Teacher-Student Representational Alignment for Reinforcement Learning-driven Imitation Learning

Anonymous Authors
Paper under double-blind review

***Abstract*—Imitation learning (IL) from a state-based reinforcement learning (RL) policy is a common approach to overcome the curse of dimensionality in complex and high-dimensional observation spaces prevalent in robotics. This paper addresses the irreducible imitation gap that emerges when teacher and student are learned in isolation, and the teacher policy has the liberty to rely on privileged state information that the student cannot infer from its observations. Instead of improving poor student performance with RL finetuning after IL, which often requires a whole new training setup, we propose a novel algorithm which learns a shared embedding space that hides agent-specific observations and thus trains imitable teacher policies by construction. We train the shared embedding space with self-supervised contrastive learning in parallel to the teacher policy and prevent it from extracting private information by limiting its gradients from updating the encoder networks. We perform evaluations on several example domains and compare to state-of-the-art baseline showing that our algorithm enables higher student performance with substantially reduced imitation gap.**

***Index Terms*—imitation learning, imitation gap, contrastive learning, reinforcement learning**

I. Introduction

Learning behavior policies in high-dimensional observation spaces using reinforcement learning (RL) suffers from high sample complexity [1], while more efficient approaches such as imitation learning (IL) require access to costly expert demonstrations. Recent work shows the effectiveness of first training an RL-based teacher on low-dimensional analytical states and then using IL to distill the behavior to a student policy operating on high-dimensional observation space [2]. For this to be efficient, the teacher is often given access to highly informative and task-relevant state variables and dense reward signals. While providing more information to the teacher agent leads to higher teacher performance and faster convergence, it also enables the teacher to rely on information that is private to it and cannot be derived from student observations (See Fig. 1 for an illustration). This can lead to a non-reducible imitation gap [3] in student performance during distillation, which is often addressed by an additional RL finetuning phase on the student policy [4], [5].

As the additional RL finetuning step typically involves the original difficulties with RL training, it is more desirable to train the teacher policy in a way that the student can imitate it from the beginning. To this end, we propose a combined RL and IL learning approach that automatically hides the teacher-specific private information from the teacher's observations and enables the student to imitate it by construction. Unlike previous work that constrains the teacher's action space [6] or modify its already sensitive reward signal [7], [8], we pose a representation learning problem: Our goal is to learn a low-dimensional common representation space that only retains the shared part of the teacher and student observations while removing agent-specific state variables. The learned joint representation is used by the teacher during its online RL training on the task and by the student during the imitation phase. The embedding space and the policy are trained in alternating phases with separate objectives to prevent the teacher from extracting observations specific to its observation space. We apply additional alignment and stability losses to the embedding space to encourage representational similarity between aligned observations and enable stable policy learning.

The advantage of our approach is that the teacher learns a policy that can be directly imitated by the student without changing the teacher's reward structure, which can break task learning or lead to reward hacking by the agent. The change

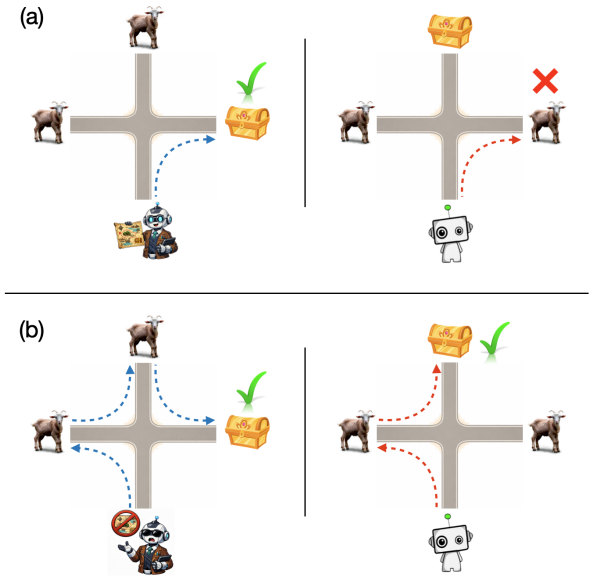


Fig. 1: Illustration of the imitation gap: (a) Training the teacher (left) and student (right) in isolation leads to non-imitable teacher behavior in presence of private information, such as the goal location. (b) Our approach automatically hides the private information from the teacher, making the learned teacher behavior imitable by the student.

