EVIA: Entropic Variational Inference Auto-encoding

Yunfei Teng[1,2] **Zhichao Chen**[1] **Xinyu Chen**[2] **Lulu Tang**[2] **Sixin Zhang**[3] **Zhouchen Lin**[*1]

[1]State Key Laboratory of General AI, School of Intelligence Science and Technology, Peking University
[2]Beijing Academy of Artificial Intelligence
[3]Univ Toulouse, Toulouse INP, CNRS, IRIT, Toulouse, France

Abstract

We present Entropic Variational Inference Auto-encoding (EVIA), a framework that extends classical variational approaches—including WAEs [Tolstikhin et al., 2018] and AAEs [Makhzani et al., 2015]—by incorporating entropy regularization [Cuturi, 2013]. This regularization yields an entropic objective with a closed-form Gibbs posterior, established via the Donsker–Varadhan representation [Donsker and Varadhan, 1975], which admits efficient sampling [Robert and Casella, 2004] and lets the model reconstruct data while aligning the aggregated posterior with the prior. Empirical results show that EVIA performs well across generative modeling tasks, including posterior estimation, variational inference and image inpainting.

1 INTRODUCTION

Variational autoencoding seeks a mapping between the data distribution p_x and an aggregated latent distribution q_z—a marginal on the latent space, defined in Equation (3) below—which is typically constrained to approximate a predefined prior p_z. Many methods have been proposed for this alignment, most prominently the Variational Autoencoder (VAE) [Kingma and Welling, 2014] and the Wasserstein Autoencoder (WAE) [Tolstikhin et al., 2018].

Since exact Bayesian marginalization is often computationally intractable, variational inference (VI) treats the task as an optimization problem over a variational distribution, minimizing a functional that balances reconstruction accuracy against a regularization term:

$$\mathcal{L}^{\mathrm{VI}}(q_{z,x}) \triangleq \mathbb{E}_{x\sim p_x}\big[\mathrm{KL}\big(q_{z|x}(\cdot \mid x) \,\|\, p_z\big)\big] + \frac{\lambda}{2}\,\mathbb{E}_{q_{z,x}}\big[-\log p_{\mathcal{A}}(x \mid z)\big], \qquad (1)$$

where $\mathcal{Q} \triangleq \{q \in \mathcal{P}(\mathcal{Z}\times\mathcal{X}) : (\pi_{\mathcal{X}})_{\#}q = p_x\}$ is the set of joint probability measures whose x-marginal is the data distribution, $\pi_{\mathcal{X}}$ denoting the coordinate projection. Every $q_{z,x} \in \mathcal{Q}$ disintegrates as $q_{z,x}(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x)\, q_{z|x}(\mathrm{d}z \mid x)$ for a Markov kernel $q_{z|x}$ that is unique up to a p_x-null set, so the objectives below do not depend on the chosen version of the kernel (Section A, Supplement). Finally, $p_{\mathcal{A}}(\cdot \mid z)$ is the density, with respect to a fixed reference measure on $\mathcal{X}$, of the observation law given the latent z.

The loss in Equation (1) balances two competing terms: the negative log-likelihood encourages accurate reconstruction, while the Kullback–Leibler (KL) divergence pulls the posterior toward the prior. This approach has widely recognized drawbacks. The KL divergence is prone to numerical instabilities and tends to be mode-seeking, which often results in poor coverage of the true posterior support. In addition, a parametric likelihood model restricts the generative process, limiting its ability to model the non-Euclidean data manifolds common in practice.

An alternative class of variational techniques, which we term Wasserstein inference (WI) [Tolstikhin et al., 2018], is formulated as follows:

$$\mathcal{L}^{\mathrm{WI}}(q_{z,x}) \triangleq \mathbb{D}(q_z \,\|\, p_z) + \frac{\lambda}{2}\mathbb{E}_{q_{z,x}}\big[c\big(x, \mathcal{A}(z)\big)\big], \qquad (2)$$

where $q_{z,x} \in \mathcal{Q}$ and

$$q_z \;\triangleq\; (\pi_{\mathcal{Z}})_{\#}q_{z,x} \;=\; \int_{\mathcal{X}} q_{z|x}(\cdot \mid x)\, p_x(\mathrm{d}x) \qquad (3)$$

is the *aggregated posterior*: the mixture of all conditional posteriors under the data distribution.

The structural difference between Equation (1) and Equation (2) lies in where the prior acts. VI regularizes *each*

[*]Corresponding author. Email: zlin@pku.edu.cn

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6701–6717.

conditional $q_{z|x}(\cdot \mid x)$ toward p_z through the per-sample KL term, so every posterior must individually overlap the prior—a requirement that pushes the conditionals toward one another and is a well-known driver of over-smoothed generations and *posterior collapse* (the encoder ceasing to use z). WI penalizes only the aggregated posterior of Equation (3): individual conditionals may remain well separated, preserving information about x, provided their *mixture* matches p_z [Makhzani et al., 2015, Tolstikhin et al., 2018]. In Equation (2), $\mathbb{D}$ is a chosen divergence enforcing this distribution-level alignment, and $c\big(x, \mathcal{A}(z)\big)$ is a transport cost measuring the discrepancy between an observation x and its reconstruction through the operator $\mathcal{A}$. Up to an overall rescaling, Equation (2) matches the WAE objective of Tolstikhin et al. [2018], in which the reconstruction term is unweighted and the penalty coefficient multiplies the latent divergence; we instead place the weight $\frac{\lambda}{2}$ on the reconstruction term to match the likelihood convention of Equation (1). For a deterministic decoder, minimizing the reconstruction term over Markov kernels $q_{z|x}$ whose aggregated posterior is held fixed at q_z recovers the optimal transport cost between p_x and $\mathcal{A}_{\#}q_z$ by Theorem 1 of Tolstikhin et al. [2018]; the theorem itself, however, is stated under the hard constraint $q_z = p_z$, which Equation (2) replaces with the divergence penalty $\mathbb{D}$. Once the entropic term is introduced below, we no longer invoke this equivalence and instead define all objectives directly on the joint distribution $q_{z,x}$.

However, standard implementations of WAE and AAE commonly use deterministic amortized encoders, which map each observation to a single latent code. Such point-valued inference cannot directly represent a genuinely multimodal conditional posterior, as may arise when multiple latent codes are consistent with the same observation. WIs also enforce distribution alignment through adversarial critics or kernel-based penalties, which brings optimization instability, sensitivity to hyperparameters, and reduced flexibility in high dimensions, particularly when the critic induces a highly nonconvex objective.

In this work, we extend the WI framework to the entropic case, proposing a method we term Entropic Wasserstein Inference (EWI). Formulated specifically for autoencoding, EWI augments the WI objective with an entropic penalty on the joint variational distribution:

$$\mathcal{L}^{\mathrm{EWI}}(q_{z,x}) \triangleq \mathcal{L}^{\mathrm{WI}}(q_{z,x}) + \gamma\,\mathrm{KL}(q_{z,x} \,\|\, \kappa_{z,x})\,, \quad (4)$$

where $\gamma > 0$ is a temperature parameter and $\kappa_{z,x} \in \mathcal{Q}$ is a reference joint distribution, $\kappa_{z,x}(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x)\,\kappa_{z|x}(\mathrm{d}z \mid x)$ with $\kappa_{z|x}$ a Markov kernel. In practice, we take $\kappa_{z|x}$ to be an isotropic Gaussian centered at a running latent estimate; see Equation (14). By convention, the penalty is set to $+\infty$ whenever $q_{z,x} \not\ll \kappa_{z,x}$. The additional KL term serves as an entropic regularizer, analogous to that in entropy-regularized optimal transport [Cuturi, 2013]. It smooths the inference distribution toward the reference coupling $\kappa_{z,x}$ and improves numerical stability.

Conceptually, this construction differs from Sinkhorn-based autoencoders [Patrini et al., 2020], which employ entropic optimal transport *in the latent space* as a drop-in replacement for the adversarial prior-matching term and evaluate it on minibatches through Sinkhorn iterations. In contrast, EWI regularizes the *joint* data–latent coupling itself: the entropic term acts directly on the inference distribution $q_{z,x}$, which, as we show below, yields a closed-form Gibbs posterior amenable to efficient gradient-based sampling rather than a fixed-point matrix-scaling procedure.

Motivated by this formulation, we introduce **E**ntropic **V**ariational **I**nference **A**uto-encoding (EVIA), which yields a *generalized posterior* determined by the optimal transport cost and entropy regularization. The entropic term is what turns inference from point estimation into distribution estimation: the optimal conditional is a closed-form Gibbs posterior, the encoder only supplies an amortized anchor—the reference-kernel center that tethers it—and the critic shapes an energy landscape over the latent space.

In summary, we propose an entropic variational inference framework for autoencoding that extends the WAE formulation [Tolstikhin et al., 2018] through a proper design of the latent discrepancy term $\mathbb{D}$. Our main contributions are summarized as follows:

1. We expand the WAE framework with an entropic regularization term, so that the optimal inference distribution is a continuous Gibbs posterior rather than a deterministic point. When several latent explanations are consistent with the same observation—the typical situation for a many-to-one decoder—EVIA recovers the set of plausible codes where a deterministic encoder must commit to one (Section C, Supplement).
2. We propose a training strategy driven by Stochastic Gradient Langevin Dynamics (SGLD) sampling. Trained on these samples, the critic learns an energy landscape that captures the structure of the prior rather than a mere decision boundary.
3. We further show that taking expectations over the sampling dynamics induces an objective closely related to Entropy-SGD [Chaudhari et al., 2017], encouraging the generative model to avoid sharp or suboptimal minima and improving the stability of adversarial training.

Notation. Throughout the paper, $\mathcal{X}$ is a Borel subset of $\mathbb{R}^D$ and $\mathcal{Z} = \mathbb{R}^d$; all distributions ($p_x$, p_z, and the variational objects of this paper) are Borel probability measures, rather than identified with densities, and conditional distributions such as $q_{z|x}$ are Markov kernels. We write $f_{\#}\mu$ for the pushforward of a measure μ by a Borel map f, and set $\mathrm{KL}(\mu \,\|\, \nu) \triangleq \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu$ if $\mu \ll \nu$ and $+\infty$ otherwise. The decoder $\mathcal{A} : \mathcal{Z} \to \mathcal{X}$ is a fixed Borel map. In

expressions such as $\mathcal{G}(z; x)$, the semicolon separates the varied argument from arguments held fixed. The complete measure-theoretic conventions (kernel measurability, expectation conventions, disintegration, the entropy chain rule, and the Gibbs variational formula) are collected in Section A, Supplement. Readers primarily interested in the algorithm may informally treat all measures as densities.

2 RELATED WORK

2.1 VARIATIONAL AUTOENCODERS

VAEs introduced a framework for deep latent variable modeling by maximizing a variational lower bound on the log-likelihood of the data [Kingma and Welling, 2014, Rezende et al., 2014, Xu et al., 2026c,b]. They regularize the approximate posterior via a Kullback–Leibler divergence toward a prior, typically Gaussian. While this enables efficient amortized inference, it often leads to overly smooth generations and posterior collapse with high-capacity decoders. Several extensions have attempted to improve posterior flexibility and generative quality, including β-VAE [Higgins et al., 2017] and InfoVAE [Zhao et al., 2017].

2.2 OPTIMAL TRANSPORT

Optimal Transport (OT) provides a geometry-aware distance between probability distributions, with the Wasserstein distance being particularly attractive due to its weak topology and meaningful gradients [Villani, 2009, Wang et al., 2025a,b, Pan et al., 2026]. For a cost c, the OT cost between P and Q is $\inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(x,y)\sim\pi}[c(x, y)]$—the cheapest coupling of the two distributions—and the Wasserstein distances arise from $c = \|x - y\|^p$. The application of OT to generative modeling gained prominence with Wasserstein GAN (WGAN) [Arjovsky et al., 2017], which replaces Jensen–Shannon divergence with the Earth Mover's distance to stabilize adversarial training. Subsequent improvements such as WGAN-GP [Gulrajani et al., 2017] enforce Lipschitz constraints via gradient penalties. OT-based objectives provide better behaved gradients when distributions have low support overlap.

