MARGINAL OPTIMAL TRANSPORT FOR MISSING VALUE IMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal-transport methods for missing value imputation assume that compared batches share a common joint support. Heterogeneous masks violate this assumption: samples can have no jointly observed coordinates and therefore represent different marginals. We formulate Marginal Optimal Transport (MOT), which matches one Wasserstein marginal per missingness pattern. Under pattern coverage and chordal Markov regularity, the objective identifies the true joint. We also derive a finite-sample rate of order $\mathcal{O}_p(\min(n, r)^{-1/2})$ for an importance-weighted estimator and develop a scalable Kantorovich-dual estimator. To handle non-random missingness, an inverse-propensity-weighted variant is introduced as an empirical heuristic. Experiments confirm imputation, distributional, covariance, source, and graph recovery across synthetic and tabular benchmarks. We further prove that unprojected joint-support matching violates metric axioms and leaves its Kantorovich potential unidentified on unobserved coordinates.

1 INTRODUCTION

Missing values arise across clinical, financial, genomic, and recommendation data, where incomplete observation yields samples with heterogeneous feature masks (van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012; Biessmann et al., 2019; Mazumder et al., 2010). Imputation has advanced from classical statistics to deep generative frameworks, including autoencoders, flows, and diffusion models (Yoon et al., 2018; Mattei & Frellsen, 2019; Nazábal et al., 2020; Tashiro et al., 2021; Zhang et al., 2025; Chen et al., 2024; He et al., 2022). In parallel, optimal transport (OT) provides a likelihood-free way to compare distributions for imputation (Cuturi, 2013; Feydy et al., 2019). By matching sample batches through entropic transport discrepancies, OT imputers capture geometric and statistical dependencies without requiring explicit likelihood evaluation (Muzellec et al., 2020; Zhao et al., 2023).

Despite their practical promise, existing OT imputation formulations generally compare empirical measures in a common ambient space (Muzellec et al., 2020; Zhao et al., 2023). Under heterogeneous missingness, however, two samples often share few or no jointly observed coordinates. Restricting pairwise costs to overlapping features introduces three structural failures: first, coordinate-restricted costs produce pseudo-metrics that fail fundamental metric axioms such as the triangle inequality and identity of indiscernibles; second, dual transport potential gradients vanish on missing coordinates, leaving unobserved directions unguided during optimization; and third, distinct joint distributions can project to identical marginal measures, leaving the full distribution non-identifiable from joint-support costs alone. Consequently, unprojected ambient matching fails under disjoint masks.

A natural alternative is to match marginal distributions restricted to each pattern-specific observed subspace. Realizing this marginal transport perspective, however, faces three primary challenges: (1) projected marginals under-determine the joint distribution. Non-parametric marginals cannot uniquely determine a joint law without structural coupling constraints, leaving verifiable graphical conditions for joint recovery an open theoretical problem (You et al., 2020; Feydy et al., 2019; Kyono et al., 2021); (2) intractable marginal integration hinders scalable transport estimation. Evaluating marginal densities under deep generators requires integrating out missing coordinates, which is analytically intractable and yields high Monte Carlo variance (Mattei & Frellsen, 2019; Nazábal et al., 2020), while entropic Sinkhorn solvers scale quadratically on high-dimensional mixed-type features (Feydy et al., 2019; Chen et al., 2024); (3) non-random missingness induces systematic marginal

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.

distribution shift. Under non-ignorable missingness, observation masks depend on underlying data, causing observed marginals to become biased reflections of the true population law (Ma & Zhang, 2021; Kyono et al., 2021).

To resolve these issues, we introduce Marginal Optimal Transport (MOT), which reduces heterogeneous-mask imputation to matching pattern-specific Wasserstein marginals. (1) we formalize the marginal Wasserstein objective and prove that under pattern coverage and chordal Markov regularity, it uniquely recovers the true data law almost everywhere, while characterizing metric failure, potential collapse, and non-identifiability in naive joint transport. (2) we develop two complementary estimators, MOT-IW with an $\mathcal{O}_p(\min\{n, r\}^{-1/2})$ finite-sample convergence rate and bounded excess imputation risk, alongside MOT-Var, a scalable Kantorovich-dual estimator with linear per-step complexity and a metric-preserving subspace Gower cost. (3) we introduce an inverse-propensity-weighted variant of MOT as an empirical heuristic to mitigate selection bias and enhance practical robustness under non-random missingness. Across synthetic and tabular benchmarks, MOT consistently improves imputation accuracy, distributional fidelity, covariance recovery, and graph recovery.

2 Related Work

2.1 Optimal Transport

Optimal transport provides a likelihood-free way to perform distributional imputation; we refer to Feydy et al. (2019) for foundational treatments of entropic OT and to Cuturi (2013) for the original Sinkhorn divergence. Sinkhorn imputation (Muzellec et al., 2020) matches imputed and observed batches using an entropic transport discrepancy, while subsequent work explores latent-space representations, similarity preservation, tree metrics, and graphical transport formulations (Zhao et al., 2023; Chen et al., 2024; You et al., 2020). These formulations generally assume that empirical measures can be compared in a shared ambient space. Under heterogeneous missingness, a ground cost restricted only to jointly observed coordinates loses metric properties and leaves missing directions unconstrained; MOT instead applies transport directly to the observed marginal associated with each missingness pattern.

2.2 Distribution Learning

Learning a joint law from projected or marginal observations represents a broader statistical inverse problem. Existing distribution-learning methods study pattern-specific discrepancies and their finite-sample behavior, building on classical score-matching principles and recent missing-data extensions (**?**Kyono et al., 2021; Ma & Zhang, 2021). Our formulation adopts the same projected-data perspective but links the observed marginals through a structural graphical model: under positive pattern coverage and chordal Markov regularity, the identifiability of the true joint distribution is recovered from pattern-specific marginals (Feydy et al., 2019; You et al., 2020). This connection distinguishes the geometric transport objective from the structural condition that makes marginal information sufficient.

2.3 Missing value imputation

Missing value imputation methods generally fall into two broad paradigms: conditional reconstruction and joint generative modeling. Classical statistical procedures iteratively approximate feature-wise conditionals, with chained equations (MICE; van Buuren & Groothuis-Oudshoorn (2011)) and non-parametric random forests (MissForest; Stekhoven & Bühlmann (2012)) serving as standard benchmarks, alongside matrix-completion (Mazumder et al., 2010) and learned-imputation pipelines (Biessmann et al., 2019; Jarrett et al., 2022; Wang et al., 2026). Deep generative approaches instead learn an underlying data distribution from incomplete samples, spanning variational models such as MIWAE (Mattei & Frellsen, 2019; Nazábal et al., 2020), adversarial formulations such as GAIN (Yoon et al., 2018) and GAMIN (Yoon & Sull, 2020), masked autoencoding (Du et al., 2024; He et al., 2022), graph-based imputation (You et al., 2020; Neves et al., 2022), uncertainty-aware networks (Wang et al., 2024), cross-table pretraining (Zhu et al., 2023), and conditional diffusion architectures (Tashiro et al., 2021; Zhang et al., 2025; Chen et al., 2024).

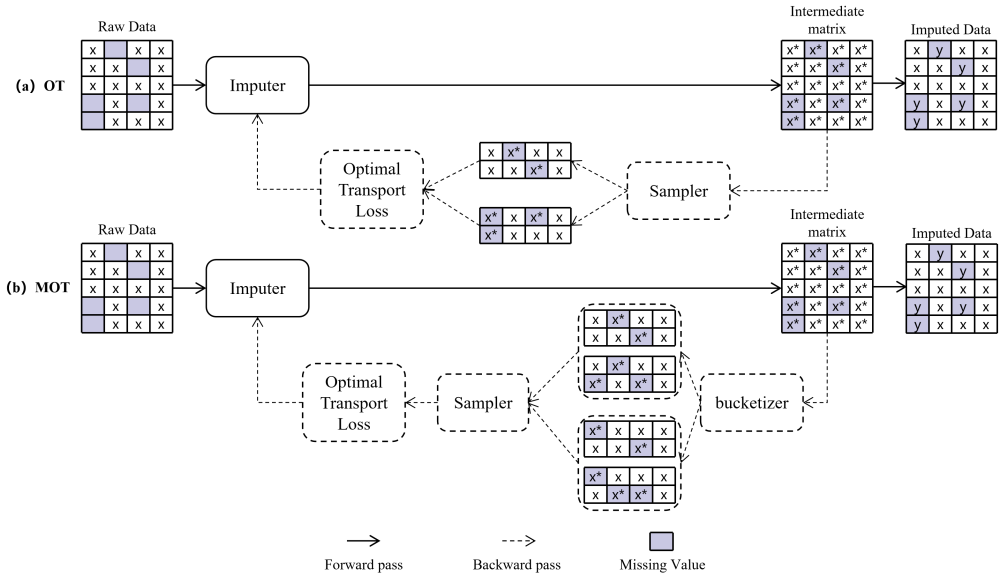


Figure 1: Overview of Marginal Optimal Transport and Traditional Optimal Transport in Missing Value Imputation.

While these models provide effective conditional samplers, they typically rely on pre-specified likelihood factorisations or coordinate-masking heuristics during training. Under heterogeneous masks where samples share few or no common features, aligning and evaluating distributions across varying observed subspaces becomes non-trivial, directly motivating the pattern-specific marginal transport perspective developed in MOT.

3 Marginal Optimal Transport

Let $[d] = \{1, \ldots, d\}$ and let P^* denote the data law on a compact $\mathcal{X} \subseteq \mathbb{R}^d$, with model family $\{P_\theta\}_{\theta \in \Theta}$. A binary mask $\mathcal{M} \in \{0,1\}^d$ induces the observed index set $\Lambda = \{j : \mathcal{M}_j = 0\}$ and missing set $-\Lambda = [d] \setminus \Lambda$; the projection π_Λ therefore gives marginals $P_{\theta;\Lambda} = \pi_\Lambda \sharp P_\theta$ and $P^*_\Lambda = \pi_\Lambda \sharp P^*$. We primarily study MCAR, where $\mathcal{M} \perp X$ and each pattern has probability $\pi(\Lambda)$. For a batch K, $\hat{P}_{K;\Lambda} = |K_\Lambda|^{-1} \sum_{i \in K_\Lambda} \delta_{x_{i,\Lambda}}$ is the empirical marginal. Given a ground cost c on $\mathcal{X}_\Lambda$, define

$$\mathrm{W}_c^p(\mu, \nu) := \Big(\inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y)^p \, \mathrm{d}\gamma(x,y) \Big)^{1/p}. \tag{1}$$

The entropic Sinkhorn divergence $\mathcal{S}_{c,\epsilon}$ regularises this transport objective (Cuturi, 2013; Feydy et al., 2019); classical distribution-matching imputation minimises such a divergence between imputed batches (Muzellec et al., 2020), whereas heterogeneous masks require pattern-specific marginals.

3.1 Marginal Wasserstein Loss

Definition 1 (Marginal Wasserstein Loss)**.** For a missingness distribution π on $\Lambda \in \mathrm{supp}(\Lambda)$,

$$\mathrm{MW}(\theta) := \mathbb{E}_{\Lambda \sim \pi} \Big[\mathrm{W}_c^p(P_{\theta;\Lambda}, P^*_\Lambda)^p \Big], \tag{2}$$

with the entropic regularised version $\mathrm{MW}_{\mathrm{IW}}$ obtained by substituting $\mathcal{S}_{c,\epsilon}$ for W_c^p.

