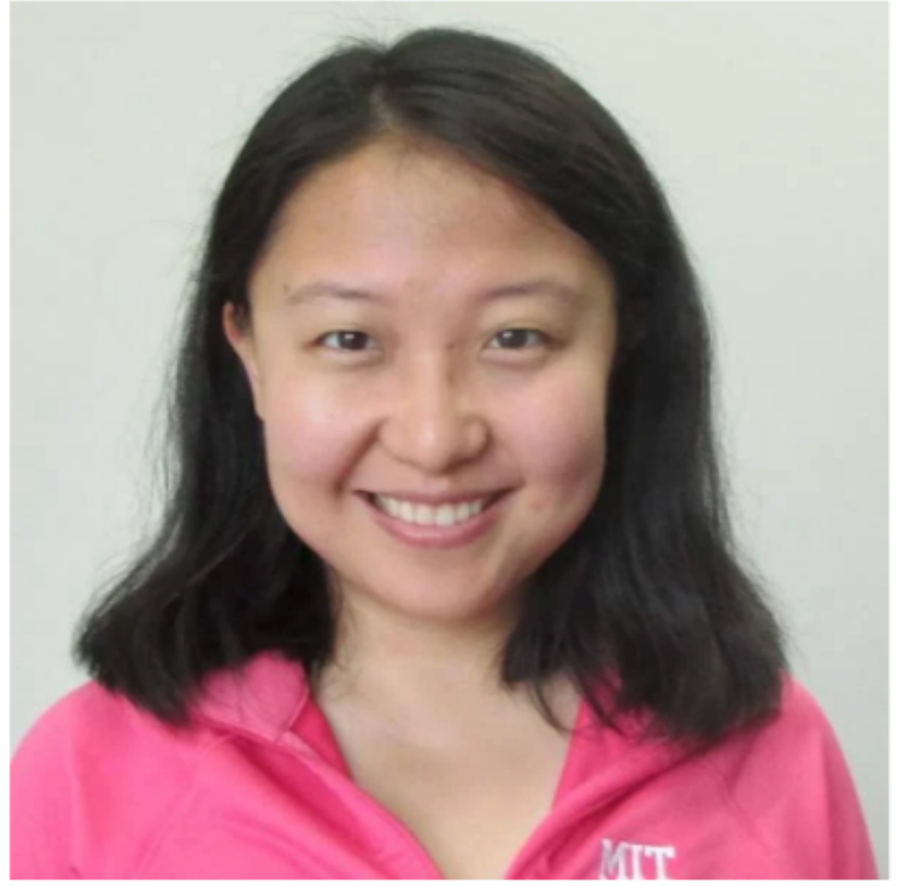
Representation Matters: Offline Representation Learning for Sequential Decision Making

Sherry Yang — sherryy@

Ofir Nachum — ofirnachum@



Paper: https://arxiv.org/abs/2102.05815
Code: https://github.com/google-research/google-research/tree/master/rl_repr

Representation Learning on Offline Data

Given a fixed set of experience $\mathcal{D}$, what can we do?

Representation Learning on Offline Data

Given a fixed set of experience $\mathscr{D}$, what can we do?

- Offline reinforcement learning: $\mathscr{D} \longrightarrow \pi(\,\cdot \mid s)$

Representation Learning on Offline Data

Given a fixed set of experience $\mathcal{D}$, what can we do?

- Offline reinforcement learning: $\mathcal{D} \longrightarrow \pi(\cdot \mid s)$
- Offline representation learning: $\mathcal{D} \longrightarrow \phi(s)$

$$\phi(s) \downarrow \text{Downstream tasks}$$

Representation Learning on Offline Data

What kind of downstream tasks might benefit from representation learning?

Representation Learning on Offline Data

What kind of downstream tasks might benefit from representation learning?

- Imitation in low-data regime: $\mathcal{D}_{expert} \xrightarrow[\text{learning}]{\text{Supervised}} \pi(\cdot \mid \phi(s))$

  Limited expert demonstration, much undirected experience

Representation Learning on Offline Data

What kind of downstream tasks might benefit from representation learning?

- Imitation in low-data regime: $\mathcal{D}_{expert} \xrightarrow[\text{learning}]{\text{Supervised}} \pi(\cdot \mid \phi(s))$

  Limited expert demonstration, much undirected experience

- Offline RL: $\mathcal{D} \longrightarrow \pi(\cdot \mid \phi(s))$

  Expensive/unavailable environments (e.g., recommendation systems)

Representation Learning on Offline Data

What kind of downstream tasks might benefit from representation learning?

- Imitation in low-data regime: $\mathcal{D}_{expert} \xrightarrow[\text{learning}]{\text{Supervised}} \pi(\cdot \mid \phi(s))$

  Limited expert demonstration, much undirected experience

- Offline RL: $\mathcal{D} \longrightarrow \pi(\cdot \mid \phi(s))$

  Expensive/unavailable environments (e.g., recommendation systems)

- Online RL: Env $\pi(\cdot \mid \phi(s))$

  Potentially in partially observable environments

Representation Learning Objectives

Inverse model: $-\log P(a_t | f(\phi(s_{t:t+1})))$ predict action

Representation Learning Objectives

Inverse model: $-\log P(a_t|f(\phi(s_{t:t+1})))$ predict action

Forward raw model: $||r_t - g(\phi(s_t), a_t)||^2 - \log P(s_{t+1}|f(\phi(s_t), a_t))$

Forward latent model (DeepMDP): $||r_t - g(\phi(s_t), a_t)||^2 - \log P(\phi(s_{t+1})|f(\phi(s_t), a_t))$

Forward energy model: $\frac{\rho(s_{t+1})\exp\{\phi(s_{t+1})^\top W f(\phi(s_t), a_t)\}}{\mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W f(\phi(s_t), a_t)\}]}$

predict reward & future state

Representation Learning Objectives

Inverse model: $-\log P(a_t | f(\phi(s_{t:t+1})))$ predict action

Forward raw model: $||r_t - g(\phi(s_t), a_t)||^2 - \log P(s_{t+1} | f(\phi(s_t), a_t))$

Forward latent model (DeepMDP): $||r_t - g(\phi(s_t), a_t)||^2 - \log P(\phi(s_{t+1}) | f(\phi(s_t), a_t))$

Forward energy model: $\frac{\rho(s_{t+1}) \exp\{\phi(s_{t+1})^\top W f(\phi(s_t), a_t)\}}{\mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W f(\phi(s_t), a_t)\}]}$ predict reward & future state

Value prediction network (VPN): $\phi(s_t), a_{t:t+k} \rightarrow r_{t+k}, V_{t+k}$ predict future reward & value function

Representation Learning Objectives

Inverse model: $-\log P(a_t | f(\phi(s_{t:t+1})))$ predict action

Forward raw model: $||r_t - g(\phi(s_t), a_t)||^2 - \log P(s_{t+1} | f(\phi(s_t), a_t))$

Forward latent model (DeepMDP): $||r_t - g(\phi(s_t), a_t)||^2 - \log P(\phi(s_{t+1}) | f(\phi(s_t), a_t))$

Forward energy model: $\frac{\rho(s_{t+1}) \exp\{\phi(s_{t+1})^\top W f(\phi(s_t), a_t)\}}{\mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W f(\phi(s_t), a_t)\}]}$ predict reward & future state

Value prediction network (VPN): $\phi(s_t), a_{t:t+k} \to r_{t+k}, V_{t+k}$ predict future reward & value function

Deep bisimulation: $\phi(s_1) - \phi(s_2) \leftrightarrow d(s_1, s_2)$ state distance ↔ bisimulation distance

Representation Learning Objectives

Inverse model: $-\log P(a_t | f(\phi(s_{t:t+1})))$ predict action

Forward raw model: $||r_t - g(\phi(s_t), a_t)||^2 - \log P(s_{t+1} | f(\phi(s_t), a_t))$

Forward latent model (DeepMDP): $||r_t - g(\phi(s_t), a_t)||^2 - \log P(\phi(s_{t+1}) | f(\phi(s_t), a_t))$

Forward energy model: $\dfrac{\rho(s_{t+1}) \exp\{\phi(s_{t+1})^\top W f(\phi(s_t), a_t)\}}{\mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W f(\phi(s_t), a_t)\}]}$ predict reward & future state

Value prediction network (VPN): $\phi(s_t), a_{t:t+k} \rightarrow r_{t+k}, V_{t+k}$ predict future reward & value function

Deep bisimulation: $\phi(s_1) - \phi(s_2) \leftrightarrow d(s_1, s_2)$ state distance ↔ bisimulation distance

Temporal contrastive learning (TCL): $-\phi(s_{t+1})^\top W \phi(s_t) + \log \mathbb{E}_\rho[\exp\{\phi(\tilde{s})^\top W \phi(s_t)\}]$

Attentive contrastive learning (ACL): BERT-style contrastive learning of $\phi(s_t)$

contrast two state representations

Carles Gelada, et al., Deepmdp: Learning continuous latent space models for representation learning. In International Conference on Machine Learning, pages 2170–2179. PMLR, 2019.
Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning, 2020.
Amy Zhang, et al., Learning invariant representations for reinforcement learning without reconstruction. arXiv preprint arXiv:2006.10742, 2020.
Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. arXiv preprint arXiv:1707.03497, 2017.
Deepak Pathak, et al., Curiosity-driven exploration by self-supervised prediction. In International Conference on Machine Learning, pages 2778–2787. PMLR, 2017.
Evan Shelhamer, et. al., Loss is its own reward: Self-supervision for reinforcement learning. CoRR, abs/1612.07307, 2016. URL http://arxiv.org/abs/1612.07307.