2.3 WASSERSTEIN AUTOENCODERS

WAEs [Tolstikhin et al., 2018] were introduced to connect VAEs [Kingma and Welling, 2014, Rezende et al., 2014] with OT. In contrast to VAEs, which optimize a KL-regularized evidence lower bound, WAE formulates generative modeling as minimizing a relaxed OT distance between the data distribution and the model distribution. Concretely, WAE replaces the per-sample KL penalty with a divergence penalty between the aggregated posterior q_z and a prescribed prior. This shift yields a transport-based objective that prioritizes distribution-level matching and often produces more geometrically structured latent representations.

2.4 GENERALIZED VARIATIONAL INFERENCE VIA OPTIMAL TRANSPORT

Sinkhorn AutoEncoders [Patrini et al., 2020] extend the WAE framework by substituting adversarial or MMD-based latent regularization with an explicit entropically regularized optimal transport objective, efficiently computed via the Sinkhorn algorithm [Cuturi, 2013].

Chi et al. [2024] propose a generalized variational inference framework grounded in optimal transport rather than KL divergence. In contrast to classical VI—which minimizes KL and is therefore susceptible to mode-seeking behavior and support mismatch—their OT-based approach quantifies discrepancies through transport geometry.

More recently, the E-SUOT framework [Chen et al., 2026b] revisits this line of work from a semi-dual optimal transport perspective. By incorporating entropic regularization, it derives an entropic Wasserstein objective under which, when coupled with an f-divergence, the optimal solution admits a closed-form Gibbs posterior. However, the approach is not specifically designed for autoencoding, suffers from considerable computational overhead due to its sampling procedure, and remains confined to f-divergence–based discrepancy measures, which can limit numerical stability.

3 FORMULATION

To formalize the procedure, we restrict the divergence $\mathbb{D}(\cdot \,\|\, \cdot)$ to a specific class, which we refer to as an *adversarial divergence*. In particular, we focus on the family of (f, Γ)-divergences [Birrell et al., 2022], and accordingly denote it by $\mathbb{D}_{f,\Gamma}(\cdot \,\|\, \cdot)$.

We then show that the resulting objective can be optimized using an SGLD scheme [Welling and Teh, 2011, Xu et al., 2026a, Chen et al., 2025, 2026a, Guo et al., 2026, Teng et al., 2025], leading to an entropy-regularized variational adversarial divergence. The resulting framework differs from prior approaches in its variational structure and in the use of entropy as a regularizer.

3.1 (f, Γ)-DIVERGENCE

The expression $\mathbb{D}_{f,\Gamma}(\cdot \,\|\, \cdot)$ quantifies the statistical mismatch between the model's latent marginal distribution and the prior. This formulation includes both Γ-divergences and f-divergences. Since direct minimization of this term is typically computationally intractable, we adopt the variational dual formulation proposed by Farnia and Tse [2018]. For

$P, Q \in \mathcal{P}(\mathcal{Z})$,

$$\mathbb{D}_{f,\Gamma}(P \,\|\, Q) \triangleq \sup_{w \in \mathcal{W}} \left\{ \mathbb{E}_{z \sim P}[w(z)] - \mathcal{F}_Q(w) \right\}, \quad (5)$$

where $\mathcal{W}$ is a convex, symmetric ($w \in \mathcal{W} \Rightarrow -w \in \mathcal{W}$) class of bounded Borel functions on $\mathcal{Z}$, each acting as a latent-space discriminator (also called a *critic* or *witness*), and $\mathcal{F}_Q$ characterizes the divergence type. This variational representation includes two widely used families of distribution discrepancies: f-divergences [Csiszár, 1967] and integral probability metrics [Müller, 1997]. The integral probability metrics (IPMs) are also commonly referred to as Γ-divergences. Both f-divergences and Γ-divergences can be recovered from Equation (5):

- f-divergences: Recovered when $\mathcal{F}_Q$ relates to the Fenchel conjugate f^*, appropriate for density matching (e.g. χ^2 and KL divergences).
- Γ-divergences: Recovered by setting $\mathcal{F}_Q(w) = \mathbb{E}_Q[w(z)]$. If we restrict $\mathcal{W}$ to 1-Lipschitz functions, $\mathbb{D}_{f,\Gamma}$ becomes the 1-Wasserstein distance (W_1), which underlies WGAN [Arjovsky et al., 2017].

f-divergence Classically, for a convex generator f with $f(1) = 0$ and $P \ll Q$, the f-divergence is $\mathbb{D}_f(P \,\|\, Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q$; the generator $f(t) = t \log t$ yields KL. By Fenchel–Legendre duality [Rockafellar, 1970], with $f^*(s) \triangleq \sup_t \{st - f(t)\}$ the convex conjugate of f, we work directly with the variational characterization, a supremum over bounded Borel witness functions $w : \mathcal{Z} \to \mathbb{R}$ with $\mathbb{E}_Q[f^*(w)]$ understood in $(-\infty, +\infty)$:

$$\mathbb{D}_f(P \parallel Q) \triangleq \sup_w \left\{ \mathbb{E}_P[w(z)] - \mathbb{E}_Q[f^*(w(z))] \right\}. \quad (6)$$

By invoking this variational representation and substituting the marginal penalty into our primal objective, we derive the semi-dual result established below.

Γ-divergence The class of Γ-divergences (or IPMs) [Müller, 1997] is defined by:

$$\mathbb{D}_\Gamma(P \parallel Q) \triangleq \sup_{w \in \mathcal{W}} \left\{ \mathbb{E}_P[w(z)] - \mathbb{E}_Q[w(z)] \right\}, \quad (7)$$

where $\mathcal{W}$ is a prescribed class of Borel functions integrable with respect to both P and Q; throughout we take it to be the witness class of Equation (5). Unlike f-divergences, which typically require density ratios and absolute continuity, Γ-divergences depend only on expectations and remain well-defined even when the supports of P and Q do not overlap. This property is often associated with more stable gradients in generative modeling.

3.2 ENTROPIC REGULARIZATION

Entropic regularization provides smoother gradients, robustness against outliers, and better sample complexity in high-dimensional settings [Genevay et al., 2019].

For a fixed critic w, each latent code z trades off the critic reward $w(z)$ against the reconstruction cost. We encode this trade-off in a utility function, in the spirit of the optimal information processing principle [Zellner, 1988]:

$$\mathcal{G}_w^{\mathcal{A}}(z; x) \triangleq w(z) - \frac{\lambda}{2} \|x - \mathcal{A}(z)\|_2^2. \quad (8)$$

Fix a temperature $\gamma > 0$ and a Markov kernel $\kappa(\cdot \mid x)$ from $\mathcal{X}$ to $\mathcal{Z}$, and assume the integrability condition

$$\int_{\mathcal{Z}} \exp\left(\frac{\mathcal{G}_w^{\mathcal{A}}(z; x)}{\gamma} \right) \kappa(\mathrm{d}z \mid x) < \infty \qquad \text{for } p_x\text{-a.e. } x. \quad (9)$$

We define the conditional supremum variational functional:

$$\Psi_w^{\mathcal{A}}(x; \gamma, \kappa) = \sup_{q \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{z \sim q}[\mathcal{G}_w^{\mathcal{A}}(z; x)] - \gamma \, \mathrm{KL}(q \,\|\, \kappa(\cdot \mid x)) \right\}. \quad (10)$$

By our KL convention, only $q \ll \kappa(\cdot \mid x)$ contribute to the supremum. By the Gibbs variational formula of Donsker–Varadhan type (Theorem 2; Donsker and Varadhan, 1975), under (9) the supremum equals $\gamma \log \int_{\mathcal{Z}} e^{\mathcal{G}_w^{\mathcal{A}}(z;x)/\gamma} \, \kappa(\mathrm{d}z \mid x)$ and is attained uniquely at the Gibbs measure whose density with respect to $\kappa(\cdot \mid x)$ is proportional to $e^{\mathcal{G}_w^{\mathcal{A}}(\cdot;x)/\gamma}$— the reference exponentially tilted toward high utility, the familiar softmax trade-off at temperature γ; in particular, $x \mapsto \Psi_w^{\mathcal{A}}(x; \gamma, \kappa)$ is Borel measurable. For bounded w, condition (9) in fact holds for every x, since $\mathcal{G}_w^{\mathcal{A}} \le \|w\|_\infty$.

3.3 EVIA OBJECTIVE

EVIA optimizes the joint variational distribution $q_{z,x} \in \mathcal{Q}$, which trades off reconstruction accuracy against consistency with the prior. The prior term is imposed through a general (f, Γ)-divergence regularizer [Birrell et al., 2022]. Combining this discrepancy with the entropically regularized reconstruction cost gives the EVIA primal objective:

$$\mathcal{L}^{\text{EVIA-Primal}}(q_{z,x}) \triangleq \mathbb{D}_{f,\Gamma}(q_z \| p_z) + \frac{\lambda}{2} \mathbb{E}_{q_{z,x}}\left[\|x - \mathcal{A}(z)\|_2^2 \right] + \gamma \, \mathrm{KL}(q_{z,x} \,\|\, \kappa_{z,x}), \quad (11)$$

where $q_z \triangleq (\pi_{\mathcal{Z}})_{\#} q_{z,x}$ is the aggregated posterior and $\kappa_{z,x}(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x) \, \kappa(\mathrm{d}z \mid x)$ is the reference joint distribution built from the kernel of Equation (9). Optimizing this primal objective directly is intractable, since the adversarial divergence couples all conditional distributions $q_{z|x}$ through the marginal q_z. We instead pass to its semi-dual formulation.

Theorem 1 (Unified Semi-Dual of EVIA). *Let $\mathcal{Q} = \{q \in \mathcal{P}(\mathcal{Z} \times \mathcal{X}) : (\pi_{\mathcal{X}})_{\#} q = p_x\}$ and let $\mathcal{W}$ be a convex, symmetric class of bounded Borel functions on $\mathcal{Z}$ with $\mathcal{F}_{p_z}^*(w) < \infty$ for every $w \in \mathcal{W}$. Under the compactness*

assumption (A1) stated in Section A, strong duality holds for the EVIA primal objective of Equation (11)*:*

$$\inf_{q_{z,x}\in\mathcal{Q}} \mathcal{L}^{\text{EVIA-Primal}}(q_{z,x}) = \sup_{w\in\mathcal{W}} \Big[-\mathcal{F}^*_{p_z}(w) - \mathbb{E}_{x\sim p_x}\Psi^{\mathcal{A}}_w(x;\gamma,\kappa)\Big], \quad (12)$$

where $\mathcal{F}^*_{p_z}$ *is the* variational dual functional *unifying the discrepancy families:*

$$\mathcal{F}^*_{p_z}(w) \triangleq \begin{cases} \mathbb{E}_{z\sim p_z}[f^*(-w(z))], & \text{if } f\text{-divergence}, \\ -\mathbb{E}_{z\sim p_z}[w(z)], & \text{if } \Gamma\text{-divergence}, \end{cases} \quad (13)$$

and the regularized potential $\Psi^{\mathcal{A}}_w(x;\gamma,\kappa)$ *is defined as in Equation* (10)*. The star and negated argument in* $\mathcal{F}^*_{p_z}$ *merely record the substitution* $T=-w$ *unifying the two branches (Section A, Step 2)—it is not a Fenchel conjugate—and the representation is a* semi-*dual because only the divergence is dualized, the transport–entropy part being solved in closed form by* $\Psi^{\mathcal{A}}_w$*. In the* Γ*-branch, for instance,* $-\mathcal{F}^*_{p_z}(w) = \mathbb{E}_{p_z}[w]$*, so maximizing over* w *pushes the critic up on prior samples and (through* $\Psi^{\mathcal{A}}_w$*) down on Gibbs samples. We write* $\mathcal{L}^{\text{EVIA-SemiDual}}$ *for the right-hand side of* (12)*. The complete proof is deferred to Section A, Supplement.*