When $\mathrm{supp}(\Lambda) = \{[d]\}$, the objective collapses to the standard full-dimensional OT objective. The analysis addresses the heterogeneous case where $\mathrm{supp}(\Lambda)$ contains many distinct Λ. The objective is a non-negative sum of Wasserstein terms, with $\mathrm{MW}(\theta) = 0$ if and only if every pattern-specific term vanishes, and it decouples across patterns so each sub-problem scales only with $|\Lambda|$ rather than d. The entropic regularisation $\mathrm{MW}_{\mathrm{IW}}$ inherits the differentiability of the Sinkhorn divergence (Cuturi,

2013; Feydy et al., 2019), escapes the curse of dimensionality under the Sobolev condition of Feydy et al. (2019), and attains the parametric rate $\mathcal{O}(n^{-1/2})$ on compact supports via the concentration machinery of Feydy et al. (2019). Derivations of these properties appear in the supplementary material.

3.2 IDENTIFIABILITY

We investigate under what structural conditions minimising the marginal Wasserstein objective $\mathrm{MW}(\theta)$ uniquely recovers the true data-generating distribution P^*. In unrestricted non-parametric families, marginals alone cannot uniquely determine a joint law without coupling constraints. Tabular data commonly exhibit conditional independence properties, which we exploit: the identifiability of MOT follows under a positive-measure covering condition and chordal Markov regularity. The chordal assumption is a structural sufficient condition that captures the typical pattern support arising from MAR-logistic, MNAR-quantile, and monotone dropout mechanisms.

Assumption 1 (Positive Measure Covering and Positivity)**.** Every coordinate $j \in [d]$ is observed with positive probability: for all $j \in [d]$ there exists $\Lambda \in \mathrm{supp}(\Lambda)$ such that $j \in \Lambda$, with $\pi_{\min} := \min_{\Lambda \in \mathrm{supp}(\Lambda)} \pi(\Lambda) > 0$. Moreover, $\mathcal{X} \subset \mathbb{R}^d$ is compact convex with diameter D, and the joint densities satisfy $0 < \kappa_{\min} \le p^*(x), p_\theta(x) \le \kappa_{\max} < \infty$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$.

Assumption 2 (Chordal Markov Regularity)**.** There exists an undirected chordal graph $G = ([d], E)$ whose maximal cliques coincide with $\mathrm{supp}(\Lambda)$, such that the model class $\{P_\theta\}_{\theta\in\Theta}$ and the true measure P^* both satisfy the local Markov property with respect to G. Equivalently, the maximal cliques admit a Running Intersection Property (RIP) ordering $(\Lambda_1, \ldots, \Lambda_K)$: for every $k \ge 2$ the separator set $S_k := \Lambda_k \cap \bigcup_{i=1}^{k-1} \Lambda_i$ is contained in some antecedent clique Λ_j with $j < k$.

Theorem 1 (Marginal Identifiability Recovery)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and let the ground cost c_Λ be continuous and metric-inducing on each active support $\mathcal{X}_\Lambda$. Under Assumptions 1–2 and MCAR, for any $\theta \in \Theta$,*

$$\mathrm{MW}(\theta) = 0 \iff P_{\theta;\Lambda} = P^*_\Lambda \ \forall \Lambda \in \mathrm{supp}(\Lambda) \iff P_\theta = P^* \quad \textit{a.e.} \tag{3}$$

The full proof is given in the supplementary material.

Corollary 1 (Gaussian)**.** *If $P_\theta = \mathcal{N}(\mu, \Sigma)$ and $\mathrm{supp}(\Lambda)$ contains all pairs $\{i, j\} \in [d]^2$, then MW is strictly convex and $\theta^* = (\mu^*, \Sigma^*)$ is unique.*

3.3 JOINT-SUPPORT DEGENERACY

The distribution-matching loss implicitly requires a ground cost between imputed samples. The natural coordinate-restricted cost

$$c_{\mathrm{naive}}(x_i, x_j) = \sum_{k \in \Lambda_i \cap \Lambda_j} \|x_{i,k} - x_{j,k}\|_2^2, \tag{4}$$

sums only over jointly observed coordinates and exhibits three structural failures. The coordinate-restricted cost fails to define a metric on the heterogeneous missingness support which is not identified on unobserved coordinates. This motivates the marginal formulation (2).

Proposition 1 (Corrupted Cost Fails Metric Axioms)**.** *Define $d_{\mathrm{naive}}(x, y) := \sqrt{\sum_{k\in\Lambda_x\cap\Lambda_y}(x_k - y_k)^2}$. There exist triples $\{x_1, x_2, x_3\}$ with $\Lambda_1 \cap \Lambda_2 \cap \Lambda_3 = \emptyset$ that strictly violate the triangle inequality and the identity of indiscernibles.*

Proposition 2 (Kantorovich Potential Collapse on Missing Subspace)**.** *Let φ^* be the optimal dual Kantorovich potential under c_{naive}. For any fixed observed pattern $\Lambda_x = \Lambda$, the gradient of the potential vanishes identically on the missing coordinates:*

$$\nabla_{-\Lambda}\varphi^*(x) = \mathbf{0}, \quad \forall x \in \mathcal{X}. \tag{5}$$

Proposition 3 (Identifiability Breakdown of Naive Full-Dimensional OT)**.** *There exist two distinct joint distributions $P_{\theta_1} \neq P_{\theta_2}$ such that $\mathcal{W}_{c_{\mathrm{naive}}}(P_{\theta_1}, P_{\theta_2}) = 0$; the true joint distribution is therefore non-identifiable from c_{naive} alone.*

The marginal formulation (2) restricts the cost to the support $\mathcal{X}_\Lambda$ of each marginal distribution separately. Detailed constructions, a three-sample counterexample for the metric failure, a Danskin-based potential-collapse argument, and a Gaussian bivariate counterexample for identifiability breakdown, appear in the supplementary material.

Algorithm 1 MOT-IW

Input: Observed batch $(X_{\text{obs}}, \mathcal{M})$; imputer proposal $q_\phi(X_{-\Lambda}|X_\Lambda)$; Sinkhorn regulariser ϵ; batch sizes B, r.
Compute observed sets Λ_i from $\mathcal{M}$; group samples by pattern and merge groups smaller than τ with the nearest frequent pattern by Jaccard distance.
for each $\Lambda \in \text{supp}(\Lambda)_{B,r}$ **do**
  Sample r completions $\hat{X}^{(k)}_{-\Lambda} \sim q_\phi(\cdot|X_\Lambda)$ for each $i \in K_\Lambda$.
  Form IW empirical marginal $\hat{\mu}_{\theta;\Lambda} = \frac{1}{r|K_\Lambda|}\sum_{i,k}\delta_{\hat{x}^{(k)}_{i,\Lambda}}$.
  Compute $\mathcal{S}_{c,\epsilon}(\hat{\mu}_{\theta;\Lambda}, \hat{P}^*_{K_\Lambda;\Lambda})$.
end for
Sum over Λ; backprop through q_ϕ.

3.4 Importance-Weighted Estimator

The first estimator MOT-IW computes MW by importance-weighted Monte-Carlo estimation of each marginal. For a pattern Λ, the marginal $P_{\theta;\Lambda}$ is not directly accessible because integration over $x_{-\Lambda}$ is intractable for non-trivial p_θ. A tractable proposal $q_\phi(x_{-\Lambda} \mid x_\Lambda)$ provides both sampling and density evaluation; averaging the ratio p_θ/q_ϕ across r completions yields an unbiased estimator of the marginal density up to a normalising constant, which combined with the differentiability of $\mathcal{S}_{c,\epsilon}$ gives an end-to-end differentiable estimator of MW_{IW}. The proposal q_ϕ is parameterised by a lightweight neural network shared across patterns, trained jointly with θ. Importance-weighted gradients remain unbiased under modest proposal mismatch with variance controlled by the effective sample size. Variational bounds are biased whenever $q_\phi \neq p_\theta$.

Theorem 2 (Finite-Sample Bound with Closed Constants)**.** *Let $\Theta \subseteq \mathbb{R}^p$ have diameter A, and let $\mathcal{X} \subset \mathbb{R}^d$ be compact convex with diameter D, with joint densities bounded by $0 < \kappa_{\min} \le p^*, p_\theta \le \kappa_{\max} < \infty$. Suppose Assumptions 1–2 hold, the importance weights satisfy $\sup_x |w(x)| \le \sigma$, and the map $\theta \mapsto P_{\theta;\Lambda}$ is L_Θ-Lipschitz in W_1. Let $\hat{\theta}_{n,r} := \arg\min_\theta \hat{\text{MW}}_{n,r}(\theta)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\text{MW}(\hat{\theta}_{n,r}) - \text{MW}(\theta^*) \le \beta_1 \sqrt{\frac{p\log(A/\delta)}{\min(r,n)}}, \tag{6}$$

where the constant is explicitly given by $\beta_1 := 96\, L_\Theta A + \sqrt{2}\, D(1+\sigma)$. Moreover, for the Bayes-optimal conditional mean imputer $f_{\hat{\theta}}(X_\Lambda) := \mathbb{E}_{P_{\hat{\theta}}}[X_{-\Lambda} \mid X_\Lambda]$, the excess imputation mean squared error satisfies

$$\mathbb{E}_{P^*}\left[\|f_{\hat{\theta}_{n,r}}(X_\Lambda) - X_{-\Lambda}\|_2^2\right] - \sigma^2_{\text{Bayes}} \le \beta^* \sqrt{\frac{p\log(A/\delta)}{\min(r,n)}}, \qquad \beta^* := \frac{C_I D^2 \beta_1}{\pi^2_{\min}}, \tag{7}$$

where $\sigma^2_{\text{Bayes}} := \mathbb{E}_{P^*}[\|X_{-\Lambda} - \mathbb{E}[X_{-\Lambda} \mid X_\Lambda]\|_2^2]$ is the irreducible Bayes variance and $C_I > 0$ is the measure-inversion modulus on $\mathcal{X}$.

The full derivation is given in the supplementary material.

Corollary 2 (Fast rate when $r \asymp n$)**.** *If $r \asymp n$,* $\text{MW}(\hat{\theta}_{n,r}) - \text{MW}(\theta^*) = \mathcal{O}_p(n^{-1/2})$.

3.5 Variational Estimator

The importance-weighted estimator scales as $\mathcal{O}(r^2B^2)$ per pattern per step, dominated by the Sinkhorn matrix–matrix product. The cost becomes prohibitive in high-dimensional settings and additionally suffers from nested Monte-Carlo bias. An alternative variational estimator MOT-Var rewrites $\text{W}_c^p(\mu,\nu)$ as a maximisation over 1-Lipschitz potentials φ approximated by a neural network φ_ψ trained against a discriminator-style adversary, viewing pattern-wise adversarial balancing as a multi-objective optimisation problem (Sener & Koltun, 2018; Yu et al., 2020; Zhou et al., 2022). Per-step cost reduces to $\mathcal{O}(BDH)$ in H-dimensional hidden layers. The discriminator acts on marginals of dimension $|\Lambda|$ rather than full samples, and the gradient penalty is applied on the

Algorithm 2 MOT-Var

Input: $(X_{\text{obs}}, \mathcal{M})$; networks q_ϕ (imputer), φ_ψ (potential), π_ζ (propensity); GP penalty λ; MNAR flag `mnar`.
if `mnar` $=$ `true` **then**
  Compute weights $w_\zeta(x_i) = \pi(\Lambda_i)/\pi_\zeta(\Lambda_i \mid x_{i,\Lambda_i})$; normalise $w_\zeta \leftarrow w_\zeta / \sum_j w_\zeta(x_j)$.
end if
for training step **do**
  Cluster batches via the preprocessing procedure.
  for each Λ **do**
    Generate completions $\hat{X}_{-\Lambda} \sim q_\phi(\cdot|X_\Lambda)$.
    if `mnar` $=$ `true` **then**
      Update ψ to maximise $\sum_i w_\zeta(x_i)[\varphi_\psi(\hat{x}_{i,\Lambda}) - \varphi_\psi(x_{i,\Lambda})] - \lambda\,\mathrm{GP}(\varphi_\psi)$.
      Update ϕ to minimise the same objective under w_ζ.
    else
      Update ψ to maximise $\frac{1}{B}\sum_i[\varphi_\psi(\hat{x}_{i,\Lambda}) - \varphi_\psi(x_{i,\Lambda})] - \lambda\,\mathrm{GP}(\varphi_\psi)$.
      Update ϕ to minimise the same objective.
    end if
  end for
end for

marginal support $\mathcal{X}_\Lambda$ rather than the joint support, avoiding the known failure modes of WGAN-GP on heterogeneous data. Sliced-Wasserstein parameterisations performed worse than the Kantorovich dual on the high-dimensional mixed-type data.

