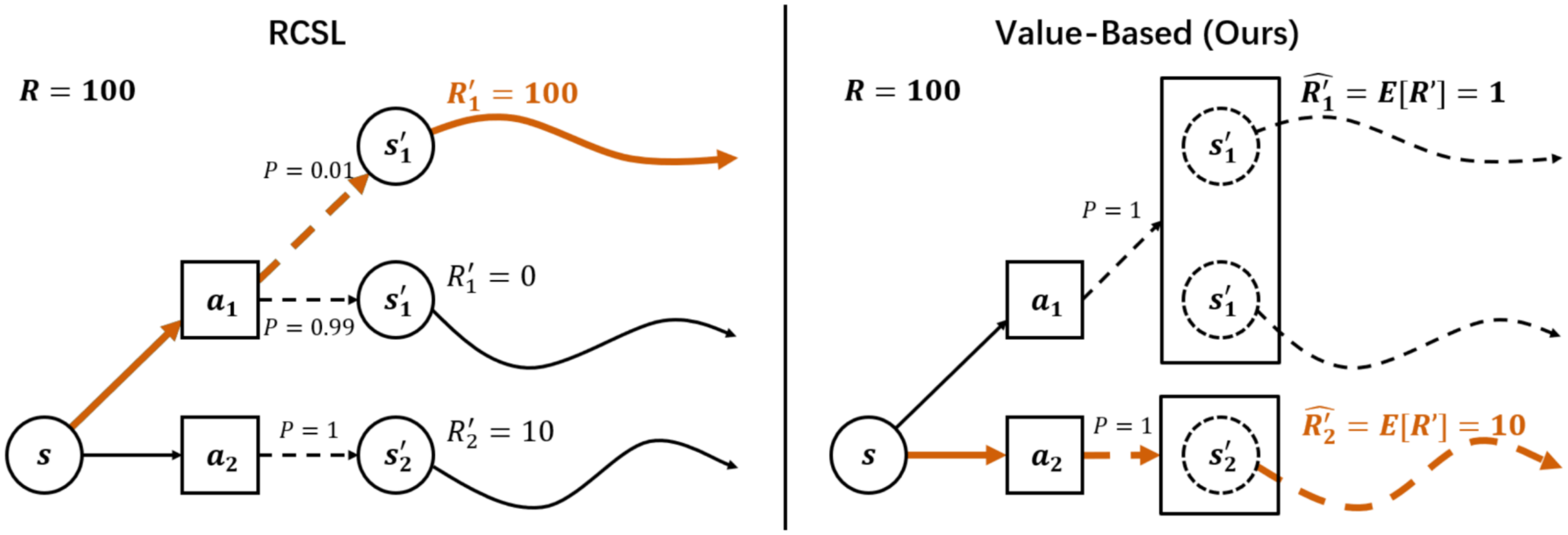
Value-Guided Decision Transformer: A Unified Reinforcement Learning Framework for Online and Offline Settings

