International Conference on Learning Representations (ICLR) 2022

Generalized Decision Transformer for Offline Hindsight Information Matching

Hiroki Furuta[1], Yutaka Matsuo[1], Shixiang Shane Gu[2]

[1]The University of Tokyo, [2]Google Brain

Contact: furuta@weblab.t.u-tokyo.ac.jp

Reinforcement Learning with Hindsight Information

Orthogonal to standard reward maximization scheme …

- ❑ Future trajectory information $\tau_{\mathbf{t:T}}$
- ❑ Context $\mathbf{z}$
- ❑ Contextual policy $\pi(\mathbf{a_t} | \mathbf{s_t}, \mathbf{z})$
- ❑ Parameterized reward function $\mathbf{r}(\mathbf{s_t}, \mathbf{a_t}, \mathbf{z})$

We derive a generic problem formulation: **Hindsight Information Matching (HIM)**.

Hindsight Information Matching

Information statistics $I(\tau_t)$: any function of a trajectory $\tau_t = \{s_t, a_t, s_{t+1}, a_{t+1}, \ldots\}$

Hindsight Information Matching

Information statistics $I(\tau_t)$: any function of a trajectory $\tau_t = \{s_t, a_t, s_{t+1}, a_{t+1}, \dots\}$

Feature Function $\Phi(\cdot, \cdot) : S \times A \rightarrow F$: e.g. reward function, some dims of state

Trajectory & Information statisctic: $\tau_t^{\Phi} = \{\phi_t, \phi_{t+1}, \dots, \phi_T\}, \phi_t = \Phi(s_t, a_t) \in F$ $I^{\Phi}(\tau_t)$

Hindsight Information Matching

Information statistics $I(\tau_t)$: any function of a trajectory $\tau_t = \{s_t, a_t, s_{t+1}, a_{t+1}, \dots\}$

Feature Function $\Phi(\cdot, \cdot) : S \times A \rightarrow F$: e.g. reward function, some dims of state

Trajectory & Information statisctic: $\tau_t^{\Phi} = \{\phi_t, \phi_{t+1}, \dots, \phi_T\}, \phi_t = \Phi(s_t, a_t) \in F$ $I^{\Phi}(\tau_t)$

Information Matching problem (inspired by moment matching)

Objective: $$\min_{\pi} \mathbb{E}_{z \sim p(z), \tau \sim \rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z) \right]$$

Hindsight Information Matching

Information statistics $I(\tau_t)$: any function of a trajectory $\tau_t = \{s_t, a_t, s_{t+1}, a_{t+1}, \dots\}$

Feature Function $\Phi(\cdot,\cdot) : S \times A \to F$: e.g. reward function, some dims of state

Trajectory & Information statisctic: $\tau_t^\Phi = \{\phi_t, \phi_{t+1}, \dots, \phi_T\}, \phi_t = \Phi(s_t, a_t) \in F$ $I^\Phi(\tau_t)$

Information Matching problem (inspired by moment matching)

Objective: $$\min_\pi \mathbb{E}_{z\sim p(z), \tau\sim\rho_z^\pi(\tau)} \left[D(I^\Phi(\tau), z) \right]$$

Hindsight Information Matching algorithms

setting desired z* as $\mathbf{z}^* = \mathbf{I}^\mathbf{\Phi}(\tau)$eans trajectory τ optimal w.r.t $\mathbf{z} = \mathbf{z}^*$

i.e. samples of $(\tau_\mathbf{i}, \mathbf{z}_\mathbf{i}^*)$n be used to accelerate RL or do BC.

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z \sim p(z), \tau \sim \rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z) \right]}$$

$I^{\Phi}(\tau)$ an be ...

(1)

(2)

(3)

(4)

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z)\right]}$$

$I^{\Phi}(\tau)$an be ...

(1) **Goal-based** ϕ_T

(2)

(3)

(4)

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z)\right]}$$

$I^{\Phi}(\tau)$ an be ...