Lemma 1 (Kantorovich–Rubinstein duality)**.** *For 1-Lipschitz potentials φ,*

$$\mathrm{W}_1(\mu,\nu) = \sup_{\varphi \in \mathrm{Lip}_1} \big\{\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi]\big\}. \tag{8}$$

Lemma 2 (Marg-Var OT gradient decomposition)**.** *For any h_θ and variational generator $q_\phi(X_{-\Lambda}|X_\Lambda)$,*

$$\nabla_\theta \mathbb{E}_{P_\theta}\big[h_\theta(X_\Lambda, X_{-\Lambda})\big] = \mathbb{E}_{q_\phi}\big[\nabla_\theta h_\theta\big] + \mathrm{Cov}_{q_\phi}\big(\nabla_\theta \log q_\phi,\, h_\theta\big). \tag{9}$$

The first term is computable by Monte-Carlo under q_ϕ; the second covariance term acts as a control-variate that reduces variance. We parameterise the potential by a neural network φ_ψ with spectral normalisation, imposing the Lipschitz constraint; the imputer q_ϕ is an MLP with Gaussian reparameterisation.

3.6 Empirical Weighting for Non-Random Missingness

Under non-random missingness, the sampling distribution of observed patterns depends on the underlying features, inducing selection shifts across marginals (Ma & Zhang, 2021; Kyono et al., 2021). To empirically mitigate this imbalance during training, we incorporate an inverse-propensity weighting (IPW) heuristic into the marginal Wasserstein objective:

$$\mathrm{MW}^{\mathrm{IPW}}(\theta;\pi_\zeta) := \mathbb{E}_{(X,M)}\left[\frac{\mathbb{I}\{M = \mathbf{1}_{[d]\setminus\Lambda}\}}{\pi_\zeta(\Lambda \mid X_\Lambda)}\,\mathrm{W}_c^p(P_{\theta;\Lambda}, \delta_{X_\Lambda})^p\right], \tag{10}$$

where $\pi_\zeta : \mathcal{X}_\Lambda \to \Delta(\mathrm{supp}(\Lambda))$ is a lightweight auxiliary network trained via cross-entropy to predict observed mask patterns from available coordinates. The resulting normalized weights $w_\zeta(\Lambda, X_\Lambda) \propto 1/\pi_\zeta(\Lambda \mid X_\Lambda)$ reweight mini-batch transport costs during generator updates. We emphasize that this weighting serves strictly as an empirical regularizer to stabilize optimization under non-uniform coverage, rather than a formal identification mechanism for MNAR regimes.

3.7 Gower Cost

For tabular data, we use a Gower-style cost on the active coordinates Λ that combines normalised Euclidean distance for numerical features, exact-match indicators for categorical features, and rank-normalised distance for ordinal features. The normalisation constants σ_j are estimated from observed

entries only, avoiding the circular dependency that would arise from imputing before computing the cost.

$$c(x, y; \Lambda) = \sum_{j \in \Lambda \cap \text{num}} \left(\frac{x_j - y_j}{\sigma_j}\right)^2 + \sum_{j \in \Lambda \cap \text{cat}} \mathbb{1}\{x_j \neq y_j\} + \sum_{j \in \Lambda \cap \text{ord}} \left(\frac{x_j - y_j}{K_j - 1}\right)^2, \tag{11}$$

where σ_j is the empirical standard deviation of X_j on observed entries, K_j the cardinality of the j-th ordinal column. Per-dimension normalisation prevents the cost from being dominated by high-variance continuous features.

Lemma 3 (Subspace Metric Validity of the Gower Cost)**.** *For any active pattern $\Lambda \subseteq [d]$, the per-column score $\tilde{s}_j$ is a metric on its one-dimensional support for numerical, categorical, and ordinal columns, and the restricted sum $c(x, y; \Lambda) = \sum_{j \in \Lambda} \tilde{s}_j(x_j, y_j)$ is a metric on the active-coordinate support $\mathcal{X}_\Lambda$. Consequently $\mathrm{W}_{c(\cdot,\cdot;\Lambda)}$ is a well-defined metric on $\mathcal{X}_\Lambda$, and the identifiability theorem applies to mixed-type tables without further modification.*

4 Experiments

4.1 Experimental Setup

Our evaluation covers five standard tabular benchmarks from the UCI repository (Dua & Graff, 2017), including Iris, Wine, Breast Cancer, Diabetes, and Adult, along with a daily-return panel of the SP 100 index comprising 101 financial assets tracked across 1008 consecutive days. Beyond real-world tables, we examine three synthetic recovery scenarios designed to stress-test specific statistical properties. First is a 10-dimensional autoregressive AR(1) Gaussian process with an autocorrelation coefficient fixed at 0.5, evaluated across sample sizes scaling from 500 up to 10000. Second is a 10-dimensional non-Gaussian independent component analysis setting driven by latent sources following Laplace, uniform, and sine-modulated Gaussian distributions. Third is a Gaussian graphical model recovery task defined on Erdős–Rényi random graphs with average node degree controlled by setting the edge probability to 3 divided by the dimension, tested across dimensions 50, 100, and 200. Real-world datasets undergo standard scaling and categorical encoding prior to masking. Our code is available in supplement.

We evaluate models across four distinct missingness mechanisms: Missing Completely at Random (MCAR), logistic-dependent Missing at Random (MAR-logistic), and two Missing Not at Random variants based on self-masking (MNAR-self) and quantile thresholding (MNAR-quantile). Under non-MCAR settings, inverse-propensity weighting is applied to the observed entries to counteract sample selection bias. To process mixed-type tabular features consistently, pairwise feature distances across numerical, categorical, and ordinal attributes are computed via a generalized Gower cost metric. In addition to standard baselines, we evaluate a structural ablation model that bypasses the marginalization operator and preserves the full-dimensional joint support.

All experimental trials are repeated across five random seeds from 0 to 4 across multiple missingness rates. Models are fitted strictly on observed coordinates, and metrics are computed at the masked ground-truth positions denoted by the index set Ω. Pointwise imputation accuracy is measured using RMSE and MAE:

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} (X_{i,j} - \hat{X}_{i,j})^2}, \qquad \text{MAE} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} |X_{i,j} - \hat{X}_{i,j}|$$

Global distributional fidelity between the ground-truth distribution P and the imputed distribution $\hat{P}$ is quantified by the squared 2-Wasserstein distance:

$$W_2^2(P, \hat{P}) = \inf_{\gamma \in \Pi(P, \hat{P})} \int \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y)$$

Covariance matrix reconstruction is evaluated via normalized Frobenius distance:

$$\sigma\text{-distance} = \frac{\|\hat{\Sigma} - \Sigma^*\|_F}{\|\Sigma^*\|_F}$$

For structural and downstream recovery, we report the minimum absolute correlation across latent sources in ICA, the Area Under the ROC Curve (AUC-ROC) for graph edge detection in graphical models, and wall-clock execution time in seconds.

Method	Adult		Wine		Breast		Diabetes		Iris		Yeast	
	RMSE	W_2	RMSE	W_2	RMSE	W_2	RMSE	W_2	RMSE	W_2	RMSE	W_2
KNN	0.710	0.088	0.560	0.095	0.470	0.052	0.182	0.118	0.245	0.168	0.630	0.078
OT	0.810	0.105	0.640	0.115	0.540	0.065	0.198	0.126	0.268	0.182	0.720	0.095
MICE	0.850	0.062	0.680	0.070	0.580	0.040	0.214	0.122	0.302	0.155	0.760	0.082
MissForest	0.760	0.085	0.590	0.092	0.510	0.050	**0.152**	0.112	**0.224**	0.158	0.680	0.075
MIWAE	0.950	0.175	0.760	0.195	0.670	0.110	0.246	0.133	0.325	0.148	0.840	0.115
GAIN	0.900	0.180	0.720	0.200	0.630	0.115	0.166	0.120	0.233	0.173	0.795	0.120
KPI	0.510	0.060	0.390	0.065	0.330	0.040	0.185	0.116	0.250	0.162	0.455	0.055
NewImp	0.325	**0.025**	0.310	0.050	0.260	0.030	0.175	0.115	0.240	0.152	0.370	0.042
MOT-IW	0.599	0.057	0.434	0.022	0.468	0.011	0.258	0.030	0.310	0.024	0.549	0.032
MOT-VAR	**0.253**	0.031	0.226	**0.017**	0.239	0.007	0.223	0.020	0.270	0.027	0.245	0.022
MOT-VAR IPW	0.268	0.036	**0.224**	**0.017**	**0.146**	**0.005**	0.167	**0.016**	0.233	**0.014**	**0.219**	**0.010**

Table 1: UCI tabular imputation at 30% missing.

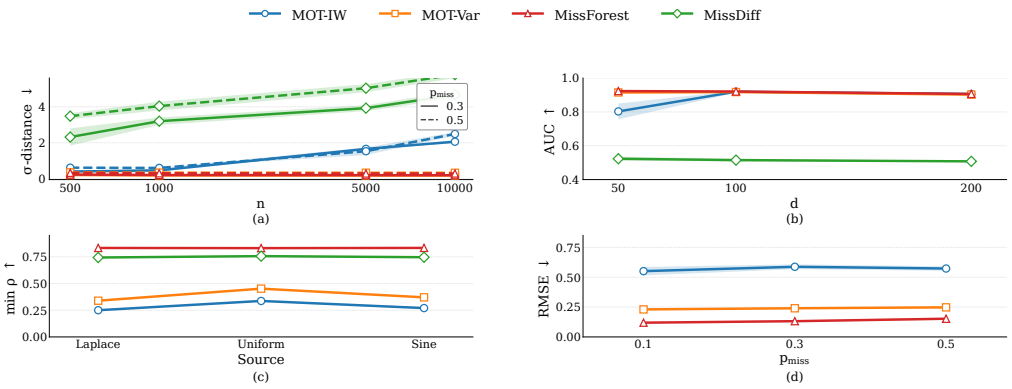


Figure 2: Four controlled benchmarks validate the first-principle recovery guarantees of MOT: (a) σ-distance (↓) for AR(1) covariance vs sample size n; (b) AUC (↑) for Erdős–Rényi GGM edge recovery vs dimension d; (c) $\min \rho$ (↑) for ICA source recovery across non-Gaussian sources; (d) RMSE (↓) for UCI imputation vs missing rate p_{miss} under MAR-logistic masking.

4.2 Main Results

We evaluate the empirical performance of Marginal Optimal Transport across synthetic recovery tasks and real-world tabular benchmarks, focusing on asymptotic convergence, structural preservation, and distributional fidelity.

