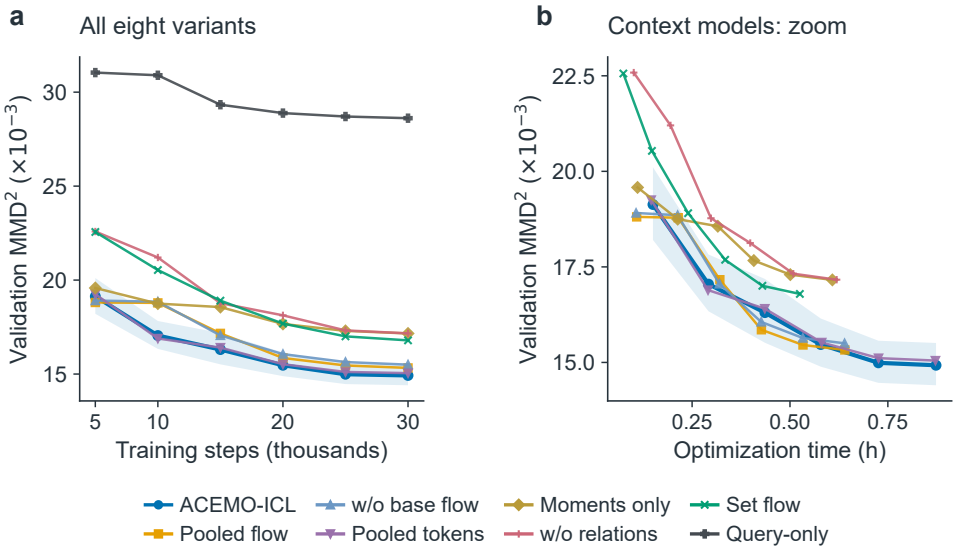

Figure 15: Validation learning curves along each training run. All formal comparisons use the final 30,000-step checkpoint.

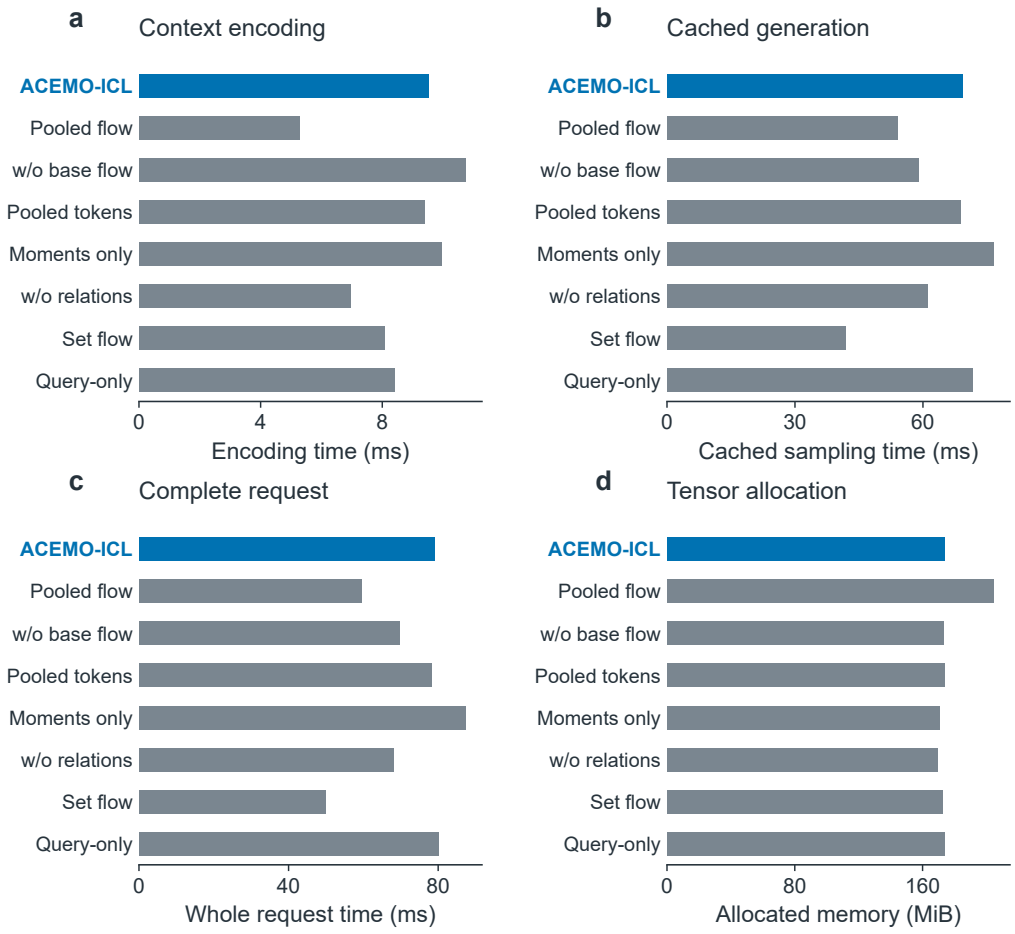

Figure 16: Measured context-encoding time, cached sampling time, complete-request time, and allocated memory on the evaluation hardware.