(1) Goal-based

ϕ_T

(2) **Multi-task**

$\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$

(3)

(4)

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z \sim p(z), \tau \sim \rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z) \right]}$$

$I^{\Phi}(\tau)$ an be ...

(1) Goal-based ϕ_T

(2) Multi-task $\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$

(3) **Return-based** $\sum_t \gamma^t r_t$

(4)

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z)\right]}$$

$I^{\Phi}(\tau)$ an be ...

(1) Goal-based ϕ_T

(2) Multi-task $\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$

(3) Return-based $\sum_t \gamma^t r_t$

(4) **Full trajectory imitation** τ

How does HIM formulation cover existing problems?

Based on the choice of information statistics $I^{\Phi}(\tau)$, all prior works can be categorized to four generic problem types

Objective

$$\boxed{\min_{\pi} \mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)} \left[D(I^{\Phi}(\tau), z)\right]}$$

$I^{\Phi}(\tau)$ an be ...

(1) Goal-based ϕ_T

(2) Multi-task $\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$

(3) Return-based $\sum_t \gamma^t r_t$

(4) Full trajectory imitation τ

(5) **Distribution-based (ours)** $I^{\Phi}(\tau) = \text{histogram}(r_t, \gamma)$

Hindsight Information Matching: Summary

Given a choice of $\mathbf{I}^{\Phi}(\tau)$ HIM algorithms consist of three components:

- **Algorithm Type** (e.g. “RL” or “BC”)
- **Training** Procedure (e.g. “online” or “offline”)
- Network **Architectures** (e.g. MLP, CNN, Transformer, etc...)

Method	Algo. Type	Training	$\mathbf{I}^{\Phi}(\tau)$	Architectures
Andrychowicz et al. (2017)	RL	Online	ϕ_T	MLP
Pong et al. (2018)	RL	Online	ϕ_T	MLP
Chebotar et al. (2021)	RL	Offline	ϕ_T	CNN
Li et al. (2020)	RL	Online	$\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$	MLP
Eysenbach et al. (2020)	BC/RL	On/Offline	$\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$	MLP
Lynch et al. (2019)	BC	Offline	ϕ_T	Stochastic RNN
Ghosh et al. (2021)	BC	Online	ϕ_T	MLP
Srivastava et al. (2019)	BC	Online	$\sum_t \gamma^t r_t$	Fast Weights
Kumar et al. (2019)	BC	Online	$\sum_t \gamma^t r_t$	MLP
Janner et al. (2021)	BC	Offline	$\sum_t \gamma^t r_t$ or ϕ_T	Transformer
Duan et al. (2017)[3]	BC	Offline	τ	MLP + LSTM
Generalized DT (ours)	BC	Offline	Any	Transformer
DT (Chen et al., 2021a)	BC	Offline	$\sum_t \gamma^t r_t$	Transformer
Categorical DT (ours)[4]	BC	Offline	histogram(r_t, γ)	Transformer
Bi-Directional DT (ours)	BC	Offline	τ	Transformer

Hindsight Information Matching: Summary

Given a choice of $\mathbf{I}^{\Phi}(\tau)$ HIM algorithms consist of three components;

