GROMOV-WASSERSTEIN INFORMATION BOTTLENECK FOR DISTANCE-PRESERVING REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Preserving useful data geometry is a fundamental challenge in representation learning for generative and predictive tasks. However, task objectives and marginal distribution matching do not explicitly constrain how an encoder transforms relationships among samples. To provide a principled basis for geometric regularization, we revisit the information bottleneck through the lens of optimal transport, connecting representation complexity with encoder-induced geometric distortion. This perspective motivates the Gromov-Wasserstein Information Bottleneck (GWIB). In principle, we derive an upper bound on an empirical kernelized information quantity in terms of the Gromov-Monge gap. A reference-distribution relaxation further bounds this gap using encoder distortion and fused Gromov-Wasserstein alignment. To evaluate GWIB in generative and predictive settings, we apply it to Wasserstein autoencoders and counterfactual regression. In Wasserstein autoencoders, a fixed prior serves as the reference, coupling geometric regularization with aggregate-posterior matching. In counterfactual regression, mutually learned group references combine within-group geometric regularization with cross-group alignment through a shared transport plan. Experiments demonstrate improved image reconstruction and generation on CelebA dataset and reduced treatment-effect estimation errors on IHDP and ACIC datasets.

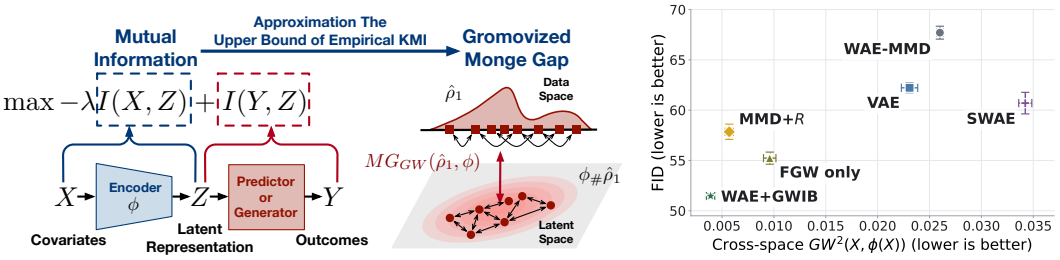
1 INTRODUCTION

Learning effective representations is crucial for many machine learning tasks, including generative modeling (Kingma & Welling, 2013; Tolstikhin et al., 2017) and predictive regression (Bengio et al., 2013; Shalit et al., 2017). A useful representation should preserve the data geometry relevant to the task, such as neighborhoods or pairwise similarities that reflect meaningful variation among observations. The challenge is to retain such useful structure while allowing the encoder to discard variation that the task does not require.

Reconstruction or prediction losses, together with marginal distribution matching, leave freedom in how a representation model preserves the data geometry in the latent space. For example, accurate reconstruction can coexist with distorted latent neighborhoods because the decoder can compensate for the encoder's transformation (Tolstikhin et al., 2017). Similarly, matching two latent distributions does not ensure that their aligned samples have corresponding local relationships (Assaad et al., 2021). Existing methods often restrict this freedom through regularizing latent distribution (Xu et al., 2020), underlying topology (Moor et al., 2020), decoder geometry (Lee et al., 2022; Nazari et al., 2023), or transport optimality (Uscidda & Cuturi, 2023; Uscidda et al., 2025), but the underlying rationale for the regularization remains unverified.

The information bottleneck (IB) (Tishby et al., 2000) provides a principle for balancing task utility and representation complexity, but does not itself prescribe how to regularize data geometry. We revisit IB through the lens of optimal transport (Peyré et al., 2019) to examine how representation complexity relates to the geometric distortion induced by an encoder. This perspective motivates the **Gromov-Wasserstein Information Bottleneck (GWIB)**, a distance-preserving representation learning method based on geometric regularization. By combining this regularization with the task objective, GWIB encourages the encoder to preserve relational geometry while learning task-

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.



(a) The Scheme of GWIB (b) GW distance vs. FID on CelebA

Figure 1: (a) The proposed regularization derives from an upper bound on empirical KMI through the Gromov–Monge gap. (b) We test WAE and its variants on CelebA face generation. WAE+GWIB obtains a smaller cross-space $GW^2(X, \phi(X))$, indicating a more distance-preserving latent representation, together with the higher model performance (a lower FID). This supports GWIB as a promising regularizer for WAE training. All GW values use a common fixed 1,024-image protocol. Error bars show three-seed sample SD; lower left is better.

relevant representations. The regularizer further extends to cross-domain distribution alignment, where representations from different domains are jointly learned through within-domain geometric regularization and cross-domain alignment.

In principle, we derive an upper bound on an empirical kernelized information quantity in terms of the Gromov-Monge gap (MG), which measures the encoder's excess structural cost relative to an optimal Gromov-Wasserstein coupling between the same input and encoded distributions. The scheme of GWIB is illustrated in Figure 1(a). We further derive a reference-distribution relaxation that upper-bounds this gap using encoder distortion and fused Gromov-Wasserstein alignment. This relaxation accommodates a fixed reference or mutually learned references, yielding geometric regularization for a single representation distribution or coupled alignment of multiple distributions with a shared transport plan. To assess this rationale, we first test WAE and its variants on CelebA face generation (Liu et al., 2015). Figure 1(b) shows that WAE+GWIB produces a latent representation with a smaller cross-space $GW^2(X, \phi(X))$ to the data distribution, while also achieving a lower FID. Thus, GWIB provides a promising WAE regularizer: distance-preserving latent representations can improve generative performance. We further evaluate the method in treatment-effect estimation (Shalit et al., 2017), which is a predictive application that couples within-group geometry with treatment–control alignment and reduces treatment-effect estimation errors on IHDP and ACIC datasets.

2 Related Work

Representation learning for generative and predictive tasks. Representation learning typically combines reconstruction or prediction objectives with latent regularization (Bengio et al., 2013). Topological autoencoders preserve input topology (Moor et al., 2020), while relationally regularized autoencoders align latent posteriors with learnable structured priors using fused Gromov–Wasserstein (FGW) discrepancies (Xu et al., 2020). Optimal transport (OT) methods regularize learned maps through the Monge gap and its extension, penalizing excess costs over optimal couplings under pointwise and relational objectives, respectively (Uscidda & Cuturi, 2023; Uscidda et al., 2025). For generation, variational autoencoders regularize approximate posteriors toward a prior (Kingma & Welling, 2013), while Wasserstein autoencoders combine reconstruction with aggregate-posterior matching (Tolstikhin et al., 2017). For treatment-effect estimation, representations support prediction while addressing treatment-group differences (Johansson et al., 2016). Counterfactual regression combines factual prediction with latent distribution balancing (Shalit et al., 2017), with extensions using reweighting (Assaad et al., 2021) and relaxed OT constraints (Wang et al., 2023).

Information bottleneck and optimal transport. The information bottleneck (IB) provides a theoretical framework for balancing predictive information and representation complexity (Tishby et al., 2000). Deep variational IB enables tractable learning for neural encoders through variational approximations (Alemi et al., 2016), with causal extensions for treatment-effect prediction (Kim

et al., 2019; Lu et al., 2022). Connecting information objectives with OT, Chen et al. (2023a) formulate IB as entropy-regularized OT, while Ozair et al. (2019) measure dependence using the Wasserstein distance between a joint distribution and the product of its marginals. InfoOT learns transport couplings by maximizing kernelized mutual information (Chuang et al., 2023). Instead, our GWIB links IB to encoder geometry through an empirical kernelized information bound involving the Gromov–Monge gap, providing geometric regularization for generative and predictive tasks.

3 Gromov-Wasserstein Information Bottleneck

3.1 Motivation

We consider an input space $(\mathcal{X}, d_{\mathcal{X}})$ and a latent representation space $(\mathcal{Z}, d_{\mathcal{Z}})$. An encoder $\phi : \mathcal{X} \mapsto \mathcal{Z}$ maps each input $x \in \mathcal{X}$ to a latent representation $z = \phi(x) \in \mathcal{Z}$, inducing a pushforward distribution $\phi_{\#}\rho$ on $\mathcal{Z}$ for $X \sim \rho$. For prediction tasks, we additionally consider an output variable Y and treat (X, Z, Y) as random variables linked through the encoder $Z = \phi(X)$. A predictor $h : \mathcal{Z} \mapsto \mathcal{Y}$ then operates on the latent representation Z to produce task-relevant outputs. Here, the encoder ϕ plays a central role: it determines what information from X is retained in Z and how the geometry of $\mathcal{X}$ is reflected in the latent space $\mathcal{Z}$. Our objective is therefore to impose appropriate constraints on ϕ so that the learned representation Z is not only predictive for the task, but also preserves meaningful structural relationships present in the input space.

From the viewpoint of the information bottleneck (IB) principle (Tishby et al., 2000), a good representation should preserve the information in X that is predictive of Y, while discarding irrelevant details in X. This trade-off is captured by the IB objective:

$$\max_{h,\phi} \underbrace{I(Z, Y; h, \phi)}_{\text{predictive term}} - \lambda \underbrace{I(Z, X; \phi)}_{\text{complexity term}}, \tag{1}$$

Here, $I(A, B; \theta) = \mathrm{KL}(p(A, B) \| p(A)p(B))$ denotes the mutual information between the random variables A and B, which can be parameterized by θ. Obviously, the first term aims to pursue the prediction power of the model, and the second term regularizes the complexity of the representation, i.e., the latent representation Z should only preserve necessary information from X for the prediction task and ignore the useless information.

For regression with a fixed-variance Gaussian conditional model, the empirical log-likelihood contribution to the predictive term in (1) is an affine function of the squared loss $\mathcal{L}(h, \phi) = \frac{1}{N}\sum_{n=1}^{N} |h(\phi(x_n)) - y_n|^2$. Specifically, under this conditional model, $p(Y \mid Z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{1}{2\sigma^2}|h(Z) - Y|^2\big)$. Based on the data $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$, we have

$$\begin{aligned} I(Z, Y; h, \phi) &= \mathbb{E}_{z,y\sim p(Z,Y)}[\log p(Y = y|Z = z) - \log p_Y(y)] \\ &\approx \underbrace{\frac{1}{N}\sum_{n=1}^{N}\big[\log p(y_n|z_n) - \log p_Y(y_n)\big]}_{\text{Empirical mutual information: } \hat{I}_N(Z,Y;h,\phi)} = -\frac{1}{2N\sigma^2}\sum_{n=1}^{N} |h(\underbrace{\phi(x_n)}_{z_n}) - y_n|^2 + C, \end{aligned} \tag{2}$$

where $C = -\log(\sqrt{2\pi}\sigma) - \frac{1}{N}\sum_{n=1}^{N} \log p_Y(y_n)$ is a constant irrelevant to the model.

Challenge. Task losses and and latent-distribution matching do not guarantee preservation of pairwise distances. As shown in Figure 1(b), even WAE-MMD and SWAE, which match encoded distributions to a prior, exhibit substantially higher pairwise-distance distortion and structural discrepancy than WAE+GWIB on CelebA. In addition, an information penalty also requires a concrete estimator and a geometric interpretation before it can guide this aspect of representation learning. We therefore examine a kernelized quantity and its relation to structural distortion.

3.2 An empirical kernelized information bound

While the complexity term $I(Z, X; \phi)$ is desirable, it is generally difficult to estimate and provides no direct control on pairwise distances. We study the following finite-sample kernelized information score; the bound below concerns this specified quantity, not general KL mutual information. Given

N covariates $\{x_n\}_{n=1}^N$, we can obtain the corresponding latent representations $\{z_n = \phi(x_n)\}_{n=1}^N$ and apply kernel density estimation to obtain their distributions (Chuang et al., 2023), i.e.,

$$\rho(X) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \kappa(X, x_n), \quad \phi_{\#}\rho(Z) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \kappa(Z, z_n),$$

where $\kappa(a,b) = \frac{1}{\sqrt{2\pi}\tau}\exp(-\frac{1}{2\tau^2}d^2(a,b))$ is the commonly-used RBF kernel with bandwidth τ and d is a valid metric. Accordingly, the joint distribution of X and Z can be estimated based on the observed covariates and their corresponding latent representations, i.e.,

$$p(X, Z) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \kappa(X, x_n)\kappa(Z, z_n). \tag{3}$$

Therefore, we can obtain the empirical kernelized mutual information between X and Z as

$$\hat{I}_{\kappa,N}(Z, X; \phi) = \frac{1}{N}\sum\nolimits_{n} \log \frac{p(x_n, z_n)}{\rho(x_n)\phi_{\#}\rho(z_n)} = \frac{1}{N}\sum\nolimits_{n} \log \frac{N\sum_m \kappa(x_n, x_m)\kappa(z_n, z_m)}{\sum_m \kappa(x_n, x_m)\sum_m \kappa(z_n, z_m)}. \tag{4}$$

However, directly applying $\hat{I}_{\kappa,N}(Z, X; \phi)$ as the regularizer is questionable in general because the kernel density estimation of $p(X, Z)$ often suffers from the curse of dimensionality. In practice, the covariates can be high-dimensional variables, so that the observed data $\{x_n, \phi(x_n)\}$ are too sparse to estimate $p(X, Z)$ with high accuracy. Specifically, when only $\{x_n, \phi(x_n)\}_{n=1}^N$ are available, the coherency of x_n and $z \in \mathcal{Z} \setminus \{\phi(x_n)\}$ and that of $x \in \mathcal{X} \setminus \{x_n\}$ and $\phi(x_n)$ are ignored. As a result, the entropy of (X, Z) is underestimated, and $\hat{I}_{\kappa,N}(Z, X; \phi)$ in (4) becomes unreliable, which prevents the utilization of the information bottleneck paradigm.

A potential solution to mitigate this problem is further optimizing the joint distribution $p(X, Z)$ to maximize $\hat{I}_{\kappa,N}(Z, X; \phi)$ and then penalizing the upper bound of $\hat{I}_{\kappa,N}(Z, X; \phi)$ in the information bottleneck paradigm. Inspired by this idea, we derive an upper bound of $\hat{I}_{\kappa,N}(Z, X; \phi)$, and accordingly, design a surrogate penalty term for (1), which leads to the proposed Gromov-Wasserstein information bottleneck learning paradigm.