The semi-dual form becomes tractable for an isotropic Gaussian reference kernel $\kappa(\cdot \mid x) = \mathcal{N}\big(\bar{z}(x), \frac{1}{\rho}I\big)$, $\bar{z}:\mathcal{X}\to\mathcal{Z}$ Borel; during training, $\bar{z}(x)$ is the running mean of the Langevin particles for x (Algorithm 1; the amortized variant uses $\bar{z}=q_\phi(x)$). We fix $\gamma=1$ (general γ merely rescales w, λ, and the potential in the Γ-branch used below; in the f-branch it additionally rescales the generator, via $\gamma^{-1}\mathbb{D}_f = D_{f/\gamma}$), so the sampling is controlled by ρ and λ. Substituting the Gaussian log-density $-\frac{\rho}{2}\|z-\bar{z}(x)\|_2^2 + \text{const}$ into Equation (10) and dropping the constant, the conditional potential reduces to a continuous *Log-Sum-Exp*:

$$\Psi(x;\rho,\bar{z}) \triangleq \log\int_{\mathcal{Z}} \exp\Big\{w(z) - \frac{\lambda}{2}\|x-\mathcal{A}(z)\|_2^2 - \frac{\rho}{2}\|z-\bar{z}\|_2^2\Big\}\mathrm{d}z, \quad (14)$$

where $\mathrm{d}z$ is Lebesgue measure on $\mathcal{Z}$ and the integral is finite for every bounded w. The integrand is, up to normalization, the Lebesgue density of the Gibbs measure $\tilde{q}(\cdot \mid x)$—equivalently, $\frac{\mathrm{d}\tilde{q}(\cdot|x)}{\mathrm{d}\kappa(\cdot|x)} \propto e^{\mathcal{G}^{\mathcal{A}}_w(\cdot;x)}$:

$$\frac{\mathrm{d}\tilde{q}(\cdot \mid x)}{\mathrm{d}z}(z) \propto \exp\Big\{w(z) - \frac{\lambda}{2}\|x-\mathcal{A}(z)\|_2^2 - \frac{\rho}{2}\|z-\bar{z}\|_2^2\Big\}, \quad (15)$$

so the critic function w can be optimized directly: the gradients of the *Log-Sum-Exp* operator in Equation (14) are expectations under $\tilde{q}(z \mid x)$ and can be estimated from *negative* samples (SGLD draws from $\tilde{q}(\cdot \mid x)$; *positive* samples are draws from the prior p_z) [Welling and Teh, 2011]—exactly the training of an energy-based model with the critic as its *negative* energy [Grathwohl et al., 2020], matching the convention $G=-\mathcal{G}^{\mathcal{A}}_w$ below.

3.4 COMPARISON OF WAE AND EVIA

Equation (12) makes the adversarial structure of EVIA explicit, with the optimal conditional potential $\Psi^{\mathcal{A}}_w(x;\gamma,\kappa)$ incorporating the decoder cost into the latent energy. From both optimization and sampling perspectives, WAE and EVIA then differ in a basic way, which we describe for a fixed critic w:

(1) WAE commonly seeks a single latent code z that maximizes the utility $\mathcal{G}^{\mathcal{A}}_w(z;x)$—equivalently, minimizes the induced energy $-\mathcal{G}^{\mathcal{A}}_w(z;x)$. EVIA instead optimizes over a distribution via $\Psi^{\mathcal{A}}_w(x;\gamma,\kappa)$, as defined in Equation (10): as $\gamma\to 0$ the Gibbs measure collapses to a point mass at $\arg\max_z \mathcal{G}^{\mathcal{A}}_w(z;x)$, recovering a deterministic, WAE-style encoder, whereas EVIA keeps $\gamma>0$.

(2) Let $\bar{z}$ denote the mean of samples drawn from optimal q at the previous iteration. The use of a local Gibbs average is conceptually related to the local-entropy smoothing employed by Entropy-SGD [Chaudhari et al., 2017], although here the smoothing acts in latent space rather than parameter space, whereas WAE reduces to standard gradient descent [LeCun et al., 2012]. Consequently, EVIA tends to favor wider minima and is able to traverse energy barriers, while WAE follows sharper descent directions (Figure 1).

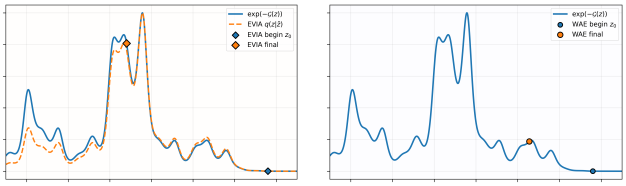


Figure 1: Comparison of EVIA (left) and WAE (right) under optimization of a fixed energy landscape $G=-\mathcal{G}^{\mathcal{A}}_w$, visualized through its Gibbs density $\exp(-G)$. WAE converges to a sharp local minimum of the energy, whereas EVIA prefers a flatter optimum. EVIA also models a conditional distribution rather than a single point.

4 ALGORITHM

In this section, we focus on variational inference under the W_1 metric. Wasserstein-based adversarial objectives commonly enforce approximate Lipschitz control through gradient penalties or spectral normalization [Gulrajani et al., 2017, Miyato et al., 2018]. In practice, these models commonly rely on WGAN [Arjovsky et al., 2017] variants that enforce the Lipschitz constraint through either gradient penalty [Gulrajani et al., 2017] or spectral normalization [Miyato et al., 2018].

We parameterize the potential $w\in\mathcal{W}$ using a neural network w_ψ, optimized within the dual framework of Equation (12). To enforce the 1-Lipschitz constraint on w_ψ, we employ a regularization strategy, denoted generally by

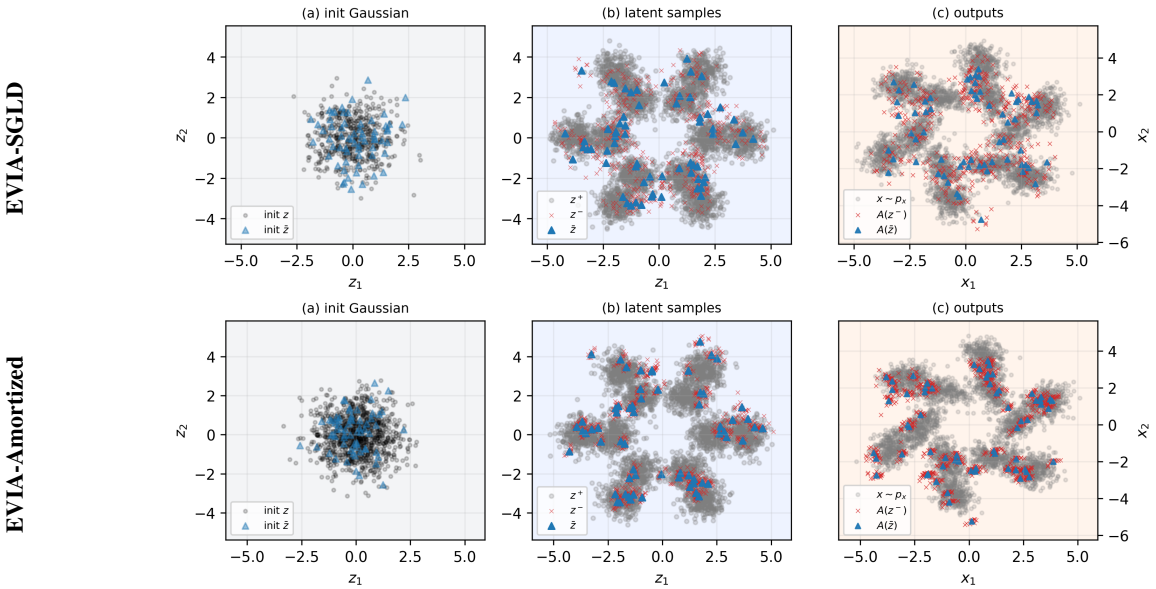


Figure 2: Latent and observation space analysis for a 6-mode Gaussian mixture. **Left:** Gaussian initialization of the latent samples z and centers $\bar{z}$. **Middle:** Alignment between Langevin samples z^-, latent centers $\bar{z}$, and the prior p_z. **Right:** Reconstructions of observations x.

$\mathcal{R}_{\mathrm{Lip}}(\psi)$. The objective with W_1 critic is:

$$\mathcal{L}_\psi^{W_1} = -\mathbb{E}_{z\sim p_z}\big[w_\psi(z)\big] + \mathbb{E}_{x\sim p_x}\big[\Psi(x;\rho,\bar{z})\big] + \mathcal{R}_{\mathrm{Lip}}(\psi), \tag{16}$$

where $\Psi(x;\rho,\bar{z})$ is the smoothed transport potential. Minimizing this loss corresponds to the maximization of the dual form in Equation (12) augmented with the Lipschitz regularization term $\mathcal{R}_{\mathrm{Lip}}$.

With gradient penalty regularization [Gulrajani et al., 2017], the Lipschitz regularizer is set as $\mathcal{R}_{\mathrm{Lip}}(\psi) = \lambda_{\mathrm{gp}}\,\mathbb{E}_{\hat{z}}\big[(\|\nabla w(\hat{z})\|_2 - 1)^2\big]$, with $\hat{z}$ drawn on random interpolations between prior and posterior samples (Section B.1); under spectral normalization [Miyato et al., 2018], we instead set $\mathcal{R}_{\mathrm{Lip}}(\psi) \equiv 0$, as the Lipschitz constraint is directly enforced via layer-wise weight normalization without explicit penalty term.

4.1 ADVERSARIAL TRAINING VIA SGLD

In practice, we estimate the objective using a mini-batch of n independent pairs $\{(x_i, z_i^+)\}_{i=1}^n$ drawn from the data distribution p_x and the prior p_z. For each observation x_i, we approximate the transport term using m samples $\{z_{ij}^-\}_{j=1}^m$, generated via SGLD from the Gibbs distribution $\tilde{q}(z \mid x_i)$ in Equation (15). Thus w learns from both positive samples drawn from p_z and negative samples produced by SGLD. The complete algorithm is in Algorithm 1; it treats the operator $\mathcal{A}$ as fixed and known (the inverse-problem setting illustrated below), while Algorithm 2 learns $\mathcal{A}_\theta$ jointly. The buffer $\mathcal{B}$ is a persistent particle store (persistent-chain MCMC): each iteration warm-starts its chains from previous round's particles, initialized from prior on the first round.

Algorithm 1 EVIA-SGLD Algorithm

Input: Batch size n, particles per sample m, SGLD steps T, sampling step size η, learning rate ϵ.
Init: parameters ψ, buffer $\mathcal{B}$.
repeat
  Sample batch $\{(x_i, z_i^+)\}_{i=1}^n$ from $p_x \times p_z$.
  Initialize particles $z_{ij} \leftarrow z_{ij}^-$ from $\mathcal{B}$ for each i, j, and set $\bar{z}_i = \frac{1}{m}\sum_{j=1}^m z_{ij}^-$ for each i.
  for $t = 1$ **to** T (in parallel for all i, j) **do**
    Sample $\xi \sim \mathcal{N}(0, I)$. ▷ Generate negative samples from the Gibbs density in Eq. (15)
    $z_{ij} \leftarrow z_{ij} + \eta\,\nabla_z \log \tilde{q}(z_{ij} \mid x_i) + \sqrt{2\eta}\,\xi$.
  end for
  Update buffer $\mathcal{B}$ by storing final $\{z_{ij}\}$ as $\{z_{ij}^-\}$.
  $\psi \leftarrow \psi - \epsilon\nabla_\psi \mathcal{L}_\psi^{W_1}$ via Eq. Equation (16).
until convergence

4.2 JOINT AMORTIZED MAPPINGS

Beyond optimizing the divergence functional, our objective includes learning the mapping $x \mapsto \mathbb{E}_{z\sim\tilde{q}(\cdot|x)}[z]$, a smoothed latent representation induced by the entropy-regularized posterior dynamics. We optimize the parameters of an amortized encoder q_ϕ by minimizing the expected squared distance between the network's predictions and the empirical targets:

$$\textit{Encoder:}\quad \mathcal{L}_\phi = \mathbb{E}_{x\sim p_x,\, z\sim\tilde{q}(\cdot|x)}\Big[\big\|q_\phi(x) - z\big\|_2^2\Big]. \tag{17}$$

When the forward operator $\mathcal{A}$ is unknown, we parameterize it with a neural network decoder $\mathcal{A}_\theta$, optimized jointly with the encoder and the energy-based prior by minimizing the