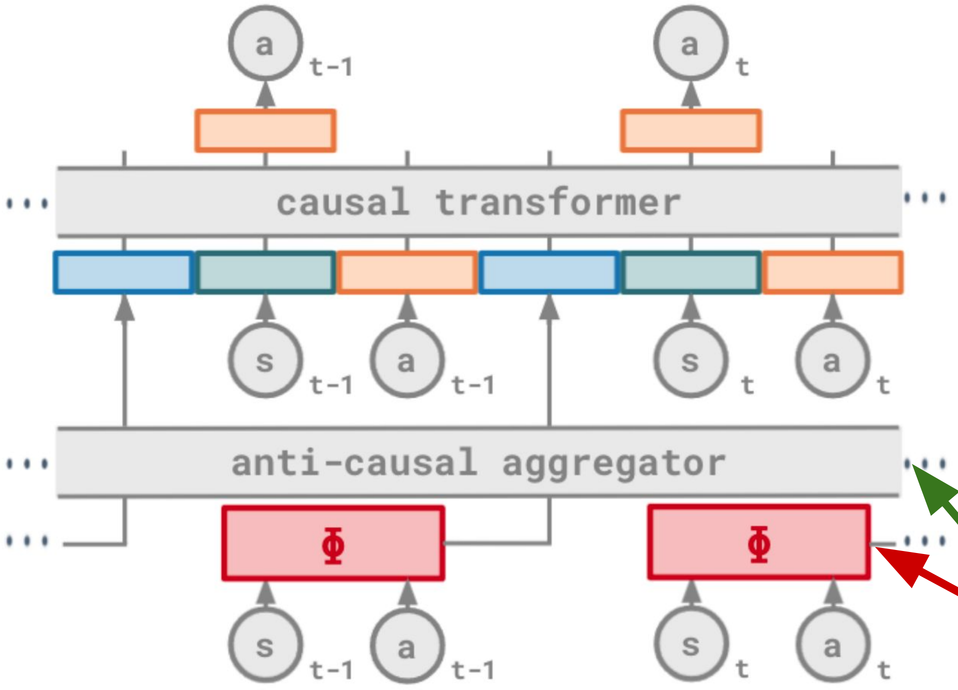
- ❏ **Algorithm Type** (e.g. “RL” or “BC”)
- ❏ **Training** Procedure (e.g. “online” or “offline”)
- ❏ Network **Architectures** (e.g. MLP, CNN, Transformer, etc…)

Method	Algo. Type	Training	$\mathbf{I}^{\Phi}(\tau)$	Architectures
Andrychowicz et al. (2017)	RL	Online	ϕ_T	MLP
Pong et al. (2018)	RL	Online	ϕ_T	MLP
Chebotar et al. (2021)	RL	Offline	ϕ_T	CNN
Li et al. (2020)	RL	Online	$\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$	MLP
Eysenbach et al. (2020)	BC/RL	On/Offline	$\arg\max \sum_t \gamma^t r(s_t, a_t, \cdot)$	MLP
Lynch et al. (2019)	BC	Offline	ϕ_T	Stochastic RNN
Ghosh et al. (2021)	BC	Online	ϕ_T	MLP
Srivastava et al. (2019)	BC	Online	$\sum_t \gamma^t r_t$	Fast Weights
[illegible]	BC	Online	$\sum_t \gamma^t r_t$	MLP
[illegible]	BC	Offline	$\sum_t \gamma^t r_t$ or ϕ_T	Transformer
Duan et al. (2017)	BC	Offline	τ	MLP + LSTM
Generalized DT (ours)	BC	Offline	Any	Transformer
DT (Chen et al., 2021a)	BC	Offline	$\sum_t \gamma^t r_t$	Transformer
Categorical DT (ours)[4]	BC	Offline	histogram(r_t, γ)	Transformer
Bi-Directional DT (ours)	BC	Offline	τ	Transformer

Our Proposals

Generalized Decision Transformer (GDT)

Generalization of Decision Transformer [Chen et al. 2021] with only small architectural changes (feature function Φ, aggregator)



Method	$\mathbf{\Phi(s, a)}$	Aggregator
DT (Chen et al., 2021a)	$r(s, a)$	Summation
DT-X (Section 5.3)	Learned	Summation
CDT (Section 5.2)	$r(s, a)$ or **any**	Binning
BDT (Section 5.4)	Learned	Transformer

Small architectural changes!

