SUPERVISION INTERFACES FOR PRIVILEGED COORDINATION ADVICE IN COOPERATIVE MARL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative multi-agent reinforcement learning commonly adopts the Centralized Training with Decentralized Execution (CTDE) paradigm, in which global information is used during training while execution relies only on local observations. However, global information typically influences the actor indirectly through a centralized critic, leaving unclear how privileged coordination advice should be delivered to decentralized actors. We propose that privileged advice can supervise behavior directly through imitation or policy regularization, or supervise representation while reinforcement learning remains responsible for action selection, allowing the actor to absorb privileged coordination structure without directly imitating the advised behavior. We introduce **Factorized Privileged Action Supervision (FPAS)**, which maps privileged coordination decisions to per-agent targets and trains each decentralized actor to predict them through a train-only auxiliary head. FPAS requires no learned teacher and adds no execution-time overhead. On SMACv2, FPAS outperforms MAPPO across all three races ($+8.2/+12.8/+5.3$ pp; paired $p < 0.03$ in each race, with 15/15 paired seed wins), with two races held out from tuning. Matched control experiments show that the gains are associated with the prediction pathway and decision-relevant target content. A loss-matched intervention further sharpens this conclusion: the same advice and cross-entropy degrade learning when applied through policy logits, but recover the FPAS gain when applied through the auxiliary prediction head. Target interventions further show that consistency, local inferability, and decision content are important properties of effective representation supervision. Across MPE, SMACv2, and Overcooked, imitation is strongest when the advice defines a strong and locally attainable policy, whereas prediction is more reliable when the advice provides only partial coordination structure or when direct execution is constrained. These results show that the supervision route through which privileged information enters the actor is itself an important design variable in CTDE.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) commonly uses Centralized Training with Decentralized Execution (CTDE): agents may use global information during training but must act from local observations at execution (Kraemer & Banerjee, 2016; Lowe et al., 2017; Yu et al., 2022; Gronauer & Diepold, 2022). This separation leaves an important design question: *how should privileged information reach the decentralized actor during learning?*

In standard actor–critic CTDE, a common route is indirect. A centralized critic uses global state for value estimation and shapes actor updates through the reinforcement-learning objective (Lyu et al., 2021; 2023). Other approaches transfer centralized values, policies, or guidance more directly into decentralized control (Zhao et al., 2024; Li et al., 2026; Pavel et al., 2026). Behavioral transfer is particularly natural when the privileged signal already specifies a useful action or policy. Coordination advice, however, is often less complete. A StarCraft focus-fire heuristic can identify which enemy should receive attention while leaving movement direction and positioning unspecified. A task-assignment solver can identify destinations while omitting collision avoidance. Such signals encode useful coordination structure while specifying only part of the decentralized policy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

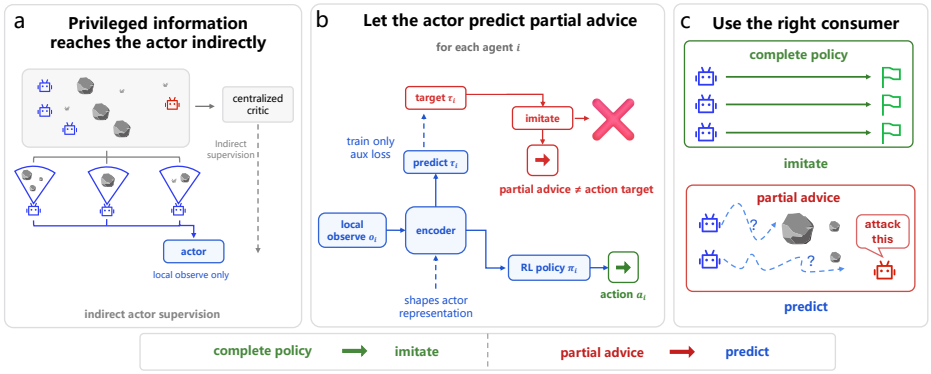

Figure 1: **Privileged coordination advice can supervise behavior or representation.** (a) Standard CTDE routes privileged state through a centralized critic. (b) FPAS uses a train-only auxiliary head to predict per-agent coordination targets from local histories while the RL policy remains responsible for action selection. (c) The relative effectiveness of behavioral and representation supervision depends on how directly the advice supports executable decentralized behavior.

This distinction motivates two ways to consume the same privileged signal. *Behavioral supervision* makes the actor follow the advice through imitation or policy regularization. *Representation supervision* instead asks the actor to predict the advice during training while leaving action selection to the reinforcement-learning objective. This makes representation supervision a natural candidate when privileged advice captures useful coordination structure without specifying a complete decentralized policy. Figure 1 summarizes this distinction.

We instantiate representation supervision with **Factorized Privileged Action Supervision (FPAS)**. A privileged rule, solver, planner, or learned evaluator produces per-agent coordination targets from global state. Each decentralized actor predicts its target from local history through an auxiliary head that shares the policy encoder. The auxiliary output is unused for action selection, so execution remains decentralized and the reinforcement-learning objective remains responsible for behavior. We therefore test whether the point at which supervision enters learning matters: coordination advice may shape the actor's representation without being imposed as the behavior that the actor must execute.

We test this hypothesis in four stages. First, we evaluate FPAS on SMACv2 and MPE. Second, matched pathway controls, target-content controls, and frozen-encoder probes isolate the mechanism. Third, interventions on target fidelity, consistency, local inferability, sharpness, and decision content identify properties associated with useful representation supervision. Finally, MPE, SMACv2, and Overcooked compare behavioral supervision, behavioral regularization, and prediction under matched advice. These comparisons support an empirical supervision regime: imitation is efficient when advice defines a strong policy that decentralized agents can attain, whereas prediction remains useful when advice captures coordination structure but leaves important behavior unspecified or difficult to reproduce locally.

Contributions. (i) We formulate privileged coordination advice as a supervision-interface problem, distinguishing behavioral and representation-level consumption within CTDE. (ii) We introduce FPAS, a one-stage auxiliary prediction interface that requires no learned teacher and adds no execution-time overhead, and show consistent gains across all three SMACv2 races. (iii) We provide matched mechanistic evidence linking the gains to the representation pathway and identify target properties that govern the usefulness of representation supervision. (iv) We map an empirical regime across MPE, SMACv2, and Overcooked that relates advice quality and local attainability to the relative effectiveness of behavioral and representation supervision.

2 Related Work

2.1 CTDE and Privileged Information

Value decomposition methods factor centralized values into decentralized utilities (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Rashid et al., 2020; Wang et al., 2021a), while policy-gradient methods pair decentralized actors with centralized critics (Foerster et al., 2018; Lowe et al., 2017; Yu et al., 2022; Kuba et al., 2022). Exploration and role-decomposition methods help agents discover coordination structure (Mahajan et al., 2019; Wang et al., 2021b). Centralized critics provide a training-time route for privileged state, while learned communication changes the information available during execution (Lyu et al., 2021; 2022; 2023; Foerster et al., 2016; Sukhbaatar et al., 2016; Das et al., 2019). FPAS uses privileged coordination outputs as train-only prediction targets for decentralized actors while leaving execution-time information unchanged.

