Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning

Cheng Lu*, Huayu Chen*, Jianfei Chen, Hang Su, Chongxuan Li, Jun Zhu

Tsinghua University

Exact Energy-Guided Diffusion Sampling

Suppose we have pretrained a diffusion model to fit data distribution $q_0(\boldsymbol{x}_0)$ .We'd like to sample from an edited distribution defined by an energy function:

$$p_0(\boldsymbol{x}_0) \propto q_0(\boldsymbol{x}_0) e^{-\mathcal{E}(\boldsymbol{x}_0)}$$

desired distribution — data distribution — energy function in data space

The form of $p_0(\boldsymbol{x}_0)$ is general and actually stems from constrained optimization:

$$\min_p \mathbb{E}_{p(\boldsymbol{x})}[\mathcal{E}(\boldsymbol{x})] + \frac{1}{\beta} D_{\mathrm{KL}}(p(\boldsymbol{x}) \parallel q(\boldsymbol{x})) \Longrightarrow p^*(\boldsymbol{x}) \propto q(\boldsymbol{x}) e^{-\mathcal{E}(\boldsymbol{x})}$$

To perform diffusion sampling, the required score function is:

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}_t) = \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t) - \nabla_{\mathbf{x}} \mathcal{E}_t(\mathbf{x})$$

desired score — pretrained score — energy guidance

where $\mathcal{E}_t(\mathbf{x})$ satisfies $p_t(\boldsymbol{x}_t) \propto q_t(\boldsymbol{x}_t) e^{-\mathcal{E}_t(\boldsymbol{x}_t)}$

Key Observation: Intermediate energy functions $\mathcal{E}_t(\mathbf{x})$ are completely determined by the data distribution $q(x)$ and the energy function $\mathcal{E}(\boldsymbol{x}_0)$ at time 0.

Theorem 1. The intermediate score functions satisfies:

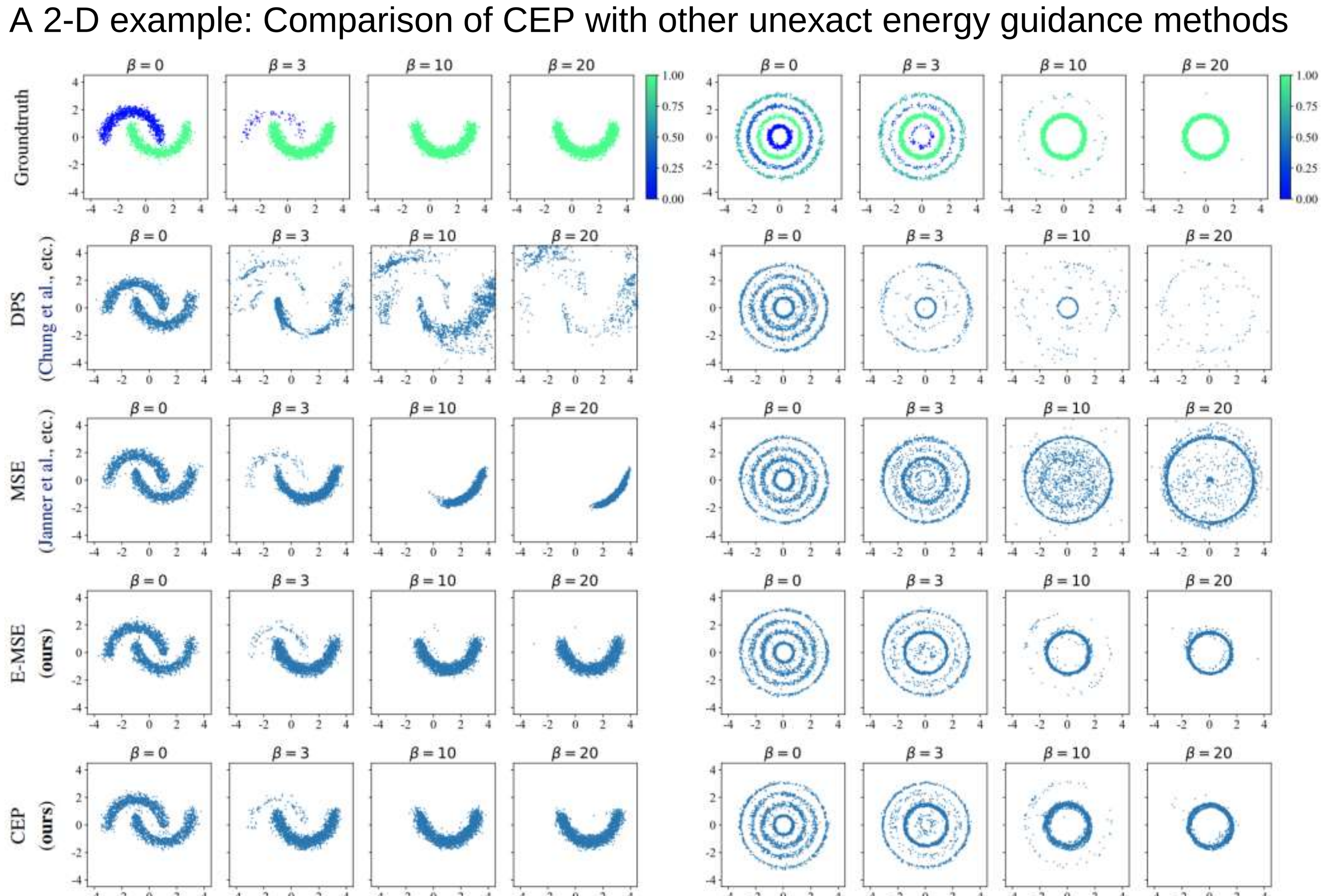
$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \underbrace{\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t)}_{\approx -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)/\sigma_t} - \underbrace{\nabla_{\boldsymbol{x}_t} \mathcal{E}_t(\boldsymbol{x}_t)}_{\text{energy guidance (intractable)}}$$