Categorical DT for State-feature Matching

Feature Function Φ: Reward or any state-features (e.g. xyz-velocities)

Input: Histogram of Φ by binning (i.e. categorical)

Metric: Empirical Wasserstein-1 distance (between target and policy)

Categorical DT can match …

- ❏ 1D reward or x-velocity distributions
- ❏ 2D xy-velocities distribution better than competitive baselines

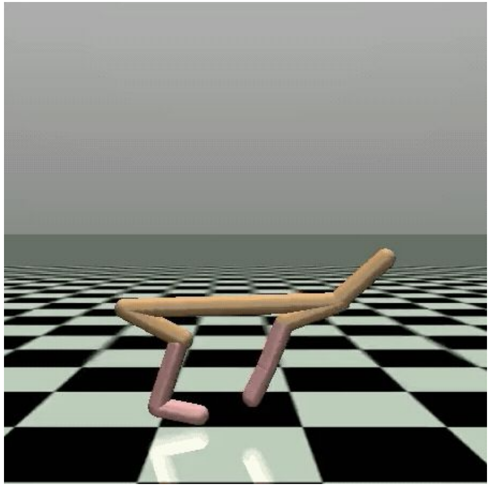
Better matching in 2D!

Method	ant		
	Expert	Medium	Average
Categorical DT	0.797 ± 0.216	0.244 ± 0.063	0.521
DT	1.714 ± 0.121	0.260 ± 0.067	0.987
Meta-BC	1.295 ± 0.708	0.351 ± 0.205	0.823
FOCAL (Li et al., 2021)	1.473 ± 0.892	0.913 ± 0.455	1.193

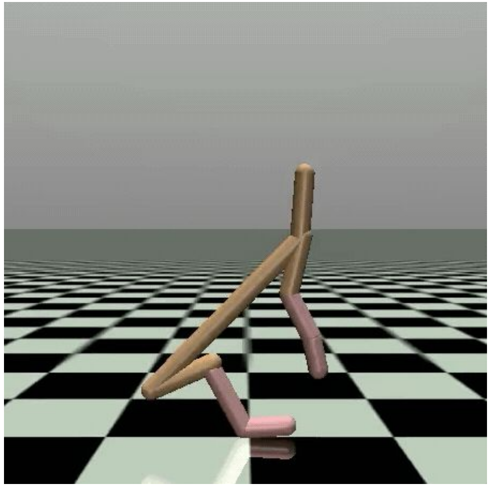
Synthesizing Unseen Bi-modal Distribution (CDT)