Our method is based on the following theorem:

Theorem 1. *Assume that $(\mathcal{X}, d_X)$ and $(\mathcal{Z}, d_Z)$ are two bounded spaces, whose diameters are denoted as $\mathrm{Diam}_{\mathcal{X}}$ and $\mathrm{Diam}_{\mathcal{Z}}$, respectively. Given samples $\boldsymbol{X} = \{x_n\}_{n=1}^N$ and corresponding $\boldsymbol{Z} = \{z_n\}_{n=1}^N$, for the empirical kernelized mutual information defined in* (4) *and using an RBF kernel with bandwidth τ, we have*

$$\hat{I}_{\kappa,N}(Z, X) \le \frac{1}{2\tau^2}\Big(\frac{1}{N^2}\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2 - \widehat{GW}_2^2(\boldsymbol{X}, \boldsymbol{Z})\Big) + C_{\kappa,N}, \tag{5}$$

where $\boldsymbol{D}_X = [d_X(x_m, x_n)] \in \mathbb{R}^{N\times N}$ and $\boldsymbol{D}_Z = [d_Z(z_m, z_n)] \in \mathbb{R}^{N\times N}$ are distance matrices,

$$\widehat{GW}_2(\boldsymbol{X}, \boldsymbol{Z}) = \min\nolimits_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)} \langle \boldsymbol{C}_{XZ} - 2\boldsymbol{D}_X\boldsymbol{T}\boldsymbol{D}_Z^\top,\ \boldsymbol{T}\rangle^{\frac{1}{2}}. \tag{6}$$

is the order-2 Gromov-Wasserstein discrepancy[1] *between $\boldsymbol{X}$ and $\boldsymbol{Z}$, where $\boldsymbol{C}_{XZ} = \frac{1}{N}((\boldsymbol{D}_X \odot \boldsymbol{D}_X)\mathbf{1}_{N\times N} + \mathbf{1}_{N\times N}(\boldsymbol{D}_Z \odot \boldsymbol{D}_Z)^\top)$. $C_{\kappa,N} = \frac{N^4(2\pi\tau^2-\alpha)^2}{8(N^4-1)\alpha^2}$, where $\alpha = \exp(-\frac{\mathrm{Diam}_{\mathcal{X}}^2+\mathrm{Diam}_{\mathcal{Z}}^2}{2\tau^2})$.*

Sketched Proof. For the empirical kernelized mutual information $\hat{I}_{\kappa,N}(Z, X)$, we first apply Jensen's inequality to the term $-\frac{1}{N}\sum_n \log \rho(x_n)\phi_{\#}\rho(z_n)$, leading to the first term of the upper bound. Optimizing the joint coupling as in (Chuang et al., 2023) and applying a Jensen-gap estimate yield the GW term and the additive constant $C_{\kappa,N}$, which bounds the logarithmic Jensen gap on the stated bounded domain (Costarelli & Spigler, 2015). A detailed proof is provided in Appendix D.1.

3.3 Connections to Monge map approximation

In theory, the upper bound in (5) corresponds the Gromovization of the Monge gap in (Uscidda & Cuturi, 2023).

[1]The formulation in (6) is based on the Proposition 1 in Peyré et al. (2016), which is applicable when applying the RBF kernel in the derivation of (5).

Definition 1 (Gromov-Monge gap)**.** Let $(\mathcal{X}, d_X)$ and $(\mathcal{Z}, d_Z)$ be two compact metric spaces, respectively. ρ is a probability measure defined on $(\mathcal{X}, d_X)$, and $\phi : \mathcal{X} \mapsto \mathcal{Z}$ is a mapping. The order-p Gromov-Monge gap associated with ϕ and ρ, denoted as $MG(\rho, \phi)$, is defined as

$$\mathbb{E}_{x,x' \sim \rho^2}[|d_X(x, x') - d_Z(\phi(x), \phi(x'))|^p] - GW_p^p(\rho, \phi_{\#}\rho),$$

where the first term measures the transport cost of pushing ρ to $\phi_{\#}\rho$ through ϕ, and the second term $GW_p(\rho, \phi_{\#}\rho)$ is the order-p GW distance between ρ and $\phi_{\#}\rho$, i.e., $\inf_{\pi \in \Pi(\rho, \phi_{\#}\rho)} (\int_{\mathcal{X}^2 \times \mathcal{Z}^2} |d_X(x, x') - d_Z(z, z')|^p \mathrm{d}\pi(x, z) \mathrm{d}\pi(x', z'))^{\frac{1}{p}}$.

The Gromov-Monge gap replaces the Wasserstein distance used in the original Monge gap (Uscidda & Cuturi, 2023) with the GW distance, making this concept applicable for the Monge map[2] across different spaces. It measures the discrepancy between the transport cost of pushing ρ to $\phi_{\#}\rho$ and the GW distance (Mémoli, 2011) between ρ and $\phi_{\#}\rho$. According to the definition, we can rewrite (5) as

$$\hat{I}_{\kappa,N}(Z, X; \phi) = \mathcal{O}(MG(\hat{\rho}_N, \phi)) + \mathcal{O}(1), \tag{7}$$

where $\hat{\rho}_N$ denotes the empirical distribution of the observed covariates $\{x_n\}_{n=1}^N$. When $N \to \infty$, we have $\widehat{GW}_2(\boldsymbol{X}, \boldsymbol{Z}) \to GW_2(\rho, \phi_{\#}\rho)$, $\frac{1}{N^2}\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2 \to \mathbb{E}_{x,x' \sim \rho \times \rho}[|d_X(x, x') - d_Z(\phi(x), \phi(x'))|^2]$, and $C_{\kappa,N} \to \frac{(2\pi\tau^2 - \alpha)^2}{8\alpha^2}$. Thus the right-hand side has the corresponding population geometric term plus the limiting additive constant, under the stated bounded-space conditions. The constant depends on the kernel and diameter bounds and cannot be discarded to infer a vanishing information score from a vanishing gap.

Similar to (Uscidda & Cuturi, 2023), we can prove that

Theorem 2. *$MG(\rho, \phi) \geq 0$, with equality if and only if the encoder-induced coupling $(\mathrm{id}, \phi)_{\#}\rho$ is GW-optimal, i.e., ϕ is a GW-optimal Monge map.*

For fixed marginals and fixed diameter bounds, a GW-optimal encoder-induced coupling minimizes the gap-dependent part of the bound. The proof of Theorem 2 is in Appendix D.

3.4 A REFERENCE-DISTRIBUTION RELAXATION

Let $q = \phi_{\#}\rho$ and let ν be a reference in the same latent metric space. Use the order-2 GW and gap formulas on complete separable metric spaces, with measurable ϕ and finite second moments for ρ, q, ν. Write $R^2 = \mathbb{E}_{\rho \otimes \rho}|d_X(x, x') - d_Z(\phi(x), \phi(x'))|^2$, so empirically $R = \|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F / N$. For $0 < \beta < 1$, define the fused cost and reference bound

$$\begin{aligned} F_\beta^2(q, \nu) &= \inf_{\pi \in \Pi(q,\nu)} \big[(1 - \beta)\mathcal{C}_W(\pi) + \beta \mathcal{C}_{GW}(\pi)\big], \\ \mathcal{U}_\beta(\rho, \phi; \nu) &= \left(R + \frac{F_\beta(q, \nu)}{\beta}\right)^2 - GW_2^2(\rho, \nu). \end{aligned} \tag{8}$$

where $F_\beta^2(q, \nu)$ is the fused Gromov-Wasserstein distance (Titouan et al., 2019), and $\mathcal{C}_{GW}$ are the squared feature and structural costs under the same coupling and metric d_Z (Appendix D.3).

Proposition 1 (Reference-distribution upper bound)**.** *Under these conditions,*

$$MG(\rho, \phi) \leq \mathcal{U}_\beta(\rho, \phi; \nu). \tag{9}$$

The fused term combines feature matching to the reference with relational alignment. Appendix D.3 gives the proof. In particular, a fixed prior gives the WAE application, and a learned group references give Coupled GWIB for CFR. We detail these two applications in the following section.

4 APPLICATIONS

4.1 GWIB FOR WASSERSTEIN AUTOENCODERS

Wasserstein autoencoders (WAEs) learn a decoder g and encoder ϕ by minimizing reconstruction cost while matching the aggregate posterior $q = \phi_{\#}\rho$ to the prior p_Z, so that $g_{\#}p_Z$ models the data

[2]As proved by Uscidda & Cuturi (2023), optimizing the Monge gap for ϕ helps approximate ϕ as a Monge map, which leads to the minimized transport cost from ρ to $\phi_{\#}\rho$.

distribution (Tolstikhin et al., 2017). For fixed g, Theorem 1 of Tolstikhin et al. (2017) gives

$$W_c(\rho, g_{\#}p_Z) = \inf_{q:\, q_Z = p_Z} \mathbb{E}_{x\sim\rho}\mathbb{E}_{z\sim q(\cdot|x)} c(x, g(z)), \tag{10}$$

where $q_Z = \int q(\cdot \mid x)\, \mathrm{d}\rho(x)$ and the infimum ranges over stochastic encoders satisfying the aggregate constraint. For squared Euclidean cost, $W_c = W_{2,X}^2$ in image space; this equality does not identify an arbitrary deterministic encoder's reconstruction loss with that distance. A standard finite-penalty objective for deterministic ϕ is

$$\min_{\phi,g} \underbrace{\mathbb{E}_{x\sim\rho} c(x, g(\phi(x)))}_{\mathcal{L}_{\mathrm{rec}}(\phi,g)} + \lambda_{\mathrm{WAE}} D_Z(q, p_Z), \qquad \lambda_{\mathrm{WAE}} > 0. \tag{11}$$

This objective generally differs from exact OT; our baseline uses squared IMQ-MMD for D_Z.

GWIB replaces this MMD penalty with the reference bound. With $\nu = p_Z$ and fixed ρ, d_X, p_Z, d_Z, the subtracted $GW_2^2(\rho, p_Z)$ in $\mathcal{U}_\beta$ is independent of (ϕ, g):

$$\min_{\phi,g} \mathcal{L}_{\mathrm{rec}}(\phi, g) + \lambda_{\mathrm{GWIB}} \mathcal{U}_\beta(\rho, \phi; p_Z), \qquad \lambda_{\mathrm{GWIB}} > 0. \tag{12}$$

Here $\mathcal{U}_\beta$ couples input–code distortion R with FGW alignment to the prior, rather than acting as a divergence of the two marginals alone.

4.2 Coupled GWIB for Counterfactual Regression

Treatment-effect estimation asks how an intervention changes the expected outcome for a population with covariates x. Each individual has an observed treatment $t \in \{0, 1\}$ and factual outcome y, leaving the other potential outcome unobserved. Under the identification assumptions in Appendix I, CFR (Shalit et al., 2017) estimates the conditional mean difference $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ using a shared encoder ϕ and two outcome heads, $\hat{\tau}(x) = h_1(\phi(x)) - h_0(\phi(x))$. Nonrandom treatment assignment can create group differences, leading to selection bias, which then motivates representation balancing alongside within-group geometry. C-GWIB applies two coupled groupwise regularizers with this shared encoder.

For continuous outcomes, squared loss fits the conditional means. For groupwise observations $\mathcal{D}_t = \{(x_{n,t}, y_{n,t})\}_{n=1}^{N_t}$, classical CFR averages factual squared errors and adds a discrepancy D between the latent distributions:

$$\min_{\phi,h_0,h_1} \underbrace{\sum\nolimits_{t=0,1} \frac{1}{N_t} \sum\nolimits_{n=1}^{N_t} |h_t(\phi(x_{n,t})) - y_{n,t}|^2}_{\mathcal{L}(h_0,h_1,\phi)} + \lambda D(\boldsymbol{Z}_0, \boldsymbol{Z}_1). \tag{13}$$

The fixed-variance Gaussian interpretation in (2) connects the predictive part of (1) to the factual loss $\mathcal{L}$. The groupwise bound (5) motivates $MG(\hat{\rho}_t, \phi)$ penalties, with fixed scale absorbed into λ_{GWIB}. Using each learned latent distribution as the other's reference also supplies cross-group alignment.

Coupled relaxation. Let $q_t = \phi_{\#}\hat{\rho}_t = N_t^{-1} \sum_{n=1}^{N_t} \delta_{\phi(x_{n,t})}$ be the empirical latent probability measure, with $\boldsymbol{a}_t = \mathbf{1}_{N_t}/N_t$ and $R_t = \|\boldsymbol{D}_{X_t} - \boldsymbol{D}_{Z_t}\|_F / N_t$. Applying Proposition 1 to $(\hat{\rho}_t, q_t, q_{1-t})$ and summing gives the exact-distance regularizer

$$\begin{aligned} \sum\nolimits_{t=0,1} MG(\hat{\rho}_t, \phi) \le \mathcal{R}_{\mathrm{CFR}}(\phi) &:= \sum\nolimits_{t=0,1} \mathcal{U}_\beta(\hat{\rho}_t, \phi; q_{1-t}) \\ &= \sum\nolimits_{t=0,1} \Big[\Big(R_t + \frac{F_\beta(q_t, q_{1-t})}{\beta} \Big)^2 - GW_2^2(\hat{\rho}_t, q_{1-t}) \Big]. \end{aligned} \tag{14}$$

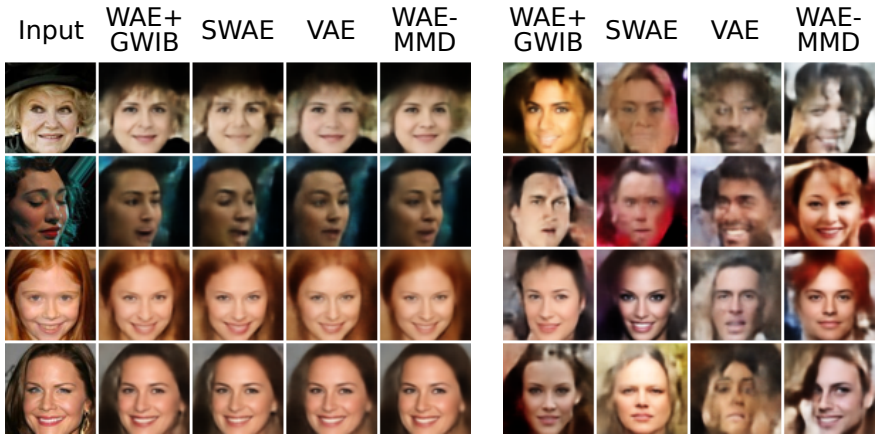
For $0 < \beta < 1$, the fused feature cost encourages cross-group alignment alongside the structural term. The negative GW terms depend on the other group's learned representation and must be retained. At $\beta = 1$, the feature term vanishes while the structural bound remains valid. For empirical groups, $F_\beta(q_0, q_1) = \widehat{FGW}_{2,\beta}(\boldsymbol{Z}_0, \boldsymbol{Z}_1)$; Appendix I.1 gives its matrix form.