2.2 Behavioral Transfer and Privileged Guidance

Policy distillation transfers behavior from teacher to student (Hinton et al., 2015; Rusu et al., 2016; Czarnecki et al., 2019), with DAgger-style corrections for the student's state distribution (Ross et al., 2011); guided policy search, kickstarting, and action advising likewise inject expert behavior into learning (Levine & Koltun, 2013; Schmitt et al., 2018; Torrey & Taylor, 2013). In MARL, CTDS transfers centralized action values (Zhao et al., 2024), MAGPO trains an autoregressive guider with KL-aligned decentralized policies (Li et al., 2026), and KD-MARL distills policy, advantage, and coordination information (Pavel et al., 2026). Work on imperfect demonstrations and privileged experts shows that transfer depends on teacher quality and learner observability (Hester et al., 2018; Gao et al., 2018; Weihs et al., 2021; Walsman et al., 2023). Song et al. (2025) characterize privileged expert distillation versus end-to-end RL in single-agent POMDPs. Here the comparison holds privileged coordination advice fixed and varies whether it supervises decentralized behavior or the actor representation.

2.3 Auxiliary Representation Learning

Learning with privileged information predates MARL (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). In reinforcement learning, privileged information can enter through asymmetric critics (Pinto et al., 2018), privileged teachers (Chen et al., 2020; Lee et al., 2020), reward shaping (Ng et al., 1999; Devlin & Kudenko, 2011), or auxiliary objectives (Jaderberg et al., 2017). Scaffolder combines privileged models with auxiliary representation objectives (Hu et al., 2024); MARL methods also use message aggregation and prediction of unobservable state, dynamics, or reward quantities to shape latent state (Guan et al., 2022; Huang et al., 2026). Dec-ADQN regresses decentralized action values toward a privileged centralized evaluator (Bhati et al., 2026). FPAS supervises the actor encoder with the output of a coordination computation: a per-agent decision signal whose behavioral scope need not specify the complete decentralized policy. This formulation enables a matched comparison between behavior-level and representation-level consumption of the same privileged advice.

3 Method

3.1 Problem Setting and Coordination Targets

In a cooperative Dec-POMDP (Oliehoek & Amato, 2016) with global state s_t, local histories h_i^t, and decentralized policies $\pi_i(a \mid h_i^t)$, a target generator g maps privileged state to a per-agent coordination output $z_i^t = g_i(s_t) \in \mathcal{Z}_i$. The generator may be a rule, solver, planner, or learned evaluator, and its output may be an action-preference vector, coordinate, slot-indexed feature vector, or categorical subtask. When the target admits a behavioral interpretation, the greedy follower π_g selects an available maximizer of the target, with seeded uniform tie breaking. We use the follower's closed-loop return to characterize the behavioral value available from directly following the advice: high-return advice is policy-like, while low-return advice can still encode task-relevant coordination structure.

3.2 Factorized Privileged Action Supervision

Each actor contains a recurrent encoder shared by its policy and a small target head f_D. Over valid agent-time pairs $\mathcal{V}$, FPAS optimizes

$$\mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{PPO}} + \lambda \, \mathbb{E}_{(t,i)\in\mathcal{V}} \left[\ell\big(f_D(h_i^t), z_i^t\big) \right]. \tag{1}$$

For the action-space instantiation, $\mathcal{Z}_i = [0,1]^{|\mathcal{A}_i|}$, $\tau_i^t = g_i(s_t)$, and ℓ is mean squared error over action slots (cross-entropy for categorical targets). Dead or padded pairs are excluded by $\mathcal{V}$; availability masks are not applied to the auxiliary regression targets. The auxiliary head is train-only and never selects actions; the centralized critic receives no target block. A modulation variant adds $\beta f_D(h_i)$ to policy logits (Silver et al., 2018; Johannink et al., 2019; Teh et al., 2017) and is used as a pathway control in Section 4.2.

3.3 Task Instantiations and Evaluation Protocol

In SMACv2, g marks the attack action of every enemy whose remaining-health fraction lies within 0.02 of the current minimum among living enemies, and marks the four move actions whenever the agent's own health fraction falls below 0.3. In MPE simple_spread, g solves the global agent–landmark assignment and emits the direction toward the assigned landmark. Appendix B specifies every target generator in closed form. We also evaluate planner-derived MPE targets and critic-derived SMACv2 targets to test alternatives to hand-specified coordination semantics.

Per-seed scores use the mean over the final 10% of evaluation points (last 200 of 2,000 evaluations on MPE; last 50 of 500 on SMACv2), fixed before held-out experiments were scored. Matched comparisons share the same environment and training configuration except for the supervision intervention. Appendix A gives training hyperparameters and statistics, and Appendix F documents configuration-parity and target-leakage audits.

4 Results

4.1 Cross-Race Efficacy

Table 1: **Cross-race SMACv2 efficacy.** 5v5 win rate (5 seeds, 10M steps). FPAS uses hard targets with auxiliary supervision only. Paired t-tests use matched seeds and all three races survive Holm correction. Terran is the development race; Protoss and Zerg are held out with settings frozen.

Race	MAPPO	FPAS	Δ	p (paired)	d
Terran (dev)	0.529 ± 0.027	$\mathbf{0.611} \pm 0.023$	$+0.082$	0.011	2.02
Protoss (held-out)	0.427 ± 0.056	$\mathbf{0.555} \pm 0.019$	$+0.128$	0.015	1.83
Zerg (held-out)	0.513 ± 0.025	$\mathbf{0.566} \pm 0.022$	$+0.053$	0.029	1.49

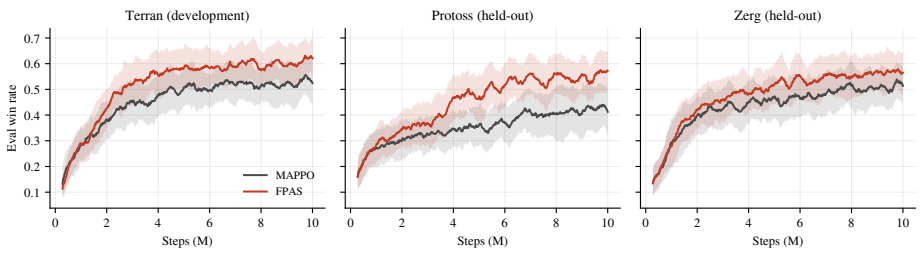

Figure 2: **Cross-race efficacy on SMACv2.** Evaluation win rate over training (mean $\pm$ s.d. over 5 seeds; curves are smoothed for visualization, while reported scores and tests use the unsmoothed evaluation points). The FPAS configuration is selected on Terran and transferred unchanged to held-out Protoss and Zerg.

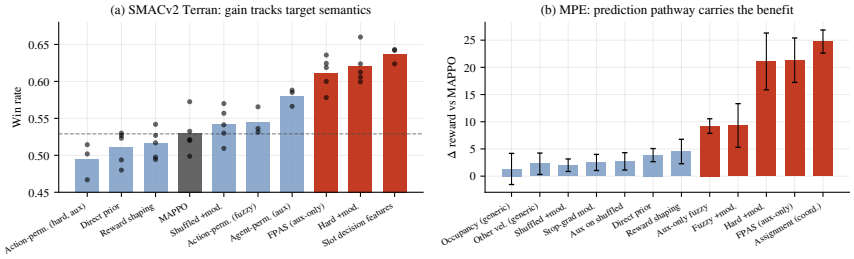

Figure 3: **Mechanism controls under matched configurations.** (a) SMACv2 Terran win rate (dots: seeds; dashed line: MAPPO). (b) MPE gain over MAPPO (error bars: s.d.). Red conditions preserve the coordination-decision target; blue controls preserve the training recipe while changing the target content or pathway. Full tables appear in Appendix C.