Paper under double-blind review.

from teacher observations to joint representation requires the teacher to maximize the original reward function by exploiting only the information that is shared between teacher and student which generates behaviors that the student can imitate. Furthermore, our choice of positive and negative samples for the contrastive learning objective preserves small variations in the consecutive states and enables the agent to learn more refined policies. Our evaluation on a series of challenging environments shows that our approach can distill successful student policies. Comparisons to state-of-the-art baseline approaches show that our approach compares favorably in terms of student performance. An ablation study shows the benefits of the different loss terms.

Our contributions can be summarized as follows:

- We propose a novel, task-agnostic method to bridge the imitation gap between RL-based teacher and IL-based student policies.
- The proposed method is a plug-and-play adaptation to existing frameworks with no modification to the reward function and minimal hyperparameter tuning.
- We evaluate our approach in two environments designed to expose the imitation gap and show that it consistently outperforms strong baseline methods.

II. Related Work

Imitation learning has been widely adopted in robotics to overcome the curse of dimensionality and accelerate policy learning. Works such as [9] cut computational costs by mimicking demanding planning algorithms using neural networks while [10] imitates human experts to learn fine-grained manipulation tasks. Despite their success, the traditional IL algorithms rely on access to an interactive expert algorithm or the existence of collected expert demonstrations. A recent paradigm in IL called "learning by cheating" [2] overcomes this limitation by obtaining an interactive teacher policy by applying RL on low-dimensional analytical states, which is later combined with IL methods to learn high-dimensional policies.

An irreducible performance gap arises between the teacher and student agents when the teacher policy makes its decisions based on state variables that are unobservable to the student. [11] recovers the student's performance by additional RL fine-tuning on the distilled student policy, however bootstrapping RL policies without a coupled critic network is shown to be ineffective and possibly lead to inferior performance [12]. [4] instead regards the teacher signals as a guidance during the student's own RL training and combines it with the original task objective. [3] trains a value function for the distilled suboptimal policy and uses it to guide the exploration of a student policy trained from scratch.

Instead, our work aims at training the teacher policies in an imitable way from construction. [7] designs dense reward signals by exploiting the environment knowledge to motivate imitable teacher behaviors. However, such knowledge is generally not available and require heavy feature engineering to achieve intended policies. A work closely related to our setup is SITT [8], which trains the teacher policy in parallel with the student policy to avoid learning behaviors that the student cannot imitate. It requires no prior knowledge of the environment and avoids post-hoc fine-tuning of the student, but modifies the original reward function and introduces a trade-off coefficient that is difficult to tune and requires careful balancing. Our method avoids feature engineering and modifications to the reward function by constructing a common representation space with contrastive learning for both policies to operate on.

Similar ideas in contrastive learning has been previously used in imitation learning settings to enable policy transfer under various distributional shifts. [13] uses contrastive learning to construct an embedding space that is invariant to different augmentations of the input space caused by changing background and lighting conditions. An adversarial IL policy trained on this latent space shows substantially reduced imitation gap compared to direct baselines. Similarly, [14] learns an embedding space under which the real and simulated versions of an environment scene look similar, which is later used to transfer skills learned in one domain to another. Their method relies on the domain knowledge to select the contrastive pairs and uses a heuristic-based similarity measure to construct the embedding space. Our algorithm however is fully task-agnostic and requires no domain knowledge, which enables it to be easily integrated into existing IL frameworks.