Task Setups

Imitation			
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset:	{domain}-{data}-v0
Choose data $\in$ {medium, medium-replay}	$\rightarrow$	Downstream task:	Behavioral cloning (BC) on first N
Choose $N \in \{10000, 25000\}$			transitions from {domain}-expert-v0

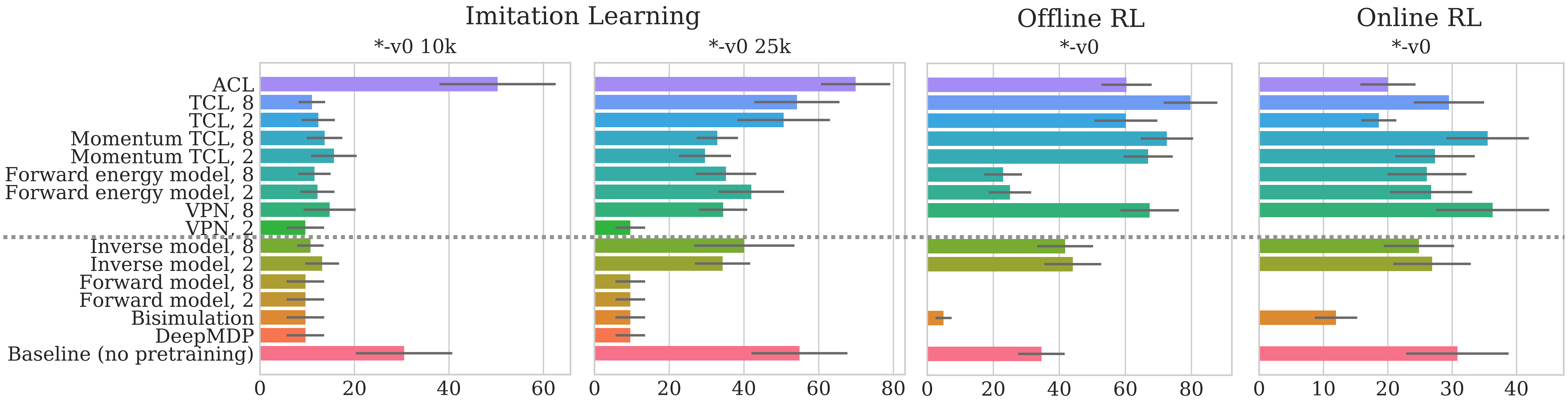
Task Setups

Imitation		
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset: {domain}-{data}-v0
Choose data $\in$ {medium, medium-replay}	$\rightarrow$	Downstream task: Behavioral cloning (BC) on first N transitions from {domain}-expert-v0
Choose $N \in \{10000, 25000\}$		
Offline RL		
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset: {domain}-{data}-v0
Choose data $\in$ {expert, medium-expert, medium, medium-replay}	$\rightarrow$	Downstream task: Behavior regularized actor critic (BRAC) on data from {domain}-{data}-v0

Task Setups

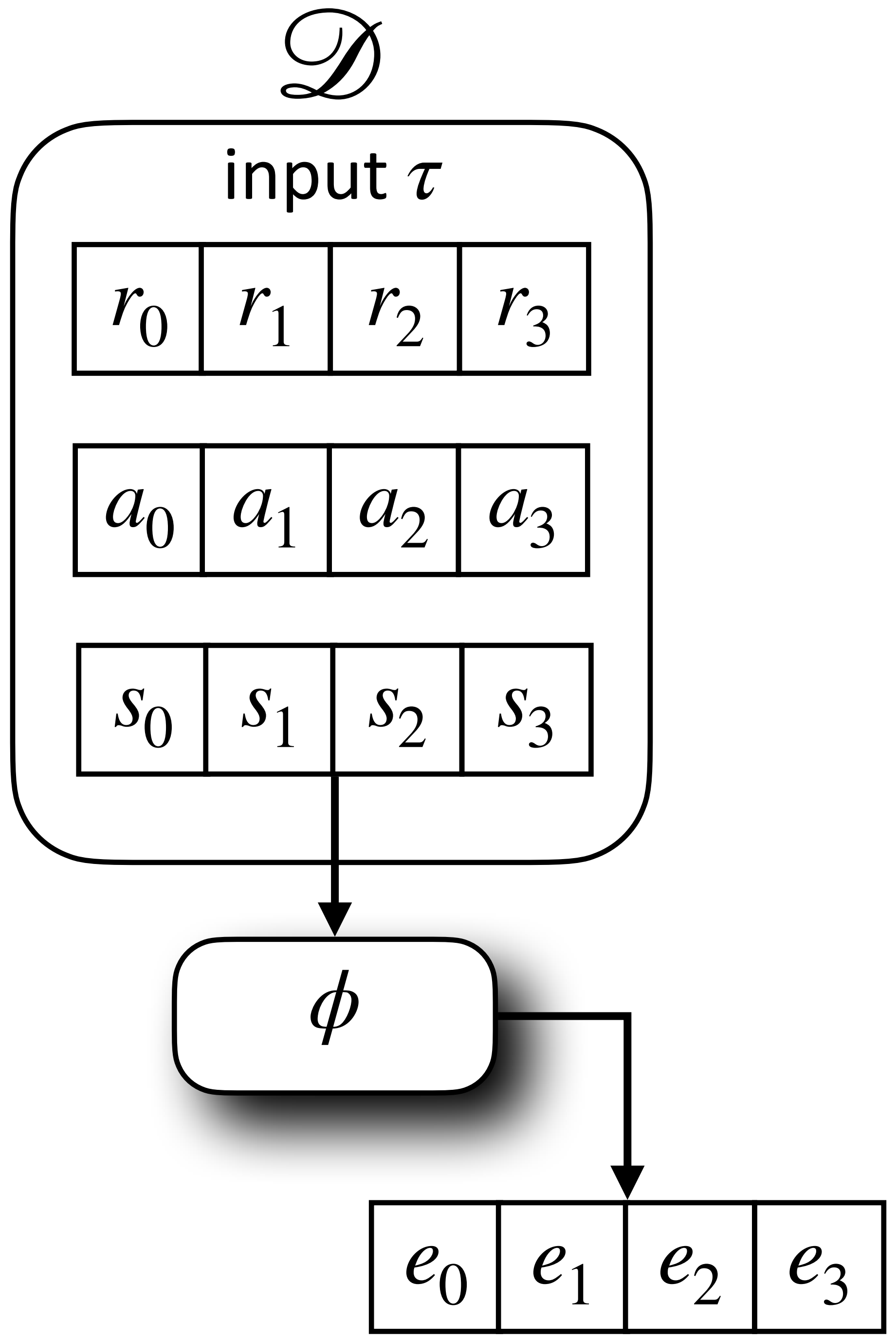
Imitation			
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset:	{domain}-{data}-v0
Choose data $\in$ {medium, medium-replay}	$\rightarrow$	Downstream task:	Behavioral cloning (BC) on first N
Choose $N \in \{10000, 25000\}$			transitions from {domain}-expert-v0
Offline RL			
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset:	{domain}-{data}-v0
Choose data $\in$ {expert, medium-expert, medium,	$\rightarrow$	Downstream task:	Behavior regularized actor critic (BRAC)
medium-replay}			on data from {domain}-{data}-v0
Online RL			
Choose domain $\in$ {halfcheetah, hopper, walker2d, ant}		Offline dataset:	{domain}-{data}-v0 with random masking
Choose data $\in$ {expert, medium-expert, medium}	$\rightarrow$	Downstream task:	Soft actor critic (SAC) on randomly
medium-replay}			masked version of {domain}