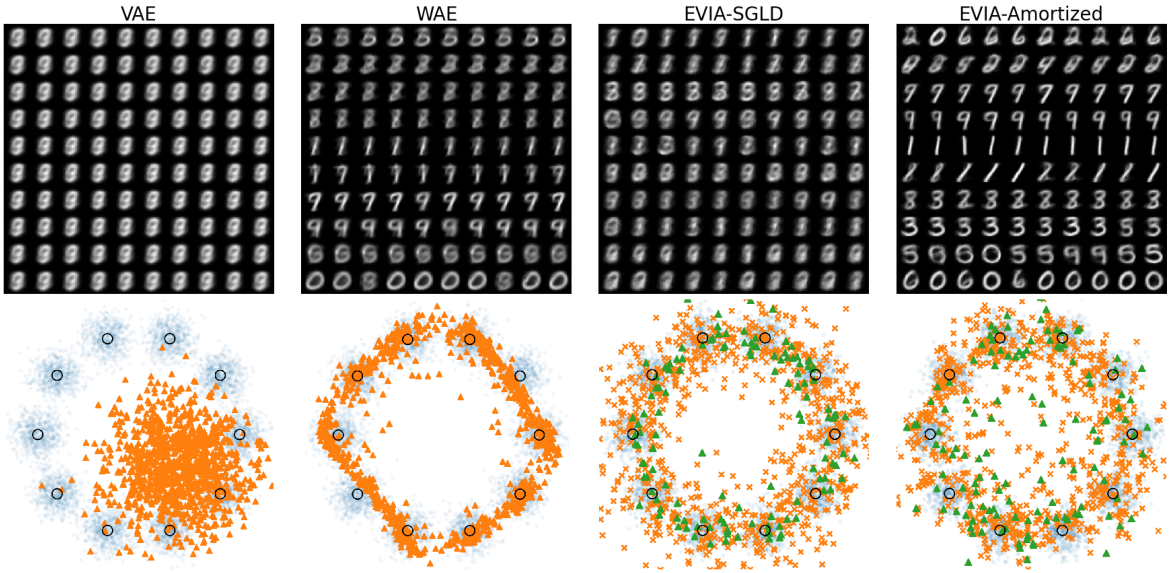


Figure 3: Comparison of generated samples from VAE, WAE and EVIA on MNIST. The orange markers denote sampled latent codes, while the green markers indicate the corresponding latent means. EVIA-Amortized learns a structured posterior with clear mode separation, avoids evident collapse, and produces higher-quality generated images.

expected reconstruction error:

$$\textit{Decoder:} \quad \mathcal{L}_\theta = \mathbb{E}_{x\sim p_x,\, z\sim\tilde{q}(\cdot|x)}\left[\|\mathcal{A}_\theta(z) - x\|_2^2\right], \quad (18)$$

where $\tilde{q}(\cdot \mid x)$ is the Gibbs measure of Equation (15). To enforce consistency between the encoder q_ϕ and the decoder $\mathcal{A}_\theta$, we combine these terms into a single objective:

$$\mathcal{L}_{\phi,\theta} = \mathbb{E}_{x\sim p_x}\left[\|\mathcal{A}_\theta(q_\phi(x)) - x\|_2^2\right] + \alpha\mathcal{L}_\phi + \beta\mathcal{L}_\theta. \quad (19)$$

In this case, $\Psi(x;\rho,\bar{z})$ can be written as

$$\Psi_{\phi,\theta}(x;\rho) = \log \int_{\mathcal{Z}} \exp\Big\{ w(z) - \frac{\lambda}{2}\|x - \mathcal{A}_\theta(z)\|_2^2 - \frac{\rho}{2}\|z - q_\phi(x)\|_2^2 \Big\} \mathrm{d}z, \quad (20)$$

with the corresponding Gibbs density $\tilde{q}_{\phi,\theta}(z \mid x)$ defined, as in Equation (15), by the integrand. The complete algorithm is shown in Algorithm 2.

We consider an inverse inference task in which the forward operator $\mathcal{A}$ is fixed and known, given by a smooth rotational mapping in $\mathbb{R}^2$. Observations are generated by pushforward sampling, $x = \mathcal{A}(z)$ with $z \sim p_z$, inducing the observation distribution $p_x = \mathcal{A}_\#(p_z)$. Given i.i.d. samples $\{x_i\} \sim p_x$, our goal is to learn a conditional inference model $q(\cdot \mid x)$ that (i) produces latent codes whose forward images reconstruct the observations (i.e., $\mathcal{A}(z) \approx x$) and (ii) remains compatible with the multimodal prior p_z.

As illustrated in Figure 2 for a 6-mode Gaussian mixture, both EVIA-SGLD (top) and EVIA-Amortized (bottom) recover the expected clustered latent structure: the inferred samples z^- populate the prior modes and their per-sample centers $\bar{z}$ align with these regions, resulting in reconstructions $\mathcal{A}(z^-)$ and $\mathcal{A}(\bar{z})$ that closely match p_x.

Algorithm 2 EVIA-Amortized Algorithm

Input: Batch size n, particles per sample m, SGLD steps T, sampling step size η, learning rate ϵ.
Init: parameters (ψ, θ, ϕ), buffer $\mathcal{B}$.
repeat
  Sample batch $\{(x_i, z_i^+)\}_{i=1}^n$ from $p_x \times p_z$.
  Initialize particles $z_{ij} \leftarrow z_{ij}^-$ from $\mathcal{B}$ for each i, j.
  for $t = 1$ **to** T (in parallel for all i, j) **do**
    Sample $\xi \sim \mathcal{N}(0, I)$. ▷ Generate negative samples from the Gibbs density of Eq. (20)
    $z_{ij} \leftarrow z_{ij} + \eta\, \nabla_z \log \tilde{q}_{\phi,\theta}(z_{ij} \mid x_i) + \sqrt{2\eta}\,\xi$.
  end for
  Update buffer $\mathcal{B}$ by storing final $\{z_{ij}\}$ as $\{z_{ij}^-\}$.
  $\psi \leftarrow \psi - \epsilon \nabla_\psi \mathcal{L}_\psi^{W_1}$ via Eq. (16).
  $(\theta, \phi) \leftarrow (\theta, \phi) - \epsilon\, \nabla_{\theta,\phi} \mathcal{L}_{\phi,\theta}$ via Eq. (19).
until convergence

5 EXPERIMENTS

For MNIST, Fashion MNIST, and CelebA, WAE and EVIA use a common latent-space WGAN-GP critic. CIFAR inpainting instead uses a spectrally normalized image critic with hinge loss. We evaluate a VAE baseline together with models from the WAE family, regarding AAE as the adversarially regularized special case in which the aggregated posterior is matched to the prior. To isolate the role of latent inference, we compare the standard amortized encoder used by WAE with two EVIA variants: *EVIA-SGLD*, which refines persistent latent particles through short-run SGLD, and *EVIA-Amortized*, which trains an encoder under the critic-induced energy landscape. For a controlled comparison,

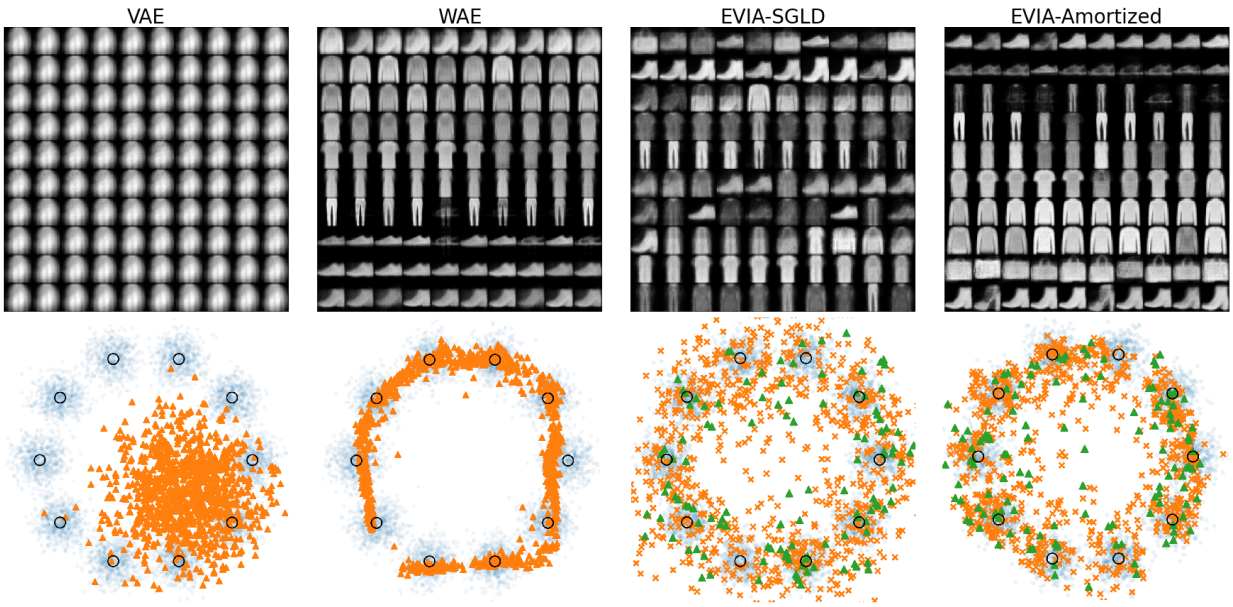


Figure 4: Comparison of generated samples from VAE, WAE and EVIA on Fashion MNIST. The orange markers denote sampled latent codes, while the green markers indicate the corresponding latent means. EVIA-Amortized learns a structured posterior with clear mode separation, avoids evident collapse, and produces higher-quality generated images.

WAE and both EVIA variants share the same convolutional encoder and decoder, latent-space MLP critic, reconstruction loss, and WGAN-GP objective. The models share network backbones and reconstruction losses, while inference procedures and critic-update schedules follow their respective training protocols. The VAE uses the same encoder and decoder backbones, with additional heads parameterizing the Gaussian mean and log-variance. Architectural and training details are provided in Section B, Supplement.

We evaluate these models on three generative benchmarks covering different data scales and inference settings. First, we study *2D manifold synthesis* on synthetic and low-dimensional datasets, enabling direct visualization of the learned latent geometry, mode coverage, and distributional fidelity. Second, we consider *unconditional image generation* on CelebA to evaluate global sample quality, visual fidelity, and diversity in facial attributes. Third, we examine *conditional image inpainting* on CIFAR-10 and CIFAR-100, which tests each inference paradigm's ability to approximate conditional distributions and preserve semantic consistency under observed-pixel constraints.

5.1 2D MANIFOLD SYNTHESIS

MNIST and Fashion MNIST. We evaluate VAE, WAE, EVIA-SGLD, and EVIA-Amortized on MNIST [LeCun and Cortes, 2010] and Fashion MNIST [Xiao et al., 2017] using a two-dimensional latent space, which enables direct visualization of the learned geometry.

Figures 3 and 4 show the resulting latent representations and decoded samples. For WAE and the EVIA variants, the architecture and WGAN-GP objective are held fixed, so the differences primarily reflect the inference procedure: WAE uses standard amortized inference, EVIA-SGLD iteratively refines latent particles, and EVIA-Amortized learns an encoder under the critic-induced energy landscape. Qualitatively, EVIA-Amortized yields the most coherent latent organization, broadest mode coverage, and greatest sample diversity, possibly because amortization encourages semantically similar samples to occupy nearby latent regions.

Table 1: Average reconstruction errors on CIFAR-10 and CIFAR-100 masked inpainting. We report the total ℓ_1 loss and sum of squared errors (SSE) computed over the masked regions **(lower is better)**.

	CIFAR-10		CIFAR-100	
Methods	$\ell_1 \downarrow$	SSE $\downarrow$	$\ell_1 \downarrow$	SSE $\downarrow$
WAE	172.386	80.549	174.148	81.873
EVIA	**170.234**	**79.805**	**170.711**	**81.351**

5.2 UNCONDITIONAL IMAGE GENERATION

CelebA. We evaluate VAE, WAE, EVIA-SGLD, and EVIA-Amortized on CelebA [Liu et al., 2015]. All images are center-cropped and resized to 64×64, and a 128-dimensional Gaussian prior is used in the latent space.

As shown in Figure 5, all four methods generate visually plausible and diverse facial images, with broadly comparable sample quality. VAE tends to produce slightly smoother outputs, while WAE and the two EVIA variants exhibit somewhat sharper details but also occasional artifacts. Overall, the qualitative differences are modest, indicating that the EVIA inference mechanisms achieve sample quality comparable to the standard amortized baselines.