Table 1 and Figure 2 show a consistent efficacy result. FPAS improves all three SMACv2 races and all 15 matched seeds. A hierarchical paired bootstrap that resamples races and then matched seeds estimates an overall improvement of $+0.088$ (95% CI $[+0.051, +0.133]$, $P(\Delta \leq 0) < 5\times10^{-5}$). The same recipe also improves MPE simple_spread, from MAPPO's -167.9 ± 2.3 to -146.6 ± 4.1 over 8 seeds ($+21.3$). Terran loss-weight sweeps place $\lambda \in \{0.1, 0.5, 1.0\}$ between 0.55 and 0.62 win rate; further robustness results appear in Appendix E. Having established the effect across races and a second benchmark, we next examine which supervision pathway and target content account for the gain.

4.2 Representation-Shaping Mechanism

The mechanism analysis separates three linked questions: which gradient pathway carries the gain, which information the target must preserve, and what representation emerges from that supervision (Figure 3).

Prediction pathway. Auxiliary prediction alone reproduces the full effect in both domains (MPE hard: $+21.3$ aux-only vs. $+21.1$ full; SMACv2: 0.611 vs. 0.620). Logit modulation with encoder gradients stopped retains $+2.5$ on MPE, comparable to the shuffled control. Direct logit priors and reward shaping recover only a small fraction on MPE and sit at or below MAPPO on SMACv2. Together, these matched interventions link the gain to the auxiliary prediction pathway through the shared encoder.

Decision-content controls. Resampling an action-dimension permutation at every step preserves each target row's entropy, norm, value distribution, and sharpness while making its coordinate semantics inconsistent. On SMACv2, performance falls to 0.544 for fuzzy targets and 0.494 for hard aux-only targets, with the latter below MAPPO (0.529). Agent-level permutation retains 0.580, consistent with focus-fire targets being largely shared across agents. A fixed invertible recoding remains learnable and preserves performance in Section 4.3. The performance ordering identifies stable decision structure as the informative target property; row-level statistics and fixed action-coordinate alignment alone provide much smaller gains.

Representation probes. Frozen-encoder probes characterize the resulting latent structure on identical Terran observation streams. For the 6-way focus-enemy variable, balanced accuracy is 0.66 ± 0.05 from the FPAS encoder and 0.30 ± 0.04 from MAPPO (paired t=11.6, p=3.2×10^{-4}); the raw 80-dimensional observation gives 0.60 ± 0.05. A multilabel focus-set probe gives 0.81 ± 0.03 versus 0.64 ± 0.02 (p=3.0×10^{-4}). For own low-health status, FPAS gives 1.00 ± 0.00 and MAPPO 0.50 ± 0.00, while the raw observation supports 0.98 ± 0.02. The auxiliary objective keeps decision-relevant variables linearly accessible in the policy representation, which is the quantity measured by these probes.

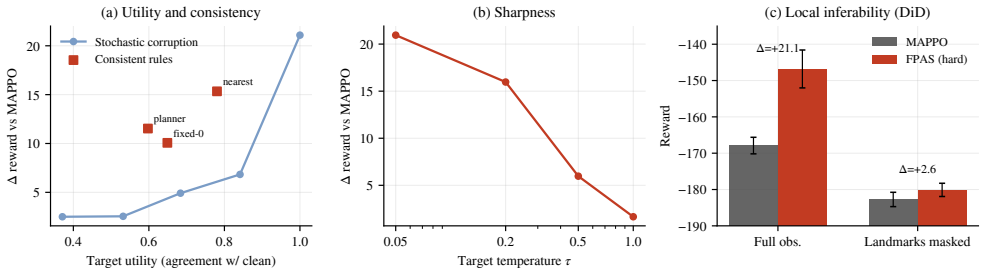

Figure 4: **Target-property interventions on MPE.** (a) Performance gain as target fidelity changes under stochastic corruption and deterministic alternative rules. (b) Gain across target temperatures τ. (c) Local inferability intervention: masking landmarks leaves target fidelity unchanged but reduces the FPAS-over-MAPPO gain; difference-in-differences $+18.5$, paired-bootstrap 95% CI $[+14.1, +22.7]$.

4.3 Properties of Useful Prediction Targets

We next vary the target while keeping the supervision recipe fixed (Figure 4). The interventions identify properties that determine whether privileged coordination advice provides useful representation supervision.

Target fidelity. Per-step stochastic corruption reduces argmax agreement with the clean hard target from 1.0 to $0.84/0.68/0.53/0.37$. The MPE gain falls monotonically from $+21.1$ to $+6.8$, $+4.9$, $+2.5$, and $+2.5$ (3 seeds each), linking target fidelity to the strength of the auxiliary signal.

Target consistency. Deterministic but suboptimal generators remain useful at lower fidelity than stochastic corruption. "Go to the nearest landmark" (fidelity 0.78) retains $+15.3$; "all agents go to landmark 0" (fidelity 0.65) retains $+10.1$; and a lookahead planner target (fidelity 0.60) retains $+11.5$. Stochastic corruption at comparable fidelity (q=0.5, fidelity 0.68) retains $+4.9$. The comparison suggests that a consistent state-to-target mapping can provide a more learnable auxiliary task than temporally resampled errors of similar aggregate fidelity.

Local inferability. A supervised probe predicts hard-target argmaxes from local observations with 0.683 accuracy (majority class 0.265). Masking the landmark block keeps target fidelity at 1.0 but reduces probe accuracy to 0.539. Across 8 matched seeds per cell, the FPAS-over-MAPPO gain changes from $+21.1$ with full observations to $+2.6$ under masking, a difference-in-differences of $+18.5$ (paired-bootstrap 95% CI $[+14.1, +22.7]$). Useful privileged targets therefore need a sufficiently learnable relationship to the decentralized actor's information.

Target sharpness. Softening the assignment target with $\tau \in \{0.05, 0.2, 0.5, 1.0\}$ produces gains of $+21.0/+16.0/+6.0/+1.7$. The decreasing curve shows that target sharpness controls how strongly the coordination decision is expressed to the encoder.

Coordination-decision content. Generic privileged features yield much smaller gains under the same auxiliary recipe. Predicting other agents' velocities yields $+2.3$ (paired p=0.06), and global landmark-occupancy flags yield $+1.3$ (p=0.39), compared with $+21.3$ for the assignment decision. The decision itself remains effective in different encodings: action-factorized targets give $+21.3$ and coordinate targets $+24.7$ ($p=1.6\times10^{-8}$). SMACv2 gives the same content-level separation. Slot-indexed decision variables plus health inputs reach 0.636 ± 0.011; health inputs alone reach 0.574 ± 0.021 (p=0.020), and decision flags alone reach 0.628 ± 0.013 (above health-only, p=0.029). A fixed invertible derangement of the action-factorized target reaches 0.631 ± 0.030 versus 0.611 ± 0.023 for FPAS (p=0.28). The useful supervision is therefore the coordination decision encoded by the target, while action-space factorization is one convenient interface for delivering it.

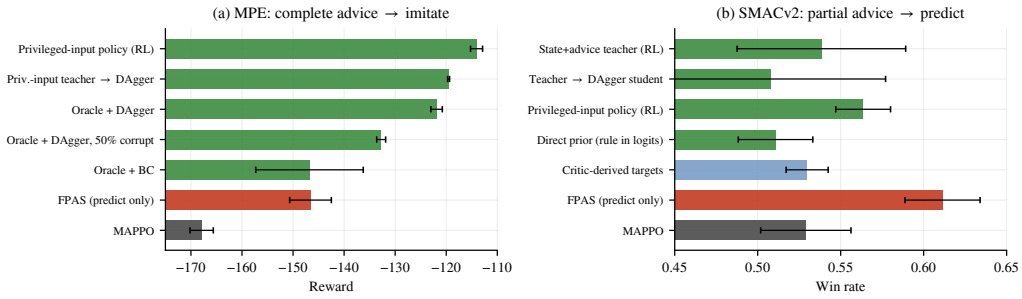

Figure 5: **Supervision regimes for privileged advice.** (a) On MPE, imitation is strongest when the advice is policy-like and locally attainable. (b) On SMACv2, representation-level prediction is strongest when the coordination advice specifies only part of the policy. Green denotes imitation-oriented routes; red denotes FPAS; error bars show s.d.