Breadth Study

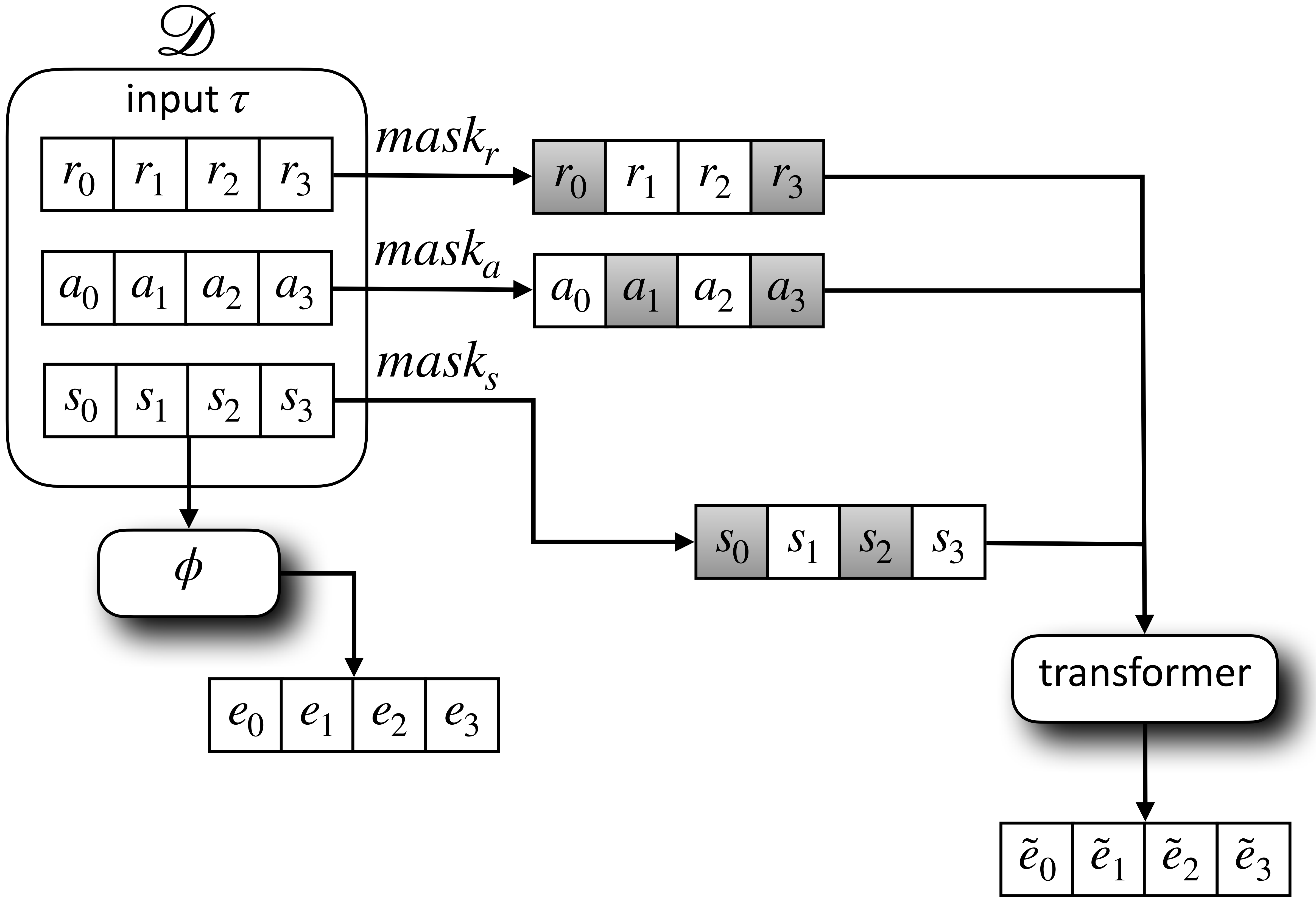


- Representation learning on average improves imitation learning, offline RL, and online RL tasks by 1.5x, 2.5x, and 15%
- Forward models of future representations (e.g., DeepMDP, Bisimulation) exhibit poor performance
- Contrastive self-prediction (e.g., ACL, TCL, VPN) works the best
- What is important in representation learning? Reward/action prediction? Direction of prediction? Momentum?

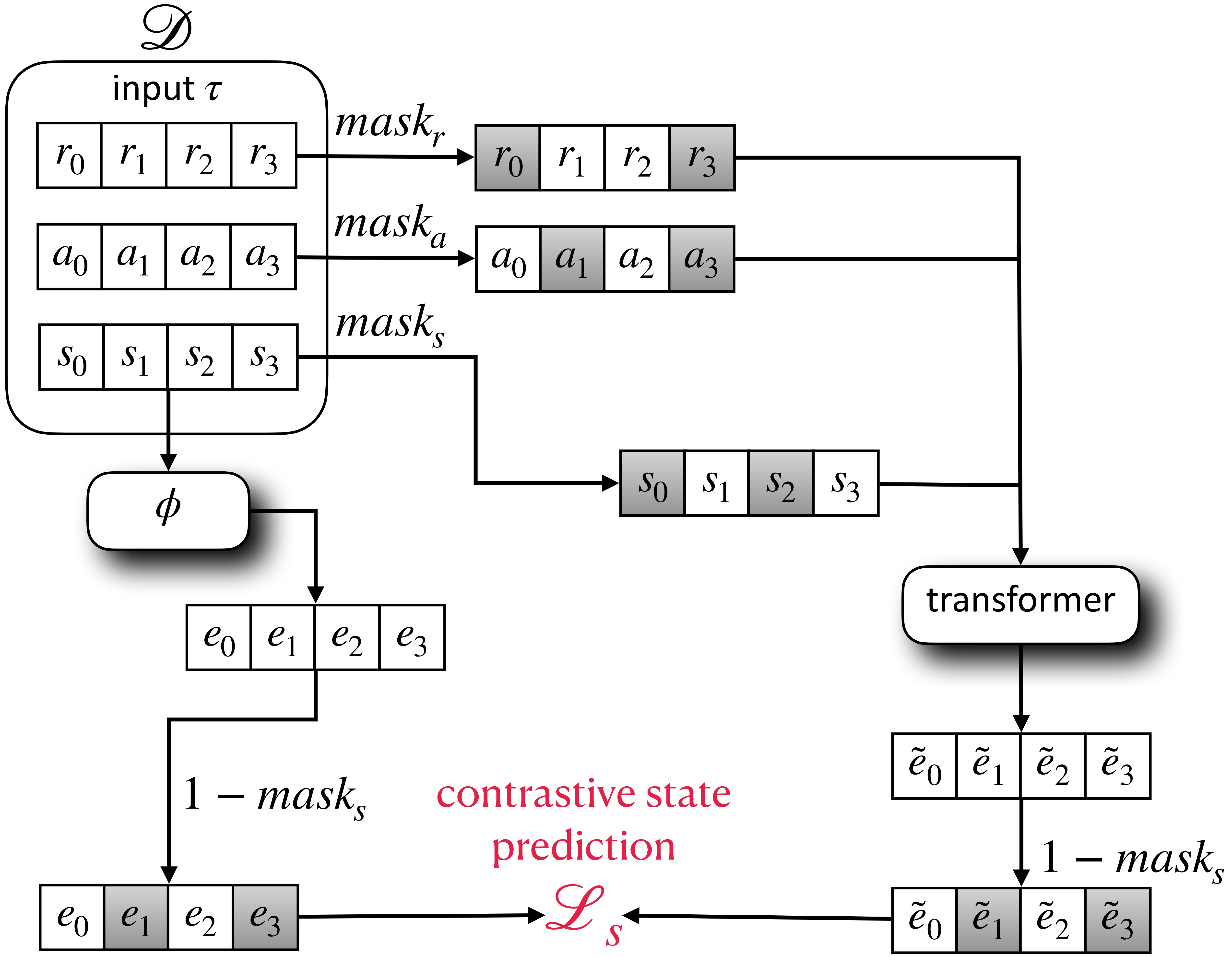
Attentive Contrastive Learning (ACL)