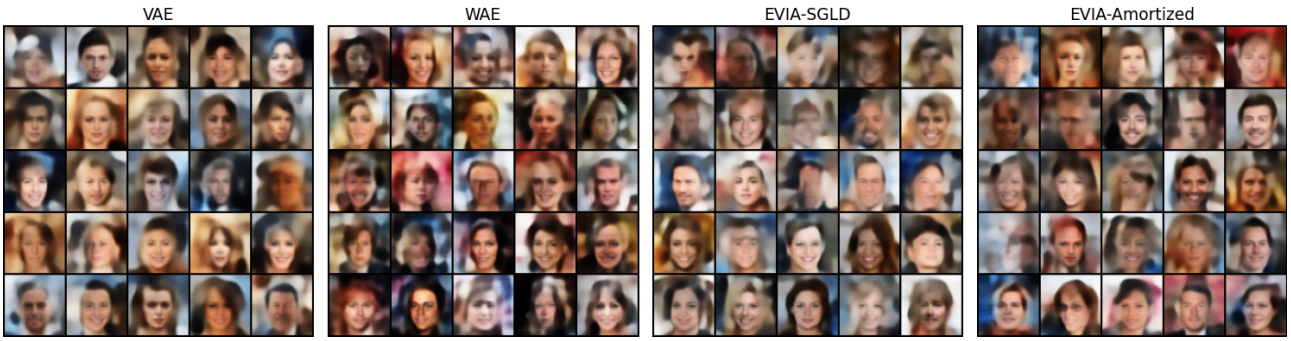
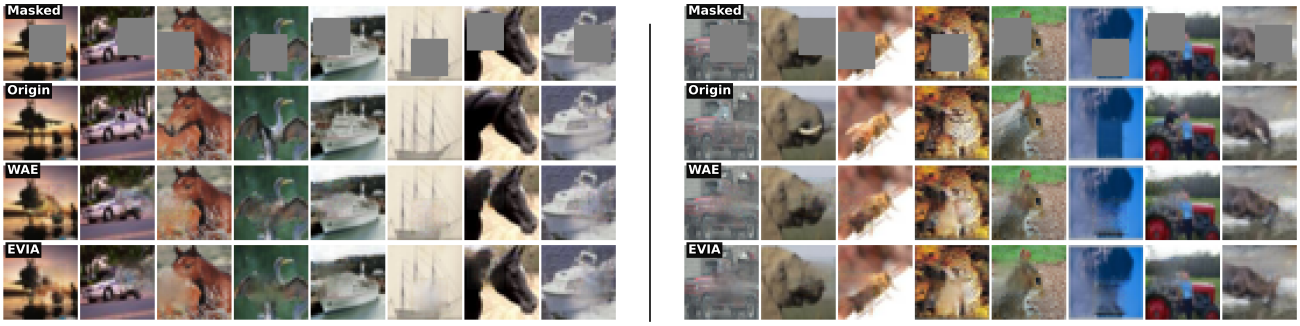


Figure 5: Comparison of generated samples from VAE, WAE, EVIA-SGLD, and EVIA-Amortized on the CelebA dataset.



(a) CIFAR-10

(b) CIFAR-100

Figure 6: Comparison of inpainted results from WAE and EVIA on the CIFAR-10 and CIFAR-100 datasets.

5.3 CONDITIONAL IMAGE INPAINTING

We evaluate WAE and EVIA-Amortized on image inpainting using CIFAR-10 and CIFAR-100 [Krizhevsky, 2009]. Because EVIA-SGLD requires maintaining persistent particle statistics for individual data points, its storage cost becomes prohibitive on larger datasets. Moreover, as observed in Section 5.1, the amortized encoder encourages observations with similar semantic structure to occupy nearby regions of the inferred space, which is particularly beneficial for conditional reconstruction. We therefore consider only EVIA-Amortized in this experiment and refer to it simply as *EVIA* throughout this subsection.

Each 32×32 image is normalized from $[0, 1]$ to $[-1, 1]$ to match the generator's `tanh` output. We then mask a randomly positioned 16×16 square region. The generator takes as input the masked image, the corresponding binary keep-mask, and Gaussian noise restricted to the missing region. Its prediction is combined with the observed pixels to form the completed image.

As shown in Figure 6, EVIA produces sharper and more structurally coherent completions than WAE, which more frequently exhibits blurred textures and inconsistent object boundaries. The quantitative results in Table 1 support this observation: EVIA achieves lower ℓ_1 and SSE reconstruction errors on both CIFAR-10 and CIFAR-100.

6 CONCLUSION AND FUTURE WORK

EVIA extends the deterministic WAE baseline considered in this work by replacing point-valued amortized inference with entropy-regularized distributional inference. Its semi-dual formulation supports variational discrepancy measures spanning f-divergences and integral probability metrics, while Langevin refinement enables the model to represent structured conditional distributions in latent space. Although our implementation uses SGLD, the framework is compatible with other MCMC and particle-based samplers. Future work will investigate more scalable alternatives, including amortized sampling [Grathwohl et al., 2021] and parallel particle-based inference [Liu and Wang, 2016].

Acknowledgements

Zhouchen Lin was supported by the Beijing Natural Science Foundation under Grant No. L257007, the National Natural Science Foundation of China under Grant No. 62276004, and the Beijing Major Science and Technology Project under Grant No. Z251100008425006. Yunfei Teng was supported by Beijing Postdoctoral Research Foundation. Zhichao Chen was supported by the China Postdoctoral Science Foundation under Grant No. 2025M781449. Sixin Zhang was supported by the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme "Investissements d'Avenir." The authors acknowledge the anonymous reviewers for their insightful comments and suggestions.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. Int. Conf. Mach. Learn.*, volume 70, pages 214–223, 2017.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *J. Mach. Learn. Res.*, 23(39):1–70, 2022.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT 2010*, pages 177–186. Physica-Verlag, 2010. doi: 10.1007/978-3-7908-2604-3_16.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *Proc. Int. Conf. Learn. Represent.*, 2017.

Zhichao Chen, Licheng Pan, Yiran Ma, Zeyu Yang, Le Yao, Jinchuan Qian, and Zhihuan Song. E^2AG: Entropy-regularized ensemble adaptive graph for industrial soft sensor modeling. *IEEE/CAA J. Autom. Sinica*, 12(4): 745–760, 2025.

Zhichao Chen, Yulong Zhang, Odin Zhang, Fangyikang Wang, Le Yao, Hao Wang, and Zhihuan Song. Blending data and knowledge for process industrial modeling under Riemannian preconditioned Bayesian framework. *IEEE Trans. Knowl. Data Eng.*, 38(1):82–95, 2026a.

Zhichao Chen, Zhan Zhuang, Yunfei Teng, Hao Wang, Fangyikang Wang, Zhengnan Li, Tianqiao Liu, Haoxuan Li, and Zhouchen Lin. Rethinking the flow-based gradual domain adaptation: A semi-dual optimal transport perspective. In *Proc. Int. Conf. Mach. Learn.*, pages 1–45, 2026b.

Jinjin Chi, Zhichao Zhang, Zhiyao Yang, Jihong Ouyang, and Hongbin Pei. Generalized variational inference via optimal transport. In *Proc. AAAI Conf. Artif. Intell.*, pages 1287–1295, 2024.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hung.*, 2:299–318, 1967.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 26, pages 2292–2300, 2013.

Monroe D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, I. *Commun. Pure Appl. Math.*, 28(1):1–47, 1975.

Paul Dupuis and Richard S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

Farzan Farnia and David Tse. A convex duality framework for GANs. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 31, pages 5254–5263, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *Proc. Int. Conf. Artif. Intell. Statist.*, volume 89, pages 1574–1583. PMLR, 2019.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy-based model and you should treat it like one. In *Proc. Int. Conf. Learn. Represent.*, 2020.

Will Sussman Grathwohl, Jacob J. Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No MCMC for me: Amortized sampling for fast and stable training of energy-based models. In *Proc. Int. Conf. Learn. Represent.*, 2021.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, pages 5767–5777, 2017.

Wei Guo, Jaemoo Choi, Yuchen Zhu, Molei Tao, and Yongxin Chen. Proximal diffusion neural sampler. In *Proc. Int. Conf. Learn. Represent.*, 2026.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proc. Int. Conf. Learn. Represent.*, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *Proc. Int. Conf. Learn. Represent.*, 2017.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, Cham, 3rd edition, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proc. Int. Conf. Learn. Represent.*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. Online database, 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, pages 9–48. Springer, 2012.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general-purpose Bayesian inference algorithm.

In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pages 2378–2386, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3730–3738, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learn. Represent.*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. Int. Conf. Learn. Represent.*, 2017.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders, 2015. URL https://arxiv.org/abs/1511.05644.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. Int. Conf. Learn. Represent.*, 2018.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Adv. Appl. Probab.*, 29(2): 429–443, 1997.

Licheng Pan, Haocheng Yang, Haoxuan Li, Yunsheng Lu, Yongqi Tong, Yinuo Wang, Shijian Wang, Zhixuan Chu, Lei Shen, Yuan Lu, and Hao Wang. Optimal transport for LLM reward modeling from noisy preference. In *Proc. Int. Conf. Mach. Learn.*, pages 1–21, 2026.

Giorgio Patrini, Rianne van den Berg, Patrick Forré, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Proc. Uncertain. Artif. Intell.*, volume 115, pages 733–743. PMLR, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. Int. Conf. Mach. Learn.*, volume 32, pages 1278–1286, 2014.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

Maurice Sion. On general minimax theorems. *Pac. J. Math.*, 8(1):171–176, 1958.

Yunfei Teng, Wenbo Gao, François Chalus, Anna Choromanska, Donald Goldfarb, and Adrian Weller. Leader stochastic gradient descent for distributed training of deep learning models. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 32, pages 9821–9831, 2019.

Yunfei Teng, Sixin Zhang, Yao Li, Kai Chen, Di He, and Qiwei Ye. Follow Hamiltonian leader: An efficient energy-guided sampling method. In *Frontiers in Probabilistic Inference: Learning Meets Sampling Workshop at ICLR*, 2025. URL https://openreview.net/forum?id=EM75aI3mAs.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *Proc. Int. Conf. Learn. Represent.*, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, Germany, 2009.

Hao Wang, Zhichao Chen, Honglei Zhang, Zhengnan Li, Licheng Pan, Haoxuan Li, and Mingming Gong. Debiased recommendation via Wasserstein causal balancing. *ACM Trans. Inf. Syst.*, 43(6):1–24, 2025a.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zhichao Chen, Haoxuan Li, Qingsong Wen, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution Wasserstein alignment. In *Proc. Int. Conf. Learn. Represent.*, 2025b.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. Int. Conf. Mach. Learn.*, pages 681–688, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Jian Xu, Shigui Li, Wei Chen, Jiacheng Li, Zhiqi Lin, Delu Zeng, Xinghao Ding, John Paisley, and Qibin Zhao. Implicit variational rejection sampling. In *Proc. Uncertain. Artif. Intell.*, pages 1–19, 2026a.

Jian Xu, Delu Zeng, and John Paisley. Sparse variational Student-t processes for heavy-tailed modeling. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–14, 2026b.

Jian Xu, Qibin Zhao, John Paisley, and Delu Zeng. Diffusion bridge variational inference for deep gaussian processes. In *Proc. Int. Conf. Learn. Represent.*, 2026c.

Arnold Zellner. Optimal information processing and Bayes's theorem. *Am. Stat.*, 42(4):278–280, 1988.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. URL https://arxiv.org/abs/1706.02262.

EVIA: Entropic Variational Inference Auto-encoding (Supplementary Material)

A PROOF OF THE MAIN THEOREM

Conventions and auxiliary results. Throughout, $\mathcal{X}$ is a Borel subset of $\mathbb{R}^D$, $\mathcal{Z} = \mathbb{R}^d$, both carry their Borel σ-algebras, and $\mathcal{P}(\mathcal{S})$ denotes the Borel probability measures on a space $\mathcal{S}$; $\mathrm{d}z$ denotes Lebesgue measure on $\mathcal{Z}$. A *Markov kernel* from $\mathcal{X}$ to $\mathcal{Z}$ is a map $x \mapsto \mu(\cdot \mid x) \in \mathcal{P}(\mathcal{Z})$ such that $x \mapsto \mu(B \mid x)$ is Borel for every Borel $B \subseteq \mathcal{Z}$. Expectations are written interchangeably as $\mathbb{E}_\mu[f] = \mathbb{E}_{s\sim\mu}[f(s)] = \int f \,\mathrm{d}\mu$, with values in $(-\infty, +\infty]$ when the integrand is bounded below (the KL integral is nevertheless always well defined in $[0, +\infty]$, its negative part being integrable). All measures below are Borel probability measures on $\mathcal{Z}$, on $\mathcal{X}$, or on their product, and $c_{\mathcal{A}}(x, z) \triangleq \|x - \mathcal{A}(z)\|_2^2$ with $\mathcal{A} : \mathcal{Z} \to \mathcal{X}$ Borel. We use the following standard facts.