$$\mathcal{E}_t(\boldsymbol{x}_t) := \begin{cases} \beta \mathcal{E}(\boldsymbol{x}_0), & t = 0, \\ -\log \mathbb{E}_{q_{0t}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[e^{-\beta \mathcal{E}(\boldsymbol{x}_0)}\right], & t > 0. \end{cases}$$

We cannot use arbitrary intermediate energy guidance!

Method	Optimal Solution of Energy	Optimal Solution of Guidance	Exact Guidance
CEP (ours)	$-\log \mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}\left[e^{-\mathcal{E}_0(\boldsymbol{x}_0)}\right]$	$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}\left[-e^{\mathcal{E}_t(\boldsymbol{x}_t)-\mathcal{E}_0(\boldsymbol{x}_0)}\nabla_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)\right]$	✓
MSE	$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}[\mathcal{E}_0(\boldsymbol{x}_0)]$	$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}\left[\mathcal{E}_0(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)\right]$	✗
DPS	$\mathcal{E}_0\left(\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}[\boldsymbol{x}_0]\right)$	$\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}\left[\left(\left(\nabla\mathcal{E}_0\left(\mathbb{E}_{q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)}[\boldsymbol{x}_0]\right)\right)^\top \boldsymbol{x}_0\right)\nabla_{\boldsymbol{x}_t} \log q_{0t}(\boldsymbol{x}_0\vert\boldsymbol{x}_t)\right]$	✗

A 2-D example: Comparison of CEP with other unexact energy guidance methods



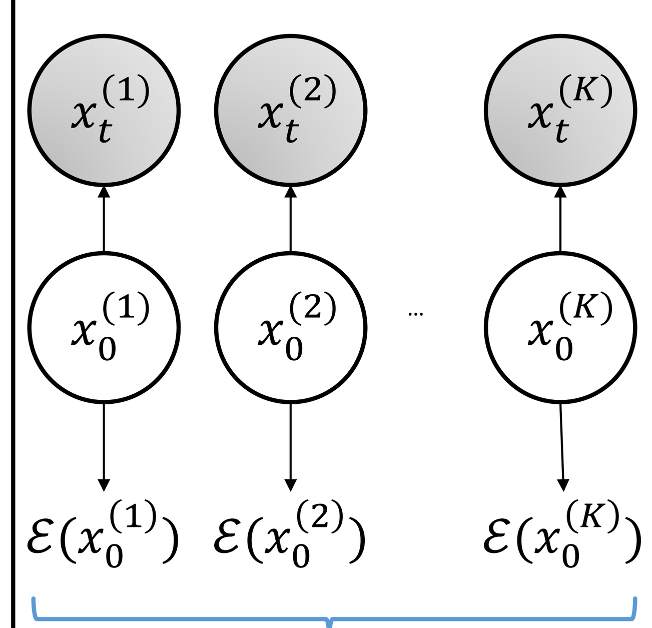
How to Estimate the Exact Energy Guidance?