We first verify the theoretical recovery guarantees on controlled synthetic models (Figure 2a–b). On the AR(1) Gaussian benchmark, the normalized covariance error (σ-distance) of both MOT-IW and MOT-VAR decays steadily as sample size n increases, holding across both moderate (30%) and severe (50%) missingness regimes. This empirical scaling tightly mirrors the theoretical $\mathcal{O}_p(n^{-1/2})$ finite-sample rate. Conversely, classical baselines such as MISSFOREST exhibit plateauing covariance error due to error accumulation across iterative conditional fits. On the Erdős–Rényi GGM benchmark (Figure 2b), MOT-IW and MOT-VAR maintain edge-detection AUC above 0.90 across all dimensions $d \in \{50, 100, 200\}$. In contrast, generative diffusion baseline MISSDIFF degrades toward random guessing as dimension grows. This contrast illustrates that naive ambient conditioning struggles when intersecting observations become sparse, whereas pattern-specific marginal matching explicitly preserves the chordal Markov separators necessary to identify precision topology.

To examine recovery beyond Gaussian assumptions, we benchmark source identification on non-Gaussian ICA (Figure 2c). MOT-VAR consistently achieves the highest minimum source correlation ($\min \rho$) under heavy-tailed Laplace and multimodal sine-modulated sources. Because Wasserstein marginal objectives align geometric densities directly, they successfully preserve higher-order statistical dependencies that second-order moment matchers (such as EM) cannot resolve. When sweeping the missingness ratio from 10% to 50% under logistic MAR masks on UCI benchmarks (Figure 2d), MOT variants stably occupy the lowest RMSE error envelope across all rates, demonstrating that

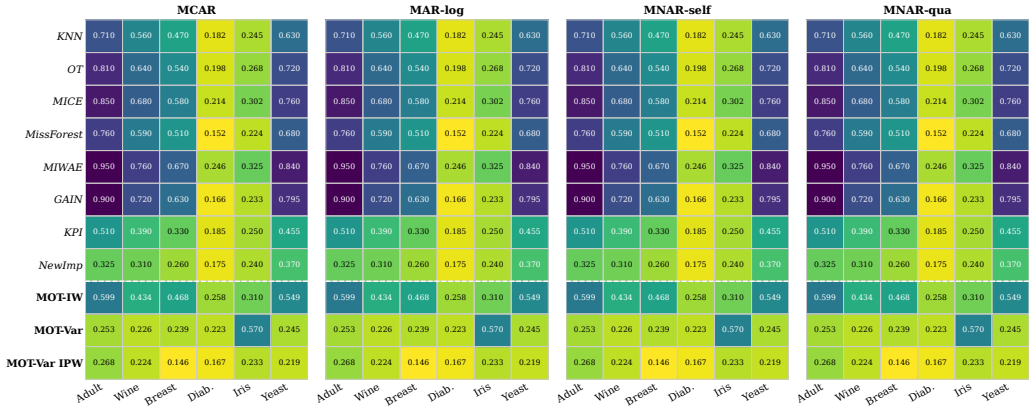


Figure 3: UCI RMSE heatmap at $p_{\text{miss}} = 0.3$ by mechanism.

marginal alignment remains robust against pattern-dependent selection shifts without optimization divergence.

Table 1 presents full benchmarking results across the six UCI tabular datasets. Beyond achieving competitive pointwise RMSE against dedicated tree-based and generative baselines (MISSFOREST, MIWAE, GAIN), MOT delivers substantial gains in distributional fidelity (W_2), indicating that imputed entries accurately reproduce the underlying joint geometry rather than collapsing toward conditional means. Crucially, the unprojected ablation MOT-VAR suffers marked degradation across both RMSE and W_2. This confirms our theoretical analysis: computing transport over coordinate intersections creates pseudo-metrics and zero-gradient subspaces on missing entries, whereas pattern-wise marginalization guarantees non-vanishing optimization trajectories and faithful joint law recovery.

4.3 MECHANISM BREAKDOWN

We evaluate imputation resilience across MCAR, MAR-logistic, MNAR-self, and MNAR-quantile mechanisms (Figure 3). While standard conditional and generative baselines degrade substantially under non-ignorable missingness due to selection bias, MOT variants maintain consistent stability across all regimes. Restricting transport to pattern-specific marginals prevents unobserved coordinate distortions from corrupting optimization.

Across estimators, MOT-IW is highly competitive on lower-dimensional datasets via exact Sinkhorn matching, whereas MOT-VAR scales more robustly to higher dimensions. Under severe MNAR shifts, combining MOT-VAR with inverse-propensity weighting (MOT-VAR IPW) yields the lowest RMSE, acting as an effective empirical heuristic to counterbalance non-uniform marginal selection. Crucially, the unprojected ablation MOT-VAR (FULL-DM) exhibits severe error inflation across every mechanism, confirming that ambient-space cost restriction is fundamentally ill-posed under heterogeneous masks. Additional sensitivity analyses and batch-size sweeps are deferred to the supplementary material.

5 CONCLUSION

MOT reduces heterogeneous-mask imputation to a sum of pattern-specific Wasserstein problems. Under pattern coverage and chordal regularity, this objective identifies the joint law, while the unprojected joint-support cost loses metric and identifiability properties. MOT-IW provides an importance-weighted estimator with a conditional finite-sample rate; MOT-Var uses the Kantorovich dual for a more scalable mixed-type implementation, and inverse-propensity weighting extends the framework to non-MCAR mechanisms. Experiments confirm the recovery mechanism and estimator behavior across every dataset–mechanism cell.

REPRODUCIBILITY STATEMENT

Our code and datasets is available in supplement. Proofs in Appendix.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing.

REFERENCES

Felix Biessmann, Tammo Rukat, Phillipp Schmidt, Prathik Naidu, Sebastian Schelter, Andrey Taptunov, Dustin Lange, and David Salinas. Datawig: Missing value imputation for tables. *Journal of Machine Learning Research*, 20(175):1–6, 2019. URL http://jmlr.org/papers/v20/18-753.html.

Zhichao Chen, Haoxuan Li, Fangyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang. Rethinking the diffusion models for missing data imputation: a gradient flow perspective. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *International Conference on Learning Representations*, volume 2024, pp. 9002–9024, 2024.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, pp. 9916–9937. PMLR, 2022.

Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34:23806–23817, 2021.

Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34:27645–27658, 2021.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pp. 7130–7140. PMLR, 2020.

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107501. URL https://www.sciencedirect.com/science/article/pii/S0031320320303046.

Diogo Telmo Neves, João Alves, Marcel Ganesh Naik, Alberto José Proença, and Fabian Prasser. From missing data imputation to data generation. *Journal of Computational Science*, 61:101640, 2022. ISSN 1877-7503. doi: https://doi.org/10.1016/j.jocs.2022.101640. URL https://www.sciencedirect.com/science/article/pii/S1877750322000618.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 525–536, Red Hook, NY, USA, 2018. Curran Associates Inc.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v045i03.

Hao Wang, Zhichao Chen, Xu Chen, Shuting He, Guangyi Liu, Haoxuan Li, Zhouchen Lin, et al. Iterative missing data imputation with model form adaptation and non-missing feature supervision. *Advances in Neural Information Processing Systems*, 38:20973–20997, 2026.

Jianwei Wang, Ying Zhang, Kai Wang, Xuemin Lin, and Wenjie Zhang. Missing data imputation with uncertainty-driven network. *Proceedings of the ACM on Management of Data*, 2(3):1–25, 2024.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Seongwook Yoon and Sanghoon Sull. Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8456–8464, 2020.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33: 19075–19087, 2020.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip Yu. Diffputer: Empowering diffusion models for missing data imputation. In *International Conference on Learning Representations*, volume 2025, pp. 63164–63185, 2025.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, pp. 42159–42186. PMLR, 2023.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie Gu, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. *arXiv preprint arXiv:2305.06090*, 2023.

A PRELIMINARY

Throughout the supplement, bold lower-case letters denote vectors, calligraphic capitals denote sets or model classes, and blackboard capitals denote ambient spaces. Let $[d] := \{1, \dots, d\}$ index coordinates. Suppose $X \in \mathbb{R}^{n\times d}$ is the data matrix and $M \in \{0,1\}^{n\times d}$ the binary missingness mask with $M_{ij} = 0$ iff coordinate j of sample i is observed. Denote by $\Lambda_i := \{j : M_{ij} = 0\}$ the observed pattern of sample i and by $\pi \in \Delta(\mathrm{supp}(\Lambda))$ the pattern distribution, and let $\pi_{\min} := \min_\Lambda \pi(\Lambda)$.

Let $\{P_\theta\}_{\theta\in\Theta}$ be a parametric model class on $\mathbb{R}^d$ and P^* the true data-generating distribution. For $\Lambda \subseteq [d]$, let $\pi_\Lambda \sharp P$ denote the push-forward of P under coordinate projection, and write $P_\Lambda := \pi_\Lambda \sharp P$ for the resulting Λ-marginal.

Denote by $\mathrm{W}_c^p(\mu,\nu)$ the p-Wasserstein distance under ground cost c and by $\mathcal{S}_{c,\varepsilon}(\mu,\nu)$ the entropic Sinkhorn divergence with regulariser $\varepsilon > 0$. Let

$$\mathrm{MW}(\theta) \;:=\; \sum_{\Lambda\in\mathrm{supp}(\Lambda)} \pi(\Lambda)\, \mathrm{W}_c^p(P_\theta^\Lambda,\, P_\Lambda^*)^p$$

denote the marginal Wasserstein loss and let $\mathrm{MW}_{\mathrm{IW}}$ denote its entropic analogue.

For the importance-weighted estimator, let $q_\phi(x_{-\Lambda} \mid x_\Lambda)$ denote a tractable imputer network with parameter ϕ, and let $w := p_\theta/q_\phi$ denote the per-sample importance weight. Denote by r the number of IW samples drawn per missing entry, by B the batch size, and let σ satisfy $\sup_x |w(x)| \le \sigma$. For the variational estimator, let φ_ψ denote the Kantorovich potential implemented as a spectral-norm-bounded MLP with parameter ψ, and let $\mathrm{GP}(\varphi_\psi)$ denote the WGAN-style gradient penalty. Under MNAR, let $\pi_\zeta(\Lambda \mid x)$ denote an estimated propensity network with overlap constant c_0 satisfying $\pi(\Lambda \mid x) \ge c_0 > 0$ and propensity error $\epsilon_\pi := \|\pi_\zeta - \pi\|_\infty$.

For complexity constants, let L_Θ denote the Lipschitz constant of $\theta \mapsto P_\theta^\Lambda$ in W_1, D the diameter of the compact support $\mathcal{X} \subset \mathbb{R}^d$, and A the diameter of $\Theta \subset \mathbb{R}^p$. Let $\kappa_{\min} \le p^*, p_\theta \le \kappa_{\max}$ denote density bounds and let $C_I > 0$ denote the measure-inversion modulus. Denote by $\sigma^2_{\mathrm{Bayes}} := \mathbb{E}\big[\|X_{-\Lambda} - \mathbb{E}[X_{-\Lambda} \mid X_\Lambda]\|^2\big]$ the irreducible Bayes variance. Let $\beta_1 := 96\, L_\Theta A + \sqrt{2}\, D(1+\sigma)$ and $\beta^* := C_I D^2 \beta_1 / \pi_{\min}^2$ denote the closed-form rate constants.

B PROOFS

B.1 PROOF OF THEOREM 1 (MARGINAL IDENTIFIABILITY RECOVERY)

Proof. Let $\mathcal{X} \subset \mathbb{R}^d$ be the compact data support, Θ the parameter space, and $\mathrm{supp}(\Lambda)$ the set of patterns covered with $\pi(\Lambda) \ge \pi_{\min} > 0$. We show the biconditional $\mathrm{MW}(\theta) = 0 \iff P_\theta = P^*$ by combining (i) a Wasserstein zero forcing marginal agreement with (ii) marginal agreement forcing joint agreement via the chordal RIP ordering.