Cheetah running forward and backflipping during a single rollout

Datasets

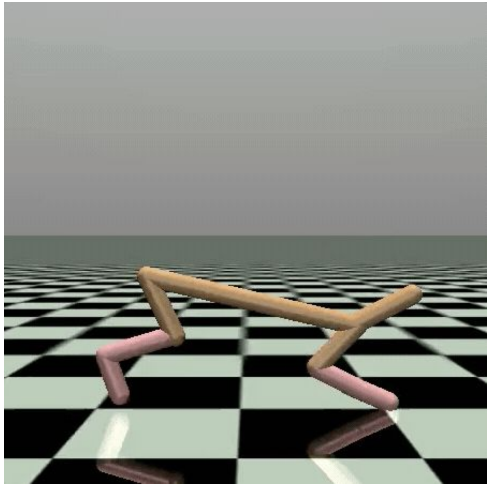


+



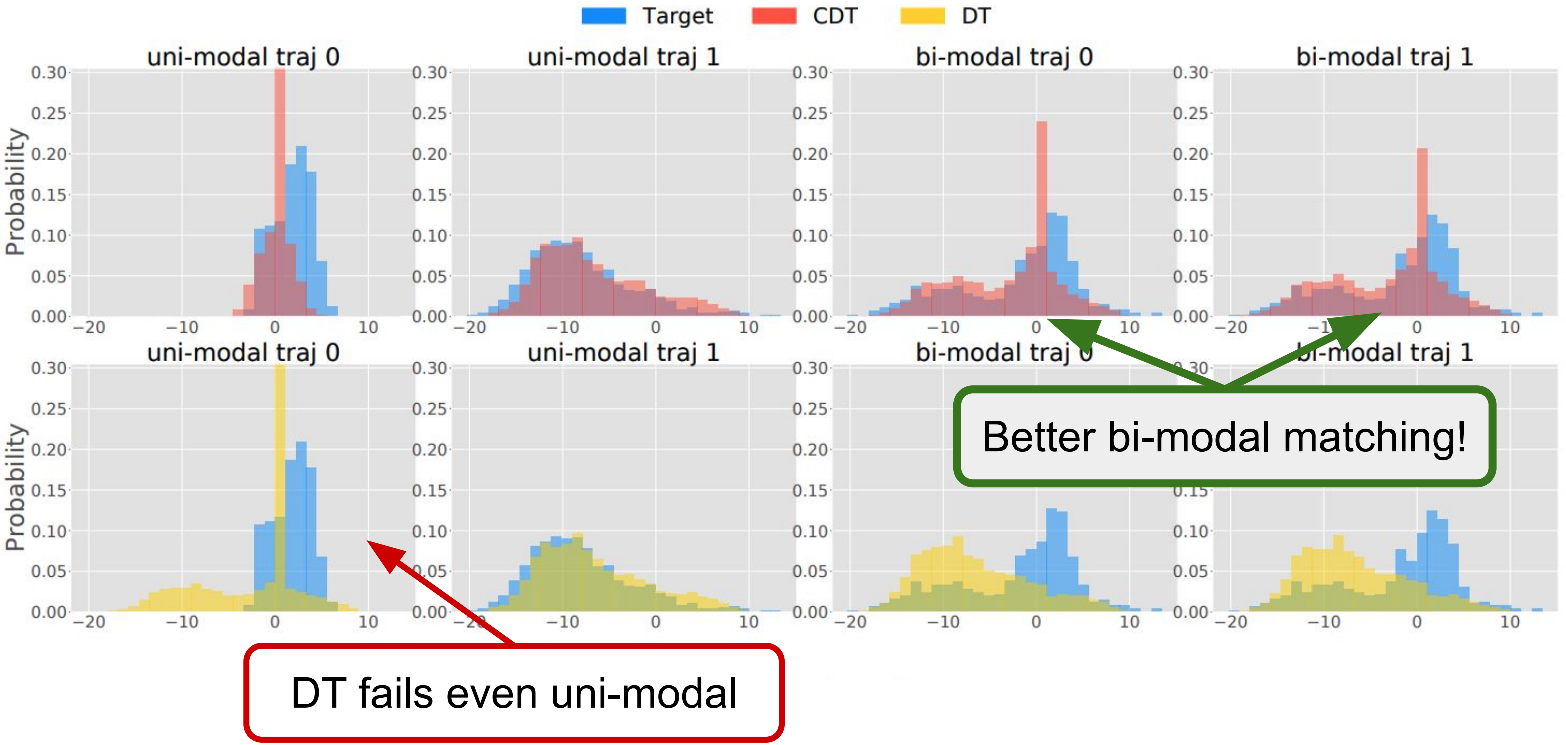
⇨

CDT outputs



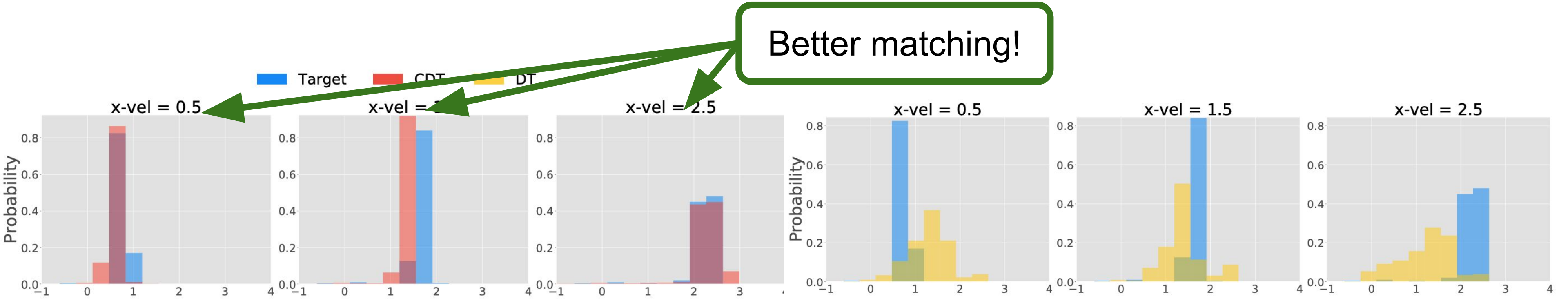
Synthesizing Unseen Bi-modal Distribution (CDT)