F Complete Grouped and Architectural Comparisons

We analyze the 672 existing task means post hoc: 117 zero-effect, 125 weak-effect, and 430 moderate-effect tasks, with 48 tasks per mechanism. Groups follow the original generator intervals. Positive gains denote baseline-minus-ACEMO MMD^2. Pointwise 95% intervals use 2,000 paired bootstrap draws over task means, stratified by mechanism within mixed groups, and quantify task uncertainty at fixed weights. The analysis uses the existing matrices and generated samples and reports no additional p-values.

Method	MMD V	MMD U	Sliced W_1	Cov. error
ACEMO-ICL	15.24	7.01	0.269	0.488
Set conditional flow	17.15	8.90	0.283	0.544
Pooled-context single flow	15.70	7.45	0.273	0.512
No base decomposition	15.78	7.53	0.273	0.513
Compressed demonstrations	15.29	7.06	0.269	0.489
Moments only	17.58	9.33	0.279	0.496
No explicit relations	17.59	9.34	0.285	0.552
Query-only model	29.01	20.69	0.359	0.553
Marginal calibration	28.97	20.80	0.365	0.621
Gaussian OT	50.49	42.11	0.520	0.856
Query resampling	35.05	27.06	0.374	0.554
Support Y resampling	133.94	125.49	0.910	0.996
Independent target reference	8.22	0.03	0.195	0.375

Table 9: All neural variants and simple baselines on the IID test. MMD values are multiplied by 10^3. Table 5 gives the corresponding parameter counts and auxiliary supervision.

The token-compression contrast has matched parameter count, source supervision, task schedule, and update count. The no-base contrast changes both architecture and auxiliary objectives. Moment-only and no-relation models have fewer parameters. The generic set flow has a similar parameter count but receives fewer auxiliary targets and uses less training time. Accordingly, the token contrast isolates evidence compression, while the other contrasts include changes in supervision or capacity.

Table 10: Complete strength breakdown. The six comparators are reported in every effect group.

Effect	Comparator	Tasks	Gain $\times 10^3$	95% interval
Moderate	Gaussian OT	430	38.42	[35.52, 41.25]
Moderate	Marginal	430	15.80	[14.69, 17.01]
Moderate	Query-only	430	20.66	[18.68, 22.57]
Moderate	Query samples	430	28.86	[26.54, 31.23]
Moderate	Set flow	430	2.17	[1.85, 2.53]
Moderate	Support Y	430	113.00	[108.43, 117.56]
Weak	Gaussian OT	125	31.38	[28.22, 34.68]
Weak	Marginal	125	10.02	[9.21, 10.97]
Weak	Query-only	125	1.71	[1.51, 1.93]
Weak	Query samples	125	3.99	[3.63, 4.33]
Weak	Set flow	125	1.43	[1.17, 1.70]
Weak	Support Y	125	132.06	[123.89, 140.08]
Zero	Gaussian OT	117	27.70	[24.74, 30.91]
Zero	Marginal	117	10.06	[8.95, 11.33]
Zero	Query-only	117	1.32	[1.15, 1.51]
Zero	Query samples	117	3.44	[2.97, 3.87]
Zero	Set flow	117	1.46	[1.18, 1.74]
Zero	Support Y	117	125.35	[117.26, 133.56]