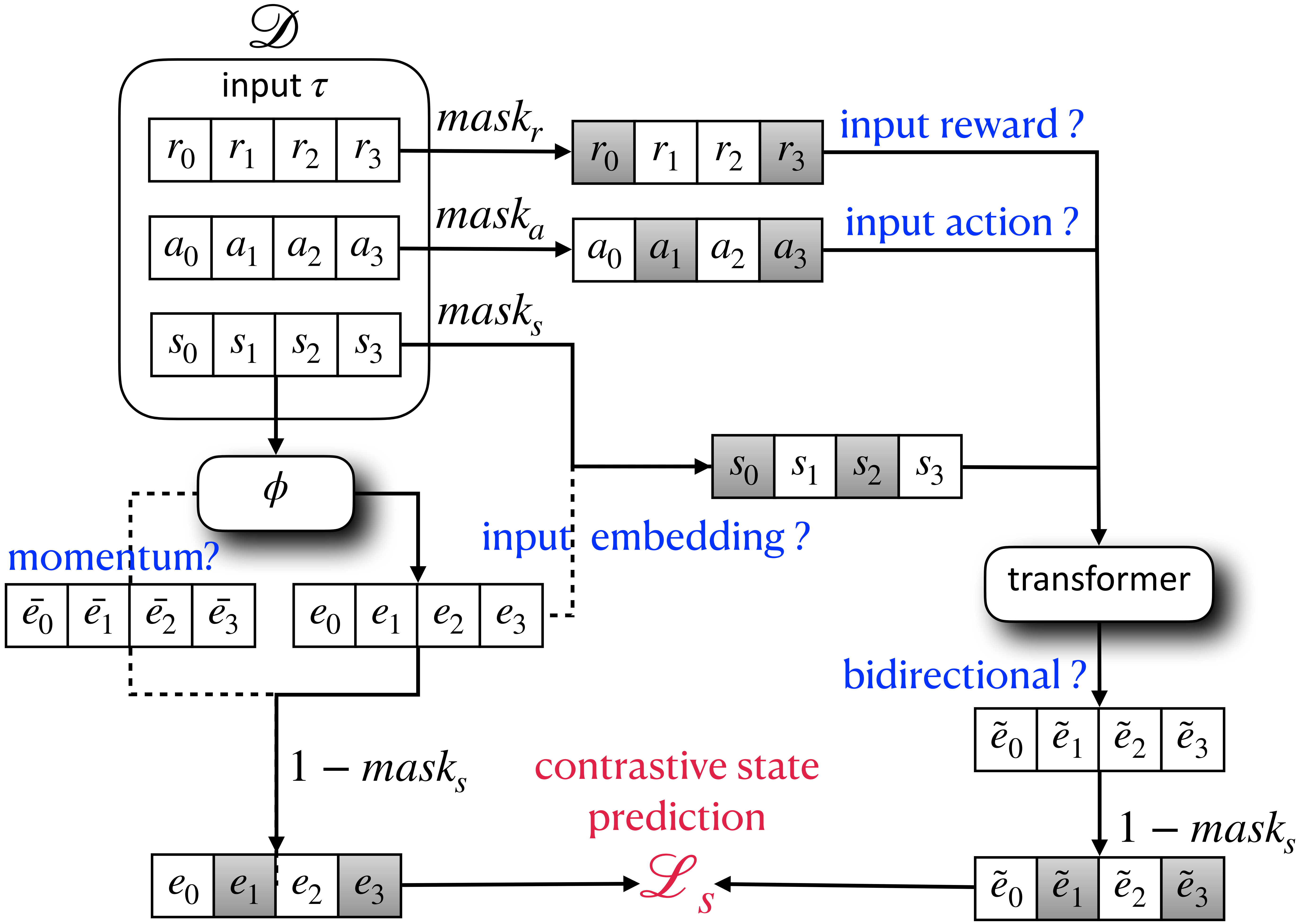
Attentive Contrastive Learning (ACL)



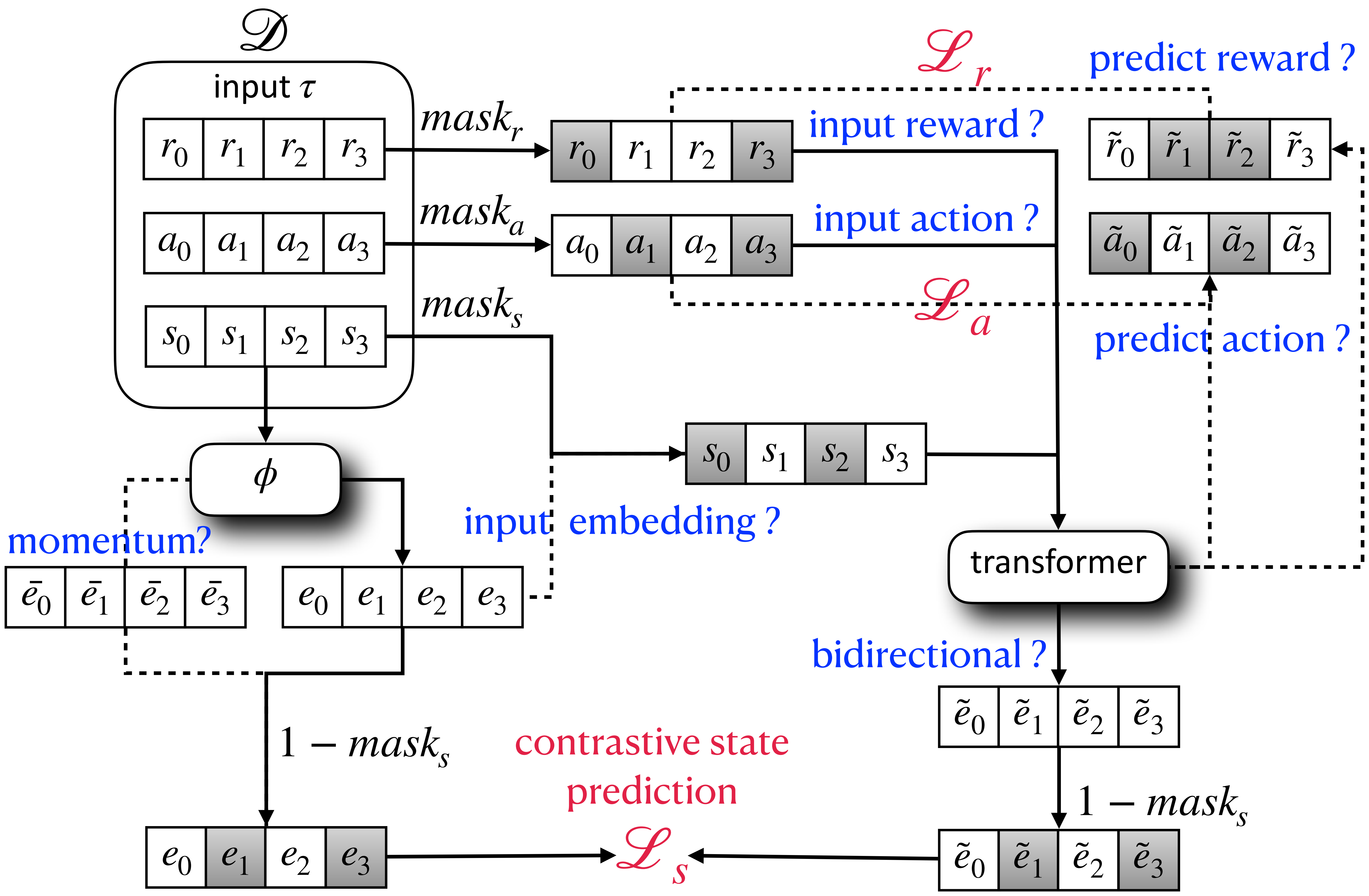
Attentive Contrastive Learning (ACL)



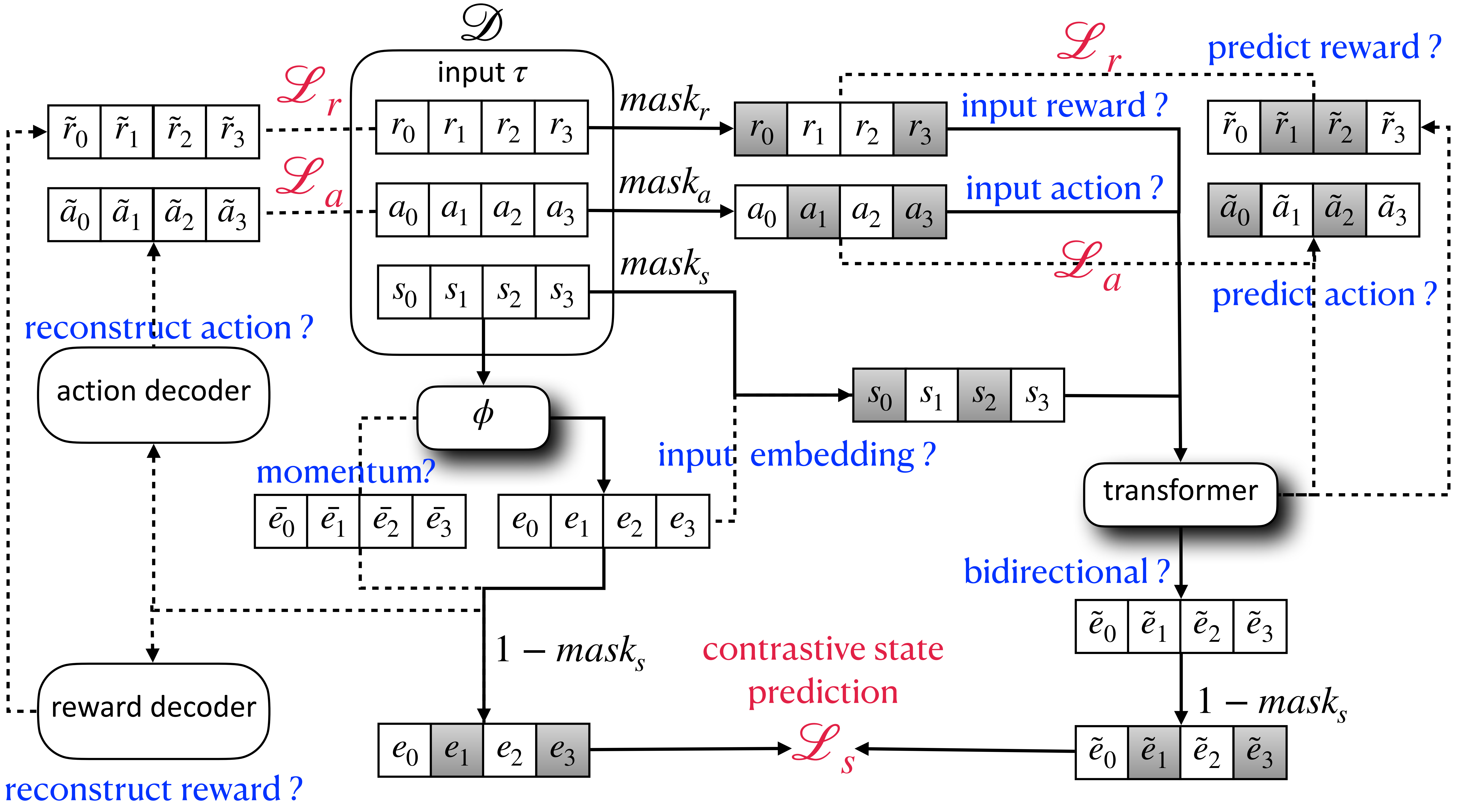
Attentive Contrastive Learning (ACL)



