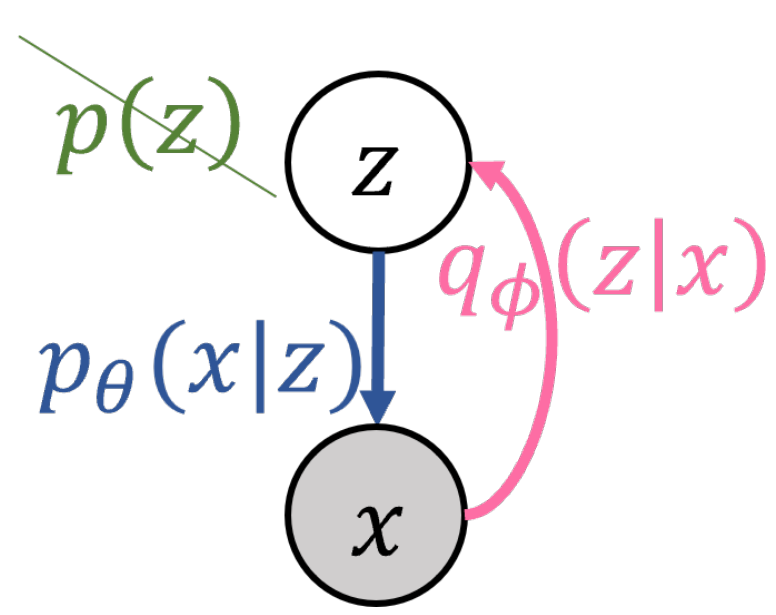


ON THE GENERATIVE UTILITY OF CYCLIC CONDITIONALS

CHANG LIU, HAOYUE TANG, TAO QIN, JINTAO WANG, TIE-YAN LIU
changliu@microsoft.com



THE QUESTION

Can we determine a **joint** distribution $p(x, z)$ only using two **conditional** distributions $p(x|z)$ and $q(z|x)$ that form a cycle?

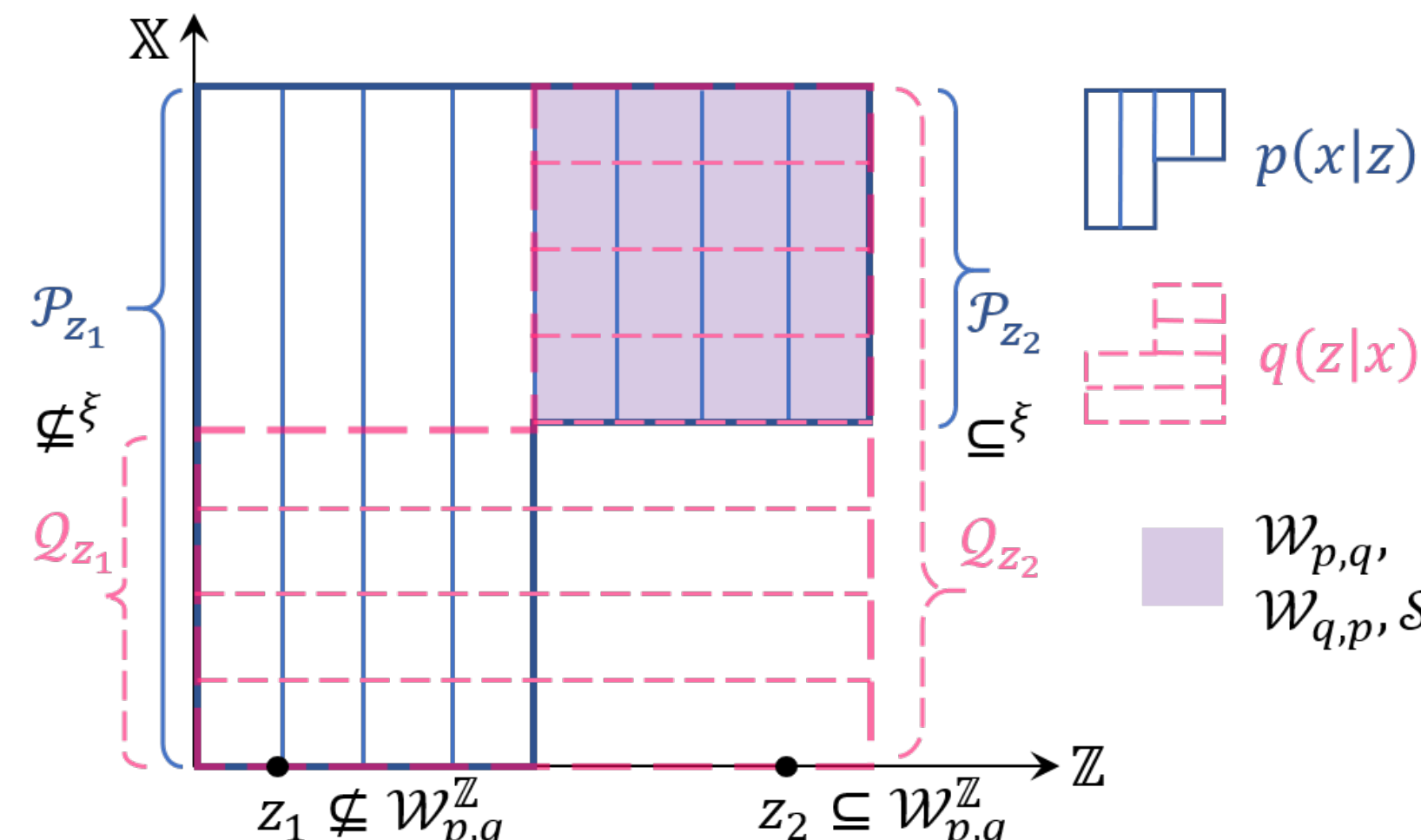
- **Compatibility:** is there a common joint that induces both conditionals?
- **Determinacy:** is the common joint unique?