III. Background

We formalize the RL task a Markov decision process (MDP) characterized by the tuple $(\mathcal{O}_T, \mathcal{A}, P, R, \gamma)$. $\mathcal{O}_T$ denotes the observation spaces of the teacher. $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition probability to state s' given action a is taken in state s, $R(s, a)$ is the reward function, and finally γ is the discount factor. The MDP for the student is identical but uses the observation space $\mathcal{O}_S$ instead. Teacher and student policies, π_T and π_S map their respective observations to a probability distribution over the action space: $a_i \sim \pi_i(\cdot \mid o_i),\ i \in \{T, S\}$. In our setup, π_T is trained with reinforcement learning to guide π_S using imitation learning.

A. Reinforcement learning

Reinforcement learning is a framework to learn optimal policies through interaction with an environment. We train the teacher policy π_T with RL to maximize the expected discounted return:

$$J(\pi_T) = \mathbb{E}_{\tau \sim \pi_T,\ a_t \sim \pi_T} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right], \tag{1}$$

where $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, and $\tau = (o_0, a_0, r_0, o_1, a_1, r_1 \dots)$ denotes a trajectory generated by the teacher policy interacting with the environment. We use Proximal Policy Optimization (PPO) [15] algorithm throughout our experiments.

B. Contrastive learning

Contrastive learning (CL) is a self-supervised representation learning method to learn rich data representations under absence of direct supervision signals. CL achieves this

by bringing similar samples closer in the embedding space while pushing dissimilar ones apart. Given an anchor sample and a corresponding positive example, along with a set of negative samples, the objective is typically formulated using a similarity-based loss such as the triplet or InfoNCE [16] losses. Let z_i and z_j denote latent representations of a positive pair, and $\mathcal{N}_i$ the set of negatives for i. The InfoNCE objective is:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{z_k \in z_j \cup \mathcal{N}_i} \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ denotes a measure of similarity between the embeddings (e.g., cosine similarity), and τ is a temperature parameter.

IV. Method

Our training setup consists of two separate modules with differing objectives. The first module uses contrastive learning to learn a mapping from each observation space to a common latent space that only contains information that is available to both agents while hiding out the agent-specific private information. The second module is trained with reinforcement learning on the learned common embedding space to maximize the task objective.

We iteratively train each module in two phases: (1) In phase one, we learn the embedding space with contrastive learning trained on aligned dataset of observations from both agents, collected by rolling out the teacher policy. (2) In the second phase, the policy network is trained with RL on the latent embeddings collected by the teacher to maximize the task objective. To prevent the policy gradients from extracting private information from the teacher observations, the policy gradients from the second phase are only allowed to update the policy network and are not passed through the embedding networks.

A. Problem setup

We adopt a multi-view approach to the imitation learning problem, where the teacher and student agents observe the environment through different modalities. We assume that their paired observations, denoted as $o_T^t \in \mathcal{O}_T$ and $o_S^t \in \mathcal{O}_S$, are two distinct views of the same underlying state $s^t \in \mathcal{S}$:

$$o_T^t = f_T(s^t), \quad o_S^t = f_S(s^t) \quad (3)$$

where f_T and f_S are observation functions that map the environment state to respective observations, which are readily available in simulation environments. Since each view captures only a subset of the state's information with a certain overlap, we model each observation as a combination of their common (c^t) and private components ($p_T^t,\ p_S^t$):

$$o_T^t = [c^t,\ p_T^t], \quad o_S^t = [c^t,\ p_S^t]. \quad (4)$$

The common component c^t contains information that is available to both modalities, e.g. the relative obstacle locations in a collision avoidance task, while the private components p_T^t and p_S^t contain modality-specific details, which are typically not retrievable from the other modality (see Fig. 2 for overview). Following the widely adopted literature in multi-view representation learning [17], [18], we assume that the private components are conditionally independent given the shared variable c^t:

$$p_T^t \perp p_S^t \mid c^t \quad (5)$$

Under these conditions, we aim to retrieve c^t from the original teacher and student observations using contrastive learning.