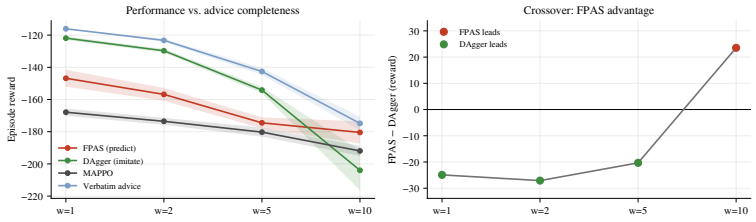

Figure 6: **Controlled crossover as omitted behavior becomes more task-relevant.** Increasing the MPE collision weight w lowers the advice-follower return and reverses the FPAS-minus-DAgger gap: DAgger leads at low w, whereas FPAS leads at $w{=}10$.

4.4 Supervision Regimes for Privileged Advice

Behavioral supervision, behavioral regularization, and representation prediction use privileged advice in distinct ways. Let $q_i^g(\cdot \mid s_t)$ denote the behavioral distribution induced by advice: a one-hot teacher action for BC/DAgger or normalized positive advice flags for the SMACv2 hybrid. Their objectives are

$$\underbrace{\min_\theta \mathrm{CE}\big(q_i^g,\, \pi_\theta(\cdot \mid h_i)\big)}_{\text{imitation (BC, DAgger)}} \qquad \underbrace{\mathcal{L}_{\mathrm{PPO}} + \lambda_{\mathrm{BC}}\, \mathrm{CE}\big(q_i^g,\, \pi_\theta(\cdot \mid h_i)\big)}_{\text{behavioral regularization}} \qquad \underbrace{\mathcal{L}_{\mathrm{PPO}} + \lambda\, \mathcal{L}_{\mathrm{aux}}\big(f_D(h_i^t),\, z_i^t\big)}_{\text{prediction (FPAS)}} \tag{2}$$

The comparisons below vary the quality and local attainability of the same underlying advice while keeping the evaluation task explicit (Figure 5).

Policy-like advice. On MPE, the assignment rule executed directly earns approximately -115, making it a strong and locally recoverable policy. DAgger therefore provides an efficient route: a student of a learned privileged-input teacher (teacher -114.1) reaches $-119.5{\pm}0.2$, and a student of the analytic oracle reaches $-121.9{\pm}1.1$, both ahead of FPAS at $-146.6{\pm}4.1$. With 50% corrupted labels, DAgger still reaches $-132.7{\pm}0.9$, while FPAS on the same corrupted signal reaches $-163.0{\pm}1.6$. When advice already specifies strong executable behavior, imitation efficiently transfers that behavior.

Partial advice. The StarCraft focus-fire target identifies attack preferences but leaves movement direction and positioning to the learner. Its greedy follower wins $3.0{\pm}4.5\%$, $9.0{\pm}2.2\%$, and $8.0{\pm}4.5\%$ on Terran, Protoss, and Zerg. A direct logit prior reaches $0.511{\pm}0.022$, and a privileged-input RL policy reaches $0.564{\pm}0.017$ (3 seeds), compared with FPAS at $0.611{\pm}0.023$ (Welch $t{=}3.44$, $p{=}0.016$). Training the privileged-input policy for 20M steps reaches $0.579{\pm}0.013$ ($p{=}0.044$ versus FPAS at 10M). We also train a centralized state+advice policy and distill it with DAgger. The teacher reaches $0.538{\pm}0.051$ under the final-10% metric; doubling hidden capacity

yields 0.536 ± 0.012, and the decentralized student reaches 0.508 ± 0.069. FPAS exceeds both in the matched pipeline (paired $p=0.049$ versus teacher; $p=0.035$ versus student). These results place partial coordination advice in a different regime from the policy-like MPE signal.

Controlled crossover. We vary the MPE collision penalty w, which increases the importance of behavior omitted by the assignment advice. The follower return declines from -116 to -175 as $w{:}1 \to 10$. Over the same sweep, DAgger changes from a 24.9-point lead over FPAS at $w=1$ to a 23.5-point deficit at $w=10$ (Figure 6). The crossover links the relative effectiveness of the two routes to how much task-relevant behavior the advice already specifies.

Local attainability in Overcooked. Overcooked further tests whether decentralized imitation can realize the behavior encoded by the advice in closed loop. On `coordination_ring`, the scripted `GreedyHumanModel` follower deadlocks and earns 0; DAgger also earns 0 on all five seeds. FPAS predicts the same six-class subtask decision and reaches 211.5 ± 10.5, with every seed above 200, while MAPPO scores 163.9 ± 92.0 (FPAS versus DAgger paired $p < 10^{-5}$). On `cramped_room`, the follower scores 184 and DAgger reaches 180.0 ± 28.3, matching its teacher at roughly $25\times$ fewer training steps than RL. MAPPO reaches 219.9 ± 0.2 and FPAS 227.6 ± 11.4 (FPAS versus DAgger $p=0.023$; FPAS versus MAPPO $p=0.20$). A same-signal control uses exactly the primitive-action labels imitated by DAgger as FPAS targets and obtains 207.8 ± 10.7 on `coordination_ring` and 236.0 ± 8.9 on `cramped_room` (versus DAgger $p = 1.7 \times 10^{-6}$ and $p=0.005$; the latter is also above MAPPO, $p=0.016$). The same label stream can therefore support different final policies depending on whether it constrains actions or shapes the representation used by RL.

Gradient-entry comparison. PPO plus cross-entropy toward the advice distribution retains the reward objective while moving the supervision into the policy head. On SMACv2, $\lambda_{\mathrm{BC}} \in \{0.1, 0.5, 1.0\}$ yields 0.121 ± 0.014, 0.105 ± 0.002, and 0.090 ± 0.004; linear annealing gives 0.122 ± 0.033. Applying the same cross-entropy to the auxiliary head gives 0.609 ± 0.003, matching squared-error FPAS at 0.611 ± 0.023 ($p=0.84$). Over the first 30 updates, the weighted behavioral term at $\lambda_{\mathrm{BC}}=0.1$ averages $1.2\times$ the PPO gradient norm, while FPAS-CE averages $1.4\times$. Restricting both routes to executable actions raises the behavioral route to 0.316 ± 0.014 but leaves it below MAPPO ($p < 10^{-4}$); the auxiliary route reaches 0.574 ± 0.011 (paired $p=0.002$). In Overcooked, where labels are close to executable behavior, annealed regularization reaches 268.6 ± 6.4 and 238.1 ± 3.2, matching FPAS-primitive on `cramped_room` ($p=0.68$) and exceeding it on `coordination_ring` ($p < 10^{-3}$). Behavioral regularization is therefore a viable third route when advice labels align closely with executable actions.