Shared correspondence and learning. Each row of $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ describes the same individual. We therefore use one cross-group plan $\boldsymbol{T} \in \Pi(\boldsymbol{a}_0, \boldsymbol{a}_1)$ for the $0 \to 1$ comparisons and its transpose for $1 \to 0$. Let $\boldsymbol{T}_0 = \boldsymbol{T}$, $\boldsymbol{T}_1 = \boldsymbol{T}^\top$, and let $F_\beta(\boldsymbol{T})$ denote the square root of the bracketed cost in Equation 32. The training regularizer and objective are

$$\mathcal{R}(\phi, \boldsymbol{T}, \boldsymbol{T}, \boldsymbol{T}^\top) = \sum\nolimits_{t=0,1} \Big[\Big(R_t + \frac{F_\beta(\boldsymbol{T})}{\beta} \Big)^2 - G(\boldsymbol{X}_t, \boldsymbol{Z}_{1-t}; \boldsymbol{T}_t) \Big], \tag{15}$$

Table 1: CelebA 64×64 test results, where mean $\pm$ sample SD over three training seeds; bold marks the best column mean.

Method	n	FID↓	KID $\times 10^3$ ↓	PSNR↑	SSIM↑
WAE-MMD	3	67.70 ± 0.64	71.70 ± 1.53	22.08 ± 0.01	0.6775 ± 0.0004
SWAE	3	60.70 ± 1.07	60.47 ± 2.79	22.00 ± 0.05	0.6709 ± 0.0012
VAE	3	62.22 ± 0.49	61.96 ± 0.69	22.37 ± 0.03	0.6862 ± 0.0018
WAE+GWIB	3	$\mathbf{51.46 \pm 0.24}$	$\mathbf{51.05 \pm 0.80}$	$\mathbf{22.50 \pm 0.05}$	$\mathbf{0.6962 \pm 0.0005}$



(a) Reconstructions (b) Prior generations

Figure 2: CelebA fixed samples. Columns denote models; rows follow the first four prescribed inputs (a) or prior vectors (b); the remaining four appear in Appendix Figure 9.

$$\min_{\phi,h_0,h_1} \mathcal{L}(h_0, h_1, \phi) + \lambda_{\mathrm{GWIB}} \mathcal{R}(\phi, \boldsymbol{T}^*, \boldsymbol{T}^*, (\boldsymbol{T}^*)^\top). \tag{16}$$

Here $\mathcal{L}$ fits factual outcomes, R_t measures within-group encoder distortion, F_β aligns latent features and relations across groups, and the subtracted costs retain the relative-gap construction. Shared sample indices define the correspondence without requiring an injective encoder.

With the encoder fixed, conditional gradient and line search approximately minimize the sum of the three alignment costs to construct $\boldsymbol{T}^*$, and the network is then updated with that plan fixed. Appendix I gives the auxiliary objective and Appendix E the algorithm.

5 Experiments

5.1 Reconstruction and Prior Generation on CelebA

Protocol and metrics. We evaluate reconstruction and prior generation on the official 64×64 CelebA splits (Liu et al., 2015). FID and KID measure generation quality (lower is better), while PSNR and SSIM measure reconstruction fidelity (higher is better) (Chen et al., 2023b). Table 1 reports mean $\pm$ sample SD over three runs, with KID scaled by 10^3; Appendix J gives the architecture, training, checkpoint, and evaluation details.

Baselines. We compare aggregate-prior matching by MMD (**WAE-MMD**) (Tolstikhin et al., 2017), sliced Wasserstein distance (**SWAE**) (Kolouri et al., 2018), and per-example KL regularization (**VAE**) (Kingma & Welling, 2013). **WAE+GWIB** instead couples encoder distortion with fused Gromov–Wasserstein alignment to the prior.

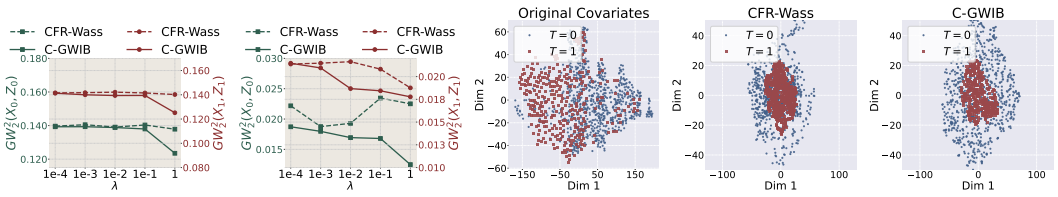
Overall results. WAE+GWIB has the best mean in all four columns of Table 1. Relative to WAE-MMD, it lowers FID/KID by 24.0%/28.8% and raises PSNR/SSIM from 22.08/0.6775 to 22.50/0.6962. Figure 2 shows similar facial content in reconstruction and fewer pronounced prior-sample distortions for GWIB.

Table 2: CelebA WAE mechanism ablation (mean ± sample SD over three seeds). R^2 and GW^2 use the common fixed 1,024-image test subset.

Method	$R^2(\phi)\downarrow$	$GW^2\downarrow$	FID↓	KID↓	PSNR↑	SSIM↑
MMD only	0.0348 ± 0.0005	0.0260 ± 0.0003	67.70 ± 0.64	71.70 ± 1.53	22.08 ± 0.01	0.6775 ± 0.0004
MMD+R	0.0060 ± 0.0005	0.0057 ± 0.0004	57.85 ± 0.76	57.19 ± 1.20	22.36 ± 0.03	0.6900 ± 0.0008
FGW only	0.0124 ± 0.0008	0.0096 ± 0.0006	55.23 ± 0.60	54.80 ± 1.00	22.30 ± 0.03	0.6865 ± 0.0010
Full WAE+GWIB	$\mathbf{0.0058 \pm 0.0002}$	$\mathbf{0.0039 \pm 0.0004}$	$\mathbf{51.46 \pm 0.24}$	$\mathbf{51.05 \pm 0.80}$	$\mathbf{22.50 \pm 0.05}$	$\mathbf{0.6962 \pm 0.0005}$

Table 3: Treatment-effect estimation on ACIC and IHDP. Entries are mean ± SD over five random-seed repetitions; both errors are lower-is-better.

Datasets	ACIC				IHDP			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
S-Learner	$.8120_{\pm.2220}$	$2.2790_{\pm.0340}$	$1.0137_{\pm.2844}$	$2.2997_{\pm.0493}$	$.4032_{\pm.5535}$	$1.2304_{\pm.4826}$	$.3629_{\pm.3130}$	$.9359_{\pm.2040}$
T-Learner	$.9086_{\pm.0779}$	$1.7873_{\pm.0131}$	$2.0749_{\pm.1887}$	$2.6557_{\pm.1270}$	$.3038_{\pm.1356}$	$1.3700_{\pm.0412}$	$.3305_{\pm.3468}$	$1.4256_{\pm.1844}$
k-NN	$.9005_{\pm.0007}$	$4.9456_{\pm.0004}$	$.7432_{\pm.0085}$	$5.0947_{\pm.0011}$	$1.1671_{\pm.0224}$	$3.3651_{\pm.0283}$	$1.3891_{\pm.0592}$	$2.9556_{\pm.0762}$
PSM	$.8712_{\pm.0000}$	$4.7940_{\pm.0003}$	$.6153_{\pm.0001}$	$4.9540_{\pm.0006}$	$1.0764_{\pm.0000}$	$3.2496_{\pm.0000}$	$.7785_{\pm.0245}$	$2.1364_{\pm.0000}$
TARNet	$.7775_{\pm.2063}$	$1.7215_{\pm.1005}$	$1.6519_{\pm.9499}$	$2.4702_{\pm.9361}$	$.1311_{\pm.0510}$	$.8711_{\pm.0387}$	$.0827_{\pm.0539}$	$.8944_{\pm.0686}$
DragonNet	$.5160_{\pm.2251}$	$1.6614_{\pm.1269}$	$.7845_{\pm.6219}$	$2.1005_{\pm.6729}$	$.1355_{\pm.0054}$	$.8752_{\pm.0020}$	$.1078_{\pm.0052}$	$.8989_{\pm.0017}$
CFR-Wass	$.4213_{\pm.0125}$	$1.5291_{\pm.0079}$	$.7470_{\pm.3123}$	$1.7889_{\pm.1082}$	$.1057_{\pm.0265}$	$.8700_{\pm.0678}$	$.0597_{\pm.0346}$	$.9142_{\pm.0529}$
CFR-MMD	$.4156_{\pm.2296}$	$1.3874_{\pm.0673}$	$.9183_{\pm.5629}$	$1.7446_{\pm.3479}$	$.1073_{\pm.0100}$	$.8650_{\pm.0843}$	$.0565_{\pm.0490}$	$.8859_{\pm.1034}$
ESCFR	$.4155_{\pm.0292}$	$1.4091_{\pm.0048}$	$1.0831_{\pm.1972}$	$1.8048_{\pm.0811}$	$.1015_{\pm.0008}$	$.8315_{\pm.0066}$	$.1135_{\pm.0026}$	$.8854_{\pm.0096}$
SITE	$.8716_{\pm.3252}$	$2.4311_{\pm.1879}$	$.7173_{\pm.2787}$	$2.3388_{\pm.0285}$	$.1132_{\pm.1044}$	$.9766_{\pm.0656}$	$.1004_{\pm.0854}$	$.8398_{\pm.1229}$
MitNet	$.3576_{\pm.0268}$	$1.3934_{\pm.0238}$	$.6942_{\pm.1586}$	$1.6045_{\pm.0442}$	$.1711_{\pm.0156}$	$.9780_{\pm.0076}$	$.2478_{\pm.0080}$	$.8197_{\pm.0033}$
CausalVIB	$.3803_{\pm.0145}$	$1.4025_{\pm.0112}$	$.6489_{\pm.3080}$	$1.6076_{\pm.0623}$	$.1480_{\pm.0111}$	$.9510_{\pm.0022}$	$.2017_{\pm.0047}$	$.8138_{\pm.0138}$
BWCFR	$.6391_{\pm.2422}$	$1.5775_{\pm.0664}$	$1.2175_{\pm.2302}$	$1.9100_{\pm.1161}$	$.1391_{\pm.0436}$	$.8592_{\pm.0911}$	$.0517_{\pm.0224}$	$.8639_{\pm.0843}$
C-GWIB	$\mathbf{.3252}_{\pm.0498}$	$\mathbf{1.2542}_{\pm.0045}$	$\mathbf{.5163}_{\pm.2786}$	$\mathbf{1.4527}_{\pm.0389}$	$\mathbf{.0506}_{\pm.0006}$	$\mathbf{.5085}_{\pm.0003}$	$\mathbf{.0503}_{\pm.0003}$	$\mathbf{.6988}_{\pm.0002}$



(a) ACIC (b) IHDP (c) Inputs (d) CFR-Wass (e) C-GWIB

Figure 3: Geometry and group balance. (a,b) Input-to-representation $\widehat{GW}_2^2$ versus weight (λ for CFR-Wass, λ_{GWIB} for C-GWIB); control/treated groups use the left/right axes. (c–e) ACIC t-SNE views; blue circles are control and red squares are treated.

Ablation studies. In this section, we conduct ablation studies to investigate the contributions of the two components in the GWIB regularizer: the input-to-code pairwise distortion R and the fused Gromov–Wasserstein term F_β/β. We compare four variants: MMD only is the WAE-MMD baseline with coefficient 100; MMD+R adds the input-to-code pairwise distortion with coefficient 10; FGW only uses $10(F_\beta/\beta)^2$ with $\beta = 0.5$ and no MMD term; and Full WAE+GWIB uses the complete $10(R + F_\beta/\beta)^2$ regularizer. Table 2 reports the common 1,024-image geometry evaluation, formal generation and reconstruction metrics. We can see that the explicit R term substantially reduces geometric distortion, the FGW term improves generation relative to MMD only, and their full combination gives the best geometric and task-specific results, while the MMD-based variants retain the closest IMQ-MMD prior match.

5.2 Treatment-effect estimation

Protocol and metrics. We evaluate on semi-synthetic IHDP and ACIC (Hill, 2011; Du et al., 2021), reporting in-sample and held-out $\epsilon_{ATE} = |N^{-1}\sum_i e_i|$ and $\epsilon_{PEHE} = N^{-1}\sum_i e_i^2$. Table 3

Table 4: Ablations on ACIC and IHDP (mean ± SD over five random seeds). Lower errors are better; bold marks the lowest column mean. PEHE follows Section 5.2.

Datasets	ACIC				IHDP			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
GWIB-FGW	$.3752_{\pm.0562}$	$1.2852_{\pm.0066}$	$1.0038_{\pm.2803}$	$1.6396_{\pm.1064}$	$.1530_{\pm.0110}$	$.6605_{\pm.0746}$	$.1580_{\pm.0067}$	$.8974_{\pm.0067}$
GWIB-RT	$.3806_{\pm.1747}$	$1.4222_{\pm.0217}$	$1.1034_{\pm.2940}$	$1.7818_{\pm.1218}$	$.1079_{\pm.0018}$	$.5681_{\pm.0108}$	$.1824_{\pm.0071}$	$.7889_{\pm.0249}$
GWIB-GW	$.4580_{\pm.0311}$	$1.2882_{\pm.0036}$	$1.0764_{\pm.3035}$	$1.6758_{\pm.1020}$	$.1304_{\pm.0032}$	$.5561_{\pm.0058}$	$.1684_{\pm.0025}$	$.8083_{\pm.0187}$
GWIB-Gap	$.3906_{\pm.0406}$	$1.2922_{\pm.0038}$	$1.2205_{\pm.2043}$	$1.7770_{\pm.0756}$	$.1248_{\pm.0016}$	$.5156_{\pm.0001}$	$.1503_{\pm.0062}$	$.7174_{\pm.0002}$
GWIB-Opt	$.5060_{\pm.0041}$	$1.2991_{\pm.0033}$	$1.3675_{\pm.0586}$	$1.8378_{\pm.0349}$	$.1326_{\pm.0043}$	$.5234_{\pm.0002}$	$.1809_{\pm.0067}$	$.7170_{\pm.0006}$
Full C-GWIB	$\mathbf{.3252}_{\pm.0498}$	$\mathbf{1.2542}_{\pm.0045}$	$\mathbf{.5163}_{\pm.2786}$	$\mathbf{1.4527}_{\pm.0389}$	$\mathbf{.0506}_{\pm.0006}$	$\mathbf{.5085}_{\pm.0003}$	$\mathbf{.0503}_{\pm.0003}$	$\mathbf{.6988}_{\pm.0002}$

gives mean ± SD over five seeds. In addition, the data splits, tuning, and evaluation details are in Appendices F.1 and I.