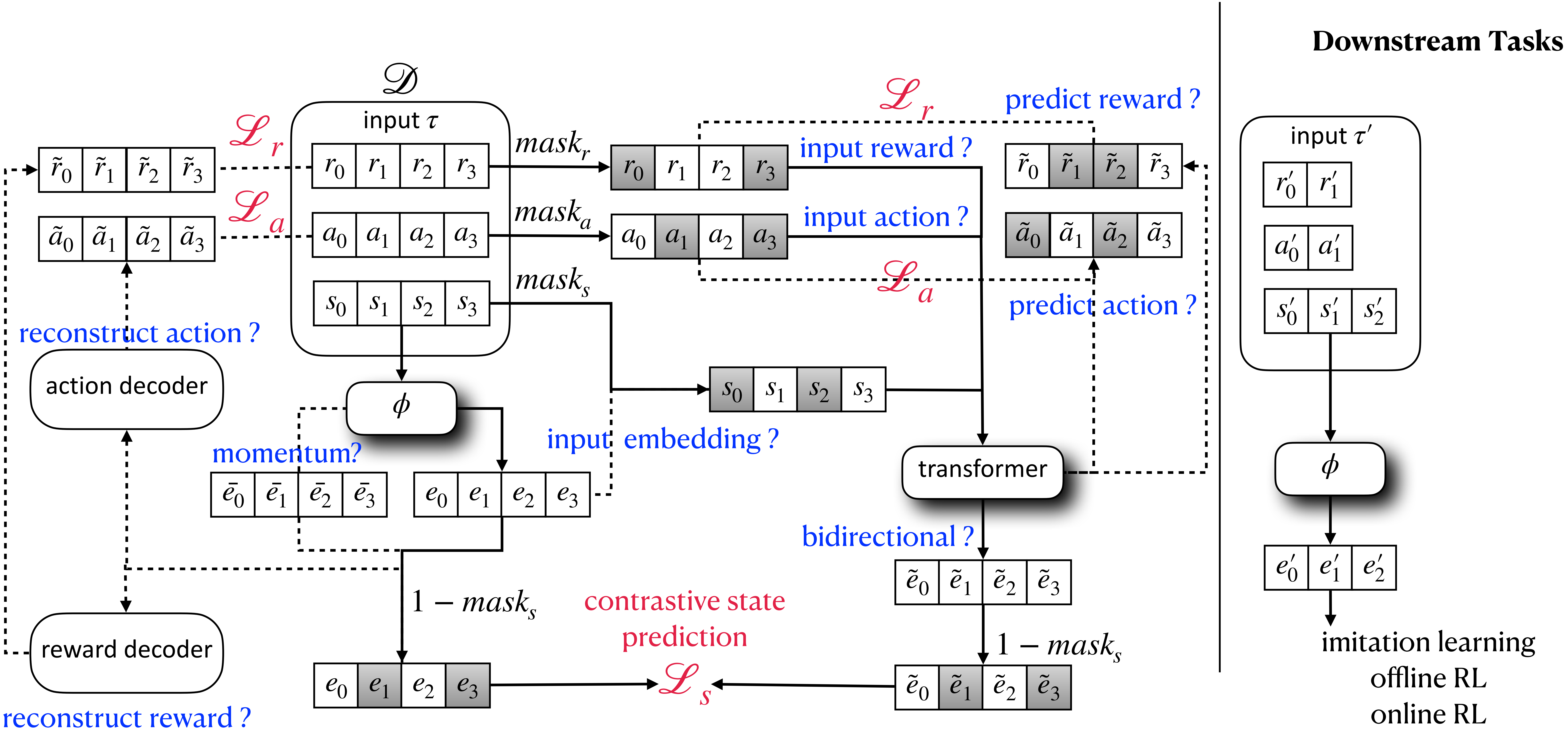
Attentive Contrastive Learning (ACL)



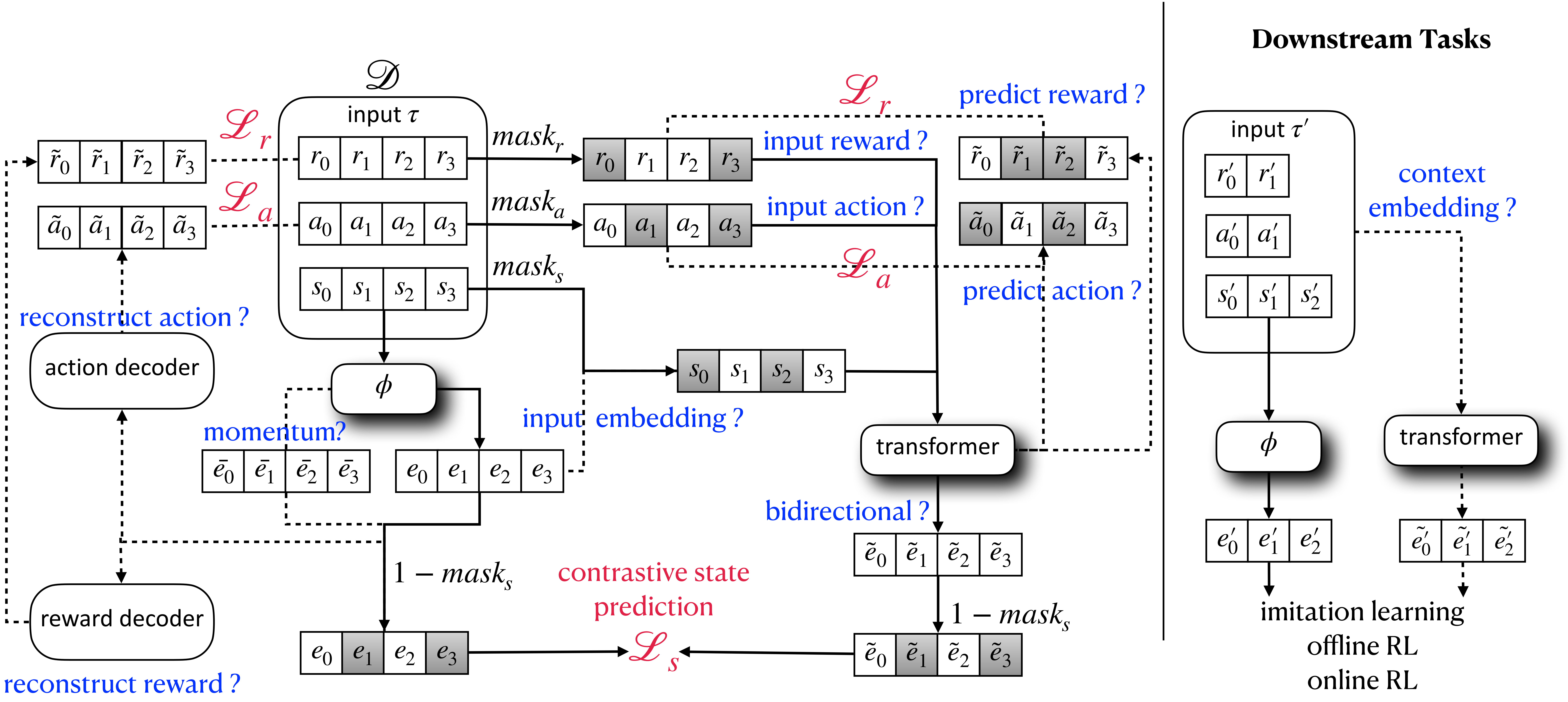
Attentive Contrastive Learning (ACL)