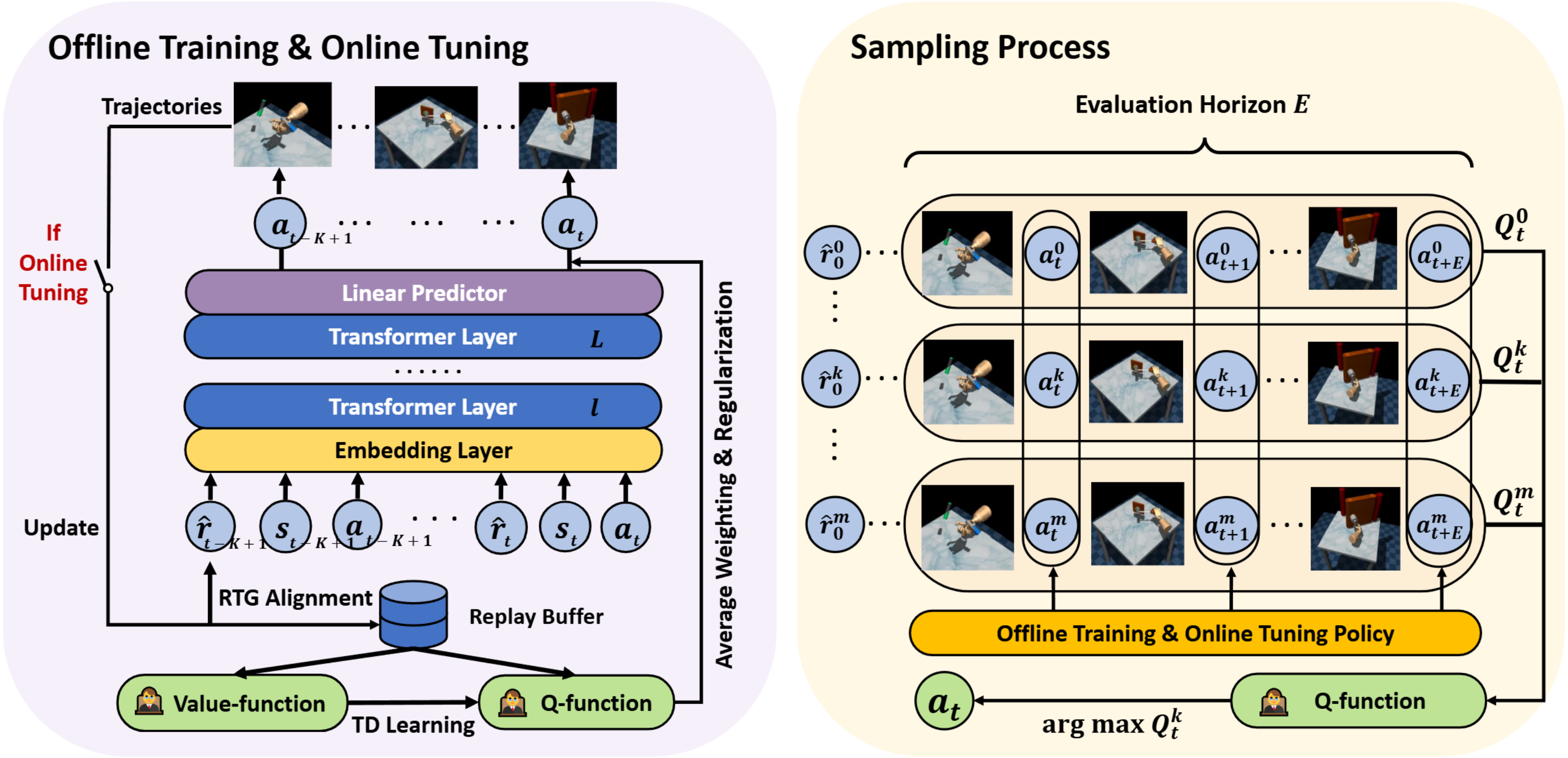
NeurIPS 2025, the Thirty-Ninth Annual Conference on Neural Information Processing Systems

Motivation



- **Insufficient Value Function Integration:** Prior methods simply relabel returns or add penalties, failing to fully harness value functions for optimization and regularization in DTs.
- **Offline-to-Online Gap:** ODT extends DTs to online RL but cannot achieve expert performance with limited or low-quality data and only improves notably after online fine-tuning.
- **Need for a Unified Framework:** There is a strong demand for a unified RL method that bridges offline and online RL, robustly handles suboptimal data, and better integrates value functions with Transformer architectures.

Method



Offline Training & Online Tuning
Trajectories
If Online Tuning
Linear Predictor
Transformer Layer L
Transformer Layer l
Embedding Layer
Update
RTG Alignment
Replay Buffer
Value-function
TD Learning
Q-function
Average Weighting & Regularization
Sampling Process
Evaluation Horizon E
Q_t^0
Q_t^k
Q_t^m
Offline Training & Online Tuning Policy
Q-function
a_t
$\arg\max Q_t^k$