The exact diffused energy is hard to estimate due to the log-expectation-exp form:

$$-\log \mathbb{E}_{q_{0t}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[e^{-\beta\mathcal{E}(\boldsymbol{x}_0)}\right]$$

Intractable: log-expectation-exp

We propose **Contrastive Energy Prediction (CEP)**: a training objective for learning the exact intermediate energy guidance



Contrastively predict the energy

Learn to predict the energy of clean data:

$$\min_\phi \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\boldsymbol{x}_0^{(1:K)})}\mathbb{E}_{p(\boldsymbol{\epsilon}^{(1:K)})}\left[-\sum_{i=1}^K \underbrace{e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(i)})}}_{\text{soft energy label}} \log \underbrace{\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},t)}}{\sum_{j=1}^K e^{-f_\phi(\boldsymbol{x}_t^{(j)},t)}}}_{\text{predicted label}}\right]$$

Contrastive (infoNCE) loss

Theorem 2. For all $K > 1$, The optimal guidance model satisfies

$$\nabla_{\boldsymbol{x}_t} f_{\phi^*}(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \mathcal{E}_t(\boldsymbol{x}_t)$$

Compute $e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(i)})}$ may be numerically unstable, so we normalize the energy for each batch:

$$\min_\phi \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\boldsymbol{x}_0^{(1:K)})}\mathbb{E}_{p(\boldsymbol{\epsilon}^{(1:K)})}\left[-\sum_{i=1}^K \underbrace{\frac{e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(i)})}}{\sum_{j=1}^K e^{-\beta\mathcal{E}(\boldsymbol{x}_0^{(j)})}}}_{\text{self-normalized energy label}} \log \underbrace{\frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},t)}}{\sum_{j=1}^K e^{-f_\phi(\boldsymbol{x}_t^{(j)},t)}}}_{\text{predicted label}}\right]$$

Connection between CEP and Classifier Guidance

If $\mathcal{E}_0(\boldsymbol{x}_0) = -\log q_0(c|\boldsymbol{x}_0)$ and $\beta = 1$: $p_0(\boldsymbol{x}_0) \propto q_0(\boldsymbol{x}_0) q(c|\boldsymbol{x}_0) \propto q(\boldsymbol{x}_0|c)$

The training objective in Theorem 2 becomes:

$$\mathbb{E}_{t,\boldsymbol{\epsilon}^{(1:K)}}\mathbb{E}_{\prod_{i=1}^K q_0(\boldsymbol{x}_0^{(i)},c^{(i)})}\left[-\sum_{i=1}^K \log \frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(i)},t)}}{\sum_{j=1}^K e^{-f_\phi(\boldsymbol{x}_t^{(j)},c^{(i)},t)}}\right]$$

Compare within data

Classifier Guidance:

$$\mathbb{E}_{t,\boldsymbol{\epsilon}^{(1:K)}}\mathbb{E}_{\prod_{i=1}^K q_0(\boldsymbol{x}_0^{(i)},c^{(i)})}\left[\sum_{i=1}^K \log \frac{e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(i)},t)}}{\sum_{j=1}^M e^{-f_\phi(\boldsymbol{x}_t^{(i)},c^{(j)},t)}}\right]$$

Classify conditions

- CEP is essentially based on info-NCE, Classifier Guidance is an cross entropy objective.
- Both can guarantee exact guidance, but CEP can be generalized to cases with no conditioning variables (energy functions).

ImageNet256
Guided sampling
Similar performance

Classifier guidance (Dhariwal & Nichol, 2021) — Energy guidance (**ours**)

Application: Offline Reinforcement Learning

The optimal policy of constrained policy optimization in offline RL satisfies:

$$\max_\pi \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}^\mu}\left[\mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}Q_\psi(\boldsymbol{s},\boldsymbol{a}) - \frac{1}{\beta}D_{\mathrm{KL}}\left(\pi(\cdot|\boldsymbol{s})||\mu(\cdot|\boldsymbol{s})\right)\right] \Longrightarrow$$

$$\pi^*(\boldsymbol{a}|\boldsymbol{s}) \propto \mu(\boldsymbol{a}|\boldsymbol{s})\, e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a})}$$

- We train a diffusion model $\mu_\theta(a|s)$ to imitate the behavior policy $\mu(a|s)$.
- We train a Q-net as an energy function to sample from the optimal policy.

$$\mathcal{T}^\pi Q_\psi(\boldsymbol{s},\boldsymbol{a}) \approx r(\boldsymbol{s},\boldsymbol{a}) + \gamma \frac{\sum_{\hat{\boldsymbol{a}}'} e^{\beta_Q Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')} Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')}{\sum_{\hat{\boldsymbol{a}}'} e^{\beta_Q Q_\psi(\boldsymbol{s}',\hat{\boldsymbol{a}}')}}$$