Empirical route-selection diagnostic. Across the pure-imitation pipelines, executed student return tracks the return that the teacher or advice follower makes available. We summarize this observation with

$$C = \frac{J_{\mathrm{follower}} - J_{\mathrm{random}}}{J_{\mathrm{RL}} - J_{\mathrm{random}}}. \tag{3}$$

We pair C with the closed-loop attainability check $J_{\mathrm{imit}} \approx J_{\mathrm{follower}}$, where J_{imit} is the return reached by the decentralized imitation pipeline. In the measured settings, FPAS leads when $C<1$; when $C \geq 1$, imitation becomes competitive when the decentralized imitator approaches the follower return. At $w=10$, for example, the follower beats MAPPO while the teacher–student gap grows to 29 points and prediction leads. Overcooked shows why attainability must be measured in closed loop: the `coordination_ring` teacher is 99.9% predictable yet earns zero, while on `cramped_room` a one-round clone earns zero before DAgger closes the state-distribution gap and reaches the follower return. Equation 3 is a post-hoc summary of the nine settings used to form it and serves as a hypothesis for prospective validation.

5 Discussion

The results support a simple view of privileged coordination advice: the same signal can help a decentralized actor either by specifying behavior worth executing or by exposing coordination structure worth representing. Behavioral supervision uses the first route by pushing the policy toward advised actions. FPAS uses the second route by predicting the privileged target during training

while leaving action selection to reinforcement learning. The matched pathway controls and frozen-encoder probes connect the gain to this representation route, while the target interventions show that effective prediction targets are sufficiently faithful, consistent, locally inferable, sharp, and tied to the coordination decision.

The next question is when each route is useful. In MPE, the assignment advice already behaves like a strong policy and can be reproduced by decentralized agents, so imitation transfers it efficiently. As the collision penalty increases, more task-relevant behavior falls outside the assignment rule and the advantage shifts toward FPAS. SMACv2 lies in a partial-advice regime because focus-fire guidance specifies whom to attack but leaves movement direction and positioning to reinforcement learning. Overcooked adds a second requirement: advice must also be realizable from decentralized observations for direct imitation to reproduce its behavior in closed loop. Behavioral regularization provides an intermediate route when the labels are already close to executable actions.

These cases are organized by two operational quantities. The follower return measures the behavioral value available from executing the advice, whereas the closed-loop imitator return measures how much of that behavior a decentralized learner can actually realize. When both are high, imitation can efficiently transfer the advised policy. When the advice covers only part of the behavior, its direct return can be low even though its coordination decision remains useful as a prediction target. When the follower is strong but decentralized imitation cannot reproduce it, local attainability becomes the limiting factor. This two-axis view explains why the same privileged signal can favor different supervision interfaces across tasks.

This perspective broadens the role of privileged information in CTDE. Partial privileged signals can remain useful because a heuristic, solver, planner, or centralized computation may expose coordination structure beyond the behavior it directly specifies. The central design problem therefore has two parts: what privileged information is available during training, and how that information is allowed to shape the actor. Treating the supervision interface explicitly connects privileged guidance, imitation, and auxiliary representation learning within one design space.

6 LIMITATIONS

The evidence spans SMACv2, MPE, and Overcooked, while the base reinforcement-learning algorithm is MAPPO. The main StarCraft targets are hand-designed coordination rules; planner-derived MPE targets retain $+11.5$ of the $+21.3$ gain, whereas critic-derived SMACv2 targets remain near MAPPO at 0.530 ± 0.013 (Appendix E). The route-selection diagnostic in Eq. 3 summarizes the same nine measured settings from which it was formed, so prospective tests on new tasks, advice generators, and learning algorithms are needed to establish how broadly its empirical thresholds transfer. These boundaries define the current scope of the supervision-regime claim.

7 CONCLUSION

This work studies how privileged coordination advice should reach decentralized actors during CTDE. FPAS provides one answer: privileged advice can supervise the actor's representation during training while reinforcement learning remains responsible for action selection. This simple interface produces consistent gains on SMACv2, and matched pathway, target-content, and representation analyses explain when and why the auxiliary signal is useful.

The cross-domain results show that the effective supervision route depends on the advice and on the decentralized learner. Imitation is efficient when advice already defines strong behavior that local actors can reproduce. Representation supervision becomes more useful when the advice captures coordination structure but leaves important behavior unspecified or difficult to realize locally. Behavioral regularization provides a third option when advice labels are already close to executable actions.

The broader implication is that the supervision interface for privileged information is itself a design variable in CTDE. Partial heuristics, solvers, planners, and centralized computations can contribute useful training structure through the coordination decisions they expose, while reward-driven learning completes the remaining behavior. This expands the usable class of training-time privileged signals beyond complete expert policies.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, provide feedback on experimental design, support result interpretation, assist with code implementation and debugging, and support manuscript drafting, translation, revision, and grammar editing. The research hypotheses and core conceptual framing were developed and refined by the authors. All AI-assisted outputs incorporated into the work were reviewed and verified by the authors, who take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Target generators, training and evaluation protocols, complete ablations, teacher-distillation procedures, and data-integrity audits are documented in the appendices. Anonymous code, raw logs, and verification scripts are included in the supplementary artifact.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *NeurIPS*, 2021.

Rupali Bhati, Anurag Kadkol, Andrea Baisero, and Christopher Amato. Decentralized asymmetric DQN: Decentralization without factorization in multi-agent reinforcement learning. *Reinforcement Learning Journal*, 7, 2026.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems*, 2019.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *CoRL*, 2020.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *AISTATS*, 2019.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. In *ICML*, 2019.

Sam Devlin and Daniel Kudenko. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *AAMAS*, 2011.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Nicolaus Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *NeurIPS Datasets and Benchmarks*, 2023.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *NeurIPS*, 2016.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*, 2018.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

Cong Guan, Feng Chen, Lei Yuan, Chenghe Wang, Hao Yin, Zongzhang Zhang, and Yang Yu. Efficient multi-agent communication via self-supervised information aggregation. In *Advances in Neural Information Processing Systems*, pp. 1020–1033, 2022.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John Agapiou, Joel Z Leibo, and Audrunas Gruslys. Deep Q-learning from demonstrations. In *AAAI Conference on Artificial Intelligence*, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EpVe8jAjdx.

Qinwei Huang, Simon Khan, Rui Zuo, Stefan Wang, Garrett E. Katz, and Qinru Qiu. Predictive auxiliary learning for belief-based multi-agent systems. In *Proceedings of the AAAI Symposium Series*, volume 8, pp. 50–58, 2026. doi: 10.1609/aaaiss.v8i1.42517.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *ICRA*, 2019.

Landon Kraemer and Bikramjit Banerjee. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190:82–94, 2016.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *ICLR*, 2022.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

Sergey Levine and Vladlen Koltun. Guided policy search. In *ICML*, 2013.

Yueheng Li, Guangming Xie, and Zongqing Lu. Multi-agent guided policy optimization. In *ICLR*, 2026.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, 2017.

Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. In *AAMAS*, 2021.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, and Christopher Amato. A deeper understanding of state-based critics in multi-agent reinforcement learning. In *AAAI*, 2022.

Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: Multi-agent variational exploration. In *NeurIPS*, 2019.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *AAAI*, 2018.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, 1999.

Frans A Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *NeurIPS Datasets and Benchmarks*, 2021.

Monirul Islam Pavel, Siyi Hu, Muhammad Anwar Masum, Mahardhika Pratama, Ryszard Kowalczyk, and Zehong Jimmy Cao. KD-MARL: Resource-aware knowledge distillation in multi-agent reinforcement learning. *arXiv preprint arXiv:2604.06691*, 2026.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.008.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *NeurIPS*, 2020.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Çaglar Gülçehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *ICLR*, 2016.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *AAMAS*, 2019.