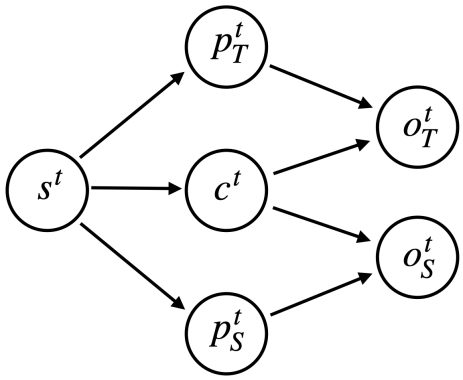


Fig. 2: We adopt a multi-view approach to the teacher and student setup where the observations from each agent (o_T^t and o_S^t) are different views of the same state with their own private (p_T^t and p_S^t) and common (c^t) components.

B. Learning common latent space

A teacher observation o_T^t at timestep t contains the same common content c^t as the student observation at the same timestep o_S^t, which in general is different from the observations collected at any other timestep $t' : t' \neq t$. Based on this knowledge, the latent embeddings from two paired observations should reside close to each other while the embeddings for the non-paired observations should be dissimilar. We construct such a latent space using the contrastive self-supervised InfoNCE loss [16], which has been shown to recover the underlying shared signals up to an invertible mapping [19] under our assumptions (4) and (5):

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_T^i, z_S^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_T^i, z_S^j)/\tau)},$$

where the unit-norm embeddings z_S^i and z_T^i are processed by separate encoder networks and their similarity is measured by a dot product: $\text{sim}(z_T^i, z_S^j) = (z_T^i)^\top z_S^j$. τ is a learnable parameter that controls the sharpness of the contrastive distribution. Note that unlike time-contrastive representation learning methods that consider all the observations from a fixed time window as positives, [14], our objective considers only the aligned student observation as a positive while any other observation in the training batch is pushed away. This distinction is critical for our method as it enables the policy network to distinguish subtle variations among consecutive states and generate distinct actions accordingly.

C. Aligning learned embeddings

The same policy network in our setup operates on the latent embeddings of the teacher observations during training and the student observations during testing. Since RL policies are known to be sensitive to small changes in the embedding space [20], we require the embeddings coming from a paired set of observations to be similar to each other. While this similarity ($\text{sim}(z_T^i, z_S^i)$) emerges implicitly at the optimum of $\mathcal{L}_{\text{contrastive}}$, we introduce an explicit alignment loss to directly maximize it and guide the training of the embedding space:

$$\mathcal{L}_{\text{alignment}} = -\frac{1}{N}\sum_{i=1}^{N} \text{sim}(z_T^i, z_S^i), \tag{6}$$

where the similarity is measured by a dot product between the embeddings.

D. Stabilizing policy learning

Since the policy gradients from the policy network are not passed through the encoders that are trained in parallel, it is essential for the encoders to learn stable embeddings throughout their training to facilitate stable training of the policy network. Since the CL objective alone is invariant under invertible transformations of the entire embedding space [19], we add a stability loss to the objective to to avoid large deviations of the latent embeddings during training:

$$\mathcal{L}_{\text{stability}} = -\frac{1}{N}\sum_{i=1}^{N} \left(\text{sim}(l_{T,\text{old}}^i, l_T^i) + \text{sim}(l_{S,\text{old}}^i, l_S^i)\right) \tag{7}$$

where l_T^i and l_S^i are the raw logits of the policy network while $l_{T,\text{old}}$ and $l_{S,\text{old}}$ are the same before the training phase started. Although this term is computed by processing the observations through both encoder and policy networks, it serves to stabilize the embedding space and thus its gradients are only used to update the encoder networks. This specific formulation encourages updates in the encoder network that will still produce similar actions under the policy network while giving enough flexibility to the embedding space to reorient itself for newly encountered observation samples.

With these terms combined, the encoder networks are trained using gradient descent to minimize the following objective:

$$\mathcal{L}_{\text{embedding}} = \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{alignment}} + \mathcal{L}_{\text{stability}} \tag{8}$$

This term is optimized using aligned trajectory dataset collected by rolling out the teacher policy, which is trained in parallel with RL to maximize the task objective.