(i) Disintegration. Since $\mathcal{Z} = \mathbb{R}^d$ is standard Borel, every $q \in \mathcal{P}(\mathcal{Z} \times \mathcal{X})$ with $(\pi_{\mathcal{X}})_{\#} q = p_x$ can be written as $q(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x)\, q(\mathrm{d}z \mid x)$ for a Markov kernel $q(\cdot \mid x)$, unique up to a p_x-null set [Kallenberg, 2021]. Consequently, functionals $\mathbb{E}_{x\sim p_x}[F(q(\cdot \mid x))]$ with F Borel and bounded below on $\mathcal{P}(\mathcal{Z})$—e.g., $F = \mathrm{KL}(\cdot \,\|\, p_z)$, which is lower semi-continuous, hence Borel—do not depend on the chosen version of the kernel.

(ii) Chain rule for relative entropy. If $q, \kappa \in \mathcal{P}(\mathcal{Z} \times \mathcal{X})$ share the x-marginal p_x and disintegrate as in (i), then

$$\mathrm{KL}(q \,\|\, \kappa) = \mathbb{E}_{x\sim p_x}\big[\mathrm{KL}\big(q(\cdot \mid x) \,\|\, \kappa(\cdot \mid x)\big)\big] \tag{21}$$

[Dupuis and Ellis, 1997, Thm. C.3.1].

Lemma 2 (Gibbs variational formula)**.** *Let $\kappa \in \mathcal{P}(\mathcal{Z})$, $\gamma > 0$, and let $G : \mathcal{Z} \to \mathbb{R}$ be Borel and bounded above, so that $Z_G \triangleq \int_{\mathcal{Z}} e^{G/\gamma} \,\mathrm{d}\kappa < \infty$. Then*

$$\sup_{q\in\mathcal{P}(\mathcal{Z})} \Big\{ \textstyle\int_{\mathcal{Z}} G \,\mathrm{d}q - \gamma\, \mathrm{KL}(q \,\|\, \kappa) \Big\} = \gamma \log Z_G, \tag{22}$$

where the objective is set to $-\infty$ whenever $\mathrm{KL}(q \,\|\, \kappa) = +\infty$ or G^- is not q-integrable, and the supremum is attained uniquely at the Gibbs measure $q^\star$ with $\frac{\mathrm{d}q^\star}{\mathrm{d}\kappa} = e^{G/\gamma}/Z_G$.

Proof. For bounded G this is the Donsker–Varadhan variational formula [Donsker and Varadhan, 1975]; see, e.g., Dupuis and Ellis [1997, Prop. 1.4.2] applied to $f = -G/\gamma$ and multiplied by $-\gamma$, which converts the infimum into a supremum. The general case follows by applying the bounded case to the truncations $G_n \triangleq G \vee (-n)$ and letting $n \to \infty$: since $G \le G_n$, the bounded case gives $\sup_q\{\int G \,\mathrm{d}q - \gamma\,\mathrm{KL}(q \,\|\, \kappa)\} \le \gamma \log \int e^{G_n/\gamma} \,\mathrm{d}\kappa \downarrow \gamma \log Z_G$ by dominated convergence ($e^{G_n/\gamma} \downarrow e^{G/\gamma}$, dominated by the constant $e^{\max(\sup G,\, -1)/\gamma} \in L^1(\kappa)$), while evaluating the objective at $q^\star$ yields the matching lower bound. Attainment and uniqueness use that G is bounded above, so that $G^+ \in L^1(q)$ for every $q \in \mathcal{P}(\mathcal{Z})$ and $G^- \in L^1(q^\star)$ since $G^- e^{G/\gamma} \le \gamma/e$. □

Assumptions. Recall that $\mathcal{W}$ is a convex, symmetric class of bounded Borel functions on $\mathcal{Z}$ with $\mathcal{F}^*_{p_z}(w) < \infty$ for every $w \in \mathcal{W}$; for the f-branch this requires f^* to be finite on $[-\sup_{w\in\mathcal{W}} \|w\|_\infty,\, \sup_{w\in\mathcal{W}} \|w\|_\infty]$ (automatic, e.g., for the KL and χ^2 generators), while it holds trivially for the Γ-branch. Since each w is bounded and $c_{\mathcal{A}} \ge 0$, we have $\mathcal{G}^{\mathcal{A}}_w \le \|w\|_\infty$, so (9) holds for every $x \in \mathcal{X}$. Theorem 1 assumes:

(A1) $\mathcal{W}$ is compact in the uniform norm; $\mathcal{Q}$ carries the total-variation topology, with $\|\mu\|_{\mathrm{TV}}$ the total mass of the variation measure $|\mu|$.

By the Arzelà–Ascoli theorem, (A1) covers, for instance, uniformly bounded and uniformly Lipschitz classes over a compact latent domain; for unbounded $\mathcal{Z}$, both the W_1 instantiation of Equation (7) and the f-branch of Equation (6) are understood with such a restricted critic class $\mathcal{W}$, so that $\mathbb{D}_{f,\Gamma}$ is the corresponding restricted divergence—a lower bound on the classical f-divergence, recovered exactly as $\mathcal{W}$ exhausts the bounded Borel functions. Compactness could instead be placed on the measure side, restricting $\mathcal{Q}$ to a weakly compact (tight) subset, at the price of verifying that the Gibbs kernels of Step 4 remain feasible; we do not pursue this here.

Proof of Theorem 1. We proceed in four steps: (1) rewriting the primal objective over the joint variational distribution, (2) applying the variational representations of the discrepancy measures, (3) invoking Sion's minimax theorem to exchange the optimization order, and (4) disintegrating the joint distribution to recover the conditional variational potential.

Step 1: Reformulating the primal objective. Recall the EVIA primal objective from Equation (11), defined for $q_{z,x} \in \mathcal{Q}$:

$$\mathcal{L}^{\text{EVIA-Primal}}(q_{z,x}) = \mathbb{D}_{f,\Gamma}(q_z \| p_z) + \frac{\lambda}{2}\,\mathbb{E}_{q_{z,x}}\big[c_{\mathcal{A}}(x,z)\big] + \gamma\,\mathrm{KL}(q_{z,x}\|\kappa_{z,x}), \tag{23}$$

with $q_z \triangleq (\pi_{\mathcal{Z}})_{\#} q_{z,x}$. Since every $w \in \mathcal{W}$ is bounded and $c_{\mathcal{A}} \geq 0$, each term is well defined in $(-\infty, +\infty]$. For $\nu \in \mathcal{P}(\mathcal{Z})$, write $\Pi(p_x, \nu) \triangleq \{q \in \mathcal{P}(\mathcal{Z} \times \mathcal{X}) : (\pi_{\mathcal{X}})_{\#} q = p_x,\ (\pi_{\mathcal{Z}})_{\#} q = \nu\}$ for the couplings of p_x and ν. Then $\mathcal{Q} = \bigcup_{\nu \in \mathcal{P}(\mathcal{Z})} \Pi(p_x, \nu)$, so minimizing over aggregated posteriors ν and, conditionally, over couplings $\pi \in \Pi(p_x, \nu)$ is the same as a single unconstrained infimum over $\mathcal{Q}$ (cf. the analogous reformulation in Theorem 1 of Tolstikhin et al. [2018]; unlike that setting, no hard marginal constraint $q_z = p_z$ is imposed here):

$$\mathcal{L}^{\text{EVIA}} = \inf_{q_{z,x} \in \mathcal{Q}} \Big\{ \mathbb{D}_{f,\Gamma}(q_z \| p_z) + \frac{\lambda}{2}\,\mathbb{E}_{q_{z,x}}\big[c_{\mathcal{A}}(x,z)\big] + \gamma\,\mathrm{KL}(q_{z,x}\|\kappa_{z,x}) \Big\}. \tag{24}$$

Step 2: Variational dual of the adversarial divergence. The adversarial divergence is *defined* variationally in Equation (5), specialized to the two branches by Equation (6) and Equation (7) with witnesses ranging over $\mathcal{W}$. Substituting $T = -w$ (using that $\mathcal{W}$ is symmetric, i.e., $w \in \mathcal{W} \implies -w \in \mathcal{W}$) unifies the two branches into

$$\mathbb{D}_{f,\Gamma}(q_z \| p_z) = \sup_{w \in \mathcal{W}} \Big\{ -\mathbb{E}_{q_z}[w(z)] - \mathcal{F}^*_{p_z}(w) \Big\}, \tag{25}$$

where $\mathcal{F}^*_{p_z}(w)$ matches the definition in Equation (13), with $\mathbb{E}_{p_z}[f^*(-w)] \in (-\infty, +\infty]$. Since $q_z = (\pi_{\mathcal{Z}})_{\#} q_{z,x}$, the change-of-variables formula for pushforwards gives $\mathbb{E}_{q_z}[w(z)] = \mathbb{E}_{q_{z,x}}[w(z)]$. Substituting back into (24) yields the saddle-point formulation:

$$\mathcal{L}^{\text{EVIA}} = \inf_{q_{z,x} \in \mathcal{Q}} \sup_{w \in \mathcal{W}} \Big\{ -\mathcal{F}^*_{p_z}(w) + \mathbb{E}_{q_{z,x}}\Big[\tfrac{\lambda}{2}\, c_{\mathcal{A}}(x,z) - w(z)\Big] + \gamma\,\mathrm{KL}(q_{z,x}\|\kappa_{z,x}) \Big\}. \tag{26}$$

Step 3: Minimax interchange via Sion's theorem. Denote by $\Phi(q_{z,x}, w)$ the bracketed functional in (26); by the standing assumptions on $\mathcal{W}$, Φ takes values in $(-\infty, +\infty]$. For each fixed w, the map $q_{z,x} \mapsto \Phi(q_{z,x}, w)$ is convex on the convex set $\mathcal{Q}$ (strictly so on its finite domain, by the Kullback–Leibler term with $\gamma > 0$) and lower semi-continuous in the total-variation topology: $q \mapsto \mathrm{KL}(q \,\|\, \kappa_{z,x})$ is lower semi-continuous; $q \mapsto \mathbb{E}_q[c_{\mathcal{A}}]$ is lower semi-continuous by monotone approximation, since $c_{\mathcal{A}} = \sup_n (c_{\mathcal{A}} \wedge n)$ with each $q \mapsto \mathbb{E}_q[c_{\mathcal{A}} \wedge n]$ total-variation continuous; and $q \mapsto \mathbb{E}_q[w]$ is continuous, as $|\mathbb{E}_q[w] - \mathbb{E}_{q'}[w]| \leq \|w\|_\infty \|q - q'\|_{\mathrm{TV}}$. For each fixed $q_{z,x}$, the map $w \mapsto \Phi(q_{z,x}, w)$ is concave on the convex set $\mathcal{W}$ (affine in the Γ-branch; concave in the f-branch since f^* is convex) and upper semi-continuous with respect to the uniform norm, by dominated convergence and Fatou's lemma. Since Φ is extended-real-valued, we apply Sion's minimax theorem [Sion, 1958] to $h \circ \Phi$ for a fixed increasing homeomorphism $h : [-\infty, +\infty] \to [-1, 1]$; this preserves quasi-convexity, quasi-concavity, and both semicontinuities, and the resulting identity transfers back through h^{-1}. As $\mathcal{W}$ is compact and convex under (A1), we obtain:

$$\mathcal{L}^{\text{EVIA}} = \sup_{w \in \mathcal{W}} \left\{ -\mathcal{F}^*_{p_z}(w) + \inf_{q_{z,x} \in \mathcal{Q}} \Big[\mathbb{E}_{q_{z,x}}\Big[\tfrac{\lambda}{2}\, c_{\mathcal{A}}(x,z) - w(z)\Big] + \gamma\,\mathrm{KL}(q_{z,x}\|\kappa_{z,x}) \Big] \right\}. \tag{27}$$