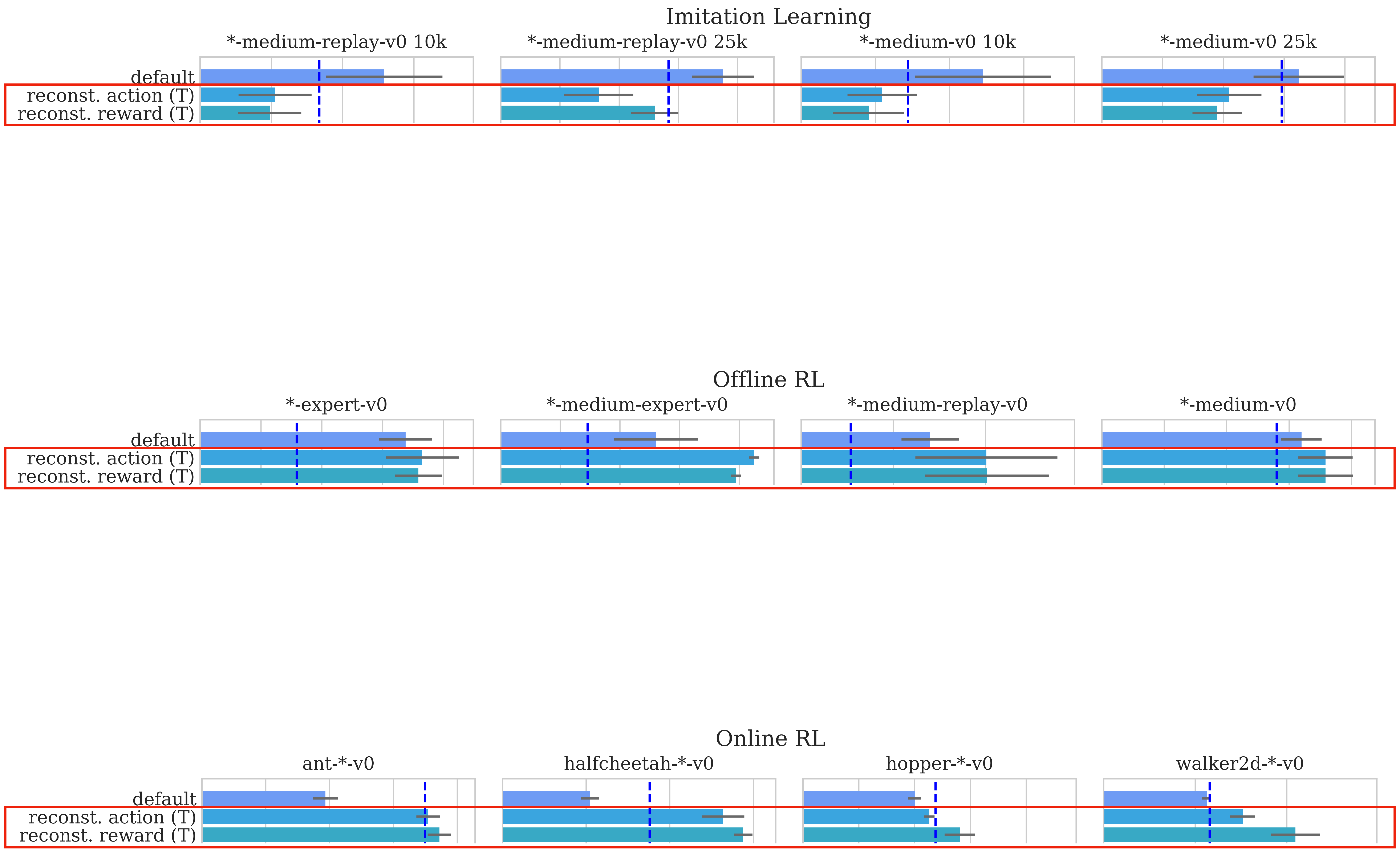
Attentive Contrastive Learning (ACL)



Attentive Contrastive Learning (ACL)



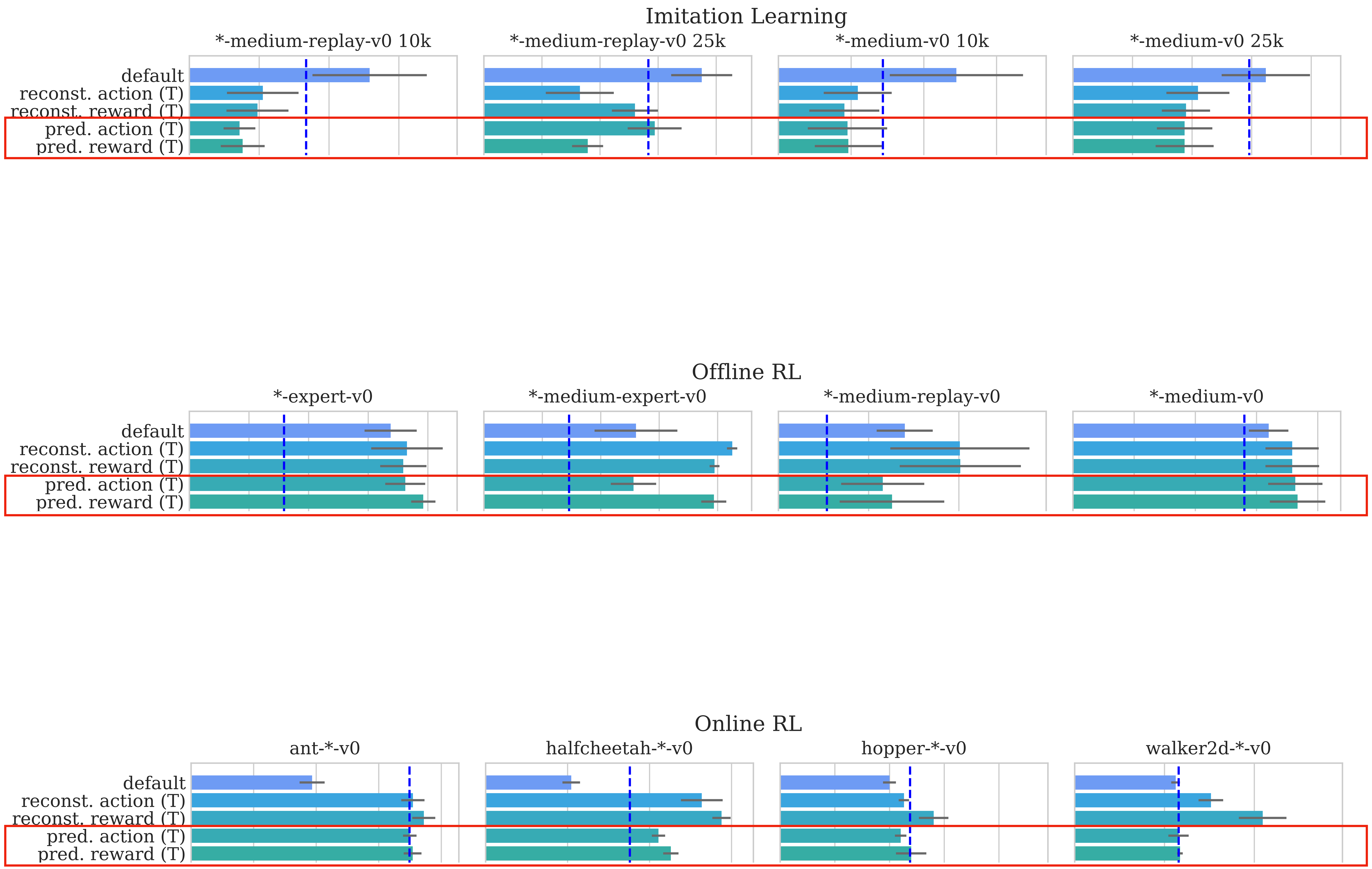
Depth Study on ACL



Depth Study on ACL

Factor	Description	Imitation	Offline	Online
reconstruct action	Add action prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
reconstruct reward	Add a reward prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$