V. Experiments

A. Environments

We conduct experiments in two different simulation environments that highlight the imitation gap between policies. In both of the environments, the optimal policy for the teacher agent trained in isolation cannot be imitated by the student due to its relatively limited information about the environment.

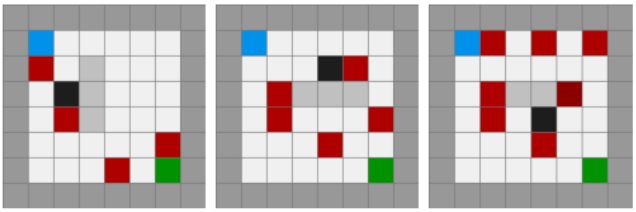


Fig. 3: TunnelVision environment with the agent (black), start (cyan) and goal (green) coordinates with randomly generated obstacle configurations (red). Teacher can observe all 8 surrounding cells while the student sees only 3 adjacent cells in front of it (highlighted in gray), direction of which is controlled by the agent's orientation.

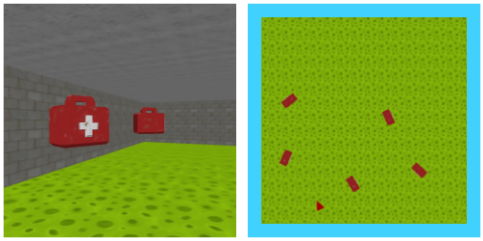
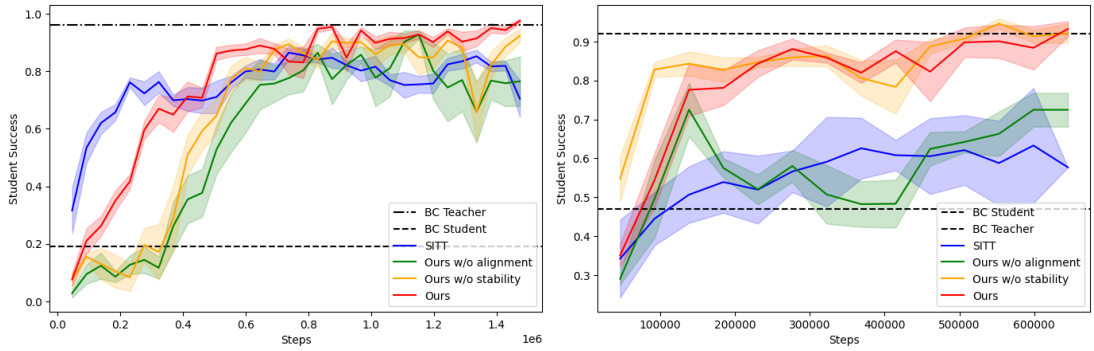


Fig. 4: CollectHealth environment from (a) the student's perspective and (b) in bird-eye view. Teacher can observe the exact box locations but needs to reorient itself towards the next target to enable the student to imitate it.

CollectHealth [21] The agent needs to collect five health boxes that are randomly spawned in a 3D 15m $\times$ 15m room (see Fig. 4). The agent can move in eight discrete directions or change the camera's orientation by $\pm 15^\circ$ ($|\mathcal{A}| = 10$). The agent is rewarded for each box it collects and punished by a small amount for every intermediary step. The episode is terminated if all the boxes are collected or the agent hits one of the walls. The teacher agent has access to the exact location of the boxes, which allows it to solve the task without changing its orientation. The student agent observes the environment through limited-FOV FPV images and needs to constantly reorient itself to attend to the next target.