Cheetah running forward and backflipping during a single rollout



Diverse Unseen Distribution from Meta-Learning Task

Categorical DT generalizes unseen target better and slightly outperforms Meta-BC



Method	x-vel: 0.5	x-vel: 1.5	x-vel: 2.5	Average
Categorical DT	0.060 ± 0.026	0.211 ± 0.022	0.149 ± 0.110	0.140
DT	1.197 ± 0.227	0.533 ± 0.105	0.861 ± 0.247	0.864
Meta-BC	0.150 ± 0.069	0.152 ± 0.127	0.167 ± 0.055	0.156
FOCAL (Li et al., 2021)	0.472 ± 0.005	0.952 ± 0.073	0.346 ± 0.186	0.590

Bi-directional DT for Distribution Matching in Full State

Feature Function Φ: Learned (not specified)

Input: Full state in target trajectories (using anti-causal Transformer)

In 1D tasks, BDT seems competitive to CDT or DT w/o state-feature specification!

Method	Average
DT-AE	0.843
DT-CPC	1.591
DT-AE (joint)	2.650
DT-CPC (joint)	1.410
DT-E2E	2.517
DT-AE (frozen)	0.916
DT-CPC (frozen)	1.405
BDT (N=20)	0.631
BDT (N=50)	0.443

Competitive performance!

Method	Average
Categorical DT	0.347
DT	0.387
BC (no-context)	1.498
Meta-BC	0.699
FOCAL (Li et al., 2021)	1.147

Summary

1. We generalize a wide range of hindsight algorithms as **Hindsight Information Matching (HIM)** problem.

Summary

1. We generalize a wide range of hindsight algorithms as **Hindsight Information Matching (HIM)** problem.
2. To solve any kind of HIM problems, we propose **Generalized Decision Transformer**, and its practical instantiations (Categorical & Bi-directional DT).

Summary

1. We generalize a wide range of hindsight algorithms as **Hindsight Information Matching (HIM)** problem.
2. To solve any kind of HIM problems, we propose **Generalized Decision Transformer**, and its practical instantiations (Categorical & Bi-directional DT).
3. **Categorical DT** can generalize even synthesized bi-modal distributions or diverse unseen distributions better.

Summary

1. We generalize a wide range of hindsight algorithms as **Hindsight Information Matching (HIM)** problem.
2. To solve any kind of HIM problems, we propose **Generalized Decision Transformer**, and its practical instantiations (Categorical & Bi-directional DT).
3. **Categorical DT** can generalize even synthesized bi-modal distributions or diverse unseen distributions better.
4. **Bi-directional DT** can solve distribution matching comparable to CDT or DT only given raw states without state-feature specification.

Summary

1. We generalize a wide range of hindsight algorithms as **Hindsight Information Matching (HIM)** problem.
2. To solve any kind of HIM problems, we propose **Generalized Decision Transformer**, and its practical instantiations (Categorical & Bi-directional DT).
3. **Categorical DT** can generalize even synthesized bi-modal distributions or diverse unseen distributions better.
4. **Bi-directional DT** can solve distribution matching comparable to CDT or DT only given raw states without state-feature specification.

We hope our proposed framework, algorithms, and benchmarks inspire more supervised sequence modeling approaches in RL beyond classic reward maximization.