THE ANSWER: THEORY

Absolutely-Continuous Case:

“smooth” distr. on continuous spaces, all distr. on discrete spaces.

- VAE, diffusion-based.
- **Thm: Compatibility** is achieved, iff on a suit-



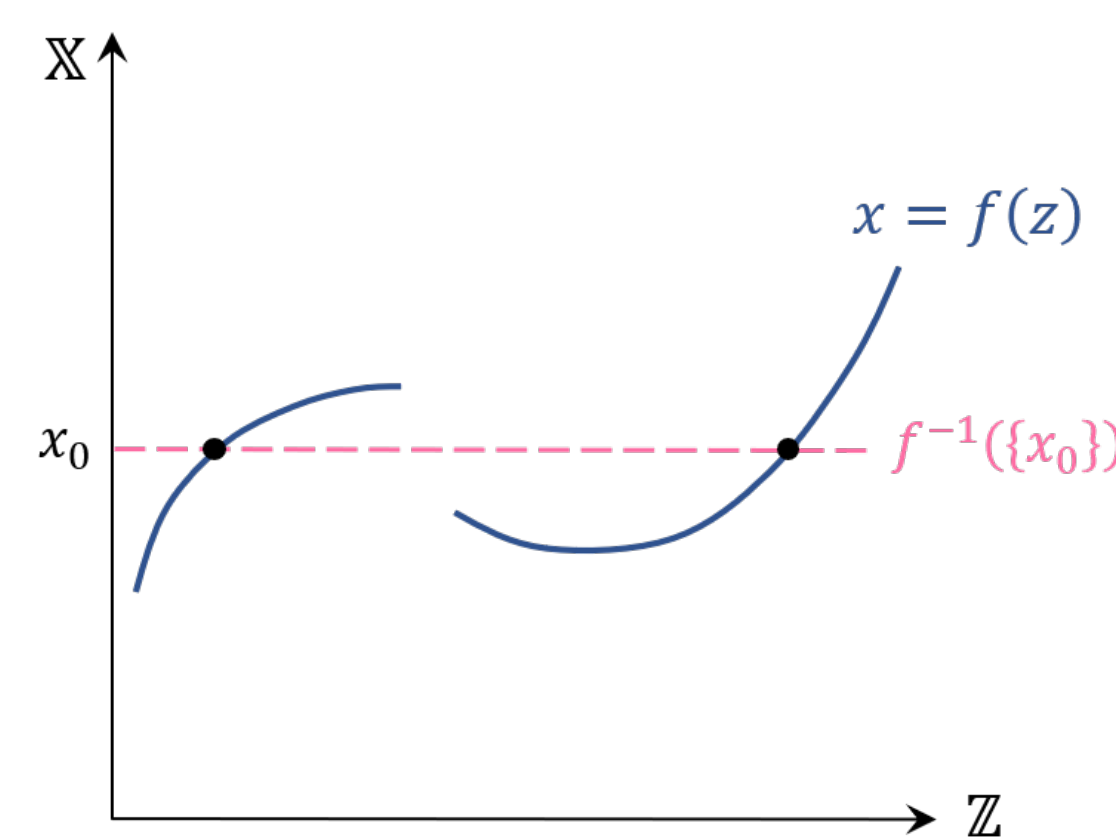
able set $S \subseteq \mathbb{X} \times \mathbb{Z}$ constructed from the conditional densities, $\frac{p(x|z)}{q(z|x)}$ a.e. **factorizes** as $a(x)b(z)$, where $a(x)$ is integrable.