Offline Training Phase

- **Value function training**

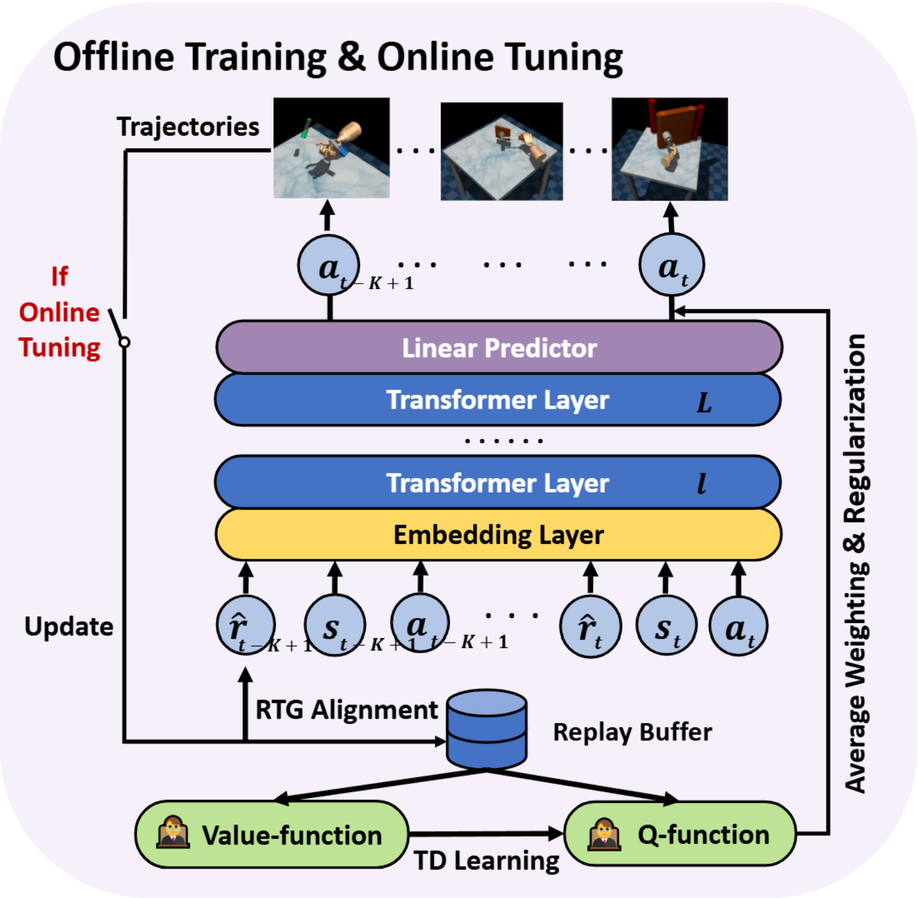
$$L_V(\psi) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[L_2^{\epsilon}\left(Q_{\hat{\theta}}(s_t,a_t) - V_\psi(s_t)\right)\right]$$

$$L_2^{\epsilon}(u) = |\epsilon - \mathbb{I}(u<0)|u^2$$

$$\mathbb{E}_{(s_t,a_t,r_t,\ldots,s_{t+n})\sim\mathcal{D}}\left[\left(\sum_{k=0}^{n-1}\gamma^k r_{t+k} + \gamma^n V_\psi(s_{t+n}) - Q_{\theta_i}(s_t,a_t)\right)^2\right]$$