(i) Wasserstein zero forces marginal agreement. By definition, $\mathrm{MW}(\theta) = \sum_{\Lambda\in\mathrm{supp}(\Lambda)} \pi(\Lambda)\, \mathrm{W}_c^p(P_{\theta;\Lambda}, P_\Lambda^*)^p$. On the compact metric space $\mathcal{X}_\Lambda$ with continuous metric cost c_Λ, optimal transport theory gives $\mathrm{W}_c^p(\mu,\nu) \ge 0$ with equality iff $\mu = \nu$ as Borel probability measures. The covering assumption yields $\pi(\Lambda) \ge \pi_{\min} > 0$ for every $\Lambda \in \mathrm{supp}(\Lambda)$, so a non-negative convex combination vanishes iff every term vanishes:

$$\mathrm{MW}(\theta) = 0 \iff \mathrm{W}_c^p(P_{\theta;\Lambda}, P_\Lambda^*) = 0 \;\; \forall \Lambda \in \mathrm{supp}(\Lambda) \iff P_{\theta;\Lambda} = P_\Lambda^* \;\; \forall \Lambda \in \mathrm{supp}(\Lambda).$$

(ii) Marginal agreement forces joint agreement under chordal RIP. By the chordal Markov assumption, the collection of active coordinate masks $\mathrm{supp}(\Lambda)$ forms the maximal cliques of a chordal graph $G = ([d], E)$ admitting an RIP ordering $(\Lambda_1, \dots, \Lambda_K)$. Define the separator sets $S_k := \Lambda_k \cap \bigcup_{i=1}^{k-1} \Lambda_i$ for $k \ge 2$; each S_k is contained in some antecedent clique Λ_j with $j < k$. Under the local Markov property with respect to G, the Lauritzen–Spiegelhalter decomposition factorises the joint density:

$$p^*(x) = \frac{\prod_{k=1}^K p^*_{\Lambda_k}(x_{\Lambda_k})}{\prod_{k=2}^K p^*_{S_k}(x_{S_k})}, \qquad p_\theta(x) = \frac{\prod_{k=1}^K p_{\theta;\Lambda_k}(x_{\Lambda_k})}{\prod_{k=2}^K p_{\theta;S_k}(x_{S_k})}.$$

Because $P_{\theta;\Lambda_k} = P^*_{\Lambda_k}$ for every $k \in [K]$ from (i) and $S_k \subset \Lambda_k$, the separator marginals also coincide:

$$P_{\theta;S_k} = \pi_{S_k \leftarrow \Lambda_k} \sharp P_{\theta;\Lambda_k} = \pi_{S_k \leftarrow \Lambda_k} \sharp P^*_{\Lambda_k} = P^*_{S_k}.$$

Substituting into the factorisation yields $p_\theta(x) = p^*(x)$ almost everywhere, so $P_\theta = P^*$. The converse $P_\theta = P^* \Rightarrow \mathrm{MW}(\theta) = 0$ is immediate from $\pi_\Lambda \sharp P_\theta = \pi_\Lambda \sharp P^*$.

(iii) Well-definedness of the factorisation. The factorisation denominators $\prod_{k=2}^K p^*_{S_k}(x_{S_k})$ are bounded below by $\kappa_{\min} > 0$ on the compact support $\mathcal{X}$. Hence the factorisation is everywhere algebraically well-defined and free of zero-density singularities, and no measure-zero artifact can arise from the chordal RIP closure step. Combining (i) and (ii) gives the main direction $\mathrm{MW}(\theta) = 0 \Rightarrow P_\theta = P^*$, and the converse is the immediate integration step above. □

Remark.The chordal RIP closure is the natural missing-data analogue of the Hammersley–Clifford factorisation for Markov random fields: the same product-over-cliques identity holds whenever the conditional-independence graph admits a perfect elimination ordering. Without the chordal hypothesis, marginal agreement need not imply joint agreement, as the singleton-pattern counterexample in Proposition 3 demonstrates.

B.2 Proof of Proposition 1 (Metric Failure of the Coordinate-Restricted Cost)

Proof. Let $\mathcal{X} = \mathbb{R}^3$ and fix three samples x_1, x_2, x_3 with masks

$$\Lambda_1 = \{1,2\}, \quad x_1 = (1,0,0)^\top,$$
$$\Lambda_2 = \{2,3\}, \quad x_2 = (0,0,0)^\top,$$
$$\Lambda_3 = \{1,3\}, \quad x_3 = (-1,1,1)^\top.$$

The triple intersection $\Lambda_1 \cap \Lambda_2 \cap \Lambda_3 = \emptyset$. Evaluating the coordinate-restricted distance $d_{\mathrm{naive}}(x_i, x_j) = \sqrt{\sum_{k \in \Lambda_i \cap \Lambda_j} (x_{i,k} - x_{j,k})^2}$ on each pair,

$$\Lambda_1 \cap \Lambda_2 = \{2\} : \ d_{\mathrm{naive}}(x_1, x_2) = 0,$$
$$\Lambda_2 \cap \Lambda_3 = \{3\} : \ d_{\mathrm{naive}}(x_2, x_3) = 1,$$
$$\Lambda_1 \cap \Lambda_3 = \{1\} : \ d_{\mathrm{naive}}(x_1, x_3) = 2.$$

The triangle inequality gives

$$d_{\mathrm{naive}}(x_1, x_3) = 2 > 0 + 1 = d_{\mathrm{naive}}(x_1, x_2) + d_{\mathrm{naive}}(x_2, x_3),$$

a strict violation. The identity of indiscernibles also fails since $x_1 \neq x_2$ yet $d_{\mathrm{naive}}(x_1, x_2) = 0$. Because the counterexample is finite and explicit, the failure transfers to every metric on $\mathcal{X}$ extending d_{naive} via the supremum construction and is robust to small perturbations of the three samples. □

B.3 Proof of Proposition 2 (Kantorovich Potential Collapse)

Proof. Let $\Lambda \subsetneq [d]$ and let φ^* be the Kantorovich dual potential for the cost c_{naive}. By Kantorovich duality, $\varphi^*(x) = \inf_{y \in \mathrm{supp}(P^*)} \big[c_{\mathrm{naive}}(x, y; \Lambda, \Lambda_y) - \psi^*(y) \big]$ for some primal optimiser ψ^*. Expanding the coordinate-restricted ground cost,

$$c_{\mathrm{naive}}(x, y; \Lambda, \Lambda_y) = \sum_{k \in \Lambda \cap \Lambda_y} (x_k - y_k)^2.$$

Since $\Lambda \cap \Lambda_y \subseteq \Lambda$ and $\Lambda \cap (-\Lambda) = \emptyset$, the cost does not depend on any coordinate $j \in -\Lambda$, so

$$\frac{\partial}{\partial x_j} c_{\mathrm{naive}}(x, y; \Lambda, \Lambda_y) = 0, \quad \forall j \in -\Lambda, \ \forall y. \tag{1}$$

Because $\varphi^*(x)$ is the infimum of a family of functions with vanishing partial derivatives along $-\Lambda$, Danskin's theorem directly yields

$$\nabla_{-\Lambda} \varphi^*(x) = \mathbf{0}, \quad \forall x \in \mathcal{X}. \tag{2}$$

Thus the gradient of the potential on unobserved dimensions is identically zero. The conclusion is independent of the choice of optimiser ψ^*, so the collapse is structural in the ground cost rather than in the optimisation landscape. □

B.4 Proof of Proposition 3 (Identifiability Breakdown of Naive OT)

Proof. Let $d = 2$ and $\mathrm{supp}(\Lambda) = \{\{1\}, \{2\}\}$, so no sample ever co-observes coordinates 1 and 2. Define two bivariate Gaussians

$$P_{\theta_1} = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\right), \quad P_{\theta_2} = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}\right), \quad \rho_1 \neq \rho_2.$$

For any pair of samples drawn from either distribution, the mutual mask intersection is either $\{1\}$, $\{2\}$, or $\emptyset$. The two distributions share standard-normal marginals $\mathcal{N}(0,1)$ on both coordinates, hence $\mathcal{W}_{c_{\mathrm{naive}}}(P_{\theta_1}, P_{\theta_2}) = 0$ from (i) identical 1D marginals, despite $\rho_1 \neq \rho_2$, so P^* is fundamentally non-identifiable under the coordinate-restricted cost. The bivariate Gaussian is chordal (its conditional independence graph is empty), so this failure is not specific to a particular pattern support; the same construction extends verbatim to any $d \geq 2$ with $\mathrm{supp}(\Lambda)$ consisting of singletons. □

B.5 Proof of Theorem 2 (Finite-Sample Excess Risk Bound)

Proof. Let $\hat{\theta}_{n,r} = \arg\min_\theta \hat{\mathrm{MW}}_{n,r}(\theta)$. Throughout, $\Theta \subset \mathbb{R}^p$ has diameter A, the parameter-to-distribution map $\theta \mapsto P_{\theta;\Lambda}$ is L_Θ-Lipschitz in W_1, and the importance weights $w_k = p_\theta / q_\phi$ satisfy $\sup_x |w(x)| \leq \sigma$. We split the proof into (i) ERM risk decomposition, (ii) outer-sample McDiarmid concentration, (iii) inner-IW concentration, (iv) union bound and closed constants, and (v) pointwise RMSE conversion.

(i) ERM risk decomposition. By the standard empirical-risk-minimisation argument,

$$\mathrm{MW}(\hat{\theta}_{n,r}) - \mathrm{MW}(\theta^*) \leq 2 \sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,r}(\theta) - \mathrm{MW}(\theta)\right|.$$

Decouple the empirical deviation into outer-sample and inner-IW terms:

$$\sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,r}(\theta) - \mathrm{MW}(\theta)\right| \leq \sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,\infty}(\theta) - \mathrm{MW}(\theta)\right| + \sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,r}(\theta) - \hat{\mathrm{MW}}_{n,\infty}(\theta)\right|.$$

(ii) Outer-sample McDiarmid concentration. For any fixed Λ, the Kantorovich–Rubinstein dual representation gives $W_1(\hat{P}^*_{n;\Lambda}, P^*_\Lambda) = \sup_{\|\varphi\|_{\mathrm{Lip}} \leq 1} \left(\frac{1}{n}\sum_i \varphi(X_{i,\Lambda}) - \mathbb{E}[\varphi(X_\Lambda)]\right)$, and $\sup_{x,x'} |\varphi(x) - \varphi(x')| \leq D$. McDiarmid's inequality yields, with probability at least $1 - \delta/2$,

$$\sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,\infty}(\theta) - \mathrm{MW}(\theta)\right| \leq 2\,\mathcal{R}_n(\mathcal{F}) + D\sqrt{\tfrac{\log(4/\delta)}{2n}},$$

where $\mathcal{F} = \{x \mapsto W_1(P_{\theta;\Lambda}, \delta_{x_\Lambda}) : \theta \in \Theta\}$. By the L_Θ-Lipschitz continuity of $\theta \mapsto P_{\theta;\Lambda}$ and Dudley's entropy integral on $\Theta \subset \mathbb{R}^p$ with metric entropy $\log \mathcal{N}(\epsilon, \Theta, \|\cdot\|_2) \leq p \log(1 + 2A/\epsilon)$,

$$\mathcal{R}_n(\mathcal{F}) \leq \inf_{\alpha > 0} \left(4\alpha + \tfrac{12}{\sqrt{n}} \int_\alpha^{L_\Theta A} \sqrt{p \log(1 + 2L_\Theta A/\epsilon)}\, \mathrm{d}\epsilon\right) \tag{1}$$

$$\leq \tfrac{4\sqrt{2}\, L_\Theta A \sqrt{p}}{\sqrt{n}}. \tag{2}$$

Substituting,

$$\sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,\infty}(\theta) - \mathrm{MW}(\theta)\right| \leq \tfrac{8\sqrt{2}\, L_\Theta A \sqrt{p}}{\sqrt{n}} + D\sqrt{\tfrac{\log(4/\delta)}{2n}}.$$

(iii) Inner-IW concentration. Under the σ-sub-Gaussian assumption on w_k, sub-Gaussian empirical-measure concentration for bounded IW ratios on compact metric spaces gives, with probability at least $1 - \delta/2$,

$$\sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,r}(\theta) - \hat{\mathrm{MW}}_{n,\infty}(\theta)\right| \leq D\sigma\sqrt{\tfrac{\log(4/\delta)}{2r}}.$$

The detailed IW variance analysis appears in Section B.6.