Simon Schmitt, Jonathan J Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M Czarnecki, Joel Z Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, and S M Ali Eslami. Kickstarting deep reinforcement learning. *arXiv preprint arXiv:1803.03835*, 2018.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *ICML*, 2019.

Yuda Song, Dhruv Rohatgi, Aarti Singh, and J. Andrew Bagnell. To distill or decide? understanding the algorithmic trade-off in partially observable RL. In *Advances in Neural Information Processing Systems*, 2025.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *NeurIPS*, 2016.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS*, 2018.

Yee Whye Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *NeurIPS*, 2017.

Lisa Torrey and Matthew E Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *AAMAS*, 2013.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

Aaron Walsman, Muru Zhang, Sanjiban Choudhury, Dieter Fox, and Ali Farhadi. Impossibly good experts and how to follow them. In *International Conference on Learning Representations*, 2023.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent q-learning. In *ICLR*, 2021a.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *ICLR*, 2021b.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. Bridging the imitation gap by adaptive insubordination. In *NeurIPS*, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *NeurIPS*, 2022.

Jian Zhao, Xunhan Hu, Mingyu Yang, Wengang Zhou, Jiangcheng Zhu, and Houqiang Li. CTDS: Centralized teacher with decentralized student for multiagent reinforcement learning. *IEEE Transactions on Games*, 16(1):140–150, 2024. doi: 10.1109/TG.2022.3232390.

A Experimental Setup and Evaluation

Environments. SMACv2 (Ellis et al., 2023) 5v5 with procedurally generated unit compositions and start positions (races: Terran, Protoss, Zerg), the successor of SMAC (Samvelyan et al., 2019); MPE simple_spread (Lowe et al., 2017; Mordatch & Abbeel, 2018) with 3 agents, 3 landmarks, 25-step episodes; and two-player Overcooked-AI (Carroll et al., 2019) on `cramped_room` and `coordination_ring`, 400-step episodes, scored by sparse return. All are standard cooperative MARL benchmarks (Papoudakis et al., 2021).

Training. All SMACv2 RL runs use recurrent MAPPO with hidden size 64, lr 5×10^{-4}, 10M environment steps, 2 rollout threads, PPO epoch 5, evaluation every 25 episodes with 32 episodes per evaluation (500 evaluation points per run), unless an intervention explicitly changes budget or capacity. MPE RL runs use recurrent MAPPO with the codebase defaults, 2M steps and 2,000 evaluation points. The headline FPAS condition uses λ=0.5 and no logit modulation (β=0) in both domains; the modulation variant uses β=2. Within each matched RL comparison, the parity audit (Appendix F) checks map, observation dimensions, step budget, evaluation cadence, and logged hyperparameters, with an explicit whitelist for intervention parameters. Imitation and teacher-student protocols are specified separately in Appendix D.

Scoring protocol. The per-seed score is the mean over the final 10% of evaluation points (50 of 500 on SMACv2; 200 of 2,000 on MPE). The protocol was fixed on the development domains before any held-out run was scored.

Statistics. Race-level comparisons use paired t-tests on matched seeds with Holm correction across the three races. Following the guidance of Agarwal et al. (2021) against overconfident point comparisons, the overall SMACv2 effect uses a hierarchical paired bootstrap (20,000 resamples): resample races with replacement, then matched seeds within each race, then average seed-level paired differences.

B Target Generator Specifications

Targets are recomputed from privileged simulator state at every step and appended to each agent's training observation; the actor removes the target block before encoding, and the centralized critic receives the unchanged MAPPO global-state input without that block. The auxiliary loss supervises all action slots without availability masking. Dead agents receive all-zero target rows.

SMACv2 hard targets (headline FPAS condition). Let $r_j = h_j/h_j^{\max}$ be enemy j's remaining-health fraction (enemies with $h_j \le 0$ are dead), $m = \min\{r_j : j \text{ alive}\}$, and ρ_i agent i's own health fraction. Action indices follow SMAC: 0 no-op, 1 stop, 2–5 the four moves, and $n_{\text{na}}{+}j$ the attack on enemy j (with n_{na}=6 for 5v5). For every living agent i,

$$\tau_i(n_{\text{na}}{+}j) = \mathbb{1}\big[\, j \text{ alive } \wedge\ |r_j - m| < 0.02 \,\big], \qquad \tau_i(d) = \mathbb{1}\big[\rho_i < 0.3\big] \ \text{ for } d \in \{2,3,4,5\}, \quad (4)$$

and all remaining entries are 0. Ties are intentional: every enemy within the 0.02 band of the minimum is marked, so the target is a focus-fire *set*, not an argmax. No noise is added.

SMACv2 fuzzy targets (modulation variant). Attack entries score $\tau_i(n_{\text{na}}{+}j) = \text{clip}\big(0.6\,(1 - r_j) + 0.3\mathbb{1}[|r_j - m| < 0.02] + 0.1,\, 0, 1\big)$ for living enemies; move entries are 0.35 if $\rho_i < 0.3$ and 0.1 otherwise; stop is 0.05. Gaussian noise $\mathcal{N}(0, 0.1^2)$ is added elementwise and the result is clipped to $[0, 1]$.

MPE targets. Let $C_{ik} = \|x_i - \ell_k\|_2$ be the agent–landmark distance matrix, solved by the Hungarian algorithm; write $\Delta = \ell_{k(i)} - x_i$ for agent i's assigned displacement. *Hard:* if $\|\Delta\| < 0.01$, the stay action gets 1; otherwise the dominant-axis signed direction gets 1 (right/left if $|\Delta_x| \geq |\Delta_y|$, else up/down). No noise. *Fuzzy:* both axis directions are scored $0.3 + 0.5\,|\Delta_{x \text{ or } y}|/(|\Delta_x| + |\Delta_y|)$ on their signed sides, stay is 0.05 (0.5 when $\|\Delta\| < 0.01$), with $\mathcal{N}(0, 0.1^2)$ noise, clipped to $[0, 1]$.

Loss weight. All headline and mechanism conditions use λ=0.5 unless the condition itself varies λ (robustness sweep: $\lambda \in \{0.1, 0.5, 1.0\}$, Appendix E).

C Full Mechanism and Efficacy Results

Tables 2 and 3 consolidate the conditions supporting the efficacy, mechanism, target-property, and held-out analyses in Sections 4.1–4.3. The headline FPAS rows are bolded to anchor the comparison; the remaining rows report matched controls and extensions.

Table 2: **Full MPE mechanism and target-property results.** Mean over the final 10% of evaluations; Δ is measured relative to MAPPO. The headline FPAS condition is bolded; the remaining rows provide the pathway, fidelity, consistency, encoding, and inferability controls used in the main text.