- We use the proposed CEP method to train another diffused Q-network to estimate the energy guidance term when performing guided sampling:

$$\nabla_{\boldsymbol{a}_t} \log \pi_t(\boldsymbol{a}_t|\boldsymbol{s}) = \underbrace{\nabla_{\boldsymbol{a}_t} \log \mu_t(\boldsymbol{a}_t|\boldsymbol{s})}_{\approx -\boldsymbol{\epsilon}_\theta(\boldsymbol{a}_t|\boldsymbol{s},t)/\sigma_t} + \underbrace{\nabla_{\boldsymbol{a}_t} \mathcal{E}_t(\boldsymbol{s},\boldsymbol{a}_t)}_{\approx f_\phi(\boldsymbol{s},\boldsymbol{a}_t,t)}$$

where $\mathcal{E}_t(\boldsymbol{s},\boldsymbol{a}_t) = \log \mathbb{E}_{\mu_{0t}(\boldsymbol{a}_0|\boldsymbol{a}_t,\boldsymbol{s})}\left[e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a}_0)}\right]$

- The training objective for the diffused Q-network is.

$$\min_\phi \mathbb{E}_{t,\boldsymbol{s},\boldsymbol{\epsilon}} - \sum_{i=1}^K \frac{e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a}^{(i)})}}{\sum_{j=1}^K e^{\beta Q_\psi(\boldsymbol{s},\boldsymbol{a}^{(j)})}} \log \frac{e^{f_\phi(\boldsymbol{s},\boldsymbol{a}_t^{(i)},t)}}{\sum_{j=1}^K e^{f_\phi(\boldsymbol{s},\boldsymbol{a}_t^{(j)},t)}}$$

- We use DPM-solver to accelerate the sampling precedure

D4RL evaluations:

Dataset	Environment	CQL	BCQ	IQL	SfBC	DD	Diffuser	D-QL	D-QL@1	QGPO (ours)
Medium-Expert	HalfCheetah	62.4	64.7	86.7	**92.6**	90.6	79.8	**96.1**	**94.8**	**93.5 ± 0.3**
Medium-Expert	Hopper	98.7	100.9	91.5	**108.6**	**111.8**	**107.2**	**110.7**	100.6	**108.0 ± 2.5**
Medium-Expert	Walker2d	**111.0**	57.5	**109.6**	**109.8**	**108.8**	**108.4**	**109.7**	**108.9**	**110.7 ± 0.6**
Medium	HalfCheetah	44.4	40.7	47.4	45.9	49.1	44.2	50.6	47.8	**54.1 ± 0.4**
Medium	Hopper	58.0	54.5	66.3	57.1	79.3	58.5	82.4	64.1	**98.0 ± 2.6**
Medium	Walker2	79.2	53.1	78.3	77.9	**82.5**	79.7	**85.1**	82.0	**86.0 ± 0.7**
Medium-Replay	HalfCheetah	**46.2**	38.2	44.2	37.1	39.3	42.2	**47.5**	44.0	**47.6 ± 1.4**
Medium-Replay	Hopper	48.6	33.1	94.7	86.2	**100.0**	**96.8**	**100.7**	63.1	**96.9 ± 2.6**
Medium-Replay	Walker2d	26.7	15.0	73.9	65.1	75.0	61.2	**94.3**	75.4	84.4 ± 4.1
Average (Locomotion)		63.9	51.9	76.9	75.6	81.8	75.3	**86.3**	75.6	**86.6**
Default	AntMaze-umaze	74.0	78.9	87.5	**92.0**	-	-	68.6	69.4	**96.4 ± 1.4**
Diverse	AntMaze-umaze	**84.0**	55.0	62.2	**85.3**	-	-	53.0	56.4	74.4 ± 9.7
Play	AntMaze-medium	61.2	0.0	71.2	**81.3**	-	-	0.0	1.0	**83.6 ± 4.4**
Diverse	AntMaze-medium	53.7	0.0	70.0	**82.0**	-	-	18.4	14.8	**83.8 ± 3.5**
Play	AntMaze-large	15.8	6.7	39.6	59.3	-	-	10.6	15.8	**66.6 ± 9.8**
Diverse	AntMaze-large	14.9	2.2	47.5	45.5	-	-	4.2	1.6	**64.8 ± 5.5**
Average (AntMaze)		50.6	23.8	63.0	74.2	-	-	25.8	26.5	**78.3**
# Action candidates		1	100	1	32	1	1	50	1	1
# Diffusion steps		-	-	-	15	100	100	5	5	15

Energy guidance demonstration on images

A toy example: color guidance

$$\mathcal{E}(\boldsymbol{x}) := -\left\|h(\boldsymbol{x}) - h_{\mathbf{tar}}\right\|_1$$

(hue value, computed by Hue-Saturation-Intensity (HSI) decomposition)