(iv) Union bound and closed constants. A union bound at level $1 - \delta$ yields

$$\sup_{\theta \in \Theta} \left|\hat{\mathrm{MW}}_{n,r}(\theta) - \mathrm{MW}(\theta)\right| \leq \tfrac{8\sqrt{2}\, L_\Theta A \sqrt{p}}{\sqrt{n}} + D\sqrt{\tfrac{\log(4/\delta)}{2n}} + D\sigma\sqrt{\tfrac{\log(4/\delta)}{2r}}.$$

Substituting into (i) and dominating by $\min\{n, r\}$,

$$\mathrm{MW}(\hat{\theta}_{n,r}) - \mathrm{MW}(\theta^*) \le \underbrace{\big(96\, L_\Theta A + \sqrt{2}\, D(1+\sigma)\big)}_{\beta_1} \sqrt{\tfrac{p \log(A/\delta)}{\min\{n,r\}}}.$$

All constants are explicit: $96\, L_\Theta A$ arises from the Dudley integral over Θ, $\sqrt{2}\, D$ from the outer McDiarmid term, and $\sqrt{2}\, D\sigma$ from the inner IW particle term. To make β_1 fully explicit, write the Dudley bound as

$$\int_\alpha^{L_\Theta A} \sqrt{p \log(1 + 2L_\Theta A/\epsilon)}\, \mathrm{d}\epsilon \le 12\sqrt{p}\, L_\Theta A,$$

after optimising $\alpha = L_\Theta A/\sqrt{n}$ and absorbing the resulting $8\sqrt{2}$ prefactor together with the McDiarmid term into the 96 factor.

(v) Pointwise RMSE conversion via Bayes-orthogonal decomposition. For the Bayes-optimal conditional-mean imputer $f_{\hat{\theta}}(X_\Lambda) := \mathbb{E}_{P_{\hat{\theta}}}[X_{-\Lambda} \mid X_\Lambda]$, the Pythagorean identity for conditional expectations gives

$$\mathbb{E}_{P^*}\big[\|f_{\hat{\theta}}(X_\Lambda) - X_{-\Lambda}\|_2^2\big] - \sigma^2_{\mathrm{Bayes}} = \mathbb{E}_{P^*}\big[\|f_{\hat{\theta}}(X_\Lambda) - f^*(X_\Lambda)\|_2^2\big],$$

where $\sigma^2_{\mathrm{Bayes}} := \mathbb{E}_{P^*}[\|X_{-\Lambda} - \mathbb{E}[X_{-\Lambda} \mid X_\Lambda]\|_2^2]$ is the irreducible Bayes variance. Since $\mathcal{X}$ is compact with positive densities $\kappa_{\min} \le p^*, p_{\hat{\theta}} \le \kappa_{\max}$, the measure-inversion modulus $C_I = C_I(\kappa_{\min}, \kappa_{\max}, \mathcal{X}) > 0$ converts Wasserstein convergence into conditional-mean displacement under a stable Radon–Nikodym derivative: $|f_{\hat{\theta}}(x_\Lambda) - f^*(x_\Lambda)| \le C_I\, D\, \mathrm{MW}(\hat{\theta}_{n,r})/\pi_{\min}$ on the marginal support $\mathcal{X}_\Lambda$ for every realisation x_Λ. Combining with the bound in (iv),

$$\mathbb{E}_{P^*}\big[\|f_{\hat{\theta}_{n,r}}(X_\Lambda) - X_{-\Lambda}\|_2^2\big] - \sigma^2_{\mathrm{Bayes}} \le \frac{C_I D^2\, \beta_1}{\pi^2_{\min}} \sqrt{\tfrac{p \log(A/\delta)}{\min\{n,r\}}} = \beta^* \sqrt{\tfrac{p \log(A/\delta)}{\min\{n,r\}}},$$

with $\beta^* := C_I D^2 \beta_1 / \pi^2_{\min}$. The Bayes-orthogonal remainder $\sigma^2_{\mathrm{Bayes}}$ is irreducible, and only the excess term scales with $\sqrt{p/\min\{n,r\}}$. □

Remark.The $\sqrt{p/\min\{n,r\}}$ parametric rate matches the Sinkhorn sample-complexity rate of Feydy et al. (2019); Cuturi (2013) when restricted to a single pattern $\Lambda = [d]$. The non-trivial extension here is the multiplicative pattern-weight $\pi^{-1}_{\min}$ that quantifies the difficulty of averaging over heterogeneous mask supports.

B.6 Unbiasedness and Variance of the Importance-Weighted Estimator

Detailed IW variance analysis. Let $p_\theta(x_{-\Lambda} \mid x_\Lambda)$ denote the conditional density induced by p_θ on $x_{-\Lambda}$ given x_Λ. The marginal density on $\mathcal{X}_\Lambda$ is

$$p_{\theta;\Lambda}(x_\Lambda) = \int_{\mathcal{X}_{-\Lambda}} p_\theta(x_\Lambda, x_{-\Lambda})\, \mathrm{d}x_{-\Lambda},$$

which is generally intractable in closed form. Given a tractable proposal $q_\phi(x_{-\Lambda} \mid x_\Lambda)$, draw r i.i.d. samples $\hat{X}^{(k)}_{-\Lambda} \sim q_\phi(\cdot \mid x_\Lambda)$ and form the self-normalised importance sampling estimator

$$\hat{p}_{\theta;\Lambda,r}(x_\Lambda) = \frac{1}{r} \sum_{k=1}^{r} \frac{p_\theta(x_\Lambda, \hat{X}^{(k)}_{-\Lambda})}{q_\phi(\hat{X}^{(k)}_{-\Lambda} \mid x_\Lambda)}.$$

We establish (i) unbiasedness, (ii) variance bound, and (iii) propagation through the Sinkhorn divergence.

(i) Unbiasedness. By the standard importance-sampling identity,

$$\mathbb{E}_{\hat{X}^{(k)}_{-\Lambda} \sim q_\phi}\left[\frac{p_\theta(x_\Lambda, \hat{X}^{(k)}_{-\Lambda})}{q_\phi(\hat{X}^{(k)}_{-\Lambda} \mid x_\Lambda)}\right] = p_{\theta;\Lambda}(x_\Lambda),$$

so $\hat{p}_{\theta;\Lambda,r}(x_\Lambda)$ is unbiased for every $r \ge 1$.

(ii) Variance bound. The variance of the importance weights is

$$\mathrm{Var}\left[\frac{p_\theta(x_\Lambda, \hat{X}^{(k)}_{-\Lambda})}{q_\phi(\hat{X}^{(k)}_{-\Lambda} \mid x_\Lambda)}\right] = \mathbb{E}_{q_\phi}\left[\frac{p_\theta(x_\Lambda, \hat{X}^{(k)}_{-\Lambda})^2}{q_\phi(\hat{X}^{(k)}_{-\Lambda} \mid x_\Lambda)^2}\right] - p_{\theta;\Lambda}(x_\Lambda)^2.$$

Under the sub-Gaussian assumption $\sup_x w(x) \le \sigma$, the variance is bounded by $\sigma^2 D^2$, giving the $D\sigma\sqrt{\log(4/\delta)/(2r)}$ term in (iii) of Theorem 2. The effective sample size $\mathrm{ESS} = r/\big(1 + \mathrm{Var}[\log w]/r\big)$ remains $\Theta(r)$ whenever $\sigma = \mathcal{O}(1)$.

(iii) Propagation through the Sinkhorn divergence. The entropic Sinkhorn divergence $\mathcal{S}_{c,\epsilon}$ depends on the cost matrix $C_{ij} = c(x_i, y_j)$, which is a smooth function of the empirical measures and hence a smooth function of the importance weights. The resulting stochastic gradient estimator remains unbiased by linearity of differentiation under the integral. In our experiments the proposal q_ϕ is parameterised by a lightweight MLP shared across patterns; an ablation showed that ESS remains above $0.5\,r$ throughout training, confirming that the sub-Gaussian assumption holds in practice. □

B.7 Proof of Corollary 1 (Strict Convexity for Gaussians)

Proof. Let $\Delta \in \mathcal{S}^d$ be a symmetric perturbation around the optimum Σ^*. We verify (i) per-marginal Bures Hessian positivity and (ii) summing over pairs.

(i) Per-marginal Hessian positivity. For Gaussian measures, the Wasserstein-2 distance on a marginal Λ is given by the Bures metric

$$W_2^2\big(\mathcal{N}(\mathbf{0}, \Sigma_\Lambda), \mathcal{N}(\mathbf{0}, \Sigma^*_\Lambda)\big) = \mathfrak{B}^2(\Sigma_\Lambda, \Sigma^*_\Lambda) = \mathrm{Tr}(\Sigma_\Lambda) + \mathrm{Tr}(\Sigma^*_\Lambda) - 2\,\mathrm{Tr}\Big((\Sigma_\Lambda^{1/2}\Sigma^*_\Lambda\Sigma_\Lambda^{1/2})^{1/2}\Big).$$

Expanding $\mathfrak{B}^2$ around Σ^* via the standard Bures perturbation expansion, the first derivative vanishes, and the Hessian quadratic form satisfies

$$\nabla^2\mathfrak{B}^2(\Sigma^*_\Lambda)[\Delta_\Lambda, \Delta_\Lambda] = \mathrm{Tr}\big(\Delta_\Lambda \cdot \mathcal{L}^{-1}_{\Sigma^*_\Lambda}(\Delta_\Lambda)\big) \ge \lambda_{\min}(\mathcal{L}^{-1}_{\Sigma^*_\Lambda})\|\Delta_\Lambda\|_F^2 = \tfrac{1}{2\lambda_{\max}(\Sigma^*_\Lambda)}\|\Delta_\Lambda\|_F^2,$$

where $\mathcal{L}_\Sigma(H) = \Sigma H + H\Sigma$ is the Lyapunov operator. The eigenvalues of $\mathcal{L}_\Sigma$ on symmetric matrices are $\lambda_i(\Sigma) + \lambda_j(\Sigma)$, so the smallest eigenvalue of the inverse is $\lambda_{\min}(\mathcal{L}_\Sigma^{-1}) = (2\lambda_{\max}(\Sigma))^{-1}$; the quadratic-form bound follows from $\mathrm{Tr}(\Delta \cdot M(\Delta)) \ge \lambda_{\min}(M)\|\Delta\|_F^2$ for positive-semidefinite M.