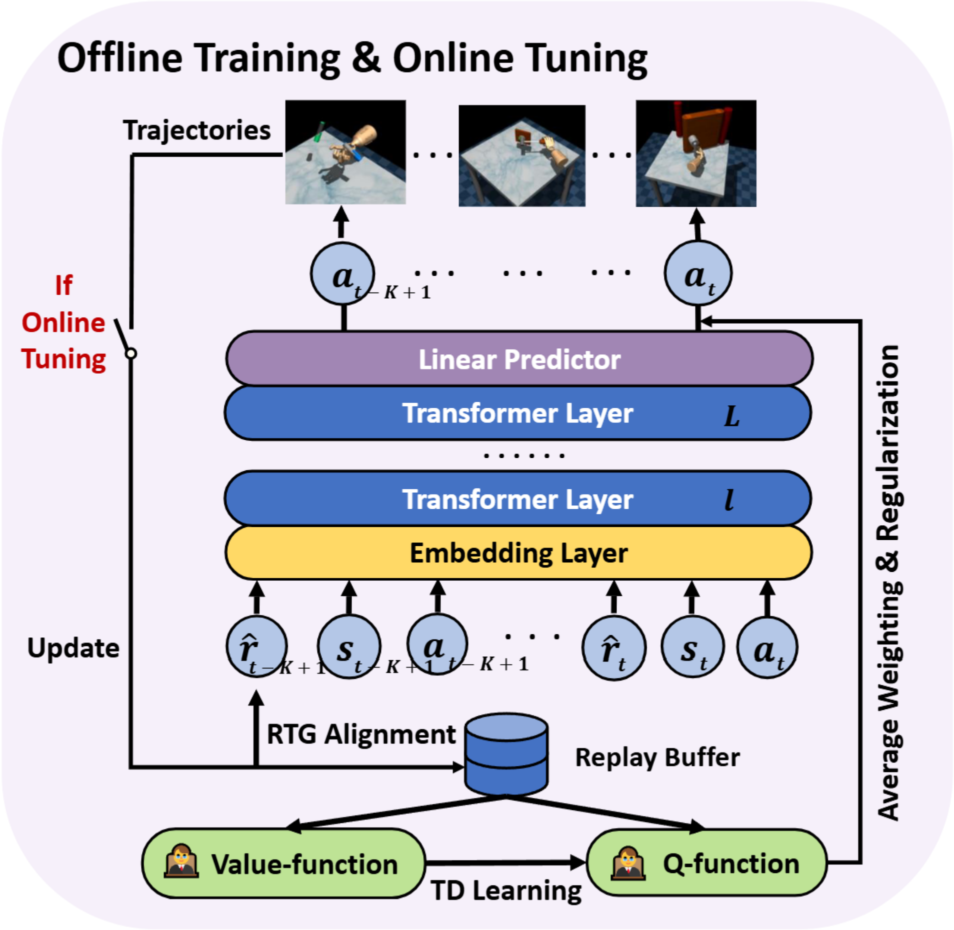
- **Loss function**

$$\mathbb{E}_{\substack{\tau_t\sim\mathcal{D} \\ (s_t,a_t)\sim\tau_t}}\left[\exp(\eta(\min_{i=1,2} Q_{\hat{\theta}_i}(s_t,a_t) - V_\psi(s_t)))\|\pi_{DT}(\tau_t) - a_t\|^2 - \lambda\cdot\min_{i=1,2} Q_{\hat{\theta}_i}(s_t,\pi_{DT}(\tau_t))\right]$$



Online Tuning Phase

- **Trajectory-Level Replay Buffer:** During the online tuning phase, VDT adopts a trajectory-level replay buffer, storing complete trajectories rather than single transitions. The buffer is first filled with the highest-return trajectories from offline data, and is updated in a first-in-first-out manner whenever the policy generates a new trajectory. A two-step sampling strategy ensures uniform sampling of sub-trajectories, improving the diversity and quality of training data during online updates.
- **Return-to-go Alignment:** VDT implements a return-to-go (RTG) alignment mechanism. Instead of conditioning on a fixed, predefined RTG as in offline training, the RTG token is dynamically updated at each timestep with the actual rewards collected by the agent during online interaction.



Sampling Process

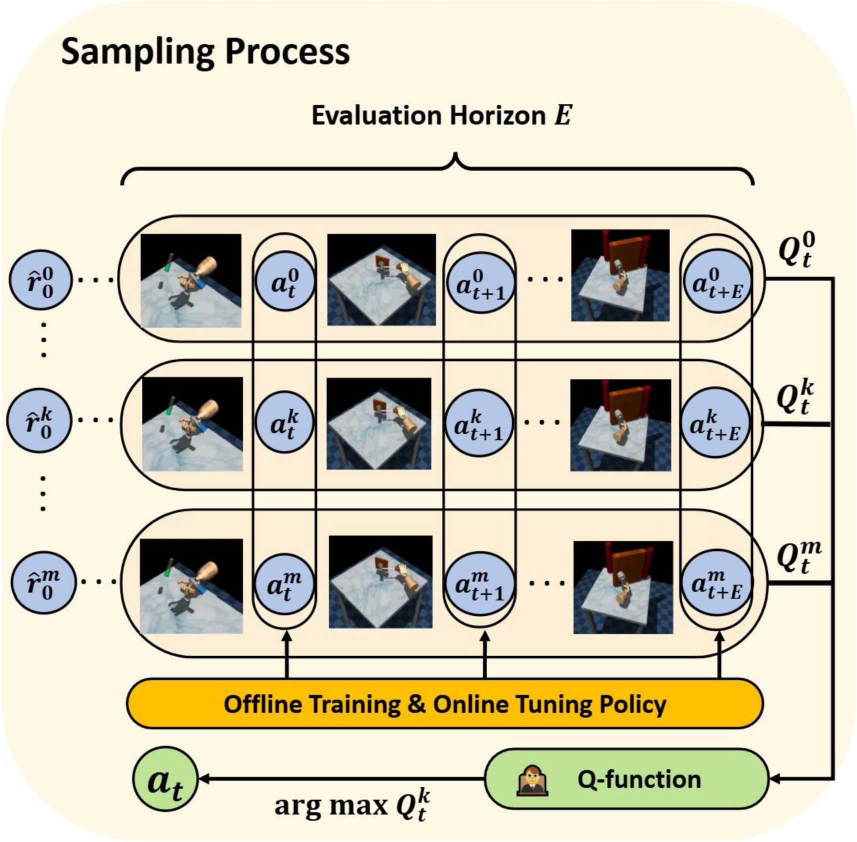
Algorithm 3 Sampling Process

Input: Initial state s_0, candidate RTGs $\{\hat{r}_0^1, ..., \hat{r}_0^m\}$, Evaluation horizon E, Discount γ, Policy π_{DT}, Q-networks $Q_{\hat{\theta}_1}, Q_{\hat{\theta}_2}$