Step 4: Disintegration and conditional potential recovery. By (i), every $q_{z,x} \in \mathcal{Q}$ disintegrates as $q_{z,x}(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x)\, q(\mathrm{d}z \mid x)$ with $q(\cdot \mid x)$ a Markov kernel, and likewise $\kappa_{z,x}(\mathrm{d}z, \mathrm{d}x) = p_x(\mathrm{d}x)\, \kappa(\mathrm{d}z \mid x)$; the chain rule (21) then decomposes the entropic term. Writing the inner objective of (27) through this disintegration and bounding the integrand from below by its pointwise infimum gives

$$\begin{aligned} &\inf_{q_{z,x} \in \mathcal{Q}} \Big\{ \mathbb{E}_{q_{z,x}}\Big[\tfrac{\lambda}{2}\, c_{\mathcal{A}}(x,z) - w(z)\Big] + \gamma\,\mathrm{KL}(q_{z,x}\|\kappa_{z,x}) \Big\} \\ &\qquad \geq \mathbb{E}_{x \sim p_x}\left[\inf_{q \in \mathcal{P}(\mathcal{Z})} \Big\{ \mathbb{E}_{z \sim q}\Big[\tfrac{\lambda}{2}\, c_{\mathcal{A}}(x,z) - w(z)\Big] + \gamma\,\mathrm{KL}\big(q \,\|\, \kappa(\cdot \mid x)\big) \Big\} \right]. \end{aligned} \tag{28}$$

The right-hand side of (28) is well defined: the inner infimum equals $-\Psi^{\mathcal{A}}_w(x; \gamma, \kappa) = -\gamma \log Z(x)$ with $Z(x) \triangleq \int_{\mathcal{Z}} e^{\mathcal{G}^{\mathcal{A}}_w(z;x)/\gamma}\, \kappa(\mathrm{d}z \mid x)$, it is bounded below by $-\|w\|_\infty$, and it is Borel in x by the measurability part of the Fubini–Tonelli

theorem for Markov kernels (see, e.g., Kallenberg, 2021), applied to the jointly Borel integrand $e^{\mathcal{G}_w^{\mathcal{A}}/\gamma}$. Conversely, by Theorem 2 (applicable since $\mathcal{G}_w^{\mathcal{A}}(\cdot; x) \le \|w\|_\infty$ for every x), the inner infimum at each x is attained by the Gibbs measure $q^\star(\cdot \mid x)$ with $\frac{\mathrm{d}q^\star(\cdot|x)}{\mathrm{d}\kappa(\cdot|x)} = e^{\mathcal{G}_w^{\mathcal{A}}(\cdot;x)/\gamma}/Z(x)$; the map $x \mapsto q^\star(\cdot \mid x)$ is a Markov kernel, since the same measurability fact applied to $\mathbf{1}_B\, e^{\mathcal{G}_w^{\mathcal{A}}/\gamma}$ shows that $x \mapsto q^\star(B \mid x)$ is Borel for every Borel $B \subseteq \mathcal{Z}$. Hence $p_x(\mathrm{d}x)\, q^\star(\mathrm{d}z \mid x) \in \mathcal{Q}$ attains the right-hand side, and (28) holds with equality. Recalling $\mathcal{G}_w^{\mathcal{A}}(z; x) = w(z) - \frac{\lambda}{2}\, c_{\mathcal{A}}(x, z)$ and factoring out a negative sign, which flips the infimum to a supremum,

$$\inf_{q \in \mathcal{P}(\mathcal{Z})} \Big\{ \mathbb{E}_{z\sim q}\big[-\mathcal{G}_w^{\mathcal{A}}(z; x)\big] + \gamma\, \mathrm{KL}\big(q \,\|\, \kappa(\cdot \mid x)\big) \Big\} = - \sup_{q \in \mathcal{P}(\mathcal{Z})} \Big\{ \mathbb{E}_{z\sim q}\big[\mathcal{G}_w^{\mathcal{A}}(z; x)\big] - \gamma\, \mathrm{KL}\big(q \,\|\, \kappa(\cdot \mid x)\big) \Big\}, \tag{29}$$

and by Equation (10) the supremum is precisely the entropically regularized conditional potential $\Psi_w^{\mathcal{A}}(x; \gamma, \kappa)$. Substituting back into (27),

$$\inf_{q_{z,x} \in \mathcal{Q}} \mathcal{L}^{\text{EVIA-Primal}}(q_{z,x}) = \sup_{w \in \mathcal{W}} \Big[- \mathcal{F}_{p_z}^*(w) - \mathbb{E}_{x\sim p_x}\big[\Psi_w^{\mathcal{A}}(x; \gamma, \kappa)\big] \Big] = \mathcal{L}^{\text{EVIA-SemiDual}}, \tag{30}$$

which is exactly Equation (12) and concludes the proof. □

Remark 3 (Choice of the reference kernel). Nothing above requires the reference kernel $\kappa(\cdot \mid x)$ to be Gaussian: Theorem 1 only asks for a Markov kernel satisfying (9), so that the Gibbs measure is well defined. We use a Gaussian reference because its log-density is smooth and cheap to differentiate, which is exactly what the SGLD updates consume. For discrete (e.g., binary) latent variables, SGLD itself is no longer the appropriate sampler; the Gibbs measure can instead be sampled by discrete MCMC such as Gibbs sampling, or the latent space can be relaxed through a Concrete/Gumbel–Softmax distribution [Maddison et al., 2017, Jang et al., 2017], with SGLD run in the relaxed space and binary codes recovered by thresholding or temperature annealing.

B EXPERIMENT DETAILS

Across all benchmarks, EVIA and its corresponding baselines use identical network architectures, reconstruction losses, and adversarial objectives. Because optimizer design is outside the scope of this study [Bottou, 2010, Kingma and Ba, 2015, Loshchilov and Hutter, 2019, Teng et al., 2019], we use Adam [Kingma and Ba, 2015] throughout for consistency. The methods differ only in their inference procedures: the baselines use the encoder output directly, whereas EVIA initializes from this prediction and applies a short Langevin refinement of the Gibbs conditional $\tilde{q}(\cdot \mid x)$ in Equation (15). For MNIST, Fashion MNIST, and CelebA, the inferred variable is the latent code z, mapped to image space by the learned decoder $\mathcal{A}_\theta$. For CIFAR inpainting, the inferred variable is the completed image x, and the forward operator is the known masking operator rather than a learned decoder.

B.1 MNIST AND FASHION MNIST

We use 28×28 grayscale images and a two-dimensional latent space. The prior p_z is a 10-component isotropic GMM with means uniformly placed on a circle of radius 3 and component standard deviation 0.35. The encoder q_ϕ has two 4×4 stride-2 convolutional layers with 128 and 256 channels, followed by fully connected layers outputting $q_\phi(x) \in \mathbb{R}^2$ (the VAE uses Gaussian mean/log-variance heads on the same trunk). The decoder $\mathcal{A}_\theta$ mirrors this with transposed convolutions and outputs logits. The critic w_ψ is a two-layer MLP with 256 hidden units. All networks use LeakyReLU(0.2).

All methods use MSE reconstruction on sigmoid outputs,

$$\mathcal{L}_{\text{rec}}(x, \hat{x}) = \|\sigma(\hat{x}) - x\|_2^2. \tag{31}$$

VAE, WAE, and EVIA-Amortized decode $q_\phi(x)$ directly at test time. EVIA-SGLD maintains a persistent center $\bar{z}_i$ per training image and m latent particles $\{z_{ij}^-\}_{j=1}^m$ (see Algorithm 1). Particles are initialized as $\bar{z}_i + \sigma_{\text{init}}\epsilon_z$ and refined by unadjusted Langevin dynamics on $\tilde{q}(\cdot \mid x_i)$,

$$z_{t+1} = z_t + \eta \nabla_z \log \tilde{q}(z_t \mid x) + \sqrt{2\eta}\, \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I), \tag{32}$$

with Gibbs density matching Equation (15),

$$\log \tilde{q}(z \mid x) = w_\psi(z) - \frac{\lambda}{2}\, \mathcal{L}_{\text{rec}}(x, \mathcal{A}_\theta(z)) - \frac{\rho}{2}\|z - \bar{z}\|_2^2, \tag{33}$$

and $\bar{z}_i \leftarrow \frac{1}{m}\sum_{j=1}^{m} z_{ij}^{-}$ after refinement. EVIA-Amortized sets the Gaussian reference center to $\bar{z} = q_\phi(x)$ (see Equation (20), Algorithm 2) and trains the encoder–decoder pair with the joint consistency objective of Equation (19), where the encoder term is the empirical particle-mean form of Equation (17),

$$\alpha \big\| q_\phi(x) - \mathrm{sg}\big(\tfrac{1}{m}\textstyle\sum_{j=1}^{m} z_j^{-}\big)\big\|_2^2. \tag{34}$$

The critic is trained with the W_1 dual of Equation (16) under gradient penalty,

$$\mathcal{L}_\psi^{W_1} = -\mathbb{E}_{z\sim p_z}[w_\psi(z)] + \mathbb{E}_{z\sim q_{\mathrm{neg}}}[w_\psi(z)] + \lambda_{\mathrm{gp}}\,\mathcal{R}_{\mathrm{Lip}}(\psi), \tag{35}$$

where $\mathcal{R}_{\mathrm{Lip}}$ is the gradient-penalty regularizer of Equation (16) and interpolants are formed between prior samples z^+ and negative samples z^-. For WAE, q_{neg} is the detached encoder distribution; for EVIA, it is the refined-particle distribution. All models are trained with Adam at learning rate 10^{-3} for 5,000 iterations with batch size 128. Each iteration performs 5 critic updates; on EVIA, critic training starts after a 200-iteration reconstruction warmup. The VAE adds a standard factorized Gaussian KL term plus a penalty $\frac{1}{2}\min_k \|\mu - \mu_k\|_2^2$ that ties the encoder mean to the nearest GMM prior mode.

B.2 CELEBA

We use the CelebA `train` split; images are resized and center-cropped to 64×64 and scaled to $[0, 1]$. The latent dimension is 128 and the prior is $p_z = \mathcal{N}(0, I)$. EVIA and WAE share a deterministic encoder–decoder backbone: a DCGAN-style encoder with channels $64 \to 128 \to 256 \to 512$ (stride 2) projecting to $q_\phi(x) \in \mathbb{R}^{128}$, and a matching transposed-convolution decoder $\mathcal{A}_\theta$ outputting logits. The VAE uses the same trunk with a Gaussian encoder head. WAE and EVIA additionally employ a critic $w_\psi : \mathbb{R}^{128} \to \mathbb{R}$: a two-layer MLP with 512 hidden units and LeakyReLU(0.2). In WAE this critic separates prior samples from encoder codes; in EVIA it is the potential w_ψ in Equations (15) and (16).

The reconstruction loss is

$$\mathcal{L}_{\mathrm{rec}}(x, \hat{x}) = \mathrm{BCEWithLogits}(\hat{x}, x) + 0.5\|\sigma(\hat{x}) - x\|_1. \tag{36}$$

We also report an MSE variant ($\|\sigma(\hat{x}) - x\|_2^2$) trained under the same schedule. Baselines decode $q_\phi(x)$ directly. EVIA initializes m particles from $q_\phi(x) + 0.1\epsilon_z$ and applies the same Langevin update as above with center $\bar{z} = q_\phi(x)$ and target

$$\log \tilde{q}(z \mid x) = w_\psi(z) - \frac{\lambda}{2}\,\mathcal{L}_{\mathrm{rec}}(x, \mathcal{A}_\theta(z)) - \frac{\rho}{2}\|z - q_\phi(x)\|_2^2. \tag{37}$$

EVIA-Amortized additionally minimizes the encoder matching term of Equation (17) (particle-mean form with weight α), as on MNIST. All models are trained with Adam at learning rate 2×10^{-4} for 10 epochs with batch size 128. WAE performs 5 critic steps per batch; on EVIA, critic training begins after a 200-step reconstruction warmup and one critic update is performed per batch thereafter. The critic uses the same W_1+gradient-penalty objective as in (35).