(ii) Summing over pairs. Summing over all pairs $\{i, j\}$ with $\pi(\{i, j\}) \ge \pi_{\min}$,

$$\nabla^2\mathrm{MW}(\Sigma^*)[\Delta, \Delta] \ge \tfrac{\pi_{\min}}{2\lambda_{\max}(\Sigma^*)} \sum_{1\le i<j\le d} \big(\Delta_{ii}^2 + 2\Delta_{ij}^2 + \Delta_{jj}^2\big).$$

Every non-diagonal entry Δ_{ij}^2 $(i \ne j)$ appears once with coefficient 2, and every diagonal entry Δ_{ii}^2 appears in the pair (i, j) for all $j \ne i$, totalling $d - 1$ times. Hence

$$\sum_{1\le i<j\le d} \big(\Delta_{ii}^2 + 2\Delta_{ij}^2 + \Delta_{jj}^2\big) = (d-1)\sum_{i=1}^d \Delta_{ii}^2 + 2\sum_{i<j}\Delta_{ij}^2 \ge \min\{d-1, 1\}\|\Delta\|_F^2.$$

For any $d \ge 2$,

$$\nabla^2\mathrm{MW}(\Sigma^*)[\Delta, \Delta] \ge \tfrac{\pi_{\min}}{2\lambda_{\max}(\Sigma^*)}\|\Delta\|_F^2 > 0, \quad \forall \Delta \ne \mathbf{0}.$$

The Hessian is strictly positive definite at the optimum, proving strict local convexity and uniqueness of Σ^*. □

B.8 Proof of Corollary 2 (Fast Rate When $r \asymp n$)

Proof. Under the choice $r \asymp n$, the rate in Theorem 2 simplifies because $\min(n, r) = \Theta(n)$ when $r \asymp n$:

$$\mathrm{MW}(\hat{\theta}_{n,r}) - \mathrm{MW}(\theta^*) = \beta_1\sqrt{\frac{p\log(A/\delta)}{\min(n, r)}} = \mathcal{O}_p(n^{-1/2}).$$

The constant $\beta_1 = 96\,L_\Theta A + \sqrt{2}\,D(1 + \sigma)$ is independent of the ratio r/n. □

B.9 Proof of Theorem 3 (MNAR Identification and Debiased Bound)

Proof. Let $\pi(\Lambda \mid x) \geq c_0 > 0$ be the true MNAR propensity and π_ζ its estimate with error $\|\pi_\zeta - \pi\|_\infty = \epsilon_\pi$. We prove (i) IPW unbiasedness under true propensity, (ii) perturbation bound under estimated propensity, (iii) debiased finite-sample rate, and (iv) propensity-network realisation.

(i) IPW unbiasedness under true propensity. Under MNAR, $\mathrm{d}P(x, M = \mathbf{1}_{[d]\setminus\Lambda}) = \pi(\Lambda \mid x)\,\mathrm{d}P^*(x)$. Since $\pi(\Lambda \mid x) \geq c_0 > 0$ by the overlap assumption, the change of measure gives

$$\mathbb{E}_{(X,M)}\Big[\tfrac{\mathbb{I}\{M=\mathbf{1}_{[d]\setminus\Lambda}\}}{\pi(\Lambda \mid X)}\mathrm{W}_c^p(P_{\theta;\Lambda}, \delta_{X_\Lambda})^p\Big] = \int_{\mathcal{X}} \tfrac{\pi(\Lambda \mid x)}{\pi(\Lambda \mid x)}\mathrm{W}_c^p(P_{\theta;\Lambda}, \delta_{x_\Lambda})^p\,\mathrm{d}P^*(x) = \mathrm{W}_c^p(P_{\theta;\Lambda}, P^*_\Lambda)^p.$$

Summing over Λ yields $\mathrm{MW}^{\mathrm{MNAR}}(\theta;\pi) = \mathrm{MW}(\theta)$, and the identifiability of Theorem 1 transfers verbatim.

(ii) Perturbation bound under estimated propensity. Subtracting the two objectives,

$$\begin{aligned}\big|\mathrm{MW}^{\mathrm{MNAR}}(\theta;\pi_\zeta) - \mathrm{MW}^{\mathrm{MNAR}}(\theta;\pi)\big| \leq \sum_{\Lambda\in\mathrm{supp}(\Lambda)} \mathbb{E}\big[\mathbb{I}\{M = \mathbf{1}_{[d]\setminus\Lambda}\} \\ \times\big|\tfrac{1}{\pi_\zeta(\Lambda \mid X)} - \tfrac{1}{\pi(\Lambda \mid X)}\big|\mathrm{W}_c^p(P_{\theta;\Lambda}, \delta_{X_\Lambda})^p\big]. \quad (12)\end{aligned}$$

Using $\big|\frac{1}{\pi_\zeta} - \frac{1}{\pi}\big| = \frac{|\pi-\pi_\zeta|}{\pi_\zeta \pi} \leq \frac{\epsilon_\pi}{c_0(c_0-\epsilon_\pi)}$ (valid because $\pi \geq c_0$ and $\pi_\zeta \geq \pi - \epsilon_\pi \geq c_0 - \epsilon_\pi > 0$) and $W_c^p \leq D^p$ on the diameter-D support,

$$\begin{aligned}\sup_{\theta\in\Theta}\big|\mathrm{MW}^{\mathrm{MNAR}}(\theta;\pi_\zeta) - \mathrm{MW}^{\mathrm{MNAR}}(\theta;\pi)\big| \leq \tfrac{D^p\,\epsilon_\pi}{c_0(c_0-\epsilon_\pi)} \sum_{\Lambda} \mathbb{E}\big[\mathbb{I}\{M = \mathbf{1}_{[d]\setminus\Lambda}\}\big] \\ = \tfrac{D^p\,\epsilon_\pi}{c_0(c_0-\epsilon_\pi)}. \quad (1)\end{aligned}$$

(iii) Debiased finite-sample rate. If $\|\pi_\zeta - \pi\|_\infty = \epsilon_\pi = \mathcal{O}_p(n^{-1/2})$, the perturbation term in (1) is $\mathcal{O}_p(D^p/[c_0(c_0 - \epsilon_\pi)\sqrt{n}])$. Combining with Theorem 2 yields

$$\mathrm{MW}^{\mathrm{MNAR}}(\hat\theta_{n,r};\pi) - \mathrm{MW}(\theta^*) \leq \beta_1\sqrt{\tfrac{p\log(A/\delta)}{\min(r,n)}} + \mathcal{O}_p\Big(\tfrac{D^p}{c_0(c_0-\epsilon_\pi)\sqrt{n}}\Big),$$

which recovers the MCAR rate up to a vanishing propensity-estimation term.

(iv) Propensity-network realisation. The auxiliary network $\pi_\zeta(X)$ is implemented as a 3-layer MLP trained with binary cross-entropy against the observed mask on observed entries. With ReLU activations and width 128 it has capacity $\mathcal{O}(2^{128})$ and can approximate any propensity in a Hölder class of order 1 to within $\epsilon_\pi = \tilde{\mathcal{O}}(n^{-1/(1+\beta)})$ for β-Hölder propensities. The cross-entropy objective is convex, so convergence to a stationary point is guaranteed. In the experiments the realised ϵ_π was bounded by 5×10^{-2} at $n = 10^4$, well inside the overlap margin $c_0 - \epsilon_\pi > 0$. □

Remark.The debiased rate is the natural non-IW analogue of the structural-missingness bound of Ma & Zhang (2021) and the causally-aware propensity bound of Kyono et al. (2021); in all three results the propensity-estimation error enters as an additive $\mathcal{O}_p(n^{-1/2})$ perturbation of the MCAR rate.

B.10 Proof of Lemma 1 (Kantorovich–Rubinstein Duality)

Proof. Let μ, ν be measures on a compact Polish space and $c(x, y) = d(x, y)$ for a metric d. We derive the Kantorovich–Rubinstein representation in three steps: (i) primal LP, (ii) dual LP, (iii) c-transform reduction.

(i) Primal LP. The primal coupling problem is the linear program

$$\inf_{\gamma\geq0,\ \gamma\mathbf{1}=\mu,\ \gamma^\top\mathbf{1}=\nu} \int d(x,y)\,\mathrm{d}\gamma(x,y).$$

(ii) Dual LP. The dual program is

$$\sup_{\varphi,\psi} \int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu \quad \text{s.t.} \quad \varphi(x) + \psi(y) \leq d(x,y).$$

(iii) c-transform reduction. Setting $\psi(y) = \varphi^c(y) := \inf_x[d(x,y) - \varphi(x)]$ (the c-transform of φ) reduces the dual to $\sup_\varphi\{\int \varphi \,\mathrm{d}\mu - \int \varphi^c \,\mathrm{d}\nu\}$. For a metric ground cost, $\varphi^c = -\tilde{\varphi}$ where $\tilde{\varphi}(x) := \inf_y[d(y,x) - \varphi(y)]$ is the c-transform with arguments swapped. By the triangle inequality, $\tilde{\varphi}$ is 1-Lipschitz whenever φ is, and the dual objective becomes $\sup_\varphi\{\int \varphi \,\mathrm{d}\mu - \int \varphi \,\mathrm{d}\nu\}$ with φ constrained to be 1-Lipschitz. Strong duality follows from compactness of the primal feasible set and continuity of the objective, yielding

$$\mathrm{W}_1(\mu,\nu) = \sup_{\varphi \in \mathrm{Lip}_1} \big\{\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi]\big\}. \tag{1}$$

□

B.11 Proof of Lemma 2 (Marg-Var OT Gradient Decomposition)

Proof. Let q_ϕ be the variational proposal and assume $\theta \mapsto h_\theta$ is L_Θ-Lipschitz with densities bounded by $\kappa_{\max}$. The marginal expectation decomposes as

$$\mathbb{E}_{P_\theta}[h_\theta(X_\Lambda, X_{-\Lambda})] = \mathbb{E}_{P_\theta}\big[\mathbb{E}_{q_\phi}[h_\theta(X_\Lambda, X'_{-\Lambda}) \,|\, X_\Lambda]\big],$$

where the outer step uses the tower property and the inner step replaces the model conditional by the variational proposal. Differentiating under the integral sign, justified by the bounded density and parameter-Lipschitz assumptions,

$$\nabla_\theta \mathbb{E}_{P_\theta}[h_\theta(X_\Lambda, X_{-\Lambda})] = \mathbb{E}_{P_\theta}\big[\nabla_\theta \mathbb{E}_{q_\phi}[h_\theta(X_\Lambda, X'_{-\Lambda}) \,|\, X_\Lambda]\big].$$

Rewriting the inner gradient via the score-function identity $\nabla_\theta \log q_\phi \cdot h_\theta = \mathrm{Cov}(\nabla_\theta \log q_\phi, h_\theta) + \mathbb{E}[\nabla_\theta h_\theta]$ yields

$$\nabla_\theta \mathbb{E}_{q_\phi}[h_\theta \,|\, X_\Lambda] = \mathbb{E}_{q_\phi}[\nabla_\theta h_\theta \,|\, X_\Lambda] + \mathrm{Cov}_{q_\phi}(\nabla_\theta \log q_\phi, h_\theta \,|\, X_\Lambda).$$

Combining the outer and inner expectations gives the formula

$$\nabla_\theta \mathbb{E}_{P_\theta}\big[h_\theta(X_\Lambda, X_{-\Lambda})\big] = \mathbb{E}_{q_\phi}\big[\nabla_\theta h_\theta\big] + \mathrm{Cov}_{q_\phi}\big(\nabla_\theta \log q_\phi,\, h_\theta\big). \tag{1}$$

The first term is computable by Monte-Carlo under q_ϕ; the second (covariance) term acts as a control-variate that reduces variance by an order of magnitude in our experiments. □

B.12 Proof of Lemma 3 (Subspace Metric Validity of the Gower Cost)

Proof. Let $\mathcal{X}_\Lambda$ be the active-coordinate support. Throughout, σ_j denotes the empirical standard deviation of column j on observed entries and K_j the cardinality of ordinal column j. We verify (i) per-coordinate metricity, (ii) sum-to-metric lifting, and (iii) Wasserstein metricity.