Mechanism	Comparator	Tasks	Gain $\times 10^3$	95% interval
Additive noise	Gaussian OT	48	46.03	[36.95, 56.20]
Additive noise	Marginal	48	18.06	[14.40, 22.10]
Additive noise	Query-only	48	3.80	[2.51, 5.20]
Additive noise	Query samples	48	11.47	[8.53, 14.53]
Additive noise	Set flow	48	1.22	[0.78, 1.69]
Additive noise	Support Y	48	101.61	[88.66, 115.27]
Affine gaussian	Gaussian OT	48	30.51	[25.38, 36.56]
Affine gaussian	Marginal	48	11.54	[9.62, 13.70]
Affine gaussian	Query-only	48	51.06	[39.08, 64.09]
Affine gaussian	Query samples	48	63.13	[49.38, 77.51]
Affine gaussian	Set flow	48	1.60	[0.91, 2.40]
Affine gaussian	Support Y	48	116.02	[104.72, 127.68]
Branching	Gaussian OT	48	35.95	[27.18, 46.91]
Branching	Marginal	48	25.32	[19.21, 32.05]
Branching	Query-only	48	31.86	[22.02, 42.72]
Branching	Query samples	48	40.83	[29.66, 53.40]
Branching	Set flow	48	2.74	[1.25, 4.48]
Branching	Support Y	48	108.49	[96.42, 120.40]
Composition	Gaussian OT	48	50.80	[38.13, 66.12]
Composition	Marginal	48	19.21	[14.90, 24.69]
Composition	Query-only	48	17.10	[11.43, 23.44]
Composition	Query samples	48	26.00	[19.26, 33.62]
Composition	Set flow	48	2.19	[1.54, 2.94]
Composition	Support Y	48	101.57	[89.90, 114.64]
Correlated noise	Gaussian OT	48	50.47	[40.58, 61.23]
Correlated noise	Marginal	48	21.19	[17.59, 25.31]
Correlated noise	Query-only	48	3.44	[2.29, 4.72]
Correlated noise	Query samples	48	11.34	[8.96, 13.79]
Correlated noise	Set flow	48	1.27	[0.93, 1.64]
Correlated noise	Support Y	48	106.15	[92.82, 120.93]
Heteroskedastic	Gaussian OT	48	40.30	[33.01, 48.31]
Heteroskedastic	Marginal	48	12.90	[11.15, 14.65]
Heteroskedastic	Query-only	48	1.20	[0.83, 1.65]
Heteroskedastic	Query samples	48	5.55	[4.40, 6.72]
Heteroskedastic	Set flow	48	1.35	[1.00, 1.76]
Heteroskedastic	Support Y	48	122.77	[109.11, 138.02]
Identity	Gaussian OT	48	26.44	[21.47, 31.95]
Identity	Marginal	48	9.53	[7.85, 11.56]
Identity	Query-only	48	1.38	[1.17, 1.61]
Identity	Query samples	48	3.46	[2.65, 4.14]
Identity	Set flow	48	1.48	[1.12, 1.94]
Identity	Support Y	48	119.74	[107.53, 131.96]
Lowrank mix	Gaussian OT	48	26.45	[21.44, 32.24]
Lowrank mix	Marginal	48	11.37	[9.33, 13.86]
Lowrank mix	Query-only	48	2.48	[1.58, 3.77]
Lowrank mix	Query samples	48	6.30	[5.18, 7.98]
Lowrank mix	Set flow	48	1.79	[1.10, 2.64]
Lowrank mix	Support Y	48	115.07	[101.03, 129.69]
Multiplicative noise	Gaussian OT	48	37.26	[30.56, 44.45]
Multiplicative noise	Marginal	48	12.13	[10.58, 13.75]
Multiplicative noise	Query-only	48	0.53	[0.35, 0.72]
Multiplicative noise	Query samples	48	5.06	[4.62, 5.52]
Multiplicative noise	Set flow	48	1.31	[1.00, 1.66]
Multiplicative noise	Support Y	48	119.07	[105.46, 133.04]
Nonlinear	Gaussian OT	48	26.00	[20.98, 31.99]
Nonlinear	Marginal	48	8.27	[6.89, 9.85]
Nonlinear	Query-only	48	13.94	[10.43, 18.04]

Table 11 continued

Mechanism	Comparator	Tasks	Gain $\times 10^3$	95% interval
Nonlinear	Query samples	48	18.68	[14.39, 22.99]
Nonlinear	Set flow	48	2.06	[1.31, 2.83]
Nonlinear	Support Y	48	124.54	[108.05, 141.65]
Reweight	Gaussian OT	48	38.57	[31.99, 45.63]
Reweight	Marginal	48	12.37	[10.05, 15.35]
Reweight	Query-only	48	1.73	[1.29, 2.16]
Reweight	Query samples	48	4.37	[3.45, 5.51]
Reweight	Set flow	48	1.34	[0.76, 1.85]
Reweight	Support Y	48	142.53	[127.49, 158.55]
Scale	Gaussian OT	48	28.42	[23.18, 34.69]
Scale	Marginal	48	9.10	[7.25, 11.27]
Scale	Query-only	48	6.17	[4.31, 8.27]
Scale	Query samples	48	9.56	[7.52, 11.88]
Scale	Set flow	48	3.69	[2.72, 4.78]
Scale	Support Y	48	135.46	[120.97, 149.83]
Shift	Gaussian OT	48	29.24	[24.11, 34.54]
Shift	Marginal	48	6.56	[5.55, 7.64]
Shift	Query-only	48	54.15	[39.66, 69.13]
Shift	Query samples	48	62.82	[46.86, 79.20]
Shift	Set flow	48	1.19	[0.50, 1.97]
Shift	Support Y	48	130.85	[115.95, 146.38]
Sparse mix	Gaussian OT	48	26.99	[21.30, 33.41]
Sparse mix	Marginal	48	14.58	[11.74, 17.81]
Sparse mix	Query-only	48	3.92	[1.90, 6.53]
Sparse mix	Query samples	48	8.77	[6.06, 12.11]
Sparse mix	Set flow	48	3.46	[1.82, 5.52]
Sparse mix	Support Y	48	117.87	[103.02, 134.99]

Table 11: Complete mechanism breakdown. All fourteen programs and six comparators are included. The underlying task differences and bootstrap outputs are provided with the analysis code.