Baselines. Table 3 covers meta- and matching learners and representative regression/balancing methods, including TARNet/CFR (Shalit et al., 2017), DragonNet (Shi et al., 2019), SITE (Yao et al., 2018), BWCFR (Assaad et al., 2021), ESCFR (Wang et al., 2023), MitNet (Guo et al., 2023), and CausalVIB (Lu et al., 2022); Appendix F.1 defines the full set.

Overall results. C-GWIB has the lowest mean in all eight columns of Table 3. Held-out PEHE falls from the strongest listed baseline from 1.6045 to 1.4527 on ACIC (9.5%) and from 0.8138 to 0.6988 on IHDP (14.1%), showing gains in heterogeneous-effect accuracy on both datasets. ATE can hide offsetting individual errors, whereas PEHE measures their squared magnitudes. BWCFR's IHDP held-out ATE (0.0517) is close to C-GWIB's (0.0503), but its PEHE is higher (0.8639 versus 0.6988). The benefit in this comparison is thus clearer at the heterogeneous-effect level than in population-average error.

Geometric diagnostics. Across tested weights, Figure 3(a,b) shows lower input-to-representation GW discrepancies for C-GWIB than CFR-Wass on both datasets. The t-SNE views in Figure 3(c–e) provide a qualitative counterpart: compared with CFR-Wass, C-GWIB aligns treatment and control groups without the same concentration into a narrow central region. Together with the lower held-out PEHE, this suggests that retaining within-group covariate geometry after alignment preserves information needed for individual, rather than only average, treatment-effect estimation.

Ablation studies. Table 4 tests GWIB-FGW/RT/GW, which remove the FGW term, within-group distortion R_t, or subtracted cross-group GW costs; GWIB-Gap, which restores the original within-group gap; and GWIB-Opt, which uses independent rather than shared cross-group plans. The full objective has the lowest mean error in every column: its ACIC held-out PEHE is 1.4527, versus $1.7770/1.8378$ for GWIB-Gap/GWIB-Opt, while on IHDP removing FGW is worst (0.8974 versus 0.6988). Thus the complete distortion, relative-gap, and coupled-alignment construction is preferred, although component effects vary by dataset.

6 Conclusion

In this paper, we revisited the information bottleneck through the lens of optimal transport and proposed GWIB for geometry-aware representation learning. By linking an empirical kernelized information bound to the Gromov-Monge gap and deriving a reference-distribution relaxation, we provide an information-theoretic basis for geometric regularization in Wasserstein autoencoders and counterfactual regression. Experiments show improved reconstruction and generation on CelebA and lower treatment-effect errors on IHDP and ACIC. One limitation is that the current conditional-gradient OT solver has cubic scaling with group size and optimizes a non-convex objective only to a local stationary point. Future work will investigate entropically regularized OT solvers for a more scalable and stable alternative to the conditional-gradient loop (Cuturi, 2013; Rigollet & Weed, 2018). It will also extend GWIB to attributed graphs and task-adaptive relational geometries, using fused Gromov–Wasserstein alignment and learned structured priors (Titouan et al., 2019; Xu et al., 2020).

AI USE STATEMENT

Generative AI assistance was used to inspect the existing references, organize the experimental protocol, help to implement and polish the manuscript. We have not used generative AI tools for the core intellectual contributions of this work, including the theoretical derivations and algorithm design. We have carefully reviewed and edited all AI-generated content to ensure accuracy and clarity.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the proposed method and experimental setup in the main text. Further details needed to reproduce the experiments, including dataset descriptions, preprocessing procedures, model configurations, hyperparameter selection, training and evaluation protocols, and computational resources, are provided in the experimental sections and the appendix. We have also provided the code and data used in our experiments in the supplementary material.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

Lingyi Chen, Shitong Wu, Wenhao Ye, Huihui Wu, Hao Wu, Wenyi Zhang, Bo Bai, and Yining Sun. Information bottleneck revisited: Posterior probability perspective with optimal transport. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1490–1495. IEEE, 2023a.

Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou. Dargs: Image inpainting algorithm via deep attention residuals group and semantics. *Journal of King Saud University-Computer and Information Sciences*, 35(6):101567, 2023b.

Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. Infoot: Information maximizing optimal transport. In *International Conference on Machine Learning*, pp. 6228–6242. PMLR, 2023.

Danilo Costarelli and Renato Spigler. How sharp is the jensen inequality? *Journal of Inequalities and Applications*, 2015:1–10, 2015.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

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

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34, 02 2019. doi: 10.1214/18-STS667.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Xingzhuo Guo, Yuchen Zhang, Jianmin Wang, and Mingsheng Long. Estimating heterogeneous treatment effects: Mutual information bounds and learning algorithms. In *International Conference on Machine Learning*, pp. 12108–12121. PMLR, 2023.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Sungyub Kim, Yongsu Baek, Sung Ju Hwang, and Eunho Yang. Reliable estimation of individual treatment effect with causal information bottleneck. *arXiv preprint arXiv:1906.03118*, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Yonghyeon Lee, Sangwoong Yoon, Minjun Son, and Frank C Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Zhenyu Lu, Yurong Cheng, Mingjun Zhong, George Stoian, Ye Yuan, and Guoren Wang. Causal effect estimation using variational information bottleneck. In *International Conference on Web Information Systems and Applications*, pp. 288–296. Springer, 2022.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Facundo Mémoli and Tom Needham. Distance distributions and inverse problems for metric measure spaces. *Studies in Applied Mathematics*, 149(4):943–1001, 2022.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2020.

Philipp Nazari, Sebastian Damrich, and Fred A Hamprecht. Geometric autoencoders–what you see is what you decode. *arXiv preprint arXiv:2306.17638*, 2023.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aäron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 15604–15614, 2019.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus. Mathématique*, 356(11-12):1228–1235, 2018.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 09–15 Jun 2019.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Théo Uscidda and Marco Cuturi. The monge gap: A regularizer to learn all transport maps. In *International conference on machine learning*, pp. 34709–34733. PMLR, 2023.

Théo Uscidda, Luca Eyring, Karsten Roth, Fabian Theis, Zeynep Akata, and marco cuturi. Disentangled representation learning with the gromov-monge gap. In *International Conference on Learning Representations*, volume 2025, pp. 79221–79252, 2025.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, pp. 37496–37510. PMLR, 2023.

Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, and Lawrence Carin. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*, pp. 10576–10586. PMLR, 2020.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

A Appendix Overview

The appendices provide causal assumptions, proofs, optimization algorithms, benchmark protocols, and supplementary results, followed by the image protocol and implementation disclosures.

B In-Depth Overview of Causal Inference

This section delves into the intricacies of causal inference, particularly for those who may be less acquainted with this field.

B.1 Problem Formulation

Here, we outline essential definitions and assumptions needed for causal inference using observational data. For an individual characterized by covariates x, there are two potential outcomes: $y(x, 1)$ if the individual is treated and $y(x, 0)$ if control. The true individual treatment effect (ITE) is the difference between these potential outcomes.

Definition 2. The individual treatment effect (ITE) for an individual with covariates x is expressed as

$$\tau(x) = y(x, 1) - y(x, 0), \tag{17}$$

The corresponding conditional average treatment effect (CATE) is instead

$$\tau(x) = \mathbb{E}[Y_1 - Y_0 \mid X = x], \tag{18}$$

where X denote the variable of covariate, Y_1 and Y_0 denote the outcome variables under the treated and control conditions, respectively.

Estimating the ITE from observational data presents two main challenges:

1. *Missing counterfactuals*: We can only observe the outcome for the treatment actually received. For example, if a patient undergoes treatment, the outcome had the patient not been treated remains unobservable.
2. *Selection bias*: Treatment decisions can introduce differences between treated and control groups. For instance, varying treatment plans based on patients' health conditions create heterogeneity, complicating the ability to generalize findings across the entire population.

Pearl & Mackenzie (2018) introduced a two-step approach to address these challenges. The first step is *identification*, which involves deriving an unbiased statistical estimand for the causal effect (e.g., $\tau(x)$) through an adjustment formula. Identification of the CATE requires assumptions; these do not generally identify each individual's missing potential outcome:

Assumption 1 (Ignorability / Unconfoundedness)**.** The treatment assignment is independent of potential outcomes given covariates x, i.e., $Y_1, Y_0 \perp\!\!\!\perp T \mid X = x$.

Assumption 2 (Positivity)**.** For any covariate set $X = x$, the probability of receiving each treatment (0 or 1) is positive, i.e., $0 < P(T = t \mid X = x) < 1$ for any t and x.

Assumption 3 (SUTVA)**.** The potential outcomes for any individual are unaffected by the treatment assignments of other individuals, and each treatment level is associated with a unique outcome.

Assumption 4 (Consistency)**.** The observed outcome corresponds to the potential outcome for each individual.

In the second step, *estimation*, the CATE is expressed as a statistical estimand that can be computed from observational data:

$$\begin{aligned}
&\mathbb{E}[Y_1 - Y_0 \mid X = x] = \mathbb{E}[Y_1 \mid X = x] - \mathbb{E}[Y_0 \mid X = x] \\
&\overset{(1)}{=} \mathbb{E}[Y_1 \mid X = x, T = 1] - \mathbb{E}[Y_0 \mid X = x, T = 0] \\
&\overset{(2)}{=} \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0],
\end{aligned} \tag{19}$$

where the first equality follows from Assumption 1, and the second from Assumptions 2, 3 and 4.

In practice, various methods have been proposed to estimate this conditional mean contrast. Early approaches include neighbor matching (Caliendo & Kopeinig, 2008) and meta-learners (Künzel et al., 2019). Neighbor matching uses outcomes from comparable individuals in the other group as proxies for potential outcomes, while meta-learners apply pre-existing models to fit data from both treatment and control groups. With recent advances in deep learning, representation-based methods have become powerful alternatives (Assaad et al., 2021; Wu et al., 2023). Notably, counterfactual regression (CFR) (Shalit et al., 2017) has gained attention for its ability to leverage regularization techniques to minimize distributional discrepancies, such as the Wasserstein distance, between treatment and control groups. In this paper, we revisit the objective of CFR from the viewpoint of information bottleneck, and propose a C-GWIB learning paradigm to address covariate imbalance and within-group structural distortion jointly.

C Broader Impacts

As our proposed C-GWIB method aims to contribute to the causal inference community, which finds extensive applications in the real world, numerous stakeholders can leverage its benefits. For instance, analysts working in online marketing companies can apply our method to estimate the uplift value for each user, thereby facilitating their decision-making process for coupon allocation. Furthermore, while the proposed C-GWIB learning paradigm provides a broad upper bound for the information bottleneck, demonstrating its potential across various supervised learning tasks, this paper specifically focuses on applying C-GWIB to the field of causal inference.

D Proofs of Theorems

D.1 Proof of Theorem 1

A generalized KDE of the joint probability $p(X, Z)$. Recall the joint probability $p(X, Z)$ defined in (3). We can find that it is actually a special case of the following generalized formulation proposed in (Chuang et al., 2023).

$$p(X, Z) = \sum\nolimits_{m,n=1}^{N} T_{mn}\kappa(X, x_m)\kappa(Z, z_n), \text{ with } \boldsymbol{T} = [T_{mn}] \in \Pi\Big(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N\Big), \tag{20}$$

where $\boldsymbol{T}$ represents the joint distribution of the sample pairs (x_m, z_n). Obviously, the joint distribution in (3) corresponds to the case with $\boldsymbol{T} = \frac{1}{N}\boldsymbol{I}_N$, and we have

$$p(X, Z) \leq \max\nolimits_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)} \sum\nolimits_{m,n=1}^{N} T_{mn}\kappa(X, x_m)\kappa(Z, z_n). \tag{21}$$

The Jensen gap for the logarithmic function defined on a closed domain. Suppose that $x \in (0, \infty)$ and obeys to a distribution p. For $\log x$, Jensen's inequality tells us that $\log \mathbb{E}_{x\sim p}[x] \geq \mathbb{E}_{x\sim p}[\log x]$, and the $\log \mathbb{E}_{x\sim p}[x] - \mathbb{E}_{x\sim p}[\log x]$ is called "Jensen gap".

The work in (Costarelli & Spigler, 2015) studies the sharpness of the Jensen's inequality, deriving the upper bound of Jensen gap for various functions, including $\log x$:

Lemma 1 (The upper bound of Jensen gap for logarithmic function in (Costarelli & Spigler, 2015))**.** *Suppose that $\log x$ is defined on a closed set $[a, \infty)$, where $a > 0$, and $x \sim p$, whose mean is μ and variance is σ^2. We have*

$$\log \mathbb{E}_{x\sim p}[x] - \mathbb{E}_{x\sim p}[\log x] \leq \frac{1}{2a^2} \min_{c\geq a}[\mathbb{E}_{x\sim p}[(x - c)^2] + (\mu - c)^2] = \frac{\sigma^2}{2a^2}. \tag{22}$$

Obviously, this lemma is held when $\log x$ is defined in $[a, b]$, $b < \infty$.