TunnelVision This is a grid world environment where the teacher needs to deliberately change its behavior at each timestep to enable the student to imitate it. The agent starting from a fixed position needs to navigate to a fixed goal position while avoiding obstacles that are randomly generated at every episode (see Fig. 3). The agent can independently move or change its orientations in four directions ($|\mathcal{A}| = 8$). The agent is rewarded for finishing the task and punished for collisions, upon which the episode is terminated. A small step penalty is added to motivate exploration. The teacher agent observes all eight surrounding cells while the student can only observe three cells in the direction of its orientation. The time-optimal solution for the teacher is to directly move to the goal without changing its orientation, while the student needs to reorient itself frequently to avoid moving into unobservable obstacles.



(a) TunnelVision (b) CollectHealth

Fig. 5: Success rate curves for the distilled student policies in each environment. The teacher policy of the behavior cloning method (upper dashed line) is trained in isolation with private information and serves as an upper limit on all the experiments.

B. Baselines

We compare our method against two competitive imitation learning baselines and do ablation studies to evaluate the effectiveness of our alignment and stability losses:

(1) Behavior Cloning (BC) We train the teacher policy separately to maximize its own objective, then use supervised action discrepancy loss on the trajectory dataset collected by the teacher to train the student policy:

$$\mathcal{L}_{\text{BC}} = -\frac{1}{N} \sum_{(z_T, z_S) \sim \mathcal{D}_{\mathcal{T}}}^{N} \|z_T - z_S\|^2$$

(2) SITT [8] We jointly train teacher and student and penalize the teacher for visiting states where the student cannot imitate its actions. The exact penalty is computed by a KL divergence between the action distributions and is applied both on the reward signal and the policy gradients:

$$\begin{aligned} \tilde{J}(\pi_T) = \mathbb{E}_{s \sim d^{\pi_T},\, a \sim \pi_T(\cdot|s)} \big[r(s,a) \big] \\ - \alpha \mathbb{E}_{s \sim d^{\pi_T}} \big[D_{\text{KL}} \big(\pi_T(\cdot|s) \,\|\, \pi_S(\cdot|s) \big) \big], \end{aligned} \tag{9}$$

where α is a hyperparameter controlling the tradeoff between task performance and imitation loss. We experiment with various α values and choose the one that yields the highest performance. We use the original implementation of the algorithm.

	TunnelVision			CollectHealth		
Method	T	S	Δ	T	S	Δ
BC	0.96	0.19	0.77	0.92	0.47	0.45
SITT	0.91	0.72	0.19	0.88	0.68	0.20
Ours w/o alignment	0.84	0.81	0.03	0.81	0.70	0.11
Ours w/o stability	0.94	0.93	0.01	0.95	**0.92**	0.03
Ours	0.98	**0.97**	**0.01**	0.90	0.88	**0.02**

TABLE I: Success rates of the teacher and student policies in different environments and the imitation gap between them.

For a fair comparison, all methods are trained using identical network architectures and the same number of environment interactions. Each experiment is repeated with at least five random seeds, and we report the average performance. The results are summarized in Table I. The success rate of the student policies over the training period is shown in Fig. 5.

As expected, the teacher policy trained in isolation exploits private information available to it and hence provides weak imitation signals to the BC-based student policy, which results in high imitation gap. SITT improves upon BC by explicitly penalizing the teacher for visiting states where the student cannot match its behavior, which in turn reduces the imitation gap. However, this comes at the cost of degraded teacher performance, as the modified reward introduces a trade-off between task optimality and imitability.

In contrast, our method enforces imitability at the representation level by restricting both policies to operate on a shared latent space that removes teacher-specific private information. As a result, the teacher is naturally constrained to learn behaviors that are reproducible by the student, which leads to consistently high student performance and a substantially reduced imitation gap without sacrificing the original task objective.

We also observe that removing the alignment loss leads to a noticeable degradation in student performance, which indicates that explicitly enforcing similarity between paired embeddings is important for effective imitation. The stability loss, on the other hand, improves the performance in the TunnelVision environment while slightly degrading it in the CollectHealth environment. This suggests that the stability objective can help regularize training in settings with higher sensitivity to representation shifts (a single mistake in TunnelVision can easily fail the episode), but may introduce conservative constraints that are not always beneficial across all environments. Overall, the full model achieves a strong balance and maintains high teacher performance while keeping the imitation gap minimal.