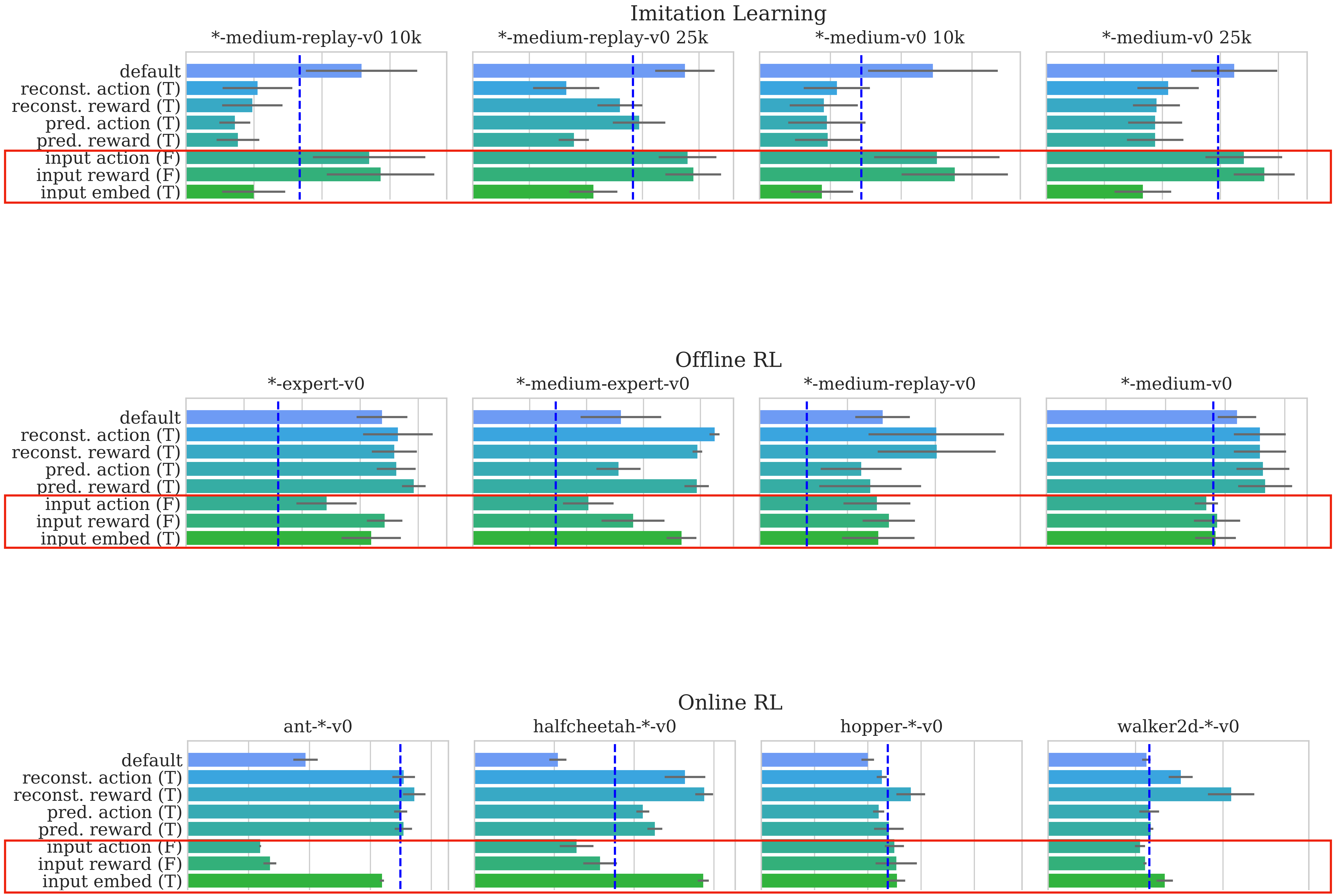
Depth Study on ACL



Depth Study on ACL

Factor	Description	Imitation	Offline	Online
reconstruct action	Add action prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
reconstruct reward	Add a reward prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
predict action	Add an action prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
predict reward	Add a reward prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$

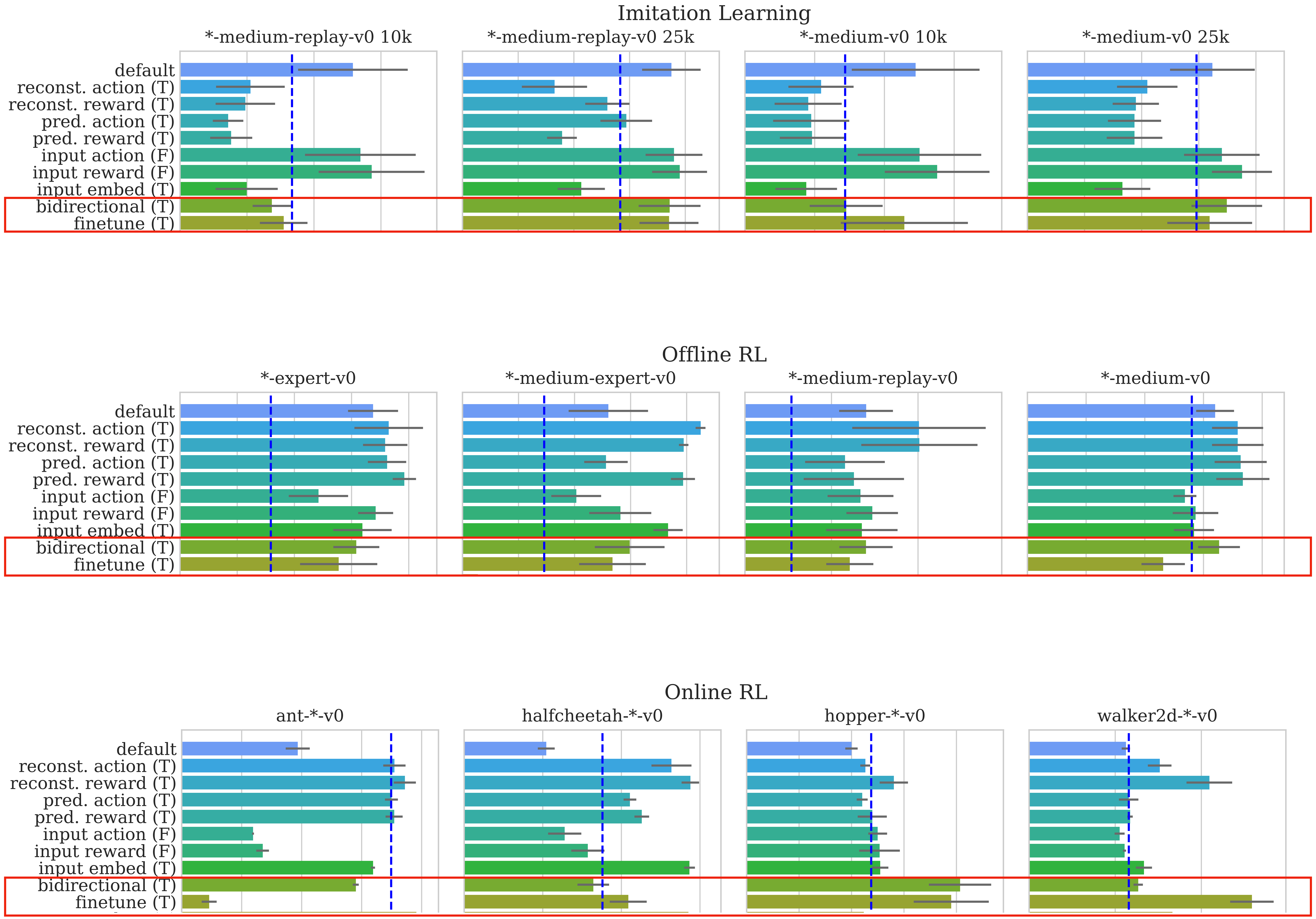
Depth Study on ACL



Depth Study on ACL

Factor	Description	Imitation	Offline	Online
reconstruct action	Add action prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
reconstruct reward	Add a reward prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
predict action	Add an action prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
predict reward	Add a reward prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
input action	Include actions in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input reward	Include rewards in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input embed	Use representations $\phi(s)$ as input to transformer, as opposed to raw observations.	$\downarrow$	$=$	$\uparrow$