Initialize: Current state $s_t \leftarrow s_0$, Active trajectories $\{\tau^k\}_{k=1}^m \leftarrow \{(s_0, \hat{r}_0^k)\}_{k=1}^m$, Target Q-networks $Q_{\hat{\theta}_i}$

while not termination condition **do**
  // Parallel candidate action generation
  for $k = 1$ **to** m **in parallel do**
    Sample action $a_t^k \sim \pi_{DT}(\tau^k)$
  end for
  // Batched trajectory prediction
  for $k = 1$ **to** m **in parallel do**
    Initialize predicted trajectory $\tau_{\text{pred}}^k \leftarrow (s_t, a_t^k)$
    Initialize cumulative Q-value $Q_{\text{total}}^k \leftarrow 0$
    for $i = 0$ **to** $E - 1$ **do**
      Predict next state: $s_{t+i+1}^k \leftarrow \text{EnvModel}(\tau_{\text{pred}}^k)$
      Sample next action: $a_{t+i+1}^k \sim \pi_{DT}(\tau_{\text{pred}}^k)$
      Compute Q-value: $q_i^k = \min_{j=1,2} Q_{\hat{\theta}_j}(s_{t+i}^k, a_{t+i}^k)$
      Accumulate: $Q_{\text{total}}^k \leftarrow Q_{\text{total}}^k + \gamma^i q_i^k$
      Append $(a_{t+i+1}^k, s_{t+i+1}^k)$ to τ_{pred}^k
    end for
  end for
  // Optimal action selection
  Select optimal index: $k^* \leftarrow \arg\max_{1 \le k \le m} Q_{\text{total}}^k$
  Execute action: $a_t \leftarrow a_t^{k^*}$
  // Environment interaction & trajectory update
  Observe reward r_t, next state s_{t+1} from environment
  for $k = 1$ **to** m **do**
    Update RTG: $\hat{r}_{t+1}^k \leftarrow \hat{r}_t^k - r_t$
    Append transition: $\tau^k \leftarrow \tau^k \oplus (a_t, r_t, s_{t+1}, \hat{r}_{t+1}^k)$
  end for
  $t \leftarrow t + 1$
end while