Given $\{x_m, z_n\}_{m,n=1}^{N} \in \mathcal{X} \times \mathcal{Z}$, where $(\mathcal{X}, d_X)$ and $(\mathcal{Z}, d_Z)$ are bounded metric spaces with $\text{diam}_{\mathcal{X}} = \sup_{x,x'} d_X(x, x')$ and $\text{diam}_{\mathcal{Z}} = \sup_{z,z'} d_Z(z, z')$. For $N \geq 2$ and $2\pi\tau^2 \geq 1$, the range of $f(x, x', z, z') = \kappa(x, x')\kappa(z, z')$ is contained in $[\frac{\alpha}{2\pi\tau^2}, \frac{1}{2\pi\tau^2}] \subseteq [\frac{\alpha}{2\pi\tau^2}, 1]$, where $\alpha = \exp(-\frac{1}{2\tau^2}(\text{diam}_{\mathcal{X}}^2 + \text{diam}_{\mathcal{Z}}^2))$.

In (20), the nonnegative weights $T_{m,n}T_{k,l}$ sum to one. With $a = \alpha/(2\pi\tau^2)$, their weighted mean and variance satisfy

$$\begin{aligned}\mu_T &= \sum_{m,n,k,l=1}^{N} T_{m,n}T_{k,l} f(x_m, x_k, z_n, z_l),\\ \sigma_T^2 &= \sum_{m,n,k,l=1}^{N} T_{m,n}T_{k,l}(f(x_m, x_k, z_n, z_l) - \mu_T)^2 \le \frac{(1-a)^2}{4} \le \frac{N^4(1-a)^2}{4(N^4-1)}.\end{aligned} \tag{23}$$

The interval bound on the variance holds for these arbitrary probability weights; the factor $N^4/(N^4-1) \ge 1$ is a conservative enlargement that preserves the original constant below.

Based on the upper bound of the variance and Lemma 1, we have

$$\begin{aligned}&\log\Big(\sum_{m,n,k,l} T_{m,n}T_{k,l}\kappa(x_m, x_k)\kappa(z_n, z_l)\Big)\\ &\le \sum_{m,n,k,l} T_{m,n}T_{k,l}\log(\kappa(x_m, x_k)\kappa(z_n, z_l)) + \underbrace{\frac{N^4(2\pi\tau^2-\alpha)^2}{8(N^4-1)\alpha^2}}_{C_{\kappa,N}},\end{aligned} \tag{24}$$

where the last term is the proposed Jensen gap in this work.

According to the above information, we have

$$\begin{aligned}&\hat{I}_{\kappa,N}(Z, X)\\ =&\frac{1}{N}\sum_n\Big(\log\Big(\frac{1}{N}\sum_m \kappa(x_n, x_m)\kappa(z_n, z_m)\Big) - \log\Big(\frac{1}{N^2}\sum_m \kappa(x_n, x_m)\sum_m \kappa(z_n, z_m)\Big)\Big)\\ \le&\log\Big(\frac{1}{N^2}\sum_{m,n}\kappa(x_m, x_n)\kappa(z_m, z_n)\Big) - \frac{1}{N^2}\sum_{m,n}\Big(\log\kappa(x_m, x_n) + \log\kappa(z_m, z_n)\Big)\\ =&\log\Big(\frac{1}{N^2}\sum_{m,n}\kappa(x_m, x_n)\kappa(z_m, z_n)\Big) + \frac{\sum_{m,n}(d_X^2(x_m, x_n) + d_Z^2(z_m, z_n))}{2N^2\tau^2} + \log 2\pi\tau^2\\ =&\log\Big(\frac{1}{N^2}\sum_{m,n}\kappa(x_m, x_n)\kappa(z_m, z_n)\Big) + \frac{\sum_{m,n} d_X(x_m, x_n)d_Z(z_m, z_n)}{N^2\tau^2} +\\ &\frac{\sum_{m,n}(d_X(x_m, x_n) - d_Z(z_m, z_n))^2}{2N^2\tau^2} + \log 2\pi\tau^2\\ \le&\max_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)}\Big[\log\Big(\sum_{m,n,k,l=1}^{N} T_{m,n}T_{k,l}\kappa(x_m, x_k)\kappa(z_n, z_l)\Big) +\\ &\frac{\sum_{m,n,k,l} T_{m,n}T_{k,l}d_X(x_m, x_k)d_Z(z_n, z_l)}{\tau^2}\Big] + \frac{1}{2N^2\tau^2}\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2 + \log 2\pi\tau^2\\ \le&\max_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)}\sum_{m,n,k,l=1}^{N} T_{m,n}T_{k,l}\Big(\log\Big(\kappa(x_m, x_k)\kappa(z_n, z_l)\Big) +\\ &\frac{d_X(x_m, x_k)d_Z(z_n, z_l)}{\tau^2}\Big) + \frac{\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2}{2N^2\tau^2} + \log 2\pi\tau^2 + C_{\kappa,N}\\ =&\max_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)}\frac{-1}{2\tau^2}\sum_{m,n,k,l} T_{m,n}T_{k,l}(d_X(x_m, x_k) - d_Z(z_n, z_l))^2\\ &+ \frac{\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2}{2N^2\tau^2} + C_{\kappa,N}\\ =&\frac{1}{2\tau^2}\Big(\frac{\|\boldsymbol{D}_X - \boldsymbol{D}_Z\|_F^2}{N^2}\\ &- \min_{\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)}\sum_{m,n,k,l} T_{m,n}T_{k,l}(d_X(x_m, x_k) - d_Z(z_n, z_l))^2\Big) + C_{\kappa,N}\end{aligned}$$

$$=\frac{1}{2\tau^2}\Big(\frac{1}{N^2}\|\boldsymbol{D}_X-\boldsymbol{D}_Z\|_F^2-\widehat{GW}_2^2(\boldsymbol{X},\boldsymbol{Z})\Big)+C_{\kappa,N},$$

which completes the proof. Here, the first inequality is derived based on Jensen's inequality. The second inequality is based on the upper bound of the generalized KDE shown in (21), which generalizes $\boldsymbol{T}=\frac{1}{N}\boldsymbol{I}_N$ to $\boldsymbol{T}\in\Pi(\frac{1}{N}\mathbf{1}_N,\frac{1}{N}\mathbf{1}_N)$. The third inequality is based on the upper bound of Jensen gap shown in (24).

D.2 Proof of Theorem 2

Specifically, the GM distance is defined as follows:

Definition 3 (Gromov-Monge distance)**.** Let μ and ν be two probability measures defined on two compact metric spaces $(\mathcal{X},d_X)$ and $(\mathcal{Z},d_Z)$, respectively. The order-p Gromov-Monge distance between μ and ν is

$$\begin{aligned}GM_p(\mu,\nu):&=\inf_{\psi_\#\mu=\nu}\Big(\int_{\mathcal{X}^2}r^p(x,x',\psi(x),\psi(x'))\mathrm{d}\mu(x)\mathrm{d}\mu(x')\Big)^{\frac{1}{p}}\\&=\inf_{\psi_\#\mu=\nu}\mathbb{E}^{\frac{1}{p}}_{x,x'\sim\mu\times\mu}[r^p(x,x',\psi(x),\psi(x'))],\end{aligned}\tag{25}$$

where the relational distance $r(x,x',\psi(x),\psi(x'))=|d_X(x,x')-d_Z(\psi(x),\psi(x'))|$. The infimum is over all functions $\psi:\mathcal{X}\mapsto\mathcal{Z}$ that push μ to ν, i.e., $\psi_\#\mu=\nu$; a function ψ^* attaining it is a Gromov–Monge minimizer.

Similar to the Monge gap in (Uscidda & Cuturi, 2023), the positivity of Gromov-Monge gap is also based on the fact (Mémoli & Needham, 2022) that

$$GM_p(\mu,\nu)\geq GW_p(\mu,\nu),\ \forall\ \mu\in\mathbb{P}(\mathcal{X}),\ \nu\in\mathbb{P}(\mathcal{Z}).\tag{26}$$

For the given encoder ϕ, $(\mathrm{id},\phi)_\#\rho$ is a feasible GW coupling. Its cost is at least the GW infimum, proving $MG(\rho,\phi)\geq 0$; equality holds precisely when this encoder-induced coupling is GW-optimal.

D.3 Proof of Proposition 1

Costs and assumptions. Let $(\mathcal{X},d_X)$ and $(\mathcal{Z},d_Z)$ be complete separable metric spaces, $\phi:\mathcal{X}\to\mathcal{Z}$ measurable, and $q=\phi_\#\rho$. Assume ρ has finite second moment on $\mathcal{X}$ and q,ν have finite second moments on $\mathcal{Z}$, so the following costs are finite. For $\pi\in\Pi(q,\nu)$, define

$$\begin{aligned}\mathcal{C}_W(\pi)&=\int_{\mathcal{Z}\times\mathcal{Z}}d_Z(z,u)^2\,\mathrm{d}\pi(z,u),\\\mathcal{C}_{GW}(\pi)&=\iint\big|d_Z(z,z')-d_Z(u,u')\big|^2\,\mathrm{d}\pi(z,u)\,\mathrm{d}\pi(z',u').\end{aligned}\tag{27}$$

Thus $GW_2^2(q,\nu)=\inf_\pi\mathcal{C}_{GW}(\pi)$, with no extra $1/2$ factor in this cost or in R^2. Empirically the encoder-induced coupling gives $R^2=N^{-2}\|\boldsymbol{D}_X-\boldsymbol{D}_Z\|_F^2$. Intrinsic matrices contain metric distances; their differences are squared in the cost. Squared Euclidean intrinsic matrices require a separate argument.

From the gap to the relaxation. Set $a=GW_2(\rho,q)$, $b=GW_2(q,\nu)$, $c=GW_2(\rho,\nu)$, and $u=F_\beta(q,\nu)/\beta$. The encoder-induced coupling is feasible, so $a\leq R$. The GW triangle inequality gives $c\leq a+b$; it follows here by gluing couplings, applying the L^2 Minkowski inequality, and taking infima. The nonnegative feature cost in every feasible fused plan gives

$$F_\beta^2(q,\nu)\geq\beta b^2,\qquad u\geq b/\sqrt{\beta}\geq b.\tag{28}$$

Consequently,

$$\begin{aligned}MG(\rho,\phi)&=R^2-a^2\\&\leq(R+b)^2-c^2\\&\leq(R+u)^2-c^2=\mathcal{U}_\beta(\rho,\phi;\nu).\end{aligned}\tag{29}$$

Algorithm 1 Bi-level optimization of Coupled GWIB

Require: $\{x_m, y_m(x_m,0)\}_{m=1}^{N_0}$, $\{x_n, y_n(x_n,1)\}_{n=1}^{N_1}$, $\hat{\rho}_0 = \{x_m\}_{m=0}^{N_0}$, $\hat{\rho}_1 = \{x_n\}_{n=1}^{N_1}$, $\beta, \lambda_{\text{GWIB}}$, learning rate α, number of iterations K, number of epochs M.

for $m = 1, \ldots, M$ **do**

UPDATE OT PLAN $\boldsymbol{T}$:

  Compute cost matrices $\boldsymbol{D}_{X_0} = [\|x_m - x_n\|_2^2] \in \mathbb{R}^{N_0 \times N_0}$, $\boldsymbol{D}_{X_1} = [\|x_m - x_n\|_2^2] \in \mathbb{R}^{N_1 \times N_1}$.

  Compute latent cost matrices $\boldsymbol{D}_{Z_0} = [\|\phi(x_m) - \phi(x_n)\|_2^2] \in \mathbb{R}^{N_0 \times N_0}$, $\boldsymbol{D}_{Z_1} = [\|\phi(x_m) - \phi(x_n)\|_2^2] \in \mathbb{R}^{N_1 \times N_1}$, $\boldsymbol{D}_{Z_{0,1}} = [\|\phi(x_m) - \phi(x_n)\|_2^2] \in \mathbb{R}^{N_0 \times N_1}$.

  $\boldsymbol{T}^0 \leftarrow \hat{\rho}_0 \hat{\rho}_1^\top$.

  for $k = 1, \ldots, K$ **do**

    $\nabla_{\boldsymbol{T}} \mathcal{L} \leftarrow$ Compute gradient w.r.t. $\boldsymbol{T}^{(k-1)}$.

    $\widetilde{\boldsymbol{T}}^{(k)} \leftarrow$ Solve OT problem with gradient $\nabla_{\boldsymbol{T}} \mathcal{L}$.

    $\tau^{(k)} \leftarrow$ Line-Search$(\beta, \widetilde{\boldsymbol{T}}^{(k)}, \boldsymbol{T}^{(k-1)}, \boldsymbol{D}_{X_0}, \boldsymbol{D}_{X_1}, \boldsymbol{D}_{Z_0}, \boldsymbol{D}_{Z_1}, \boldsymbol{D}_{Z_{0,1}})$ in Alg. 2.

    $\boldsymbol{T}^{(k)} \leftarrow (1 - \tau^{(k)}) \boldsymbol{T}^{(k-1)} + \tau^{(k)} \widetilde{\boldsymbol{T}}^{(k)}$.

  end for

  $\widehat{\mathcal{R}}(\phi, \boldsymbol{T}^*) \leftarrow$ Compute loss of (15) w.r.t. $\boldsymbol{T}^{(K)}$.

UPDATE PARAMETERS $\{\phi, h_0, h_1\}$:

  $Y_0 \leftarrow \{y_m(x_m,0)\}_{m=1}^{N_0}$, $Y_1 \leftarrow \{y_n(x_n,1)\}_{n=1}^{N_1}$.

  $\hat{Y}_0 \leftarrow h_0 \circ \phi(\hat{\rho}_0)$, $\hat{Y}_1 \leftarrow h_1 \circ \phi(\hat{\rho}_1)$.

  $\mathcal{L}(\phi, h_0, h_1) \leftarrow$ Compute squared loss of (13) w.r.t. $\hat{Y}_0, \hat{Y}_1, Y_0, Y_1$.

  $\phi \leftarrow \phi - \alpha \frac{\partial (\mathcal{L}(\phi, h_0, h_1) + \lambda_{\text{GWIB}} \mathcal{R}(\phi, \boldsymbol{T}^*))}{\partial \phi}$.

  $h_0 \leftarrow h_0 - \alpha \frac{\partial \mathcal{L}(\phi, h_0, h_1)}{\partial h_0}$.

  $h_1 \leftarrow h_1 - \alpha \frac{\partial \mathcal{L}(\phi, h_0, h_1)}{\partial h_1}$.

end for

return $\{\phi, h_0, h_1\}$.

For the first inequality, the difference is $2Rb + b^2 + a^2 - c^2 \geq 2b(R - a) \geq 0$; the second follows from $u \geq b$. This proves the proposition, and the structural bound also holds at $\beta = 1$.