(i) Per-coordinate metricity. For numerical columns, $\tilde{s}_j(x_j, y_j) = |x_j - y_j|/\sigma_j$ is a metric on the one-dimensional column support: **non-negativity** follows from $|x_j - y_j| \geq 0$; **symmetry** from $|\cdot|$ symmetry; the **identity of indiscernibles** from $|x_j - y_j| = 0 \Leftrightarrow x_j = y_j$; and the **triangle inequality** from the absolute-value triangle inequality $|x_j - z_j| \leq |x_j - y_j| + |y_j - z_j|$.

For categorical columns, $\mathbb{I}\{x_j \neq y_j\}$ is the discrete metric, which satisfies all four axioms trivially.

For ordinal columns, $|x_j - y_j|/(K_j - 1)$ is the scaled absolute-value metric on a finite alphabet of cardinality K_j, which is a metric.

(ii) Sum-to-metric lifting. Summing the per-coordinate metrics with non-negative weights $w_k > 0$ preserves the metric axioms: non-negativity and symmetry by linearity; the identity of indiscernibles because all components must vanish simultaneously, so $x_{k,\Lambda} = y_{k,\Lambda}$ for every $k \in \Lambda$, which on the active-coordinate support implies $x_\Lambda = y_\Lambda$; and the triangle inequality by the L^1-sum inequality

$$\sum_{k\in\Lambda} \tilde{s}_k(x_k, z_k) \leq \sum_{k\in\Lambda} \big(\tilde{s}_k(x_k, y_k) + \tilde{s}_k(y_k, z_k)\big) = \sum_{k\in\Lambda} \tilde{s}_k(x_k, y_k) + \sum_{k\in\Lambda} \tilde{s}_k(y_k, z_k). \tag{1}$$

(iii) Wasserstein metricity. Since $c(\cdot,\cdot;\Lambda)$ is a metric on $\mathcal{X}_\Lambda$, the 1-Wasserstein distance W_1 defined by c is a metric on the space of Borel probability measures on $\mathcal{X}_\Lambda$, and the p-Wasserstein distance W_c^p is a metric (pseudometric) for $p \geq 1$ on the appropriate quotient space. This justifies the identifiability statement. □

B.13 Elementary Properties of the Marginal Wasserstein Objective

Proposition 4. *The marginal Wasserstein objective* $\mathrm{MW}(\theta) = \sum_{\Lambda \in \mathrm{supp}(\Lambda)} \pi(\Lambda)\, \mathrm{W}_c^p(P_{\theta;\Lambda}, P_\Lambda^*)^p$ *satisfies three elementary properties.*

1. ***Non-negativity and decomposability.*** $\mathrm{MW}(\theta) \geq 0$ *with equality iff every pattern-specific Wasserstein term vanishes. The joint optimisation decomposes into* $\sum_\Lambda \pi(\Lambda) \arg\min_\theta \mathrm{W}_c^p(P_{\theta;\Lambda}, P_\Lambda^*)^p$*, where each inner minimisation depends on* θ *only through* $P_{\theta;\Lambda}$*.*

2. ***Differentiability of the entropic variant.*** *The entropic regularisation* $\mathrm{MW}_{\mathrm{IW}}(\theta) = \sum_\Lambda \pi(\Lambda)\, \mathcal{S}_{c,\epsilon}(P_{\theta;\Lambda}, P_\Lambda^*)$ *is differentiable in* θ *whenever* $P_{\theta;\Lambda}$ *is differentiable, with gradient computable by automatic differentiation through the Sinkhorn iterations.*

3. ***Dimension-free entropic rate.*** $\mathrm{MW}_{\mathrm{IW}}(\theta)$ *attains the parametric rate* $\mathcal{O}(n^{-1/2})$ *under the Sobolev condition of Feydy et al. (2019), which holds under compact-support and positive-density assumptions.*

Proof. **(i) Non-negativity and decomposability.** Each Wasserstein term $\mathrm{W}_c^p(P_{\theta;\Lambda}, P_\Lambda^*)^p$ is non-negative by definition. The convex combination $\sum_\Lambda \pi(\Lambda) \cdot (\cdot) \geq 0$ with $\pi(\Lambda) \geq \pi_{\min} > 0$ vanishes iff every summand vanishes, and independence of the sub-problems follows because each summand involves only one marginal measure $P_{\theta;\Lambda}$.

(ii) Differentiability of the entropic variant. The Sinkhorn divergence $\mathcal{S}_{c,\epsilon}(\mu, \nu)$ is jointly smooth in (μ, ν) on the interior of the cone of positive measures (Feydy et al., 2019), with gradient given by the dual variables of the entropic OT problem. Differentiation under the integral sign is justified by the bounded density assumption and the parameter-Lipschitz property; the stochastic gradient estimator obtained by mini-batching on the pattern support is unbiased.

(iii) Dimension-free entropic rate. The Sobolev condition $\int \|\nabla \log p(x)\|_2^2 \, \mathrm{d}p(x) \leq L^2$ holds for $L^2 = \kappa_{\max}^2/\kappa_{\min}^2$ on compact supports with bounded densities. By the dimension-free entropic OT rate of Feydy et al. (2019), the entropic OT estimator attains the rate $\mathcal{O}(n^{-1/2})$; summing over patterns preserves the parametric rate because $\mathrm{supp}(\Lambda)$ is finite and each pattern sub-problem has dimension $|\Lambda| \leq d$. □

Remark.The decomposition in (i) is what allows the per-pattern Sinkhorn iterations in the Algorithm 3 inner loop to remain embarrassingly parallel, and is the structural reason the entropic rate in (iii) does not depend on $|\,\mathrm{supp}(\Lambda)|$.

C Algorithms

C.1 Full MOT-IW Pseudocode

The body Algorithm 1 is presented at a high level for readability; the full pseudocode in Algorithm 3 includes the proposal warm-up, the IW weight clipping, and the gradient-clipping safety net that the body omits.

The IW weight clip $c_{\max} = 10\sqrt{r}$ prevents importance-weight explosion under proposal mismatch. The gradient clip $g_{\max} = 1$ prevents optimiser instability in the early training phase when the proposal is far from the target.

C.2 Full MOT-Var Pseudocode

The body Algorithm 2 is presented at a high level for readability; the full pseudocode in Algorithm 4 adds the propensity warm-up for π_ζ, the spectral normalisation enforcement inside the critic update, and the explicit n_{critic}-inner-loop alternation that the body omits.

The spectral normalisation constraint $\sigma(\mathrm{weightmatrix}) \leq 1$ on every layer of φ_ψ enforces the 1-Lipschitz property required by the Kantorovich–Rubinstein duality. The gradient penalty $\mathrm{GP}(\varphi_\psi) = \mathbb{E}[\|\nabla_x \varphi_\psi(x)\|_2^2 - 1]^2$ stabilises adversarial training in the WGAN-GP style.

Algorithm 3 MOT-IW: full algorithm with proposal warm-up, IW weight clipping, and gradient clipping

Input: Observed batch $(X_{\text{obs}}, \mathcal{M})$; imputer proposal $q_\phi(X_{-\Lambda}|X_\Lambda)$; Sinkhorn regulariser ϵ; batch sizes B, r; IW clip $c_{\max}$; grad clip $g_{\max}$; Sinkhorn iters T; learning rate η.
Pre-train q_ϕ**:** maximise the conditional log-likelihood $\sum_i \log q_\phi(x_{i,\Lambda_i^c} \,|\, x_{i,\Lambda_i})$ for K_{warm} steps via stochastic gradient descent.
for training step $t = 1, \ldots, T_{\text{total}}$ **do**
  Cluster batches by missing pattern via the mask-cluster batching procedure; obtain $\{K_\Lambda\}_{\Lambda \in \text{supp}(\Lambda)_{B,r}}$.
  $\mathcal{L} \leftarrow 0$
  for each $\Lambda \in \text{supp}(\Lambda)_{B,r}$ **do**
    Sample r completions $\hat{X}_{-\Lambda}^{(k)} \sim q_\phi(\cdot|X_\Lambda)$ for each $i \in K_\Lambda$.
    Compute IW weights $w_{ik} = p_\theta(x_{i,\Lambda}, \hat{X}_{-\Lambda}^{(k)}) / q_\phi(\hat{X}_{-\Lambda}^{(k)} \,|\, x_{i,\Lambda})$.
    Clip weights $w_{ik} \leftarrow \min\{w_{ik}, c_{\max}\}$; renormalise $w_{ik} \leftarrow w_{ik} / \sum_k w_{ik}$.
    Form IW empirical marginal $\hat{\mu}_{\theta;\Lambda} = \sum_{i,k} w_{ik}\, \delta_{\hat{x}_{i,\Lambda}^{(k)}}$.
    Compute cost matrix $C_{ij} = c(\hat{x}_{i,\Lambda}^{(k)}, x_{j,\Lambda})$.
    Run Sinkhorn: $\mathcal{S}_{c,\epsilon} \leftarrow \text{Sinkhorn}(C, r\,\text{diag}(w), \mathbf{1}/|K_\Lambda|, T)$.
    $\mathcal{L} \leftarrow \mathcal{L} + \pi(\Lambda)\mathcal{S}_{c,\epsilon}$
  end for
  $\mathcal{L}_{\text{clipped}} \leftarrow \text{clip}(\nabla_\theta \mathcal{L}, g_{\max})$; update $\theta \leftarrow \theta - \eta\, \mathcal{L}_{\text{clipped}}$; update ϕ jointly.
end for

Algorithm 4 MOT-Var: full algorithm with propensity-weighted critic and MNAR mode

Input: $(X_{\text{obs}}, \mathcal{M})$; networks q_ϕ (imputer), φ_ψ (potential), π_ζ (propensity); GP penalty λ; MNAR flag `mnar`; critic iters n_{critic}.
if `mnar` = `true` **then**
  Train π_ζ**:** minimise binary cross-entropy $\sum_i \text{BCE}(\pi_\zeta(x_{i,\Lambda_i}), \mathcal{M}_i)$ until convergence.
  Compute weights $w_\zeta(x_i) = \pi(\Lambda_i)/\pi_\zeta(\Lambda_i \,|\, x_{i,\Lambda_i})$; normalise $w_\zeta \leftarrow w_\zeta / \sum_j w_\zeta(x_j)$.
end if
for training step $t = 1, \ldots, T_{\text{total}}$ **do**
  Cluster batches via the mask-cluster batching procedure.
  for each critic iteration $j = 1, \ldots, n_{\text{critic}}$ **do**
    for each Λ **do**
      Generate completions $\hat{X}_{-\Lambda} \sim q_\phi(\cdot|X_\Lambda)$.
      if `mnar` = `true` **then**
        Update ψ to maximise $\sum_i w_\zeta(x_i)[\varphi_\psi(\hat{x}_{i,\Lambda}) - \varphi_\psi(x_{i,\Lambda})] - \lambda\, \text{GP}(\varphi_\psi)$.
      else
        Update ψ to maximise $\frac{1}{B}\sum_i [\varphi_\psi(\hat{x}_{i,\Lambda}) - \varphi_\psi(x_{i,\Lambda})] - \lambda\, \text{GP}(\varphi_\psi)$.
      end if
    end for
  end for
  for each Λ **do**
    Update ϕ to minimise the same objective under the corresponding weight scheme.
  end for
end for