$$\boxed{Q_t^k = \sum_{i=0}^{E} \gamma^i \cdot Q(s_{t+i}^k, a_{t+i}^k)}$$



Experiment

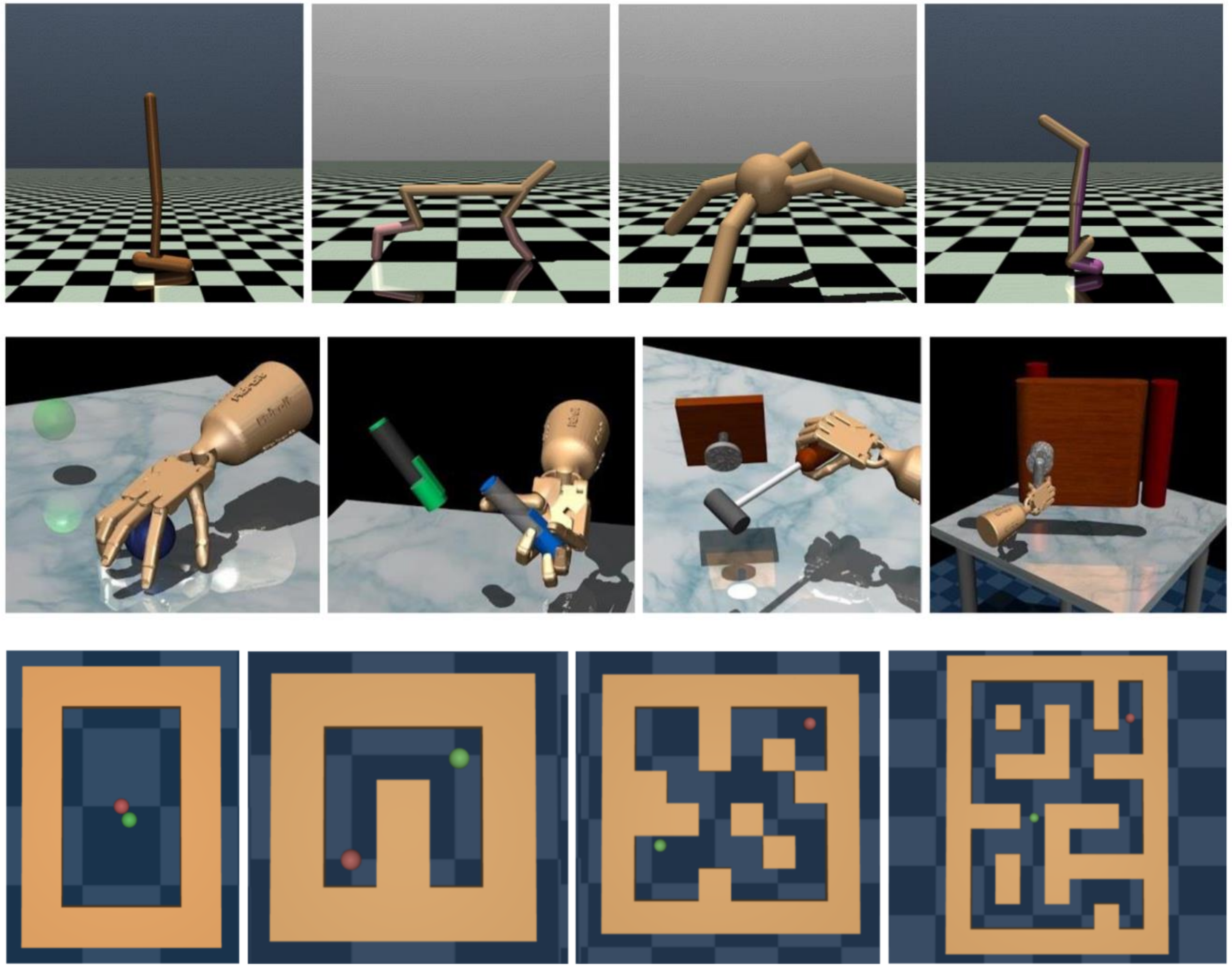


Table 1: Offline training performance of VDT and state-of-the-art baselines on D4RL tasks. For VDT, results are reported as the mean and standard error of normalized rewards over 30 random rollouts (3 independently trained models with 10 trajectories each), generally showing low variance.

Dataset	Value-Based Methods					Conditional Sequence Modeling Methods						
Gym Tasks	**BEAR**	**BCQ**	**CQL**	**IQL**	**MoRel**	**BC**	**DT**	**StAR**	**GDT**	**CGDT**	**DC**	**VDT**
halfcheetah-medium-replay-v2	38.6	34.8	37.5	**44.1**	40.2	36.6	36.6	36.8	40.5	40.4	41.3	$39.4_{\pm 2.0}$
hopper-medium-replay-v2	33.7	31.1	95.0	92.1	93.6	18.1	82.7	29.2	85.3	93.4	94.2	$\mathbf{96.0}_{\pm 1.9}$
walker2d-medium-replay-v2	19.2	13.7	77.2	73.7	49.8	32.3	79.4	39.8	77.5	78.1	76.6	$\mathbf{82.3}_{\pm 2.1}$
halfcheetah-medium-v2	41.7	41.5	44.0	**47.4**	42.1	42.6	42.6	42.9	42.9	43.0	43.0	$43.9_{\pm 0.7}$
hopper-medium-v2	52.1	65.1	58.5	63.8	95.4	52.9	67.6	59.5	77.1	96.9	92.5	$\mathbf{98.3}_{\pm 0.1}$
walker2d-medium-v2	59.1	52.0	72.5	79.9	77.8	75.3	74.0	73.8	76.5	79.1	79.2	$\mathbf{81.6}_{\pm 1.7}$
halfcheetah-medium-expert-v2	53.4	69.6	91.6	86.7	53.3	55.2	86.8	93.7	93.2	93.6	93.0	$\mathbf{93.9}_{\pm 0.1}$
hopper-medium-expert-v2	96.3	109.1	105.4	91.5	108.7	52.5	107.6	111.1	111.1	107.6	110.4	$\mathbf{111.5}_{\pm 3.8}$
walker2d-medium-expert-v2	40.1	67.3	108.8	109.6	95.6	107.5	108.1	109.0	107.7	109.3	109.6	$\mathbf{110.4}_{\pm 0.9}$
Average	48.2	53.8	77.6	76.5	72.9	52.6	76.2	66.2	79.1	82.4	82.2	**84.1**
Adroit Tasks	**BEAR**	**BCQ**	**CQL**	**IQL**	**MoRel**	**EDAC**	**BC**	**DT**	**D-QL**	**StAR**	**GDT**	**VDT**
pen-human-v1	-1.0	66.9	37.5	71.5	-3.2	52.1	63.9	79.5	72.8	77.9	92.5	$\mathbf{126.7}_{\pm 4.3}$
hammer-human-v1	2.7	0.9	4.4	1.4	2.3	0.8	1.2	3.7	0.2	3.7	**5.5**	$3.2_{\pm 0.3}$
door-human-v1	2.2	-0.05	9.9	4.3	2.3	10.7	2.0	14.8	0.0	1.5	18.6	$\mathbf{19.7}_{\pm 0.5}$
pen-cloned-v1	-0.2	50.9	39.2	37.3	-0.2	68.2	37.0	75.8	57.3	33.1	86.2	$\mathbf{145.6}_{\pm 4.0}$
hammer-cloned-v1	2.3	0.4	2.1	2.1	2.3	0.3	0.6	3.0	3.1	0.3	8.9	$\mathbf{19.6}_{\pm 1.6}$
door-cloned-v1	2.3	0.01	0.4	1.6	2.3	9.6	0.0	16.3	0.0	0.0	19.8	$\mathbf{30.6}_{\pm 0.7}$
Average	1.0	19.8	15.6	19.7	1.0	23.6	17.5	32.2	22.2	19.4	38.9	**57.6**
Kitchen Tasks	**BEAR**	**BCQ**	**CQL**	**IQL**	**O-RL**	**BC**	**DT**	**DD**	**StAR**	**GDT**	**DC**	**VDT**
kitchen-complete-v0	0.0	8.1	43.8	62.5	2.0	65.0	50.8	65.0	40.8	43.8	40.9	$\mathbf{65.9}_{\pm 0.2}$
kitchen-partial-v0	13.1	18.9	49.8	46.3	35.5	33.8	57.9	57.0	12.3	73.3	66.8	$\mathbf{76.1}_{\pm 10.8}$
Average	6.6	13.5	46.8	54.4	18.8	51.5	54.4	61.0	26.6	58.6	58.7	**71.0**
Maze2D Tasks	**BEAR**	**BCQ**	**CQL**	**IQL**	**COMBO**	**BC**	**MPPI**	**DT**	**QDT**	**GDT**	**DC**	**VDT**
maze2d-umaze-v1	65.7	49.1	86.7	42.1	76.4	85.7	33.2	31.0	57.3	50.4	20.1	$\mathbf{88.0}_{\pm 4.6}$
maze2d-medium-v1	25.0	17.1	41.8	34.9	38.5	38.3	10.2	8.2	13.3	7.8	38.2	$\mathbf{60.3}_{\pm 0.5}$
Average	45.35	33.1	64.3	38.5	72.5	63.6	21.7	19.6	35.3	29.1	57.6	**74.2**
AntMaze Tasks	**BEAR**	**BCQ**	**CQL**	**IQL**	**O-RL**	**BC**	**DT**	**RvS**	**StAR**	**GDT**	**DC**	**VDT**
antmaze-umaze-v0	73.0	78.9	74.0	87.1	64.3	54.6	59.2	65.4	51.3	76.0	85.0	$\mathbf{100.0}_{\pm 5.5}$
antmaze-umaze-diverse-v0	61.0	55.0	84.0	64.4	60.7	45.6	66.2	60.9	45.6	69.0	78.5	$\mathbf{100.0}_{\pm 4.7}$
antmaze-medium-diverse-v0	8.0	0.0	53.7	**70.0**	0.0	0.0	7.5	67.3	0.0	0.0	0.0	$30.0_{\pm 2.8}$
Average	47.3	44.6	70.6	73.8	41.7	33.4	44.3	75.0	32.3	48.3	54.5	**76.7**