VI. Future Work and Conclusions

In this work, we proposed a task-agnostic approach to reduce the imitation gap in RL-driven imitation learning by learning a shared latent representation between teacher and student observation spaces. Our method combines contrastive

learning with reinforcement learning to ensure that the teacher policy relies only on the information that is also accessible to the student, which in turn enables the student to imitate it. We also introduce stability and alignment objectives to the embedding space to facilitate steady policy training and tighter imitation gap. Unlike previous approaches, our method is task-agnostic and does not require modifications to the reward function.

Experimental results demonstrate that our approach consistently outperforms baselines and achieves high student performance with minimal imitation gap. Our ablation studies further highlight the effectiveness of our alignment and stability objectives in improving the performance of the distilled student policy. Future work includes extending the method to environments with more complex and high-dimensional continuous control tasks.

REFERENCES

[1] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2022.

[2] D. Chen, B. Zhou, V. Koltun, and P. Kr, "Learning by cheating," in *Conference on Robot Learning (CoRL)*, 2019.

[3] A. Walsman, M. Zhang, S. Choudhury, D. Fox, and A. Farhadi, "Impossibly good experts and how to follow them," in *The Eleventh International Conference on Learning Representations*, 2023.

[4] L. Weihs, U. Jain, I.-J. Liu, J. Salvador, S. Lazebnik, A. Kembhavi, and A. Schwing, "Bridging the imitation gap by adaptive insubordination," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[5] S. Schmitt, J. J. Hudson, A. Žídek, S. Osindero, C. Doersch, W. M. Czarnecki, J. Z. Leibo, H. Küttler, A. Zisserman, K. Simonyan, and S. M. A. Eslami, "Kickstarting deep reinforcement learning," *ArXiv*, vol. abs/1803.03835, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3883991

[6] G. Monaci, M. Aractingi, and T. Silander, "Dipcan: Distilling privileged information for crowd-aware navigation," in *Robotics: Science and Systems XVIII*. Robotics: Science and Systems Foundation, Jun. 2022.

[7] Y. Song, K. Shi, R. Penicka, and D. Scaramuzza, "Learning perception-aware agile flight in cluttered environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, United Kingdom: IEEE, May 2023, pp. 1989–1995.

[8] N. Messikommer, J. Xing, E. Aljalbout, and D. Scaramuzza, "Student-informed teacher training," in *The Thirteenth International Conference on Learning Representations (ICLR)*, Feb. 2025.

[9] J. Tordesillas and J. P. How, "Deep-PANTHER: Learning-based perception-aware trajectory planner in dynamic environments," *IEEE Robotics and Automation Letters*, 2023.

[10] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 1910–1924. [Online]. Available: https://proceedings.mlr.press/v270/zhao25b.html

[11] J. Xing, A. Romero, L. Bauersfeld, and D. Scaramuzza, "Bootstrapping reinforcement learning with imitation for vision-based agile flight," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=bt0PX0e4rE

[12] M. Nakamoto, Y. Zhai, A. Singh, M. S. Mark, Y. Ma, C. Finn, A. Kumar, and S. Levine, "Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=GcEIvidYSw

[13] V. Giammarino, J. Queeney, and I. C. Paschalidis, "Visually robust adversarial imitation learning from videos with contrastive learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, 2025, pp. 15 642–15 648.

[14] J. Xing, L. Bauersfeld, Y. Song, C. Xing, and D. Scaramuzza, "Contrastive learning for enhancing robust scene transfer in vision-based agile flight," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Vienna, Austria, 2020.

[17] D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello, "Multi-view causal representation learning with partial observability," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[18] Q. Lyu, X. Fu, W. Wang, and S. Lu, "Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[19] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, "Quantifying generalization in reinforcement learning," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1282–1289. [Online]. Available: https://proceedings.mlr.press/v97/cobbe19a.html

[21] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2407.17032*, 2024.