Finite second moments allow a Gaussian reference without compact support. They do not extend Theorem 1's bounded-space or kernel conditions, its factor $1/(2\tau^2)$, or its constant. For the displayed $C_{\kappa,N}$ in Appendix D.1, the sufficient conditions $N \geq 2$ and $2\pi\tau^2 \geq 1$ ensure the kernel-product interval used there. The geometric proposition itself is independent of N, τ and makes no claim about general KL mutual information.

Exact quantities and feasible plans. The proposition concerns exact infima. For a plan $\tilde{\pi}$ satisfying both marginals, its unregularized fused cost $\widetilde{F}_\beta^2 = (1 - \beta)\mathcal{C}_W(\tilde{\pi}) + \beta \mathcal{C}_{GW}(\tilde{\pi})$ is at least F_β^2. Substituting its nonnegative square root into the positive term preserves a possibly looser upper bound. An entropy-regularized objective or debiased divergence is not this raw cost, even if its solver supplies the plan. Replacing the subtracted $GW_2^2(\rho, \nu)$ by an uncertified upper estimate can invalidate the bound. WAE+GWIB drops this term only for a fixed input distribution, fixed prior, and fixed metrics; CFR's learned references require retaining it. A shared feasible plan need not optimize each constituent cost, so the CFR shared-plan numerical objective does not automatically inherit the exact-distance guarantee.

E OPTIMIZATION

E.1 ALGORITHMS AND CONVERGENCE

The algorithmic scheme of C-GWIB is outlined in Alg. 1. Each epoch involves a bi-level optimization process: first, we solve the optimal-transport based regularizer using the conditional gradient (CG) algorithm to iteratively obtain an optimal transport plan T^*. Then, we update the model parameters $\{\phi, h_0, h_1\}$ using the classic stochastic gradient descent (SGD) method.

Algorithm 2 Line-Search

Require: $\beta, \widetilde{\boldsymbol{T}}^{(k)}, \boldsymbol{T}^{(k-1)}, \boldsymbol{D}_{X_0}, \boldsymbol{D}_{X_1}, \boldsymbol{D}_{Z_0}, \boldsymbol{D}_{Z_1}, \boldsymbol{D}_{Z_{0,1}}$.
Ensure: $\tau^{(k)}$.
1: $\boldsymbol{C}_{X_0Z_1} = \frac{1}{N_0}(\boldsymbol{D}_{X_0} \odot \boldsymbol{D}_{X_0})\mathbf{1}_{N_0\times N_1} + \frac{1}{N_1}\mathbf{1}_{N_0\times N_1}(\boldsymbol{D}_{Z_1} \odot \boldsymbol{D}_{Z_1})$.
2: $\boldsymbol{C}_{X_1Z_0} = \frac{1}{N_1}(\boldsymbol{D}_{X_1} \odot \boldsymbol{D}_{X_1})\mathbf{1}_{N_1\times N_0} + \frac{1}{N_0}\mathbf{1}_{N_1\times N_0}(\boldsymbol{D}_{Z_0} \odot \boldsymbol{D}_{Z_0})$.
3: $\boldsymbol{C}_{Z_0Z_1} = \frac{1}{N_0}(\boldsymbol{D}_{Z_0} \odot \boldsymbol{D}_{Z_0})\mathbf{1}_{N_0\times N_1} + \frac{1}{N_1}\mathbf{1}_{N_0\times N_1}(\boldsymbol{D}_{Z_1} \odot \boldsymbol{D}_{Z_1})$.
4: $a = -2\langle \beta \boldsymbol{D}_{Z_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{Z_1} + \boldsymbol{D}_{Z_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{X_1} + \boldsymbol{D}_{X_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{Z_1}, \widetilde{\boldsymbol{T}}^{(k)}\rangle$.
5: $b = \langle (1-\beta)\boldsymbol{D}_{Z_{0,1}} + \beta \boldsymbol{C}_{Z_0Z_1} + \boldsymbol{C}_{X_0Z_1} + \boldsymbol{C}_{X_1Z_0}^{\top}, \widetilde{\boldsymbol{T}}^{(k)}\rangle - 2(\langle \beta \boldsymbol{D}_{Z_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{Z_1} + \boldsymbol{D}_{Z_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{X_1} + \boldsymbol{D}_{X_0}\widetilde{\boldsymbol{T}}^{(k)}\boldsymbol{D}_{Z_1}, \boldsymbol{T}^{(k-1)}\rangle) - 2(\langle \beta \boldsymbol{D}_{Z_0}\boldsymbol{T}^{(k-1)}\boldsymbol{D}_{Z_1} + \boldsymbol{D}_{Z_0}\boldsymbol{T}^{(k-1)}\boldsymbol{D}_{X_1} + \boldsymbol{D}_{X_0}\boldsymbol{T}^{(k-1)}\boldsymbol{D}_{Z_1}, \widetilde{\boldsymbol{T}}^{(k)}\rangle)$.
6: **if** $a > 0$ **then**
7: $\quad \tau^{(k)} \leftarrow \min(1, \max(0, \frac{-b}{2a}))$.
8: **else**
9: $\quad \tau^{(k)} \leftarrow 1$ if $a + b < 0$ else $\tau^{(k)} \leftarrow 0$.
10: **end if**
11: **return** $\tau^{(k)}$.

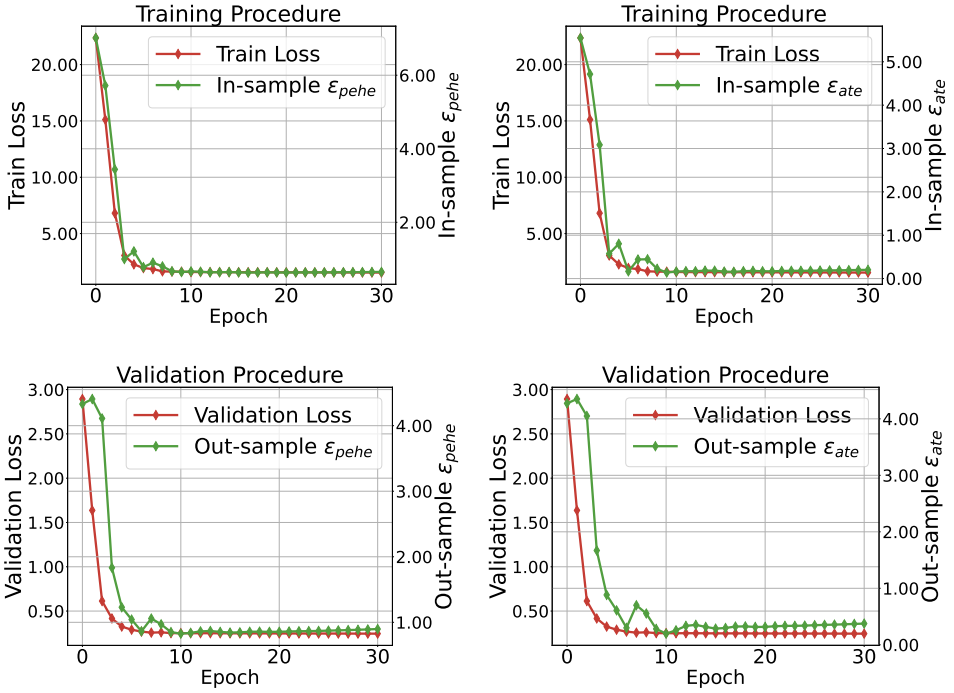
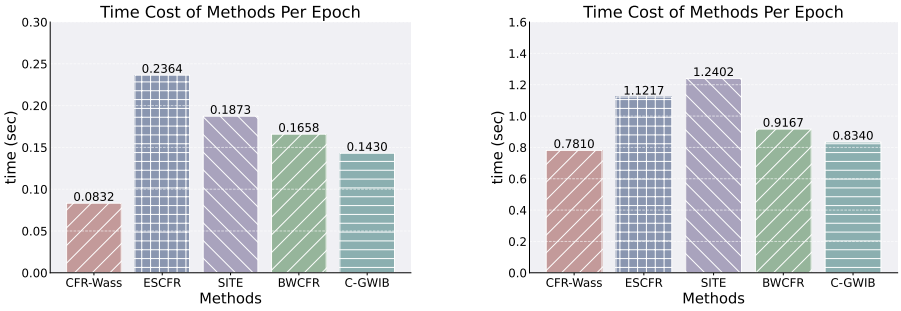


Figure 4: Visualizations on training procedure and validation procedure of C-GWIB on IHDP dataset.

Specifically, in order to update the transport plan $\boldsymbol{T}$, the CG algorithm only requires leveraging the gradient $\nabla_{\boldsymbol{T}}\mathcal{L}$ computed in Section E and then adapting a classic solver for Earth Mover's distance (EMD) such as a Lagrangian mass transport (Bonneel et al., 2011) algorithm to solve the optimal transport problem. Additionally, we propose a linesearch algorithm in Alg. 2 to search for the constrained minimization of the second-degree regularization terms in Gromov-Wasserstein distances. In practice, we repeat the above procedure for adequate iterations to obtain an optimal transport plan. Although the above quadratic problem can be provably non-convex, it can converge to a local stationary point using the CG method (Lacoste-Julien, 2016). Finally, given the optimal transport plan $\boldsymbol{T}^{(K)}$ after K iterations, we then update the model parameters $\{\phi, h_0, h_1\}$ with classic SGD approach.



(a) IHDP dataset (b) ACIC dataset

Figure 5: The comparison of computational time cost per epoch among different methods.

To evaluate the convergence of our proposed algorithms, we have visualized the training and validation results at each epoch. Figure 4 illustrates this analysis. It shows that the loss values for both training and validation decrease rapidly towards zero, indicating that our method achieves convergence swiftly. Concurrently, the performance in the ITE estimation task improves as the number of epochs increases, further demonstrating the effectiveness of our approach.

E.2 Time Complexity Comparison

To better understand the efficiency of our proposed C-GWIB, we examine its time complexity from both theoretical and empirical perspectives. Theoretically, the time complexity of our regularizer is dominated by the (F)GW distance, which is $\mathcal{O}(nm^2 + mm^2)$, where n and m denote the sizes of different treatment groups. Thus, with M training epochs, the total time complexity of C-GWIB is $\mathcal{O}(Mn^3)$. Empirically, as shown in Figure 5, while C-GWIB is slower than the original CFR-Wass method due to the additional regularizer, it remains faster than several other prominent baselines, including the OT-based ESCFR and the methods addressing the over-enforcing balance issue, SITE and BWCFR.

In technique, the information bottleneck-based framework of C-GWIB is particularly well-suited for handling scenarios with extremely high feature dimensionality. This dimensional challenge can be naturally mitigated by utilizing a deeper encoder to reduce dimensionality while still effectively preserving essential information, thanks to the information bottleneck applied to this encoder.

F Reproducibility

In this section, we present additional details regarding the experiment setup and model implementation to ensure better reproducibility.

F.1 Details of Experiment Setup

F.1.1 Datasets

We conduct experiments on two semi-synthetic datasets, namely IHDP and ACIC, to validate our proposed C-GWIB method. The IHDP dataset (Hill, 2011) focuses on the impact of special home visits for infants (treatment) on their future cognitive test scores (outcome). This dataset comprises 747 samples with 25-dimensional covariates collected from a real-world randomized experiment. Selection bias is manually introduced between the treatment and control groups by removing individuals with non-white mothers from the treated population. The potential outcomes are generated using the NPCI package following (Yao et al., 2018). The ACIC dataset (Dorie et al., 2019) is obtained from the Collaborative Perinatal Project and serves as the dataset for the competition. Specifically, ACIC is a collection of semi-synthetic datasets whose covariates are taken from a large study conducted on pregnant women and their children to identifying causal factors leading to developmental disorders. It includes data for 4802 patients with 58 covariates, encompassing both continuous and categorical variables. For both datasets, we randomly shuffle them and allocate 63% for training, 27% for validation, and 10% for testing.

Table 5: The configuration of hyperparameters in this paper.

Hyperparameter	Tuning range	Description
lr	$[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$	learning rate
bs	$[16, 32, 64, 128]$	batch size
λ_{GWIB}	$[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$	weight of the proposed regularization in (15)
β	$[0.1, 0.3, 0.5, 0.7, 0.9]$	weight inside $FGW_{2,\beta}(\boldsymbol{Z}_t, \boldsymbol{Z}_{1-t})$ in (32)
d_ϕ	$[16, 32, 64]$	the dimension of hidden layer in ϕ
d_h	$[16, 32, 64]$	the dimension of hidden layer in h_0 and h_1

F.1.2 BASELINES

In our experiments, we compare various methods to demonstrate the effectiveness of our proposed C-GWIB, including i) Meta-learners: single network for outcome prediction without giving the treatment a special role (**S-Learner**) (Künzel et al., 2019), two separate networks for the treatment group and control group, respectively (**T-Learner**) (Künzel et al., 2019); ii) Matching based methods: k-nearest neighbors (k**-NN**) (Crump et al., 2008), propensity score match with treatment regression (**PSM**) (Caliendo & Kopeinig, 2008); iii) Representation-based methods: two networks for predicting counterfactual outcomes with shared representation layer to alleviate the sample imbalance (**TARNet**) (Shalit et al., 2017), applying additional loss of discrepancy metric (e.g., Wasserstein distance and Maximum Mean Discrepancy distance) to balance the representation distributions between treatment group and control group (**CFR-Wass** and **CFR-MMD**) (Shalit et al., 2017), adjust treatment effect with propensity score layer and targeted regularization (**DragonNet**) (Shi et al., 2019), preservation of local similarity in balancing (**SITE**) (Yao et al., 2018), combination of propensity score-based re-weighting and representation balancing (**BWCFR**) (Assaad et al., 2021), achieving outlier-free balance at mini-batch level with unbalanced optimal transport technique (**ESCFR**) (Wang et al., 2023), conducting mutual information between treatment and covariates to bound treatment effect estimation (**MitNet**) (Guo et al., 2023), employing a variational information bottleneck paradigm to model causal effects (**CausalVIB**) (Lu et al., 2022). Note that propensity scores in DragonNet, SITE, PSM, and BWCFR methods are estimated using logistic regression models following (Caliendo & Kopeinig, 2008).