Experiment

Table 2: Offline-to-online performance of each method, with average rewards reported before (left of arrow) and after (right of arrow) online tuning.

Dataset	TD3+BC	AWAC	CQL	IQL	PDT	ODT	VDT
halfcheetah-medium-replay-v2	44.6 → 48.1	24.3 → 39.0	45.5 → 44.3	44.1 → 44.0	31.4 → 42.8	39.9 → 40.4	39.4 → **49.2**
hopper-medium-replay-v2	60.9 → 90.7	77.3 → 79.6	95.0 → 95.3	92.1 → 93.5	84.5 → 94.8	86.6 → 88.9	96.0 → **119.2**
walker2d-medium-replay-v2	81.8 → 82.0	63.8 → 44.0	77.2 → 78.0	73.7 → 60.9	54.5 → 79.0	68.9 → 76.9	82.3 → **95.5**
halfcheetah-medium-v2	48.3 → 50.9	37.4 → 41.1	44.0 → 29.1	47.4 → 48.0	39.4 → **69.5**	42.7→ 42.2	43.9 → 53.5
hopper-medium-v2	59.3 → 64.6	72.0 → 91.0	58.5 → 95.7	63.8 → 44.3	74.4 → 100.2	66.9 → 97.5	98.3 → **108.1**
walker2d-medium-v2	83.7 → 85.2	30.1 → 79.1	72.5 → 89.4	79.9 → 68.9	63.4 → 88.1	72.2 → 76.8	81.6 → **89.8**
halfcheetah-medium-expert-v2	90.7 → 92.1	36.8 → 41.0	91.6 → 99.9	86.7 → 95.3	82.6 → 93.3	36.8 → 100.9	93.9 → **101.7**
hopper-medium-expert-v2	98.0 → 110.2	80.9 → 111.9	105.4 → 106.3	91.5 → 92.9	77.0 → 80.0	74.3 → 99.1	111.5 → **117.8**
walker2d-medium-expert-v2	110.1 → 110.1	42.7 → 78.3	108.8 → 110.1	109.6 → 109.6	99.1 → 108.9	62.0 → 78.7	110.4 → **112.7**
antmaze-umaze-v0	78.6 → 79.1	56.7 → 59.0	70.1 → 99.4	86.7 → 96.0	48.6 → 66.8	53.1 → 88.5	100.0 → **110.0**
antmaze-umaze-diverse-v0	71.4 → 78.1	49.3 → 49.0	31.1 → 99.4	75.0 → 84.0	72.7 → 79.3	50.2 → 56.0	100.0 → **100.0**
antmaze-medium-diverse-v0	0.0 → 56.7	0.7 → 0.3	23.0 → 32.3	68.3 → 72.0	8.0 → 63.4	0.8 → 55.6	20.0 → **75.0**
Average	79.0	59.4	81.6	75.78	80.51	75.13	**94.38**