- Equivalence condition, operable, self-contained.
- Why complicated: conditionals can be arbitrary on a set of marginal measure zero (e.g., outside the marginal support).
- **Thm: Determinacy** is achieved on S , if S is “rectangular”: $S_z \stackrel{\text{a.s.}}{=} S^{\mathbb{X}}, \forall \text{ a.e. } z \in S^{\mathbb{Z}}$ and $S_x \stackrel{\text{a.s.}}{=} S^{\mathbb{Z}}, \forall \text{ a.e. } x \in S^{\mathbb{X}}$.
- Determinacy is only possible on each S .
- If both densities have *full support*, $\mathbb{X} \times \mathbb{Z}$ is the only S .

Dirac Case:

$p(\mathcal{X}|z) = \delta_{f(z)}(\mathcal{X}) := \mathbb{I}[f(z) \in \mathcal{X}]$.

- Incl. BiGAN, flow-based.
- **Thm: Compatibility** is achieved, iff $\exists x_0$ s.t. $q(f^{-1}(\{x_0\})|x_0) = 1$: q puts mass only on the pre-image.



- Only one x_0 suffices: $\delta_{(x_0, f(x_0))}$ is a common joint.
- When $q(\mathbb{Z}|x) = \delta_{g(x)}(\mathbb{Z})$, min. **cycle-consistency loss** $\mathbb{E}_{p_{\text{ref}}(x)} \ell(x, f(g(x)))$ is sufficient.
- Flow-based models are naturally compatible.
- **Determinacy:**
 - On each $\{x_0\}$, the joint $\delta_{(x_0, f(x_0))}$ is unique.
 - If such x_0 is not unique, the joint is not unique on $\mathbb{X} \times \mathbb{Z}$: compatible p, q only determine a curve in $\mathbb{X} \times \mathbb{Z}$ but not a distribution on it.