F.1.3 TRAINING CONFIGURATION

In our experiments, we carefully tune all hyperparameters in GWIB using the validation set with grid search. First, we search for the optimal learning rate and batch size in the range of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\{16, 32, 64, 128\}$, respectively. Then, we tune the hyperparameter λ_{GWIB} that trade-offs the balancing and outcome predictive loss over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The hyperparameter β in fused Gromov-Wasserstein distance is tuned in the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The detailed configurations of hyperparameters are shown in Table 5. The number of training epochs is set as 200. For a fair comparison, all baselines are tuned within the same range as our proposed method. To prevent overfitting to the training set, we employ an early stopping mechanism with a patience of 30 epochs. This precaution is particularly important as the ground-truth outcomes are generated in both IHDP and ACIC datasets, making them more prone to overfitting. In practice, we implement our C-GWIB and all baselines using the PyTorch 1.10 framework with Adam as the optimizer. All experiments are conducted on an NVIDIA A40 GPU and an Intel(R) Xeon(R) 5318Y Gold CPU @ 2.10GHz. More results of hyperparameter optimization can be referred to Appendix G.

F.1.4 EVALUATION

For evaluating the performance of the treatment effect estimation task on IHDP and ACIC datasets, where both factual and counterfactual outcomes are generated from a known distribution, we utilize the absolute error in average treatment effect ϵ_{ATE}, and the precision in estimation of heterogeneous

effects ϵ_{PEHE} as metrics, which are denoted as:

$$\begin{aligned}\epsilon_{ATE} &= \frac{1}{|N|}\Big|\sum_{i=1}^{|N|}\left(y_i(x_i,1)-y_i(x_i,0)\right)-\left(\hat{y}_i(x_i,1)-\hat{y}_i(x_i,0)\right)\Big| \\ \epsilon_{PEHE} &= \frac{1}{|N|}\sum_{i=1}^{|N|}\left(\left(y_i(x_i,1)-y_i(x_i,0)\right)-\left(\hat{y}_i(x_i,1)-\hat{y}_i(x_i,0)\right)\right)^2,\end{aligned} \tag{30}$$

where $N = N_0 + N_1$ is the total number of individuals in control and treatment groups, $\hat{y}_i(x_i, 1)$ and $\hat{y}_i(x_i, 0)$ are the estimated potential outcomes while $y_i(x_i, 1)$ and $y_i(x_i, 0)$ are the ground-truth ones. To evaluate the generalizability of our proposed C-GWIB model, we have conducted two types of experiments, namely, in-sample experiments and out-sample experiments. In the in-sample scenario, the testing samples are drawn from the training distribution, while the out-sample evaluation uses held-out observations from the same benchmark split. This random split does not define an out-of-distribution experiment; the exact historical in-sample index set remains an implementation detail to verify (Appendix I). In practice, we run each experiment five times by varying the random seeds and report the mean and standard deviation of the performances.

F.2 Details of Model Implementation

F.2.1 Details of Prediction Model

As aforementioned, our method is based on the backbone consisting of the representation mapping ϕ as well as the regression estimators h_0, h_1 for control and treatment groups, respectively, which are all parameterized by neural networks. In specific, the backbone includes the following components:

- A two-layer multilayer perceptron (MLP) with dropout and ELU activation function for representation mapping ϕ.
- Two single-layer multilayer perceptron (MLP) with dropout and ELU activation function for regression estimators h_0 and h_1, respectively.

The above prediction model utilizes MSE loss for supervised training over observational data $\{x_m, y_m(x_m, 0)\}_{m=1}^{N_0}$ and $\{x_n, y_n(x_n, 1)\}_{n=1}^{N_1}$. Unless otherwise specified, the dimensions for the two-layer MLP in ϕ are set as $[32, 16]$, while that for the single MLP in each head h_0 and h_1 is set as 16 by default. We set the dropout rate as 0.1 in all experiments. Actually, the prediction model corresponds to the representation-based method TARNet (Shalit et al., 2017), where we can observe in Table 3 that it demonstrates poorer performance than most of the other methods, since it does not explicitly add any regularization to address the selection bias.

F.2.2 Details of C-GWIB

GWIB method is designed in a general way, which can be adaptable to different foundation model, including the aforementioned two-head prediction model adopted in this paper. In practice, our implementation is built upon the Python Optimal Transport (POT 0.9.0) package (Flamary et al., 2021). However, in the original POT, there is no available API function to fuse more than two transport plans, where we need to fuse at least three transport plans for our proposed regularizer in (16). Therefore, we manually extend the original API of fused Gromov-Wasserstein distance function based on a novel conditional gradient (CG) algorithm and a specifically designed line search (LS) approach, where we detail them in Algs. 1 and 2, respectively.

G More Experimental Results

G.1 Experiments in Real-world Dataset Twins

In this section, we conduct additional experiments on a well-known **real-world dataset, Twins**, to further verify the effectiveness of our proposed coupled GWIB regularizers. **Twins** (Almond et al., 2005) contains U.S. twin births from 1989–1991. The treatment indicates whether a baby is the heavier twin, and the outcome is one-year mortality. We retain 12003 pairs with no missing values.

Table 6: Comparison the ITE performance in the **real-world dataset Twins**. The best results are highlighted by bold font.

Test Types	In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
TARNet	$0.0435_{\pm 0.0019}$	$0.2542_{\pm 0.0003}$	$0.0600_{\pm 0.0018}$	$0.2592_{\pm 0.0005}$
CFR-Wass	$0.0370_{\pm 0.0003}$	$0.2703_{\pm 0.0001}$	$0.0461_{\pm 0.0003}$	$0.2698_{\pm 0.0001}$
DragonNet	$0.0948_{\pm 0.0123}$	$0.3699_{\pm 0.0170}$	$0.0998_{\pm 0.0142}$	$0.3818_{\pm 0.0185}$
ESCFR	$0.0437_{\pm 0.0006}$	$0.2687_{\pm 0.0002}$	$0.0544_{\pm 0.0007}$	$0.2706_{\pm 0.0001}$
BWCFR	$0.0465_{\pm 0.0005}$	$0.2600_{\pm 0.0001}$	$0.0529_{\pm 0.0006}$	$0.2666_{\pm 0.0001}$
C-GWIB	$\mathbf{0.0198}_{\pm 0.0002}$	$\mathbf{0.2460}_{\pm 0.0000}$	$\mathbf{0.0327}_{\pm 0.0002}$	$\mathbf{0.2504}_{\pm 0.0000}$

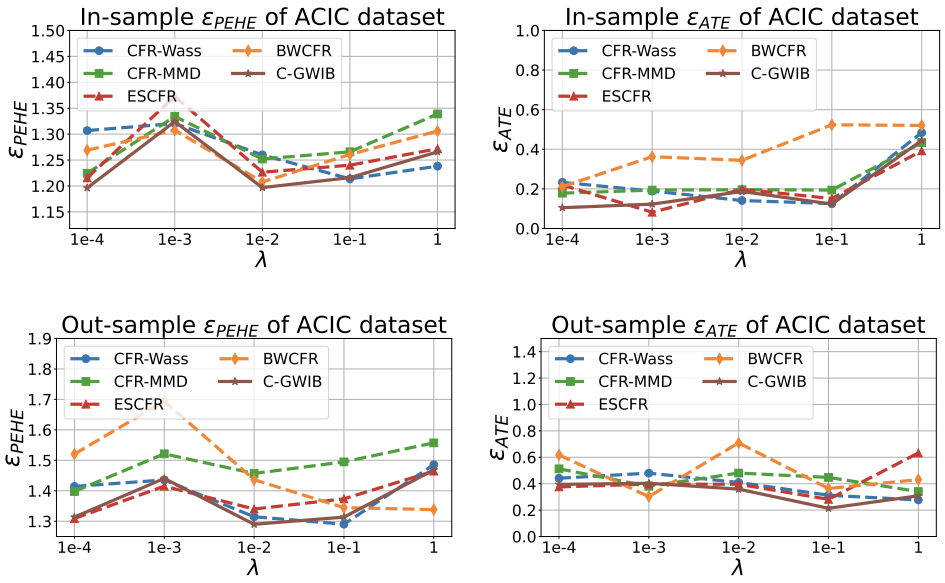


Figure 6: ACIC treatment-effect errors versus regularization weight: λ for baselines and λ_{GWIB} for C-GWIB.

Each baby has 50 covariates, relating to the parents, the pregnancy and birth. Following Louizos et al. (2017), we introduce selection bias to simulate non-random treatment assignment.

From the results reported in Table 6, we can observe that our proposed method, C-GWIB, achieves the best performance across all evaluation metrics in the real-world dataset Twins. Specifically, C-GWIB obtains the lowest estimation errors in terms of both ϵ_{ATE} and ϵ_{PEHE} for both In-sample and Out-sample scenarios. In particular, C-GWIB outperforms the closest competitor, CFR-Wass, with a significant margin, demonstrating its superior capability in capturing treatment effects accurately. These results validate the effectiveness of our method in real-world causal inference tasks.

G.2 Sensitivity to regularization weights

According to (Shalit et al., 2017), the hyperparameter λ in (13) is crucial in determining the trade-off between the outcome prediction task and the representation balancing task for CFR-based models, including CFR-Wass, CFR-MMD, ESCFR, and BWCFR in our paper. A higher value of λ often indicates more balanced representations in the latent space. Therefore, we evaluate C-GWIB using its regularization weight λ_{GWIB}, with λ reserved for the baseline coefficients in these comparisons. Similar to the main experiments, we conduct both in-sample and out-sample experiments and report ϵ_{ATE} and ϵ_{PEHE} as metrics. The respective regularization weights are varied within the

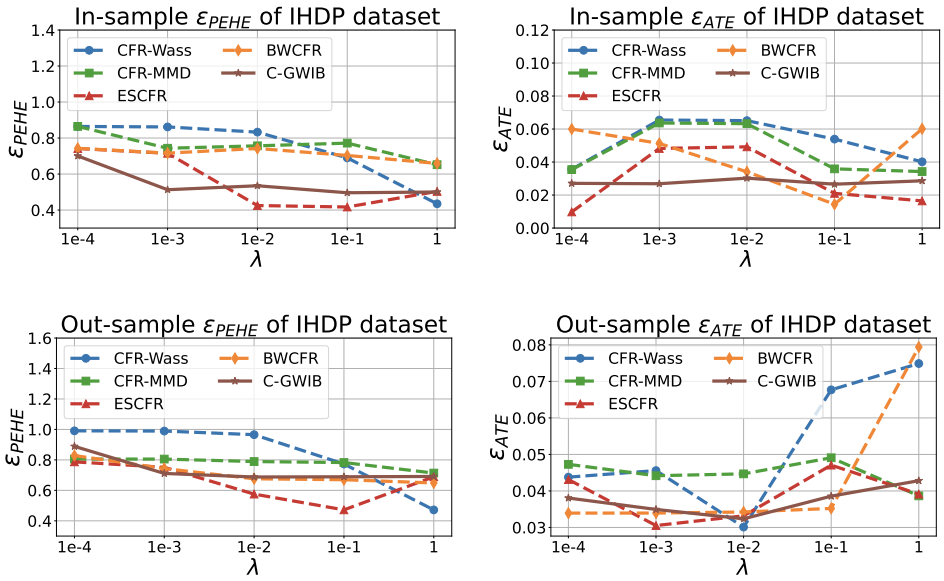


Figure 7: IHDP treatment-effect errors versus regularization weight: λ for baselines and λ_{GWIB} for C-GWIB.

range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. As demonstrated in Figure 6 and 7, we make the following observations:

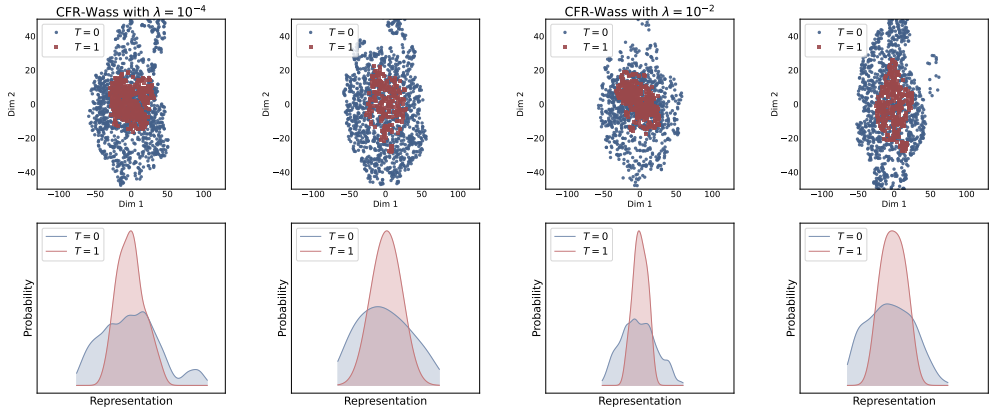
1. The impact of the regularization weight on treatment-effect estimation is evident. As the corresponding weight increases, we observe subtle differences in the tendencies of ϵ_{PEHE} and ϵ_{ATE} across all methods. Specifically, the best performance for ϵ_{ATE} is achieved at a moderate weight, while ϵ_{PEHE} often yields higher performances with extreme higher or lower weights. This is not surprising, since ϵ_{PEHE} assesses precision in predicting treatment effects at the individual level by averaging squared differences, while ϵ_{ATE} evaluates overall accuracy by averaging absolute differences in predicted and observed average treatment effects across the entire population.

2. In most cases, CFR-Wass and CFR-MMD methods, which employ pure representation balancing with different discrepancy metrics, exhibit the worst performance in both ϵ_{PEHE} and ϵ_{ATE} metrics. Besides, ESCFR method has typically obtained the sub-optimal performances across different values of λ, which introduces the unbalanced optimal transport (UOT) technique to relax marginal constraints for flexible balancing at the mini-batch level. However, all of the above baselines have overlooked the over-enforcing balance issue that would obtain undesired trivial distributions in latent space, which may inadvertently remove information that is predictive of outcomes, thus leading to deteriorated performances in the treatment effect estimation task. Although BWCFR can address the over-enforcing balance issue by reweighting the covariate a priori, its performance is significantly dependent on the accurate estimation of propensity scores, which is known to be challenging and often demonstrates high variance (Caliendo & Kopeinig, 2008).