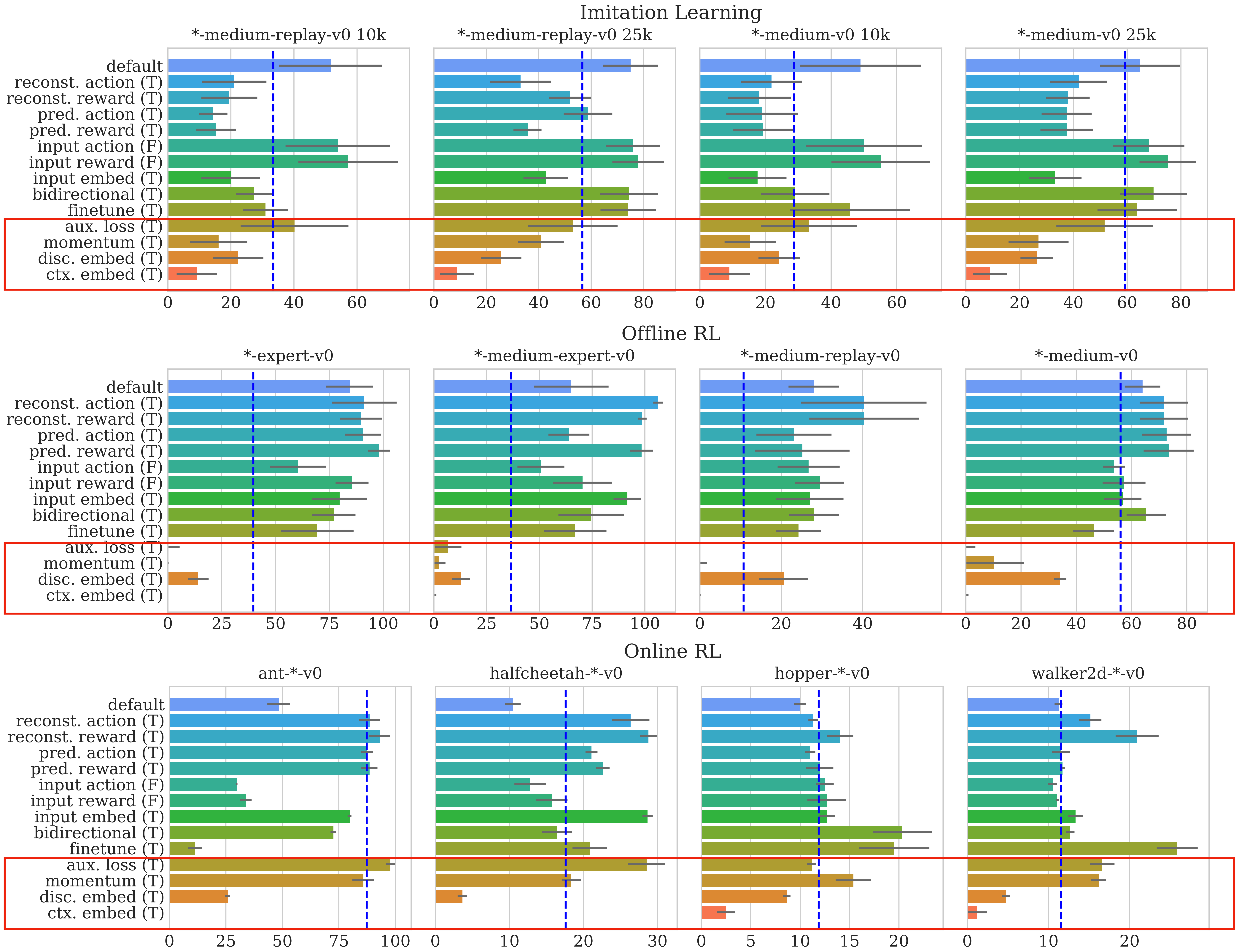
Depth Study on ACL



Depth Study on ACL

Factor	Description	Imitation	Offline	Online
reconstruct action	Add action prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
reconstruct reward	Add a reward prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
predict action	Add an action prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
predict reward	Add a reward prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
input action	Include actions in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input reward	Include rewards in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input embed	Use representations $\phi(s)$ as input to transformer, as opposed to raw observations.	$\downarrow$	$=$	$\uparrow$
bidirectional	To generate sequence output at position i, use full input sequence as opposed to only inputs at position $\leq i$.	$\downarrow$	$=$	$\uparrow$
finetune	Pass gradients into ϕ during learning on downstream tasks.	$\downarrow$	$\downarrow$	$\uparrow$

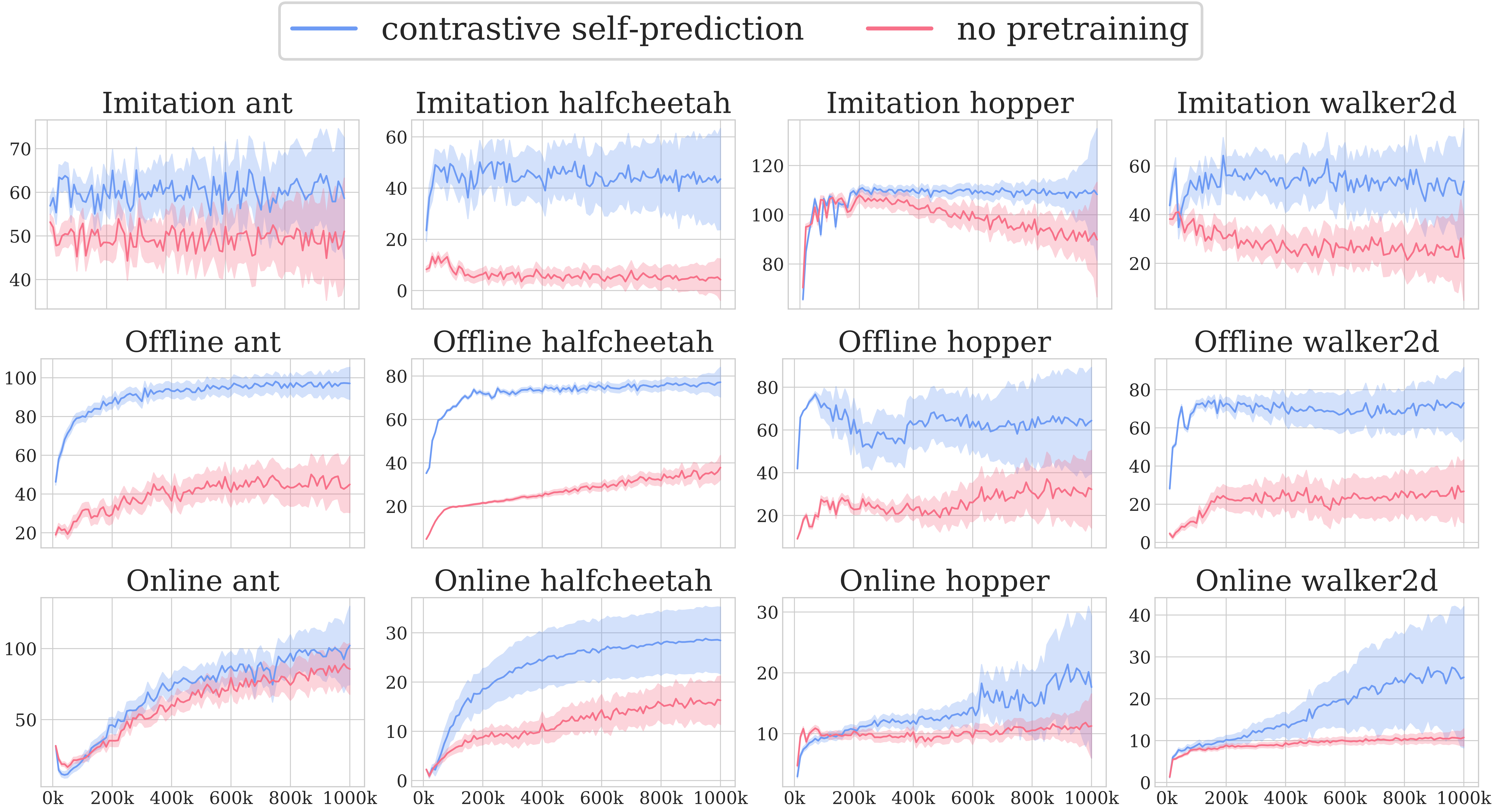
Depth Study on ACL



Depth Study on ACL

Factor	Description	Imitation	Offline	Online
reconstruct action	Add action prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
reconstruct reward	Add a reward prediction loss based on $\phi(s)$.	$\downarrow$	$\uparrow$	$\uparrow$
predict action	Add an action prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
predict reward	Add a reward prediction loss based on transformer outputs. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\uparrow$	$\uparrow$
input action	Include actions in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input reward	Include rewards in the input sequence to transformer.	$\downarrow$	$\uparrow$	$\uparrow$
input embed	Use representations $\phi(s)$ as input to transformer, as opposed to raw observations.	$\downarrow$	$=$	$\uparrow$
bidirectional	To generate sequence output at position i, use full input sequence as opposed to only inputs at position $\leq i$.	$\downarrow$	$=$	$\uparrow$
finetune	Pass gradients into ϕ during learning on downstream tasks.	$\downarrow$	$\downarrow$	$\uparrow$
auxiliary loss	Use representation learning objective as an auxiliary loss during downstream learning, as opposed to pretraining.	$\downarrow$	$\downarrow$	$\uparrow$
momentum	Adopt an additional momentum representation network. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\downarrow$	$\uparrow$
discrete embedding	Learn discrete representations. Following Hafner et al. (2020), we treat the 256-dim output of ϕ as logits to sample 16 categorical distributions of dimension 16 each and use straight-through gradients.	$\downarrow$	$\downarrow$	$\downarrow$
context embedding	Following Devlin et al. (2018), use transformer output as representations for downstream tasks. Whenever this is true, we also set 'input embed' to true.	$\downarrow$	$\downarrow$	$\downarrow$

Best ACL Configuration



Future Directions

- Combining state and action representation learning
- Other ways to apply transformer in offline RL
- Theoretical guarantees for representation learning in offline RL
- New representation learning objectives for offline RL

Future Directions

- Combining state and action representation learning
- Other ways to apply transformer in offline RL
- Theoretical guarantees for representation learning in offline RL
- New representation learning objectives for offline RL

Questions?