Table 3: Ablation study on model components during offline training. We have abbreviated some task names for simplicity, which does not affect understanding. All experiments are repeated three times, and the average value is taken.

Advantage Weighting	Regularization	Sampling	hopper-m	walker-m-e	pen-cloned	maze2d-m	antmaze-u
✓			90.3	99.9	86.1	12.1	75.1
	✓		88.9	78.1	99.3	30.5	60.9
		✓	78.6	80.3	82.0	19.3	0.0
✓	✓		95.6	103.6	131.8	40.5	95.9
✓	✓	✓	**98.3**	**110.4**	**145.6**	**60.3**	**100.0**

Experiment

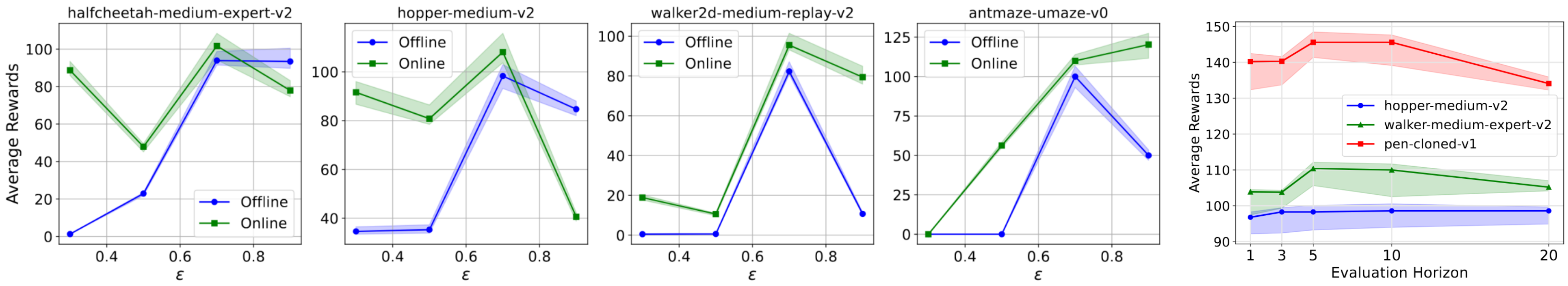


Table 4: Ablation on the computational complexity.

Complexity	Offline Training			Online Tuning		
	IQL	ODT	VDT	IQL	ODT	VDT
Memory ↓	2128 M	3968 M	4024M	2128M	3968 M	4024M
Params ↑	3.31 M	5.01M	5.24 M	3.31 M	5.01M	5.24 M
Clock Time ↓	≈ 6.0 h	≈ 9.0 h	≈ 5.0 h	≈ 10.0 h	≈ 4.5 h	≈ 4.0 h

Table 5: Offline training performance of VDT with different context lengths (K) on Gym tasks.

Datasets	VDT (8)	VDT (20)	VDT (60)	VDT(120)
halfcheetah-medium	28.6	43.9	**44.6**	43.0
hopper-medium	77.0	98.3	**99.1**	65.4
walker2d-medium	52.6	**81.6**	79.9	80.5
halfcheetah-medium-expert	89.5	**93.9**	**93.9**	77.0
hopper-medium-expert	109.3	111.5	**112.7**	111.2
walker2d-medium-expert	100.6	**110.4**	**110.4**	103.8
Average	76.3	89.9	**90.1**	80.2

Conclusion

- We incorporate the value function into the CSM architecture and enhance behavior cloning with advantage-weighted learning and regularization constraints. We further provide a theoretical guarantee of its superior performance.
- We leverage the inherent strengths of the value function to fine-tune the policy with a limited number of interactions in the online phase. By introducing the trajectory-level replay buffer and return-to-go alignment, we bridge the gap between offline training and online tuning, offering insights into the design of generalizable architectures.
- We demonstrate the effectiveness of VDT across a broad spectrum of benchmarks, exhibiting superior performance in both pure offline and offline-to-online settings.

Thanks!

hlzheng@whu.edu.cn