B.3 CIFAR-10 AND CIFAR-100

We use 32×32 RGB images scaled to $[-1, 1]$. The task is conditional inpainting with a randomly removed 16×16 block. Let $\mathrm{mask} \in \{0, 1\}^{3\times32\times32}$ be the mask ($\mathrm{mask} = 1$ on observed pixels). The observed image is $y = \mathrm{mask} \odot x_{\mathrm{real}}$. Here the inferred variable is the completed image x itself; the encoder predicts $x_{\mathrm{enc}} = q_\phi(y, \mathrm{mask})$, and the known forward (masking) operator enforces

$$x_{\mathrm{fill}} = y + (1 - \mathrm{mask}) \odot x. \tag{38}$$

All methods share the same U-Net completion network q_ϕ (base width 128, residual blocks) and spectrally normalized image critic w_ψ (so $\mathcal{R}_{\mathrm{Lip}} \equiv 0$ in Equation (16)). The reconstruction loss is the masked hole L1,

$$\mathcal{L}_{\mathrm{rec}}(x_{\mathrm{real}}, x_{\mathrm{fill}}) = \|\mathrm{mask} \odot (x_{\mathrm{fill}} - x_{\mathrm{real}})\|_1. \tag{39}$$

EVIA initializes particles from $x_{\mathrm{enc}} + 0.1\epsilon_x$ and refines by Langevin on the image-space Gibbs density

$$x_{t+1} = x_t + \eta\nabla_x \log \tilde{q}(x_t \mid y, \mathrm{mask}) + \sqrt{2\eta}\,\xi_t, \qquad \xi_t \sim \mathcal{N}(0, I), \tag{40}$$

where

$$\log \tilde{q}(x \mid y, \mathrm{mask}) = w_\psi(x_{\mathrm{fill}}) - \frac{\lambda}{2}\mathcal{L}_{\mathrm{rec}}(x_{\mathrm{real}}, x_{\mathrm{fill}}) - \frac{\rho}{2}\|x - x_{\mathrm{enc}}\|_2^2. \tag{41}$$

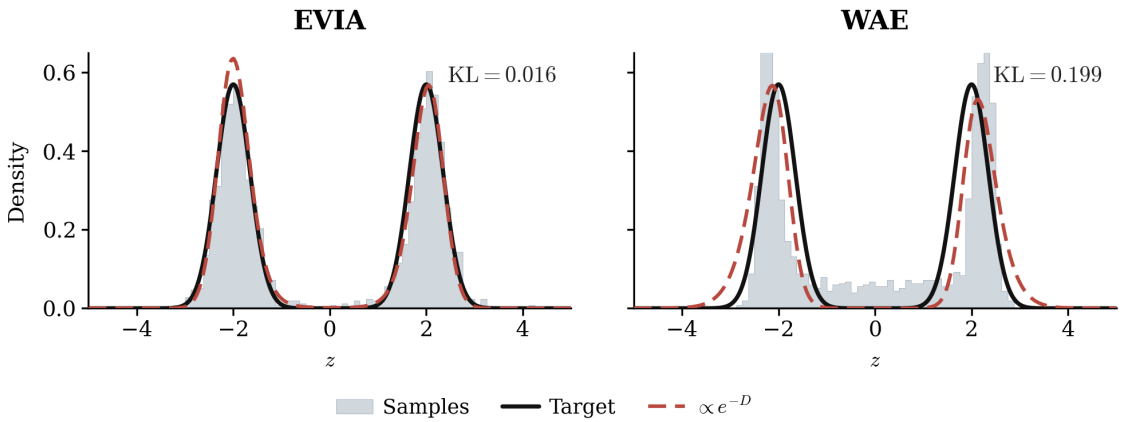


Figure 7: Critic training on the unconditional two-mode target. The same architecture is trained once with EVIA-style Langevin negatives (left) and once with WAE-style GAN fake negatives (right). Each panel shows the target density, the induced density $p_D \propto e^{-D}$, and the negative samples. EVIA reproduces the target (KL $= 0.016$); WAE separates the sample populations without recovering the mode geometry (KL $= 0.199$).

The fill x_{fill} is recomputed after each update; observed pixels remain fixed by (38). EVIA-Amortized trains q_ϕ with the particle-mean matching term of Equation (17) (weight α). The image critic uses the hinge loss

$$\mathcal{L}_D = \mathbb{E}\big[\max(0, 1 - w_\psi(x_{\text{real}}))\big] + \mathbb{E}\big[\max(0, 1 + w_\psi(x_{\text{fill}}))\big]. \tag{42}$$

All models are trained with Adam at learning rate 2×10^{-4} for $50{,}000$ iterations with batch size 128; each iteration performs one critic update followed by one encoder update. Completion-network weights are tracked with an exponential moving average (decay 0.999); reported samples use the EMA weights. The reconstruction weight in the encoder objective ramps linearly from 0 to 10 over the first $1{,}000$ iterations. Reported figures use checkpoints at iteration $4{,}000$.

B.4 HYPERPARAMETERS

Unless otherwise noted, the initialization noise scale is 0.1, $\lambda_{\text{gp}} = 10$, and the critic starts after a 200-step reconstruction warmup. MNIST/Fashion EVIA-SGLD uses $\sigma_{\text{init}} = 0.35$. In the notation of Equations (15) and (19), ρ is the tether weight, λ the reconstruction weight in the Gibbs density, and α, β weight the encoder and decoder terms $\mathcal{L}_\phi$ and $\mathcal{L}_\theta$. On MNIST and Fashion MNIST we grid-search EVIA hyperparameters and report the best configuration in each table/figure.

EVIA-SGLD sweeps $\rho \in \{0.01, 0.05, 0.1, 0.5, 1, 2\}$, $\lambda \in \{0.05, 0.1, 0.5\}$, $T \in \{4, 8\}$, $\eta \in \{10^{-3}, 10^{-2}\}$, and $m \in \{8, 16, 32\}$. EVIA-Amortized sweeps $\alpha \in \{30, 50, 100\}$, $\beta \in \{0.5, 1.0, 2.0\}$, $\rho = \lambda \in \{0.05, 0.1, 0.5\}$, $T \in \{4, 8\}$, $\eta \in \{10^{-3}, 10^{-2}\}$, and $m \in \{8, 16, 32\}$. On CIFAR-10 and CIFAR-100 inpainting we grid-search EVIA-Amortized hyperparameters and report the best configuration: $\alpha \in \{0.1, 1, 10\}$, $\beta \in \{0.5, 1.0, 2.0\}$, $\rho = \lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, $T \in \{4, 8\}$, $\eta \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$, and $m \in \{4, 8, 16\}$. On CelebA we grid-search EVIA-Amortized hyperparameters and report the best configuration: $\alpha \in \{30, 50, 100\}$, $\beta \in \{0.5, 1.0, 2.0\}$, $\rho = \lambda \in \{0.01, 0.05, 0.1, 0.5, 1\}$, $T \in \{4, 8\}$, $\eta \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$, and $m \in \{8, 16\}$.

C ADDITIONAL EXPERIMENTAL RESULTS

The following toy experiment examines the two premises underlying EVIA: (i) solutions to inverse problems are more appropriately represented by conditional distributions than by point estimates, and (ii) a critic trained on MCMC samples learns an energy landscape over the latent space rather than merely a binary decision boundary. Throughout, we sample a one-dimensional latent variable z from the equally weighted mixture

$$p_{\text{true}}(z) = \tfrac{1}{2}\mathcal{N}\big(-2,\, 0.35^2\big) + \tfrac{1}{2}\mathcal{N}\big(2,\, 0.35^2\big)\,,$$

and evaluate a learned critic D through the normalized density $p_D(z) = \frac{\exp(-D(z))}{\int \exp(-D(z'))\,\mathrm{d}z'}$ via $\mathrm{KL}(p_{\text{true}} \,\|\, p_D)$.

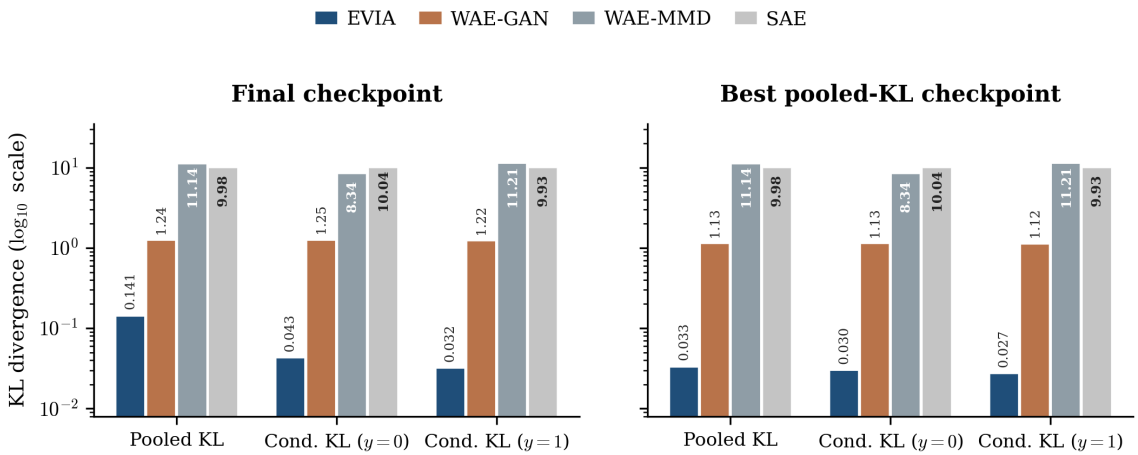


Figure 8: Conditional recovery on the two-mode Gaussian task. Each panel reports three bar groups matching the metrics above: pooled KL $\mathrm{KL}(p_{\mathrm{true}} \| p_D)$, and conditional KLs for $y{=}0$ and $y{=}1$. Left: final checkpoint; right: best pooled-KL checkpoint. EVIA reaches 0.141 (final) and 0.033 (best) on the pooled KL, with conditional KLs below 0.05. WAE-GAN stays near 1.2; WAE-MMD and SAE exceed 8 on both conditionals.

Figure 7 isolates premise (ii) on this *unconditional* target, with no classifier and no encoder. The same scalar energy network D is trained in two critic-only regimes that mirror the negative-sampling strategies used by EVIA and WAE-style methods: contrastive divergence with Langevin negatives targeting p_D, versus a GAN objective that pits real samples against generator fakes. Langevin negatives force D to shape a density over the entire latent line—an energy-based model in the sense of Grathwohl et al. [2020]—whereas adversarial fakes only require a separating surface between real and generated points. (The critic D here corresponds to $-w$ in the main-text convention.) At the best checkpoint, the EVIA-style critic reproduces the target ($\mathrm{KL} = 0.016$), while the WAE-style critic fails to capture the bimodal structure ($\mathrm{KL} = 0.199$). This is the property the SGLD refinement of Equation (15) exploits: refined particles follow an energy landscape that encodes the prior's structure.

We next turn to premise (i). We train a small classifier to identify the mixture component from which each sample is drawn, freeze it after 10^3 Adam updates [Kingma and Ba, 2015], and use it as the forward operator $\mathcal{A}$, which produces binary observations $y \in \{0, 1\}$. Since $\mathcal{A}$ is a many-to-one map, each observation corresponds to a continuum of plausible latent values rather than a unique solution. Although each label is associated with one mixture component, recovering the latent distribution across observations requires representing both components and avoiding global mode collapse.

The inference target is therefore the conditional distribution $q(z \mid y)$. We report two complementary metrics: the *pooled* KL on the full bimodal prior, $\mathrm{KL}(p_{\mathrm{true}} \,\|\, p_D)$ and the *conditional* KLs for each observation $y \in \{0, 1\}$, $\mathrm{KL}(p_{\mathrm{mode}}(\cdot \mid y) \,\|\, p_{\mathrm{guided}}(\cdot \mid y))$. Figure 8 compares EVIA—classifier-guided Langevin refinement of an EBM critic trained by contrastive divergence—with three encoder baselines that share the same $\mathrm{LabelEncoder}(y, \varepsilon)$ architecture: WAE-GAN, WAE-MMD, and Sinkhorn AE (SAE). Each encoder is trained for $3{,}000$ steps to satisfy the frozen classifier while optimizing its distributional objective; EVIA instead maintains $2{,}000$ particles per label, initialized uniformly on $[-5, 5]$ and updated with 16 inner Langevin steps per iteration. EVIA reaches a pooled KL of 0.141 at the final checkpoint and 0.033 at the best pooled-KL checkpoint, with conditional KLs below 0.05 for both $y{=}0$ and $y{=}1$. WAE-GAN remains an order of magnitude higher (pooled and conditional KLs $\sim$1.2), and WAE-MMD and SAE fare worse still (8–11), failing to match either conditional despite training on both labels. The gap is not a tuning artifact: extending *plain* WAE training to $10{,}000$ and $30{,}000$ steps leaves conditional KLs near 1.2, whereas a deterministic encoder cannot, by construction, represent the spread of latent explanations that particle-based refinement recovers.