Condition	n	Reward	Δ
MAPPO	8	-167.90 ± 2.28	–
FPAS aux-only (hard)	8	$\mathbf{-146.58} \pm 4.08$	$\mathbf{+21.32}$
Hard + modulation	8	-146.81 ± 5.22	$+21.09$
Aux-only (fuzzy)	3	-158.70 ± 1.35	$+9.20$
Fuzzy + modulation	8	-158.59 ± 4.02	$+9.31$
Modulation, stop-gradient	8	-165.39 ± 1.48	$+2.52$
Aux on shuffled targets	8	-165.18 ± 1.59	$+2.72$
Shuffled + modulation	8	-165.89 ± 1.15	$+2.01$
Direct prior	8	-164.04 ± 1.21	$+3.86$
Reward shaping (scale 2.0)	3	-163.37 ± 2.25	$+4.53$
Corrupt q=0.25	3	-161.07 ± 1.80	$+6.83$
Corrupt q=0.50	3	-162.99 ± 1.64	$+4.91$
Corrupt q=0.75	3	-165.35 ± 1.93	$+2.55$
Corrupt q=1.00	3	-165.40 ± 0.18	$+2.50$
Nearest-landmark rule	3	-152.56 ± 2.28	$+15.34$
Fixed-landmark rule	3	-157.83 ± 3.49	$+10.07$
Planner-derived	5	-156.37 ± 1.85	$+11.53$
Temperature τ=0.05/0.2/0.5/1.0	3 ea.	$+20.95/+15.97/+5.97/+1.66$	
Generic aux: other-agent velocities	8	-165.62 ± 1.97	$+2.28$
Generic aux: landmark occupancy	8	-166.58 ± 2.86	$+1.32$
Assignment target, coordinate form	8	-143.16 ± 2.13	$+24.74$
MAPPO, landmarks masked	8	-182.74 ± 1.97	–
FPAS hard, landmarks masked	8	-180.10 ± 1.82	$+2.64$

Linear probe protocol. For each of the 5 matched Terran seeds we roll out the trained MAPPO policy greedily for 50 evaluation episodes ($\sim$9–11k living-agent steps) and record, at every step, the raw observation, both frozen encoders' recurrent hidden states on the identical input stream, and labels computed from the simulator: the index of the minimum-health living enemy (6 classes including none), the own-health < 0.3 flag, and per-enemy membership in the exact focus set (within

Table 3: **Full SMACv2 mechanism and held-out results.** Win rate under the final-10% protocol. Bold rows mark the headline FPAS condition in each race; the Terran rows provide the matched pathway, target-content, teacher, and behavioral-supervision controls used in the main text.

Race	Condition	n	Win rate
Terran	MAPPO	5	0.529 ± 0.027
Terran	**FPAS aux-only (hard)**	5	$\mathbf{0.611} \pm 0.023$
Terran	Hard + modulation	5	0.620 ± 0.024
Terran	Fuzzy + modulation	5	0.582 ± 0.009
Terran	Agent-permuted aux	3	0.580 ± 0.012
Terran	Action-permuted (fuzzy)	3	0.544 ± 0.019
Terran	Action-permuted aux (hard)	3	0.494 ± 0.025
Terran	Shuffled + modulation	5	0.541 ± 0.024
Terran	Direct prior	5	0.511 ± 0.022
Terran	Reward shaping	5	0.516 ± 0.020
Terran	Privileged-input policy	3	0.564 ± 0.017
Terran	Privileged-input policy (20M)	3	0.579 ± 0.013
Terran	Centralized state+advice teacher	5	0.538 ± 0.051
Terran	Centralized teacher (hidden 128)	3	0.536 ± 0.012
Terran	Teacher $\rightarrow$ DAgger student	5	0.508 ± 0.069
Terran	Slot-indexed decision features	3	0.636 ± 0.011
Terran	Slot features, health inputs only	3	0.574 ± 0.021
Terran	Slot features, decisions only (no health)	5	0.628 ± 0.013
Terran	Fixed derangement of action target	5	0.631 ± 0.030
Terran	FPAS, cross-entropy aux loss	3	0.609 ± 0.003
Terran	PPO+BC hybrid, λ_{BC}=0.1	3	0.121 ± 0.014
Terran	PPO+BC hybrid, λ_{BC}=0.5	3	0.105 ± 0.002
Terran	PPO+BC hybrid, λ_{BC}=1.0	3	0.090 ± 0.004
Terran	PPO+BC hybrid, λ_{BC}=0.5 annealed to 0	3	0.122 ± 0.033
Terran	Critic-derived targets	3	0.530 ± 0.013
Protoss	MAPPO	5	0.427 ± 0.056
Protoss	**FPAS aux-only (hard)**	5	$\mathbf{0.555} \pm 0.019$
Protoss	Hard + modulation	5	0.590 ± 0.027
Zerg	MAPPO	5	0.513 ± 0.025
Zerg	**FPAS aux-only (hard)**	5	$\mathbf{0.566} \pm 0.022$
Zerg	Hard + modulation	5	0.559 ± 0.011

0.02 of the minimum health). Probes are standardized logistic models whose regularization strength is selected by episode-grouped 4-fold cross-validation on the first 40 episodes and evaluated on the held-out 10; we report balanced accuracy to guard against class imbalance, and the focus-set probe averages per-slot balanced accuracy over enemy slots with both classes present. The two encoders share the identical architecture and checkpointed training pipeline, differing only in λ (0 for the MAPPO control, 0.5 for FPAS), and both are evaluated on the identical observation stream generated by the MAPPO policy. Regenerate with `probe_representation.py`.

D Supervision-Regime Baselines and Diagnostics

This appendix specifies every behavioral-supervision baseline compared in Section 4.4, in the order the main text introduces them, together with the measurements the main text cites.

MPE. The learned teacher is recurrent MAPPO trained for 2M steps on observations augmented with the hard target block (final evaluation -114.1 mean over 3 seeds). Students are MLPs trained by DAgger (3 rounds, 300 episodes per round, 30 epochs, cross-entropy on teacher actions at states visited by the student), evaluated over 200 episodes. The oracle teacher replaces the learned policy with the analytic assignment rule through the identical pipeline. Behavior cloning (BC) uses oracle labels on oracle-visited states only (8 seeds). Corrupted DAgger replaces each agent's label with another agent's, independently per step with probability 0.5, mirroring the FPAS corruption inter-

vention so that imitation-based and prediction-based consumption of the same corrupted signal are compared head-to-head.

SMACv2. The centralized teacher is recurrent MAPPO under the shared 10M-step configuration. Its actor receives the 211-dim concatenation [local obs (80) | global state (120) | hard target (11)] at every step, including execution; decentralized students receive the 80-dim local observation. Per matched seed, the teacher's final-10% win rate is 0.538 ± 0.051 (5 seeds); a capacity-doubled variant (hidden size 128, 3 seeds) reaches 0.536 ± 0.012, leaving teacher performance essentially unchanged. The student shares the recurrent actor architecture of the MAPPO/FPAS runs (hidden size 64). DAgger runs 3 rounds of 300 episodes: round 0 collects teacher rollouts; later rounds roll out the current student greedily while the teacher, fed its own privileged input stream along the same trajectory, provides greedy action labels. The aggregated dataset is refit each round by truncated-BPTT cross-entropy (15 epochs, unavailable actions masked). Students are evaluated greedily over 100 episodes on the standard evaluation distribution (0.508 ± 0.069 over 5 seeds); the teachers evaluated by the same protocol score 0.560 ± 0.077. Against FPAS, paired final-10% tests give t=3.14, p=0.035 for the student and t=2.79, p=0.049 for the teacher. Regenerate with `distill_smacv2_teacher.py`.