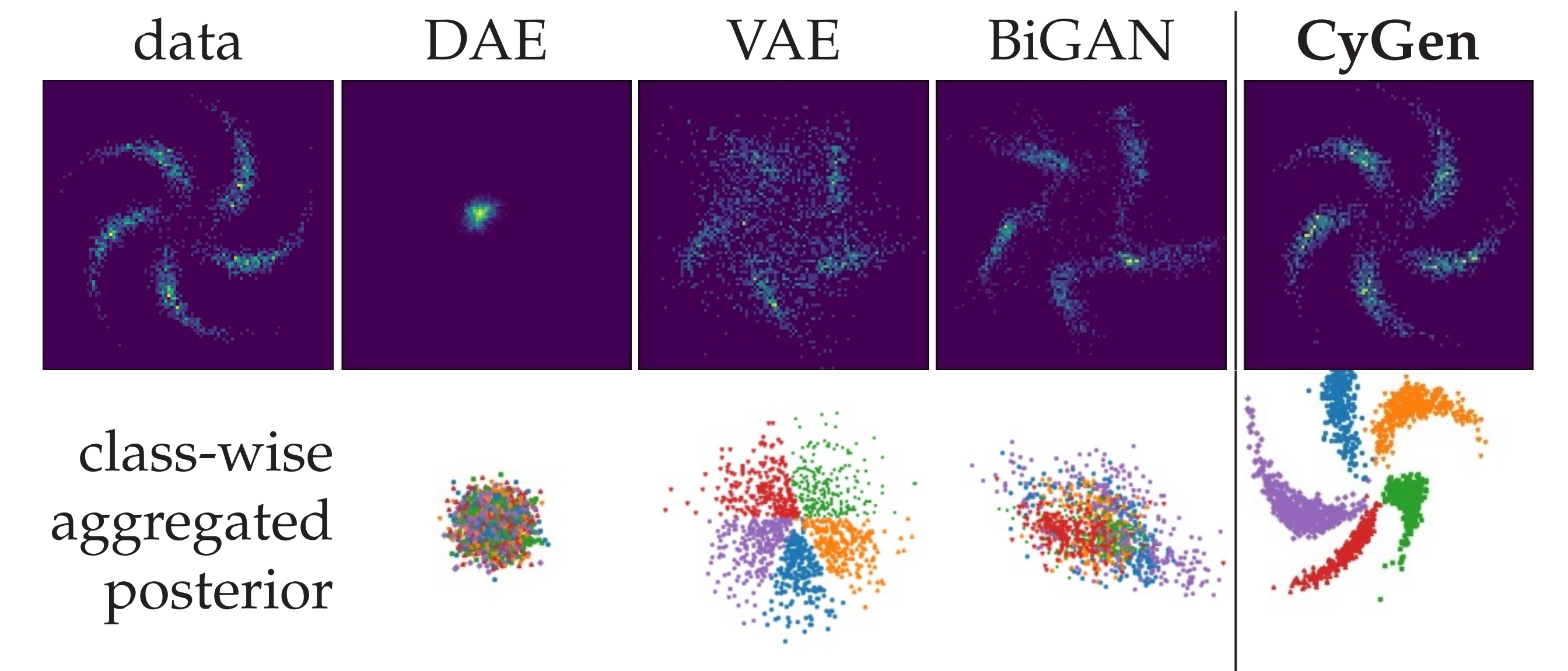
PROBLEMS OF CURRENT GENERATIVE MODELS

VAE, (Bi)GAN, flow/diffusion-based:

- Need $p(x|z)$ for generation, $q(z|x)$ for representation.
- Use a prior $p(z)$ to define joint $p(x, z) = p(z)p(x|z)$.

Specifying a prior leads to:

- **Manifold mismatch** (hinders generation): the modeled $p(x)$ is restricted to a simply-connected support.
- **Posterior collapse** (hinders representation): representations of different x are squeezed together.

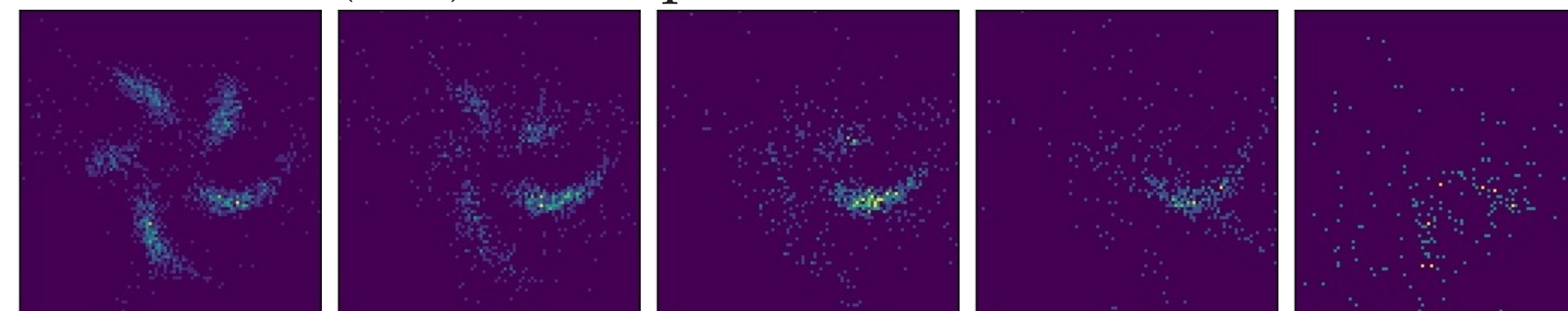


THE NEW FRAMEWORK: CYCLIC-CONDITIONAL GENERATIVE MODEL (CYGEN)