3. C-GWIB attains lower errors than the compared CFR-based baselines for most plotted settings, although performance remains sensitive to λ_{GWIB}. The regularizer controls encoder-induced structural cost relative to a GW alignment and couples this objective to cross-group correspondence. These empirical trends support the combined construction under the evaluated settings; they do not establish absolute distance preservation or information compression for every learned representation.

Table 7: Sensitive analysis of hyperparameter β on ACIC and IHDP datasets.

Datasets	ACIC				IHDP			
Test Types	In-sample		Out-sample		In-sample		Out-sample	
Methods	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}
$\beta = 0.1$	$0.3607_{\pm 0.0415}$	$1.3219_{\pm 0.0081}$	$0.7454_{\pm 0.2248}$	$1.5727_{\pm 0.0496}$	$0.1112_{\pm 0.0015}$	$0.7612_{\pm 0.0017}$	$0.2485_{\pm 0.0051}$	$0.9799_{\pm 0.0025}$
$\beta = 0.3$	$0.3236_{\pm 0.0810}$	$1.3447_{\pm 0.0074}$	$0.7042_{\pm 0.0786}$	$1.5378_{\pm 0.0061}$	$0.5165_{\pm 0.0131}$	$1.0503_{\pm 0.0257}$	$0.7804_{\pm 0.0100}$	$1.3053_{\pm 0.0434}$
$\beta = 0.5$	$0.3041_{\pm 0.0148}$	$1.3373_{\pm 0.0123}$	$0.7289_{\pm 0.1342}$	$1.5233_{\pm 0.0841}$	$0.6293_{\pm 0.0036}$	$1.1299_{\pm 0.0066}$	$0.9600_{\pm 0.0024}$	$1.3667_{\pm 0.0027}$
$\beta = 0.7$	$0.3389_{\pm 0.0310}$	$1.3668_{\pm 0.0116}$	$0.5794_{\pm 0.0388}$	$1.4795_{\pm 0.0202}$	$0.6734_{\pm 0.0072}$	$1.1555_{\pm 0.0129}$	$0.9979_{\pm 0.0058}$	$1.4051_{\pm 0.0036}$
$\beta = 0.9$	$0.3830_{\pm 0.0993}$	$1.3781_{\pm 0.0256}$	$0.7384_{\pm 0.2047}$	$1.6156_{\pm 0.0575}$	$0.6589_{\pm 0.0146}$	$1.1516_{\pm 0.0213}$	$0.9991_{\pm 0.0136}$	$1.3859_{\pm 0.0119}$



(a) $\lambda = 10^{-4}$ (b) $\lambda_{\text{GWIB}} = 10^{-4}$ (c) $\lambda = 10^{-2}$ (d) $\lambda_{\text{GWIB}} = 10^{-2}$

Figure 8: ACIC representations: CFR-Wass (columns 1/3), C-GWIB (2/4); t-SNE views (top) and distributions (bottom).

G.3 Sensitivity Analysis of Hyperparameter β

In (32), we have introduced an additional hyperparameter β in the fused Gromov-Wasserstein distance $FGW_{2,\beta}(\boldsymbol{Z}_t, \boldsymbol{Z}_{1-t})$, to control the relative importance between Gromov-Wasserstein $GW_2(\boldsymbol{Z}_t, \boldsymbol{Z}_{1-t})$ and Wasserstein distances $W_2(\boldsymbol{Z}_t, \boldsymbol{Z}_{1-t})$. Since the mechanism by which β influence the final treatment effect estimation task is not known a priori, a grid search is conducted to find the optimal value of β in the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on ACIC and IHDP datasets. The results, as shown in Table 7, reveal the following observations:

1. By adjusting the value of β, our C-GWIB method exhibits superior stability in terms of ϵ_{ATE} metric when compared to ϵ_{PEHE} across both datasets, in the reported held-out experiments. This sensitivity motivates application-specific validation; it does not guarantee performance in deployment.

2. Furthermore, the performance variations are notably distinct between the ACIC and IHDP datasets. Specifically, on the ACIC dataset, optimal performance is attained with a moderate value of β, whereas on the IHDP dataset, a smaller β value yields superior results. Under Equation 32, a smaller β places less weight on the structural term and more on latent feature matching; the observed preference does not by itself identify a dataset's intrinsic geometry. Therefore, it emphasizes the need to independently fine-tune the value of β in different applications to obtain the optimal performance.

G.4 Additional Visualizations on The Impact of The Balancing Representation

In this section, we expand our experiment to visualize the sample distributions of the control group ($T = 0$) and treatment group ($T = 1$) in the latent space. Specifically, we compare the visual outcomes between the CFR-Wass and C-GWIB methods on ACIC dataset, varying the corresponding weights λ for CFR-Wass and λ_{GWIB} for C-GWIB within $\{10^{-4}, 10^{-2}\}$. For each experiment, we present a scatter plot in the upper row, akin to the experiments in Section 5.2, accompanied by an additional probabilistic density function in the lower row for enhanced clarity.

As depicted in Figure 8, we readily observe that both CFR-Wass and C-GWIB methods effectively balance the samples in the latent space, with highly overlapping covariates between the treatment and control groups. However, without additional approximate regularization, CFR-Wass may yield trivial latent distributions where all samples are projected with similar representations, thereby potentially undermining the predictive capability for ITE estimation tasks. This issue is exacerbated with larger values of λ (e.g., the results in the third column compared to the first column in Figure 8). C-GWIB yields a different displayed distribution under these settings. These qualitative projections complement the task-error and structural-discrepancy results, but their plotted spread is not a proof of distance preservation or non-collapse.

H Limitations

Despite its computation efficiency, the performance of the conditional gradient (CG) algorithm used in the optimal transport (OT) solver in Alg. 1 highly depends on the convexity of the problem, even if it can converge to a local stationary point proved by Lacoste-Julien (2016). However, the convexity assumption does not always align with real-world applications. This discrepancy prompts us to devise our algorithm in a more adaptable fashion, taking into account other well-researched algorithms for solving optimal transport, such as entropic optimal transport (EOT), which offers more stable performance. By considering such candidate solutions, we aim to accommodate the diverse requirements encountered in practical applications.

I Additional Causal Details

The causal assumptions identify conditional average effects, rather than individual missing outcomes. Section 4.2 defines C-GWIB, Section 5.2 reports comparisons, and Appendices E and F.1 give the algorithm and protocol.

I.1 Fused-GW convention

For uniform weights $\boldsymbol{a} = \mathbf{1}_n/n$ and $\boldsymbol{b} = \mathbf{1}_m/m$, the squared structural cost at a specified plan is

$$\begin{aligned} G(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{T}) &= \sum_{i,k=1}^{n} \sum_{j,l=1}^{m} (D_{A,ik} - D_{B,jl})^2 T_{ij} T_{kl} \\ &= \langle \boldsymbol{C}_{AB} - 2\boldsymbol{D}_A \boldsymbol{T} \boldsymbol{D}_B^\top, \boldsymbol{T} \rangle, \\ \boldsymbol{C}_{AB} &= (\boldsymbol{D}_A \odot \boldsymbol{D}_A)\boldsymbol{a}\mathbf{1}_m^\top + \mathbf{1}_n \boldsymbol{b}^\top (\boldsymbol{D}_B \odot \boldsymbol{D}_B)^\top. \end{aligned} \tag{31}$$

Thus $\widehat{GW}_2^2 = \min_{\boldsymbol{T} \in \Pi(\boldsymbol{a}, \boldsymbol{b})} G(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{T})$. For Euclidean latent coordinates, set $M_{01,ij} = \|z_{i,0} - z_{j,1}\|_2^2$ and $\boldsymbol{a}_t = \mathbf{1}_{N_t}/N_t$. Equation 8 becomes

$$\widehat{FGW}_{2,\beta}^2(\boldsymbol{Z}_0, \boldsymbol{Z}_1) = \min_{\boldsymbol{T} \in \Pi(\boldsymbol{a}_0, \boldsymbol{a}_1)} \left[(1-\beta)\langle \boldsymbol{M}_{01}, \boldsymbol{T} \rangle + \beta G(\boldsymbol{Z}_0, \boldsymbol{Z}_1; \boldsymbol{T}) \right]. \tag{32}$$

The nonnegative square root gives $\widehat{FGW}_{2,\beta}$; $F_\beta(\boldsymbol{T})$ is its value before minimization. Intrinsic matrices in the metric derivation contain unsquared distances; $\boldsymbol{M}_{01}$ contains squared feature distances. FGW is defined for $\beta \in [0, 1]$, while division by β restricts the coupled objective to $0 < \beta \leq 1$.

I.2 Plan construction and scope

The shared plan used in Equation 16 is constructed by approximately solving

$$\min_{\boldsymbol{T}\in\Pi(\boldsymbol{a}_0,\boldsymbol{a}_1)} \quad G(\boldsymbol{X}_0,\boldsymbol{Z}_1;\boldsymbol{T}) + G(\boldsymbol{X}_1,\boldsymbol{Z}_0;\boldsymbol{T}^\top) \\ + (1-\beta)\langle \boldsymbol{M}_{01},\boldsymbol{T}\rangle + \beta G(\boldsymbol{Z}_0,\boldsymbol{Z}_1;\boldsymbol{T}). \tag{33}$$

This auxiliary problem constructs one correspondence for the three comparisons; it differs from minimizing $\mathcal{R}$, whose cross-group GW terms enter negatively. Conditional gradient returns a feasible approximate plan, held fixed during network updates.

J CelebA Autoencoding Protocol and Regularization Search

This appendix gives the protocol and validation search for Section 5.1.

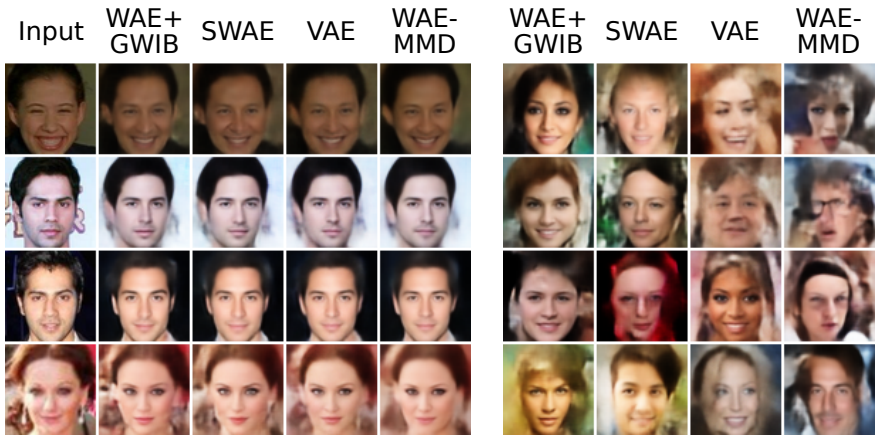
J.1 Implementation and evaluation

Data and backbone. We use ordinary aligned CelebA with the official train/validation/test split (162,770/19,867/19,962 images). RGB images are center-cropped to 140×140, resized to 64×64 with Lanczos interpolation, and mapped to $[-1, 1]$. All methods use the author-matched WAE backbone (Tolstikhin et al., 2017) and a 64-dimensional standard Normal prior. The encoder has four 5×5 stride-2 convolutions with 128/256/512/1024 channels, each followed by batch normalization and ReLU; NHWC-order flattening precedes the 64-unit linear head. The decoder projects to $8 \times 8 \times 1024$ with ReLU, then uses three 5×5 stride-2 transposed convolutions with 512/256/128 channels and batch normalization/ReLU, followed by a stride-1 RGB layer and tanh. Convolutions and transposed convolutions match TensorFlow SAME alignment. The deterministic models share 39,742,531 parameters; separate posterior mean/log-variance heads give the VAE 40,791,171. This PyTorch adaptation uses official splits and no added input noise; it is not an unchanged reproduction of the authors' experiment.

Objectives. Deterministic models use 0.05 times the batch mean of per-image summed squared pixel errors. WAE-MMD uses weight 100 and an unbiased IMQ estimator summed over scales $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. SWAE (Kolouri et al., 2018) uses weight 128 on the mean squared sorted-projection discrepancy over 50 random unit directions. WAE+GWIB replaces MMD with the complete squared regularizer in Eq. (12), with $\beta = 0.5$ and selected regularization weight $\lambda_{\mathrm{GWIB}} = 10$; this weight is distinct from the learning rate and the WAE-MMD coefficient. Geometry uses 32 examples per batch, an input divisor 75.52928161621094 fixed from training data, latent divisor $\sqrt{128}$, and an approximate feasible POT coupling (at most 50 iterations, tolerance 10^{-7}) held fixed while differentiating the raw, entropy-free cost. The VAE (Kingma & Welling, 2013) instead minimizes Gaussian negative log-likelihood with observation variance 0.3 and unit-weight KL, sampling posterior codes during training and decoding posterior means for reconstruction evaluation.

Optimization. Each run is assigned 50 epochs of full batches of 100, dropping the incomplete training batch, with Adam betas $(0.5, 0.999)$ and $\mathrm{lr}(e) = \mathrm{lr}_0\, 100^{-e/50}$ for $e = 1, \ldots, 50$. Deterministic models use $\mathrm{lr}_0 = 5 \times 10^{-4}$; the VAE uses 10^{-4}. Deterministic encoders first match sampled Normal mean/covariance using a separate Adam optimizer and batches of 256 training images, for at most 200 updates, stopping when the moment loss falls below 0.1. The VAE omits pretraining. Pretraining uses float32; main training uses bfloat16 network operations and float32 objective arithmetic. Development pauses preserve the original 50-epoch schedule and optimizer/RNG state.

Checkpoint and metric protocol. Every five epochs, fixed validation evaluation uses 2,000 real and generated images and KID with ten subsets of $1,000$. Minimum validation KID selects one checkpoint per run; exact ties select the earlier epoch. All evaluation metrics for that run use the same selected checkpoint. Test FID/KID use 10,000 generated and 10,000 fixed real images, with 100 KID subsets of $1,000$. Generation metrics share torch-fidelity 0.4.0 Inception features, uint8 conversion, resizing, and real-feature caching. PSNR/SSIM use all 19,962 test reconstructions in $[0, 1]$. The reporting protocol specifies mean $\pm$ sample SD across three training seeds, not KID subset SD.



(a) Reconstructions (b) Prior generations

Figure 9: Remaining CelebA samples from Figure 2. Columns denote models; rows follow prescribed input or prior-vector indices 4–7. Shared latent indices do not denote matching identities.