Hybrid RL+imitation (behavioral regularization). The hybrid adds $\lambda_{\mathrm{BC}}\,\mathrm{CE}\big(q_i^g, \pi_\theta(\cdot \mid h_i)\big)$ to PPO. On SMACv2 the labels are normalized multi-hot advice flags; $\lambda_{\mathrm{BC}} \in \{0.1, 0.5, 1.0\}$ is fixed or 0.5 is annealed to zero (3 seeds each). The policy-head CE condition applies the policy availability mask while the multi-hot target may include unavailable actions; a positive advice action is available in only 22–33% of Terran decisions. We therefore include an availability-matched condition (`run_sc2_executable_ce.sh`, 3 seeds) that separates supervision entry point from label executability: it restricts both targets and logits to executable actions and skips steps without an executable positive flag. The behavioral route recovers to 0.316 ± 0.014, still below plain MAPPO (Welch $p<10^{-4}$), while the identical restricted labels on the auxiliary head reach 0.574 ± 0.011 (paired p=0.002), above MAPPO (p=0.019); restricting the auxiliary target to executable slots costs a little relative to supervising all slots (0.609 ± 0.003, p=0.024). On Overcooked, fixed regularization scores 109.1 ± 71.7 and 204.2 ± 14.5 on `coordination_ring` and `cramped_room`; primitive-label FPAS is higher (paired p=0.048 and $p<10^{-3}$). Annealing reaches 268.6 ± 6.4 and 238.1 ± 3.2.

Loss-matched auxiliary controls. FPAS-CE applies cross-entropy to the auxiliary head: SMACv2 0.609 ± 0.003 (3 seeds); Overcooked 263.2 ± 28.3 / 236.0 ± 8.9 (`coordination_ring` / `cramped_room`, 5 seeds each). On Overcooked it stays above DAgger on both layouts, matches squared-error FPAS-primitive on `cramped_room` (236.0 for both), and exceeds it on `coordination_ring` (paired p=0.0031). Launchers are `run_sc2_hybrid.sh`, `run_overcooked_hybrid.sh`, and `run_stage2_controls.sh`. The availability-matched SC2 comparison uses `run_sc2_executable_ce.sh`. *Gradient-scale audit.* In the unmasked-target runs, the weighted behavioral term at λ_{BC}=0.1 averages $\lambda_{\mathrm{BC}}\|\nabla\mathcal{L}_{\mathrm{BC}}\|/\|\nabla\mathcal{L}_{\mathrm{PPO}}\| = 1.21$ (median 1.20, max 1.9), while FPAS-CE at λ=0.5 averages 1.43 (median 1.31). On this gradient-norm measure the behavioral term is smaller than FPAS-CE; the availability-matched comparison then isolates label executability separately.

Overcooked. Two-player Overcooked-AI on layouts `cramped_room` and `coordination_ring`, horizon 400, featurized 96-dim local observations. RL conditions (MAPPO and FPAS, 5 seeds each) train recurrent MAPPO for 6M environment steps with shaped reward during training and are scored by *sparse* return (soups delivered) over the final 10% of evaluation points; FPAS adds the auxiliary prediction of the six-class subtask decision as a one-hot target under the unchanged squared-error head (λ=0.5, β=0), computed from global state by the `GreedyHumanModel` decision tree in `overcooked_wrapper.py`. The imitation route trains a recurrent student pair by DAgger against scripted `GreedyHumanModel` teachers that execute the same decision tree (4 rounds of 150 episodes, 15 epochs per round, 5 seeds per layout), evaluated by sparse return over 50 greedy episodes; the cross-entropy plateaus within this budget ($\sim$240k environment steps, $\sim25\times$ fewer than the 6M-step RL runs), and on `cramped_room` the students match their teachers' returns, indicating saturation for this imitation protocol. The advice follower baseline rolls out the scripted pair directly; a uniform-random pair provides the floor and the plain-

MAPPO return the reference for the completeness normalization C of Section 4.4 (plain RL is used as the denominator so the diagnostic remains independent of the privileged-consumption methods it informs). The same-signal control (FPAS-primitive) reruns the identical FPAS recipe with the auxiliary target set to the one-hot *primitive action* the scripted teacher pair would take in the current state, exactly the label stream DAgger imitates, with a fresh teacher pair instantiated per episode (5 seeds per layout; `train_overcooked_acfd.py --oc_mode fpasprim`). Regenerate with `distill_overcooked_advice.py`, `evaluate_overcooked_advice_follower.py`, and `analyze_overcooked_results.py`. As a protocol audit, we additionally re-evaluated the saved final checkpoint of every RL run (all three conditions, both layouts, 5 seeds) under one identical protocol (sparse return, horizon 400, greedy deterministic actions, matched evaluation episodes, the same protocol used for the DAgger students), and the per-seed returns reproduce the reported training-eval numbers to within 3 points with identical rankings and paired-test conclusions (`eval_oc_final_checkpoints.py`).

Closed-loop attainability. Per-step predictability provides an incomplete measure of whether decentralized imitation can realize a teacher in closed loop. We evaluate this distinction (3 seeds per layout; 100 teacher episodes, 80/20 episode-level split, same recurrent architecture as the DAgger student). On `coordination_ring` the deadlocked teacher is almost perfectly predictable ($99.9\pm0.1\%$ held-out top-1 agreement, NLL 0.019), while its executed behavior earns 0. On `cramped_room`, $84.8\pm2.7\%$ per-step agreement coexists with a *zero-return* clone: the single-round behavior-cloned student earns 0.0 sparse return on all three seeds because small per-step errors compound off the teacher's state distribution, and it is the DAgger corrections in the main protocol that close this gap (to 180.0 ± 28.3). The closed-loop return of the imitator therefore provides the attainability measure used in Section 4.4, because it captures the compounding execution errors that are absent from per-step accuracy. Regenerate with `measure_oc_realizability.py`.

Diagnostic details. At MPE collision weight $w{=}10$, the follower scores -174.8 and MAPPO -191.9, but the teacher-student gap grows from 6 to 29 points across the sweep. On SMACv2, follower win rates of 3–9% compared with MAPPO's 0.43–0.53 give $C{\leq}0.22$. Both settings belong to the heuristic's formation set (Section 4.4).

E Target Extensions and Robustness

Automatic targets. The planner-derived MPE target obtains its semantics from simulator lookahead and retains $+11.5$ of the $+21.3$ benefit under the identical recipe. The critic-derived SC2 target takes the argmax of a learned centralized Q-head and lands at the MAPPO level (0.530 ± 0.013 over 3 seeds), indicating that this value-derived target provides little representation benefit under the tested recipe.

Hyperparameter robustness. On Terran, $\beta \in \{0.5, 1, 2, 4\}$ and $\lambda \in \{0.1, 0.5, 1.0\}$ (3 seeds each) all reach 0.55–0.62 win rate, above MAPPO's 0.529.

Modulation on held-out races. Hard targets with modulation reach 0.590 ± 0.027 (Protoss, 5 seeds) and 0.559 ± 0.011 (Zerg, 5 seeds), consistent with aux-only within noise on both held-out races.

F Reproducibility and Data Integrity

All reported values are generated by scripts that parse the raw training logs. An action-indexing error identified in the initial MPE target generator was corrected before the reported study was finalized, and every affected condition was rerun; all MPE results in the paper use the corrected generator. `build_smacv2_ground_truth.py` regenerates the benchmark table, `build_sc2_ablation_summary.py` the SC2 control table, and `analyze_mpe_mechanism_results.py` the MPE table; `make_paper_figures.py` renders all figures from the same sources. `audit_config_parity.py` asserts configuration identity within each comparison group directly from the log banners. `test_no_target_leakage.py`

verifies that target blocks appended to observations during training have zero influence on the actor's action distribution at execution. Utility and inferability measurements for every target condition are produced by `analyze_e1_conditions.py` on 5,000 sampled states. The advice-completeness diagnostic is regenerated by `evaluate_smacv2_advice_follower.py` and the frozen-encoder probes by `probe_representation.py`, both from the checkpoints included in the artifact; rerunning them requires a local StarCraft II installation.