Eligibility as a generative model:

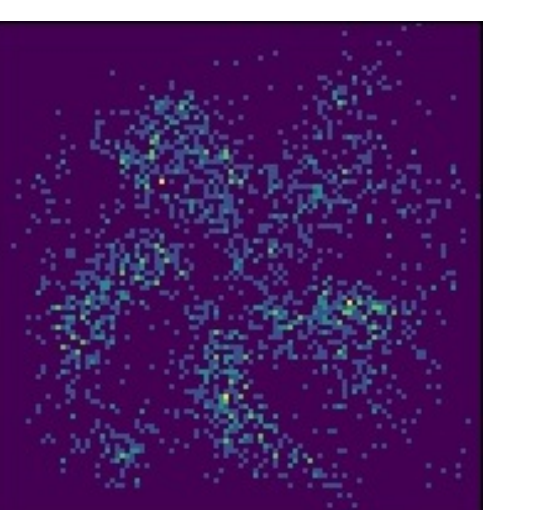
- **Determinacy:** Use absolutely-continuous conditionals, modeled by *fully-supported* densities $p_\theta(x|z), q_\phi(z|x)$ (like VAE).
- **Compatibility:**
 - $(\min_{\theta, \phi} C(\theta, \phi) := \mathbb{E}_{p_{\text{ref}}(x, z)} \|\nabla_x \nabla_z^\top \log(p_\theta(x|z)/q_\phi(z|x))\|_F^2$
 - $C(\theta, \phi) = 0 \iff p_\theta(x|z)/q_\phi(z|x)$ a.e. factorizes.
 - Generalizes cycle-consistency loss to *probabilistic* conditionals.
 - Gaussian VAE: $C(\theta, \phi) = 0 \iff$ mean fun. of p, q are affine! For nonlinear repr., one conditional must not be Gaussian.
 - **Scalable** unbiased stochastic estimate: $\mathbb{E}_{p_{\text{ref}}(x, z)} \mathbb{E}_{p(\eta): \mathbb{E}[\eta]=0, \text{Cov}[\eta]=I} \|\nabla_z (\eta^\top \nabla_x \log(p_\theta(x|z)/q_\phi(z|x)))\|_2^2$. Reduce $O(d_{\mathbb{X}}d_{\mathbb{Z}})$ to $O(d_{\mathbb{X}} + d_{\mathbb{Z}})$.

In absence of $C(\theta, \phi)$: (after pretrained as a VAE)



Usage as a generative model:

- **Fit data:** max. likelihood est. $\log p_{\theta, \phi}(x) = -\log \mathbb{E}_{q_\phi(z'|x)} [1/p_\theta(x|z')]$.
 - Est. expect. by reparameterization and logsumexp.
 - Denoising auto-encoder (DAE) objective $\mathbb{E}_{q_\phi(z'|x)} [\log p_\theta(x|z')]$ is improper: **(1)** $\geq \log p_{\theta, \phi}(x)$; **(2)** mode-collapses q thus hurts determinacy.
- **Generate data:** dynamics-based MCMC.
 - More efficient than Gibbs sampling:
 - Only requires unnormalized density: $p_{\theta, \phi}(x) \propto \frac{p_\theta(x|z)}{q_\phi(z|x)}, \forall z$.
 - E.g., Stochastic Gradient Langevin Dynamics: $x^{(t+1)} = x^{(t)} + \varepsilon \nabla_{x^{(t)}} \log \frac{p_\theta(x^{(t)}|z^{(t)})}{q_\phi(z^{(t)}|x^{(t)})} + \sqrt{2\varepsilon} \eta^{(t)}$, where $z^{(t)} \sim q_\phi(z|x^{(t)})$, $\eta^{(t)} \sim \mathcal{N}(0, I)$.



REAL-WORLD EXPERIMENTS

Data generation and classification accuracy (%) using representation:

