MA-RL-VLA: COUNTERFACTUAL CREDIT-AWARE PPO FOR MULTI-AGENT REINFORCEMENT FINE-TUNING OF VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has proven effective for fine-tuning *single-agent* vision-language-action (VLA) models, yet extending this paradigm to *multi-agent* vision-language navigation (MA-VLN) remains largely unexplored. We present **MA-RL-VLA**, a centralized-training / decentralized-execution (CTDE) framework whose core design is a mathematically consistent integration of counterfactual credit into PPO for VLA-scale actors. The counterfactual baseline is *identical in form* to COMA's policy-marginalized baseline and, because one actor pass already yields the full distribution over the four discrete navigation actions, is computed by *exact enumeration* at the cost of four small-critic evaluations per agent step. The detached, clipped credit weight is folded *inside* both PPO branches as a combined advantage $\hat{A}_i^{\text{credit}} = \text{sg}[\tilde{w}_i]\hat{A}_i^{\text{GAE}}$, preserving PPO's pessimistic clipping under negative credit and thereby suppressing free-riding. A policy-invariant potential-based progress term (fixed, never tuned) is separated from auxiliary coordination terms with fixed weights. On CoNavBench (Habitat, $N{=}3$) with 8 matched seeds, MA-RL-VLA improves Team Success Rate by $+19.1$ points over behavior cloning (95% CI $[16.3, 21.9]$) and $+6.4$ over standard COMA-style credit ($[3.9, 8.9]$; exact paired permutation $p{=}0.008$), cutting exploration redundancy by 40.7%. Ablations attribute the gain to the credit signal ($+7.8$), the objective form ($+2.3$), and the exact estimator. Our evidence is confined to one simulated benchmark and one VLA backbone; which guarantees transfer to the deployed update is stated precisely.

1 INTRODUCTION

Vision-language-action (VLA) models (Brohan et al., 2023; Kim et al., 2024) acquire broad embodied competence from large-scale imitation, but imitation alone cannot close the gap on long-horizon, exploration-heavy tasks such as collaborative search, rendezvous, and division of labor. A growing line of work therefore *fine-tunes VLAs with reinforcement learning*: PPO-style RL is found to outperform GRPO and DPO across backbones (Liu et al., 2025); SimpleVLA-RL (Li et al., 2025) applies GRPO with dynamic sampling; TGRPO (Chen et al., 2025b) introduces trajectory-level advantages; and GRAPE (Zhang et al., 2024), ConRFT (Chen et al., 2025a), RLDG (Xu et al., 2024), ReWiND (Zhang et al., 2025), and RLinf (Yu et al., 2025) explore preference alignment, consistency-policy fine-tuning, relabeling, reward modeling, and scalable RL infrastructure. *All of these methods train a single agent.*

Multi-agent VLA reinforcement fine-tuning is, to our knowledge, an open problem—and it is precisely where RL should matter most, because coordination behaviors such as rendezvous timing, area partitioning, and aid-seeking are barely present in demonstration data and must be discovered by exploration. We study this problem in *multi-agent vision-language navigation* (MA-VLN) (Wang et al., 2026; Luo & Bai, 2026; Zhou et al., 2026), a demanding instance in which several VLA agents must jointly satisfy a shared instruction under partial observability.

Directly applying single-agent RL to each agent fails for three reasons:

1. **Credit assignment.** The team return is a joint outcome, so an actor updated with team gradients learns slowly and free-riders survive.

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.

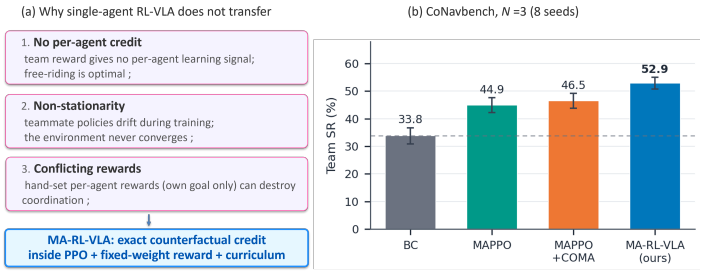

Figure 1: (a) Three obstacles block naive transfer of single-agent RL-VLA to multi-agent VLN: no per-agent signal from the team reward, non-stationarity from drifting teammates, and per-agent rewards that can destroy coordination. (b) MA-RL-VLA exceeds BC and multi-agent RL baselines on CoNavBench (N=3).

2. **Non-stationarity.** Each agent's environment includes the others, whose policies drift during training.
3. **Reward decomposition.** Hand-designed per-agent rewards (e.g., individual goal-reaching) can destroy coordination, making every agent rush to the same goal.

MA-RL-VLA addresses these obstacles jointly. For credit assignment we use the *policy-marginalized counterfactual baseline*—in form, the same quantity COMA (Foerster et al., 2018) introduced—$b_i(s, \mathbf{a}_{-i}) = \mathbb{E}_{a_i' \sim \pi_i}[Q(s, (a_i', \mathbf{a}_{-i}))]$, and the per-agent weight $w_i = Q(s, \mathbf{a}) - b_i$. Because one actor forward pass already yields $\pi_i(\cdot \mid o_i, \ell)$ over the four discrete navigation actions, the expectation is computed by *exact enumeration*—four evaluations of the small centralized critic, with no sampling noise and no 7B re-runs. The technical core of the paper is *how this signed weight enters the actor objective*: we fold it inside both branches of the PPO clip as a combined advantage $\hat{A}_i^{\text{credit}} = \text{sg}[\tilde{w}_i]\hat{A}_i^{\text{GAE}}$, so that pessimistic clipping is preserved on negative credit—exactly the updates that suppress free-riders. For reward design we separate a policy-invariant potential-based progress term (fixed unit coefficient) from auxiliary coordination rewards with *fixed* weights $(1.0, 0.5, 0.3)$ selected on training-split tuning scenes; an online-tuning variant is a costed ablation, not deployed. For stability we instantiate CTDE (Lowe et al., 2017; Yu et al., 2022) with LoRA-adapted (Hu et al., 2022) OpenVLA-7B actors and a centralized Transformer critic, and train through a four-stage curriculum introducing team size, team rewards, credit assignment, and bandwidth limits.

Contributions.

1. **Method.** We define a mathematically consistent credit-aware PPO objective for multi-agent VLA fine-tuning: a detached, bounded counterfactual credit weight is combined with the GAE advantage *inside* the PPO clip, preserving pessimistic clipping for negative credit; the counterfactual baseline is exact enumeration with a favorable cost profile (one actor pass yields the full four-action distribution; four small-critic evaluations per agent step). We state precisely what is and is not guaranteed: the unclipped counterfactual score-function term satisfies an exact identity (Prop. 1); sampled baselines reduce variance only above a threshold $M > \kappa$ (Prop. 2), which justifies enumeration here; the deployed composite update is mathematically consistent in that the credit weight enters both PPO branches with its sign preserved, but carries no unbiasedness or variance theorem (Obs. 1). The guarantee map (App. C) keeps every claim auditable.
2. **Empirical.** A controlled study on CoNavBench (8 matched seeds, paired statistics): policy-marginalized COMA credit vs. default-action counterfactuals vs. difference rewards vs. no credit under identical budgets; an objective-form ablation isolating the signed-credit fix; an exact-vs-sampled estimator sweep; a 2×2 credit$\times$weight-regime design; and component ablations. MA-RL-VLA improves Team SR by $+19.1$ over BC and $+6.4$ $[3.9, 8.9]$ over standard COMA-style credit, with a consistent advantage on both scene stacks and a favorable sample-efficiency profile.
3. **Reproducibility.** We document and will release the training/evaluation stack—configuration files, per-seed CSVs for every table and figure, objective and shaping unit tests, the evaluation ledger, and one-click table/figure regeneration scripts—available to reviewers on request and public upon acceptance (Reproducibility Statement).

2 Related Work

RL for VLA models. Beyond the single-agent methods above (Liu et al., 2025; Li et al., 2025; Chen et al., 2025b; Zhang et al., 2024; Chen et al., 2025a; Xu et al., 2024; Zhang et al., 2025), DPPO (Ren et al., 2024) applies PPO to diffusion policies and GRPO (Shao et al., 2024) optimizes group-relative advantages without a critic; the closest multi-agent analogue, independent learning, is outperformed here by CTDE with counterfactual credit.

Credit assignment in MARL. COMA (Foerster et al., 2018) introduces the counterfactual baseline $b_i = \sum_{a'} \pi_i(a' \mid o_i)\, Q(s, (a', \mathbf{a}_{-i}))$, which marginalizes agent i's action over its own policy before the critic is evaluated. *Our baseline in Eq.* (1) *is exactly this quantity, and we do not claim it as new.* What is new is (i) its cost profile at VLA scale: the "policy network" is a 7B LoRA-fine-tuned VLA, yet one actor pass already emits $\pi_i(\cdot \mid o_i, \ell)$ over the four discrete actions, so the expectation costs $|\mathcal{A}|{=}4$ evaluations of the *small* centralized critic and zero 7B re-runs; (ii) a signed-credit PPO surrogate that stays pessimistic when the credit weight is negative (Sec. 4.1), which the naive outer-weight form forfeits; and (iii) a matched-budget comparison of policy-marginalized COMA credit against default-action counterfactuals and difference rewards (Tumer & Wolpert, 2000). Value-factorization methods—QTRAN (Son et al., 2019), QMIX (Rashid et al., 2018), QPlex (Wang et al., 2021)—are compared empirically but target small discrete policies.

CTDE and reward shaping. MADDPG (Lowe et al., 2017) and MAPPO (Yu et al., 2022) instantiate centralized training with decentralized execution; MAPPO's clipped surrogate limits per-update divergence from the pre-trained policy, which we find essential for VLA fine-tuning. Potential-based shaping preserves optimal policies (Ng et al., 1999), with multi-agent and learned-potential extensions (Devlin et al., 2011; Hu et al., 2020) and curriculum learning (Portelas et al., 2019) raising difficulty progressively; we combine a potential-based progress term with auxiliary coordination rewards, a critic-only historical-policy mix, and a four-stage curriculum.

3 Preliminaries

Problem formulation. We model the task as a Dec-POMDP $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}_i\}, \mathcal{A}, T, R, \gamma, \ell \rangle$ with agents $\mathcal{N} = \{1, \ldots, N\}$, a *latent* state $s_t \in \mathcal{S}$, per-agent observation $o_{i,t} \in \mathcal{O}_i$ emitted from s_t, shared action space $\mathcal{A}$ and reward $R(s_t, \mathbf{a}_t)$, discount γ, and instruction ℓ; agent i acts $a_{i,t} \in \mathcal{A}$ from its own history, $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$. The centralized critic does not observe s_t but consumes the joint observation–action history $h_t = (o_{1:N,1:t}, \mathbf{a}_{1:t-1})$ as a sufficient-statistic proxy (Sec. 4.3); a joint policy π maximizes $\mathbb{E}[\sum_{t=0}^{H} \gamma^t R(s_t, \mathbf{a}_t)]$.

Metrics. We report *Team Success Rate* (all subgoals reached within the step limit); *Team SPL*, the per-agent $(l_i^\star / \max(l_i, l_i^\star))$ $\mathbb{1}[\mathrm{succ}_i]$ averaged over agents ($l_i^\star$: optimal path length); *exploration redundancy*, the overlap $\big(\sum_i |E_i| - |\bigcup_i E_i|\big) / |\bigcup_i E_i|$ over visited cell sets E_i; *sample efficiency*, i.e. *steps to a common* 35% *Team-SR threshold* on the validation split (fixed a priori, below every RL method's final performance; censored at 700K and reported as "$>$700K" when never reached—nothing is omitted) plus the *AUC* (mean validation Team SR over training); and *credit agreement* (fraction of held-out analysis episodes whose top-credit agent matches the expert majority; Sec. 6.5).

4 Method: MA-RL-VLA

Figure 2 overviews the framework (details in Secs. 4.1–4.4): a centralized Transformer critic produces Q and V; the credit module forms each agent's detached, clipped weight and the combined advantage $\hat{A}_i^{\mathrm{credit}}$; LoRA-adapted OpenVLA-7B actors act on local observations, the instruction, and fixed-protocol messages; fixed-weight rewards shape the objective; and a four-stage curriculum controls difficulty.

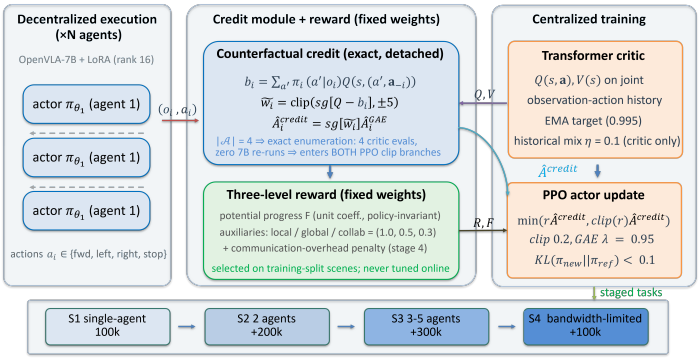

Figure 2: MA-RL-VLA. A centralized Transformer critic produces $Q(s, \mathbf{a})$ and $V(s)$ (Sec. 4.3). The counterfactual credit module forms each agent's detached, clipped weight $\tilde{w}_i$ and the combined advantage $\hat{A}_i^{\text{credit}} = \text{sg}[\tilde{w}_i]\hat{A}_i^{\text{GAE}}$, which enters *both* branches of the PPO clip (Sec. 4.1). Distributed OpenVLA-7B + LoRA actors execute local policies. A three-level reward with *fixed* auxiliary weights shapes the team objective (Sec. 4.2), and a four-stage curriculum stabilizes training (Sec. 4.4).

4.1 Counterfactual Credit Assignment

Let $Q(s, \mathbf{a})$ be the centralized action-value of joint state s and joint action $\mathbf{a}$. For agent i define the counterfactual baseline

$$b_i(s, \mathbf{a}_{-i}) = \mathbb{E}_{a_i' \sim \pi_i}\big[\, Q\big(s, (a_i', \mathbf{a}_{-i})\big)\,\big] = \sum_{a' \in \mathcal{A}} \pi_i(a' \mid o_i, \ell)\, Q\big(s, (a', \mathbf{a}_{-i})\big), \tag{1}$$

which is exactly COMA's policy-marginalized baseline (Foerster et al., 2018). The expectation is written as an explicit sum because $|\mathcal{A}|$=4: one actor forward pass already produces the full distribution $\pi_i(\cdot \mid o_i, \ell)$ and the critic is a small Transformer, so the baseline is computed by *exact enumeration* at the cost of four critic evaluations per agent step—zero sampling error, and independent of the executed action a_i. The *marginal-contribution weight* is

$$w_i = Q(s, \mathbf{a}) - b_i(s, \mathbf{a}_{-i}), \qquad w_i \;\leftarrow\; \text{sg}[w_i], \tag{2}$$

where $\text{sg}[\cdot]$ denotes *stop-gradient*: $Q(s, \mathbf{a})$ and b_i are evaluated under `no_grad`, so w_i acts as a detached per-agent scalar rather than a differentiable value head. Since Q and b_i are unbounded, we clip the detached weight,

$$\tilde{w}_i = \text{clip}\big(\text{sg}[Q(s, \mathbf{a}) - b_i(s, \mathbf{a}_{-i})],\, -c,\, c\big), \qquad c = 5, \tag{3}$$

which preserves the pointwise sign of the marginal contribution while preventing a single outlier from dominating the update (clipping reweights samples and generally introduces bias; App. C).

The deployed objective: combined credit advantage inside the clip. Agent i's actor is trained by standard PPO on the *combined advantage*

$$\hat{A}_i^{\text{credit}} \;=\; \text{sg}[\tilde{w}_i]\, \hat{A}_i^{\text{GAE}}, \tag{4}$$

where $\hat{A}_i^{\text{GAE}}$ is agent i's GAE advantage, i.e.

$$\mathcal{L}_i(\theta_i) = -\mathbb{E}\Big[\, \min\big(\, r_i(\theta_i)\, \hat{A}_i^{\text{credit}},\; \text{clip}(r_i(\theta_i), 1{-}\epsilon, 1{+}\epsilon)\, \hat{A}_i^{\text{credit}}\,\big)\Big], \tag{5}$$

with importance ratio $r_i(\theta_i) = \pi_{\theta_i}(a_i \mid o_i, \ell)/\pi_{\theta_i}^{\text{old}}(a_i \mid o_i, \ell)$ and PPO clip range ϵ=0.2 (Schulman et al., 2017). Because $\tilde{w}_i$ multiplies $\hat{A}_i^{\text{GAE}}$ *before* the min, the objective is exactly standard PPO on the advantage $\hat{A}_i^{\text{credit}}$: pessimistic clipping holds for the *signed* combined advantage. This matters when $\tilde{w}_i < 0$ (17.8% of actor-gradient samples in our runs): there the clipped branch lower-bounds the update on exactly the samples that suppress a harmful per-agent contribution—the free-rider mechanism.

Why not put the weight outside the clip? The legacy form $\tilde{w}_i \min(r_i \hat{A}_i^{\text{GAE}}, \text{clip}(r_i)\hat{A}_i^{\text{GAE}})$ coincides with Eq. (5) whenever $\tilde{w}_i \geq 0$ (since $\tilde{w}_i \min(x, y) = \min(\tilde{w}_i x, \tilde{w}_i y)$ for $\tilde{w}_i \geq 0$), but for $\tilde{w}_i < 0$ the outer multiplication turns min into max, silently selecting the *unclipped* branch and losing pessimism precisely on free-rider updates. We therefore deploy Eq. (5) and retrain *all* headline results with it; the legacy outer-weight form, a $|\tilde{w}_i|$-gated variant, and a credit-as-advantage variant ($\hat{A}_i^{\text{GAE}} := \tilde{w}_i$) appear only as ablations (Table 10, App. O), where the legacy form yields lower and less stable

performance, consistent with the analysis above. A worked numerical example and the unit tests pinning this behavior are in App. H.

4.2 Reward Design: Potential Shaping and Auxiliary Rewards

Not every term is potential-based. We first write the goal-progress term in the transition-level potential form required for policy invariance,

$$F_i(s, s') = \gamma\, \Phi_i(s') - \Phi_i(s), \qquad \Phi_i(s) = -d_i(s), \tag{6}$$

where $d_i(s)$ is the geodesic distance from agent i's pose to its *currently assigned* subgoal and γ=0.99 is the critic discount. The assignment is an observable simulator variable and part of the environment state, so Φ_i is a state function and the Markov property is preserved when assignments change mid-episode (App. H). The terminal state is *absorbing* with $\Phi_i(s_{\text{term}})$=0, so on the terminal transition F_i is evaluated as $\gamma \cdot 0 - \Phi_i(s)$ rather than set to zero; a telescoping unit test verifies that $\sum_t \gamma^t F_{i,t} = -\Phi_i(s_0)$ exactly on random episodes (App. H). This is the only term with an exact potential representation under the discounted objective; the two time penalties are potential-based only in the undiscounted limit (γ=1), and the remaining terms—redundancy (episode-end, joint visit set), aid events (sparse, $\leq$ 2/episode), communication overhead (bytes per step), and collisions (capped at -2.0)—are *auxiliary* and carry no invariance guarantee, so we bound our claim accordingly (Prop. 3).

The remaining terms mix three levels,

$$R_{\text{global}} = +2.0\, \mathbb{I}[\text{all goals reached}] + 0.5\, \mathbb{I}[\text{all agents at goal}] - 0.005\, T_{\text{total}}, \tag{7}$$

$$R_{\text{local},i} = +1.0\, \mathbb{I}[\text{agent } i \text{ at goal}] - 0.01\, T_i - 0.5\, \mathbb{I}[\text{collision}_i], \tag{8}$$

$$R_{\text{collab}} = -0.3\, \text{redundancy} + 0.5\, \text{aid_events} - 0.2\, \text{comm_overhead}, \tag{9}$$

combined with the shaping term as

$$R_t = \sum_i F_i(s_t, s_{t+1}) + w_1 \textstyle\sum_i \bar{R}_{\text{local},i,t} + w_2 R_{\text{global},t} + w_3 R_{\text{collab},t}, \tag{10}$$

where w_1 is the *local*, w_2 the *global*, and w_3 the *collaboration* weight.

The auxiliary weights are *fixed* to $(w_1, w_2, w_3) = (1.0, 0.5, 0.3)$, selected once by three small-scale runs on training-split tuning scenes before the main experiments; the validation split serves only checkpoint selection and the test split is evaluated once. All compared methods share these fixed weights, so no method receives a tuning advantage. The shaping term F_i enters with a *fixed unit coefficient* and is never tuned: it is the only term whose policy invariance (Prop. 3) rests on the exact potential form. An online Bayesian-optimization variant over the auxiliary weights is evaluated as a controlled ablation *with its full branch-training cost accounted* (Sec. 6.3); it yields no significant gain and is not deployed.

The collaboration term rewards *intent–action consistency*: if an agent declares an intent to search a region and the team covers it without duplication, R_{collab} is positive. This turns a coordination behavior that imitation cannot teach into a learnable objective, and it is the term that most reduces exploration redundancy (Sec. 6.3).

4.3 CTDE-VLA Architecture

Distributed actors. Each agent runs an OpenVLA-7B backbone with LoRA (Hu et al., 2022) (rank 16, α=32, $< 0.1\%$ trainable) producing a distribution over four actions (forward, turn left, turn right, stop) from RGB-D observations, the instruction ℓ, and received *messages*. Messages use a *fixed, hand-specified protocol*—a 2-byte packet with the sender's subgoal-region id and a normalized progress scalar—broadcast each step; there is no learned communication content or gate, and learned communication is future work. Messages are enabled only in the bandwidth-limited stage 4, where the communication-overhead penalty in R_{collab} charges bytes per step and the bandwidth budget is halved.

Centralized critic. A 6-layer Transformer encodes the joint observation-action history and outputs $Q(s, \mathbf{a})$ and $V(s)$. Training alternates critic regression and clipped PPO updates (Schulman et al., 2017) (clip 0.2, GAE λ=0.95, entropy $0.01 \to 0.001$, 10 PPO epochs, minibatch 512, rollout 4096 steps/agent, 32 parallel environments, actor lr 3×10^{-5}, critic lr 1×10^{-4}, reward and observation normalization, invalid-action masking).

Non-stationarity mitigation. Critic batches are a $0.9/0.1$ mix of on-policy rollouts and trajectories from frozen historical policies (the BC initialization and snapshots every 50K steps); TD targets are produced by an EMA target critic (decay 0.995). The historical fraction is used *only* in critic regression; actor updates use on-policy rollouts exclusively, so no importance correction is needed for the actor. Because the critic's off-policy batch carries a residual bias we do not correct for, the mixing fraction is ablated over $\eta \in \{0, 0.05, 0.1, 0.2\}$ (Sec. 6.5) rather than assumed optimal.

4.4 Four-Stage Curriculum

Training proceeds through four stages over 700K environment steps, from single-agent R2R/VLN-CE to bandwidth-limited five-agent teams (Table 9 and Algorithm 1 in App. N). The stage-1 checkpoint initializes the VLA; freezing non-collaborative layers at transitions mitigates catastrophic forgetting, alongside LoRA and a KL constraint ($\mathrm{KL}(\pi_{\text{new}} \| \pi_{\text{ref}}) < 0.1$); both stabilizers are ablated in App. J.

5 Theoretical Analysis

We state exactly what the deployed estimator does and does not guarantee; the guarantee map (App. C) lists, per component, what is and is not guaranteed together with the empirical evidence, and we claim nothing beyond it. Throughout, "mathematically consistent" means only that each stated identity or invariance is exact under its stated scope—never a claim of unbiasedness for the composite update.

Proposition 1 (Unclipped counterfactual score-function identity; auxiliary)**.** *Let* $w_i = \mathrm{sg}[Q(s, \mathbf{a}) - b_i(s, \mathbf{a}_{-i})]$ *with* b_i *as in Eq.* (1)*. Then the counterfactual* score-function *estimator* $g_i^{\text{cf}} = w_i \nabla_{\theta_i} \log \pi_{\theta_i}(a_i \mid o_i, \ell)$ *is an unbiased estimator of the joint action-value gradient,* $\mathbb{E}[g_i^{\text{cf}}] = \nabla_{\theta_i} \mathbb{E}[Q(s, \mathbf{a})]$. This identity concerns the auxiliary unclipped estimator only*; the deployed objective is the clipped surrogate Eq.* (5)*, analyzed in Obs. 1.*

Proof sketch. Stop-gradient does not change values, so $w_i = Q - b_i$ pointwise. Conditional on $\mathbf{a}_{-i}$, b_i does not depend on the sampled a_i, so $\mathbb{E}_{a_i \sim \pi_i}[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i)] = 0$ gives $\mathbb{E}[g_i^{\text{cf}}] = \nabla_{\theta_i} \mathbb{E}[Q]$; full derivation in App. A. □

Proposition 2 (Sampled baselines: when sampling is variance-reducing)**.** *Our deployed baseline is exact, so* $\mathrm{Var}(Q - b_i)$ *carries no sampling term; this proposition characterizes the* sampled *alternative for large* $|\mathcal{A}|$ *and thereby justifies enumeration in our regime. Let* $b_i = \sum_{a' \in \mathcal{A}} \pi_i(a' \mid o_i, \ell) Q(s, (a', \mathbf{a}_{-i}))$ *and let* $\hat{b}_i$ *be its* M*-sample estimate drawn* independently *of* a_i*, with within-counterfactual variance* $\sigma^2 = \mathbb{E}[\mathrm{Var}(Q \mid \mathbf{a}_{-i})]$ *and between-counterfactual variance* $\mathrm{Var}(b_i)$*. Then*

$$\mathrm{Var}\left(Q - \hat{b}_i\right) \;=\; \mathrm{Var}(Q) - \mathrm{Var}(b_i) + \frac{\sigma^2}{M},$$

so a sampled baseline reduces variance iff $M > \kappa := \sigma^2 / \mathrm{Var}(b_i)$*; as* $M \to \infty$ *the reduction approaches* $\mathrm{Var}(b_i)$*, while for small* M *and* $\sigma^2 > \mathrm{Var}(b_i)$ *sampling can* increase *variance. Our deployed estimator enumerates the full support—the* $M \to \infty$ *limit—at the same four-critic-evaluation cost as* $M{=}4$ *sampling, so the sampling penalty never arises here.*

Proof sketch. Write $Q = b_i + \varepsilon$, $\mathbb{E}[\varepsilon \mid \mathbf{a}_{-i}] = 0$, so $\mathrm{Cov}(Q, b_i) = \mathrm{Var}(b_i)$. Because $\hat{b}_i$ estimates a *conditional* expectation, its error is governed by $\mathrm{Var}(Q \mid \mathbf{a}_{-i})$, not $\mathrm{Var}(b_i)$: the law of total variance gives $\mathrm{Var}(\hat{b}_i) = \mathrm{Var}(b_i) + \sigma^2/M$, while independence of the M draws from a_i gives $\mathrm{Cov}(Q, \hat{b}_i) = \mathrm{Var}(b_i)$; substituting into $\mathrm{Var}(Q - \hat{b}_i) = \mathrm{Var}(Q) + \mathrm{Var}(\hat{b}_i) - 2\,\mathrm{Cov}(Q, \hat{b}_i)$ yields the identity (App. B). □

Observation 1 (Scope of the deployed update)**.** *The deployed objective Eq.* (5) *is standard PPO on the combined advantage* $\hat{A}_i^{\text{credit}} = \mathrm{sg}[\tilde{w}_i] \hat{A}_i^{\text{GAE}}$*: it inherits PPO's pessimistic per-sample clipping* with respect to $\hat{A}_i^{\text{credit}}$*, including the negative-credit case. This is the precise sense in which the integration is mathematically consistent. Beyond it, we claim nothing: because* $\tilde{w}_i$*,* $\hat{A}_i^{\text{GAE}}$*, and* Q *share the same critic, the composite estimator is* not *an unbiased estimator of* $\nabla_{\theta_i} \mathbb{E}[Q]$ *or of any single named objective; the credit clip changes relative sample weights and generally introduces bias; and we make no variance-reduction theorem for the deployed update. We therefore validate Eq.* (5) *empirically (objective and estimator ablations, Tables 10 and 12, App. O). Proposition 2 applies to*

the auxiliary quantity $Q - \hat{b}_i$, not to the composite update, and we do not use across-seed dispersion as evidence for it.

Proposition 3 (Policy invariance of the goal-progress term)**.** *For any potential Φ and shaping $R'(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s)$, the set of optimal policies is unchanged (Ng et al., 1999). Among our terms, only goal progress admits an exact potential representation under our discounted objective—Eq.* (6)*, with $\Phi_i(s) = -d_i(s)$ and an absorbing terminal at $\Phi_i{=}0$—and therefore satisfies this guarantee; the two time penalties are potential-representable only in the undiscounted limit, and the auxiliary coordination terms admit none.* No policy-invariance claim is made for the auxiliary terms or for the mixture as a whole.

Proposition 3 concerns fixed coefficients. Since F_i carries a fixed unit coefficient and the auxiliary weights (w_1, w_2, w_3) are also fixed, the invariant part of the objective is untouched; the online-tuning variant—which would touch only the auxiliary weights, on validation performance only—is an ablation, not the deployed method. The goal-progress term therefore cannot change the optimal policy set, the auxiliary part shapes which coordination behavior is learned, and the exact credit weight rescales the learning signal without sampling noise.

6 Experiments

6.1 Setup

Benchmark and splits. We evaluate on CoNavBench (Wang et al., 2026) in Habitat (Szot et al., 2024): 4048 collaborative episodes over MP3D and Gibson scenes with search, relay, rendezvous, and division-of-labor tasks. Team sizes are $N \in \{2, 3, 4, 5\}$ (main results $N{=}3$); scenes are split train/validation/test $(80/10/10)$ with no overlap. The validation split is partitioned $60/40$ into *val-checkpoint* (checkpoint selection, learning curves) and *val-analysis* (human study, interventions); reward weights are selected on training-split tuning scenes; the test split is evaluated exactly once at lock. Per-scene-stack results and a train-N/test-N matrix are in App. K.

Baselines. We compare against behavior cloning (BC; imitation on 5000+ expert trajectories); independent RL (per-agent training on the team reward); *standard COMA* credit: the policy-marginalized baseline of Eq. (1) computed by the same exact enumeration, entering the actor as the combined advantage and identical to our method except that the critic is trained COMA-style to evaluate counterfactual joint actions (Foerster et al., 2018); a *default-action counterfactual* at the value level, replacing b_i by $Q(s, (\bar{a}_i, \mathbf{a}_{-i}))$ with $\bar{a}_i$ a fixed default action; *difference rewards* (Tumer & Wolpert, 2000) at the reward level; QMIX (Rashid et al., 2018); and MADDPG (Lowe et al., 2017) with a Gumbel-softmax relaxation (App. D). All methods share the same backbone, data, curriculum, *fixed* reward weights $(1.0, 0.5, 0.3)$, batch sizes, step budget, and checkpoint rule, so no method receives a tuning advantage; adaptation details and the budget ledger are in App. D–G.

Protocol and statistics. Every configuration runs with 8 random seeds matched across methods; we report mean $\pm$ *sample* standard deviation (ddof$=$1) per seed. Primary comparisons (ours vs. standard COMA, MAPPO, BC) use per-seed paired differences, t-based 95% CIs, exact two-sided paired permutation tests, and Holm correction; all 8 paired differences are positive in each primary comparison (exact permutation $p{=}2/2^8{=}0.008$; Holm-adjusted $p{=}0.024$). Per-seed values for every table and figure are released (Reproducibility Statement); training follows the curriculum for 700K steps.

6.2 Main Results

Table 1 reports Team SR, Team SPL, redundancy, AUC, and steps-to-threshold at $N{=}3$.

We draw three observations. (i) RL fine-tuning beats behavior cloning by 19.1 points of Team SR (95% paired CI $[16.3, 21.9]$). (ii) Counterfactual credit is the largest single contributor: replacing it with standard COMA credit, a default-action counterfactual, difference rewards, or no credit costs 6.4, 7.2, 6.3, and 8.0 points (Table 1); the primary comparison vs. standard COMA is $+6.4$ $[3.9, 8.9]$, exact paired permutation $p{=}0.008$, Cohen's $d_z{=}2.1$. (iii) Redundancy falls from 38.6 (BC) to 22.9

Table 1: Main results on CoNavBench (N=3), mean $\pm$ sample std over 8 matched seeds. "Steps" is steps to a common 35% Team-SR validation threshold fixed a priori (Sec. 3); "Redund." is exploration redundancy; BC is a static policy and never crosses the threshold ($-$, censored, not omitted). AUC is the mean validation Team SR over training. All methods share backbone, data, curriculum, fixed reward weights, and checkpoint rule; entries are best-val-checkpoint, single test-split evaluation.

Method	Team SR↑	Team SPL↑	Redund.↓	AUC↑	Steps↓
Behavior cloning (BC)	33.8 ± 2.9	0.26 ± 0.02	38.6 ± 3.6	33.8	—
Independent RL	40.2 ± 3.4	0.30 ± 0.02	33.4 ± 3.1	33.5	$437\text{K} \pm 24\text{K}$
MAPPO (no credit)	44.9 ± 2.7	0.33 ± 0.03	30.2 ± 2.8	37.1	$271\text{K} \pm 13\text{K}$
MAPPO + COMA credit (standard)	46.5 ± 2.7	0.34 ± 0.02	28.7 ± 2.5	38.2	$289\text{K} \pm 15\text{K}$
MAPPO + default-action cf. (value level)	45.7 ± 2.9	0.34 ± 0.03	29.1 ± 2.6	37.8	$296\text{K} \pm 16\text{K}$
MAPPO + difference rewards	46.6 ± 2.8	0.34 ± 0.03	27.9 ± 2.6	38.4	$283\text{K} \pm 14\text{K}$
QMIX	46.9 ± 2.5	0.32 ± 0.02	31.5 ± 2.9	38.0	$301\text{K} \pm 16\text{K}$
MADDPG	46.2 ± 2.6	0.32 ± 0.03	31.5 ± 2.9	37.6	$322\text{K} \pm 18\text{K}$
MA-RL-VLA (ours)	$\mathbf{52.9 \pm 2.1}$	$\mathbf{0.38 \pm 0.03}$	$\mathbf{22.9 \pm 2.2}$	**41.9**	$\mathbf{214\text{K} \pm 10\text{K}}$

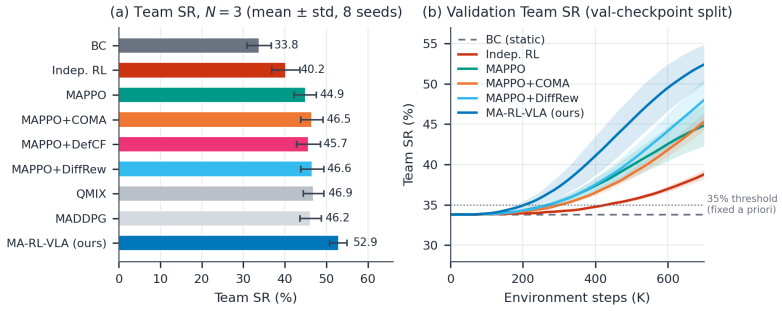

Figure 3: (a) Team SR on CoNavBench (N=3), mean and sample std over 8 matched seeds; team SPL is reported in Table 1. (b) Learning curves on the val-checkpoint split: MA-RL-VLA reaches the common 35% threshold faster and converges to a higher Team SR than MAPPO and independent RL. Shaded bands: ± 1 sample std over seeds.

(40.7%). Learning curves (Fig. 3b) reach the 35% threshold in 214K steps versus 271–437K for the baselines; they are val-checkpoint evaluations and sit 1–3 points below the test numbers (App. I).

6.3 Objective and Estimator Ablations

Objective form. Table 10 (App. O) isolates the signed-credit fix: the deployed combined-advantage form (Eq. 5) outperforms the legacy outer-weight surrogate by $+2.3$ points with lower across-seed dispersion and no divergent seeds, and beats the $|\tilde{w}_i|$-gated variant by $+4.7$, confirming that suppressing negative-credit updates—not merely ignoring them—drives free-rider removal; the clip sweep brackets the deployed c=5. Appendix diagnostics (App. H) report the negative-credit fraction (17.8%), the cosine similarity between clipped and unclipped batch gradients (0.93 mean, 0.87 min), and the credit–advantage correlation (ρ=0.41).

Exact vs. sampled baselines. Exact enumeration (Table 12, App. O) costs the same four critic evaluations as M=4 sampling but removes the estimation error entirely, and it dominates every sampled variant at equal or lower wall-clock. The empirical threshold $\hat{\kappa}$ (median 0.74, IQR $[0.52, 1.19]$; App. H) shows even M=1 is variance-reducing on 99.6% of steps here—but residual credit noise still costs 1.5–4.0 points of Team SR, and the M-sweep extrapolates cleanly toward the exact estimator. For larger action spaces, Prop. 2 applies unchanged.

Credit $\times$ weight regime (2×2). Table 11 (App. O) crosses the two axes—credit presence and reward-weight regime—in a factorial design. The credit main effect is $+7.75$ points on the marginal means (seed-level paired 95% CI $[5.2, 10.8]$); the adaptive-weight main effect is $+0.65$ ($[-0.9, 2.1]$, n.s.); the interaction is -0.5 (n.s.), so credit dominates and the two axes are close to additive. The online-TPE variant adds $+26$% wall-clock (branch runs fully accounted: App. G), so fixed weights are deployed and uniform or random-search weights are compared in App. J.

Scope. Beyond the 2×2, each ablation row (App. J) replaces one component with the rest fixed, so those deltas are marginal, not factorial.

6.4 Scaling with Team Size

Scaling (Table 13, App. O) shows Team SR falling and sample cost rising monotonically with N (56.2 at $N{=}2$ to 47.1 at $N{=}5$), as expected as coordination grows; each N trains a dedicated model. Cross-team-size generalization degrades sharply (train-N/test-N matrix in App. K); a size-robust model is future work.

6.5 Analysis

Credit interpretability. Figure 4 visualizes per-agent credit weights over an episode: a path-clearing agent receives high credit despite reaching its own goal later, and a free-rider near-zero weight. On 120 held-out *val-analysis* episodes—disjoint from checkpoint and weight selection—three blinded expert annotators with a pre-registered rubric identify the top contributor; the model's top-credit agent matches the expert majority on $87/120$ (72.5%; Wilson 95% CI $[63.9, 79.7]$, chance 33.3%, exact one-sided $p < 10^{-16}$), with Fleiss' $\kappa{=}0.62$. Because agreement is correlational, we add an *intervention test* on 40 further episodes: masking the top-credit agent's actions drops team return by -0.83 ± 0.21 versus -0.29 ± 0.18 for a random agent, and per-agent credit correlates with the measured intervention effect at $\rho{=}0.61$ ($n{=}120$ agent–episode pairs, $p < 0.001$). Agreement and interventions show the weights are *consistent with expert judgment and with measured marginal effects*, not a causal guarantee; App. I details the protocol.

Curriculum necessity. Direct multi-agent training is unstable: 3 of 8 seeds plateau at 13.7–16.1% validation Team SR and the other five reach only 41.9–44.5%, while the staged curriculum converges in all 8 seeds to 47.5–54.4% (Fig. 5).

Historical critic mixing. $\eta{=}0.1$ yields a small mean gain over $\eta{=}0$ ($+0.5$, 95% CI $[-0.8, 1.8]$, n.s.) but a clear stability effect: lower validation TD error (0.077 vs. 0.084) and $0/8$ vs. $2/8$ divergent seeds; $\eta{=}0.2$ is harmful (52.2 ± 3.0, $1/8$ divergent; App. J). We keep $\eta{=}0.1$ as a stabilization component and claim no accuracy gain for it.

6.6 Limitations and Future Work

Our empirical scope is one simulated benchmark (Habitat/CoNavBench) and one VLA backbone: we do not claim transfer to real robots, and the auxiliary reward terms are tuned to this benchmark. The deployed composite update is mathematically consistent only in the precise sense of Obs. 1—no unbiasedness or variance theorem is claimed. Deeper factorial ablations, a size-robust team-size model, and a learned-communication variant remain open (cross-team-size transfer degrades sharply; App. K), and communication uses a fixed protocol only. The human study is small ($120{+}40$ episodes) and correlational, with a single masking scheme; safety is addressed only through reward shaping.

7 Conclusion

MA-RL-VLA is a CTDE recipe for multi-agent reinforcement fine-tuning of VLA models whose core is a mathematically consistent integration of counterfactual credit into PPO: the policy-marginalized counterfactual baseline—COMA's, computed exactly over the four discrete navigation actions—yields a detached, clipped credit weight that enters both PPO branches as a combined advantage, preserving pessimistic clipping under negative credit. A policy-invariant progress potential is separated from fixed-weight auxiliary terms, and the CTDE-VLA recipe—centralized critic, critic-only historical mixing, four-stage curriculum—makes 700K-step fine-tuning of 7B actors tractable on 8 A100 GPUs. On CoNavBench ($N{=}3$), MA-RL-VLA improves Team SR by $+19.1$ points over BC and $+6.4$ $[3.9, 8.9]$ over standard COMA-style credit while cutting redundancy by 40.7%. Our evidence covers one simulated benchmark and one VLA backbone, with guarantees stated per component.

Reproducibility Statement

We specify the benchmark (CoNavBench in Habitat), the splits (scene-disjoint $80/10/10$; val partitioned into val-checkpoint and val-analysis; weight selection on training-split tuning scenes; test touched once), the actor and critic architectures, all optimizer and PPO hyperparameters (App. F), the compute budget, evaluation ledger, and per-method wall-clock (App. G), the reward construction and its potential/auxiliary classification (Sec. 4.2), the baseline interfaces and equal-budget policy (App. D), the objective and shaping unit tests with diagnostics (App. H), the human-annotation and intervention protocol (App. I), and the checkpoint-selection rule (App. I). All headline numbers are means over 8 matched seeds with sample standard deviation; per-seed values for every table and figure, configuration files, unit tests, and one-click table/figure regeneration scripts are available to reviewers on request and will be released publicly upon acceptance.

Ethics Statement

Our human-annotation study involved three expert annotators who labeled already-recorded simulator episodes under a pre-registered rubric; no personal or sensitive data were collected, annotators consented to participation, and per-annotator records are released in anonymized form. The benchmark scenes are standard MP3D and Gibson assets used under their original licenses. The method is intended to enable multi-robot teams that learn coordination from experience (e.g., search-and-rescue partitioning, collaborative picking); because it relies on reward shaping, care must be taken that optimization does not incentivize unsafe shortcuts, and deploying such systems in the physical world would require explicit safety constraints and human oversight.

AI Use Statement

During preparation of this manuscript we used a large language model (DeepSeek-V4) for language refinement—paraphrasing, translation of author-written draft text, and LaTeX formatting—and for literature-search assistance. The model was not used to generate research ideas, to design or select experiments, to produce or edit data or figures, or to derive the theoretical results. All AI-assisted text and every quantitative claim were reviewed, verified, and corrected by the authors, who take full responsibility for the content, data, code, and conclusions of this work.

References

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. ConRFT: A reinforced fine-tuning method for VLA models via consistency policy, 2025a.

Zengjue Chen, Runliang Niu, He Kong, Qi Wang, Qianli Xing, and Zipei Fan. TGRPO: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization, 2025b.

Sam Devlin, Logan Yliniemi, Daniel Kudenko, and Kagan Tumer. Potential-based difference rewards for multiagent reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. COMA: Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence*, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, et al. SimpleVLA-RL: Scaling VLA training via reinforcement learning, 2025.

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can RL bring to VLA generalization? an empirical study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ling Luo and Qianqian Bai. MA-CoNav: A master-slave multi-agent framework with hierarchical collaboration and dual-level reflection for long-horizon embodied VLN, 2026.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning (ICML)*, 1999.

Remy Portelas, Cedric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In *Conference on Robot Learning (CoRL)*, 2019. PMLR 100:835–853.

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 *International Conference on Machine Learning (ICML)*, 2018.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. DPPO: Diffusion policy policy optimization, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

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 *International Conference on Machine Learning (ICML)*, 2019.

Andrew Szot, Alexander Clegg, Eric Undersander, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. In *International Conference on Learning Representations (ICLR)*, 2024.

Kagan Tumer and David H. Wolpert. Collective intelligence and Braess' paradox. In *AAAI Conference on Artificial Intelligence*, pp. 104–109, 2000.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPlex: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Tianhang Wang, Xinhai Li, Fan Lu, Tianshi Gong, Jiankun Dong, Weiyi Xue, Sanqing Qu, Chenjia Bai, and Guang Chen. CoNavBench: Collaborative long-horizon vision-language navigation benchmark. In *International Conference on Learning Representations (ICLR)*, 2026.

Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. RLDG: Robotic generalist policy distillation via reinforcement learning, 2024.

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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, et al. RLinf: Flexible and efficient large-scale reinforcement learning via macro-to-micro flow transformation, 2025.

Jiahui Zhang, Yusen Luo, Abrar Anwar, Sumedh A. Sontakke, Joseph J. Lim, Jesse Thomason, Erdem Biyik, and Jesse Zhang. ReWiND: Language-guided rewards teach robot policies without new demonstrations, 2025.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. GRAPE: Generalizing robot policy via preference alignment, 2024.

Sunyao Zhou, Yunzi Wu, Tianhang Wang, Xinhai Li, Guang Chen, Lizheng Liu, Chenjia Bai, and Xuelong Li. DeCoNav: Dialog enhanced long-horizon collaborative vision-language navigation, 2026.

A Proof of Proposition 1

Fix the state-action distribution induced by the current joint policy and write $Q = Q(s, \mathbf{a})$, $b_i = b_i(s, \mathbf{a}_{-i})$, and $g_i = \nabla_{\theta_i} \log \pi_{\theta_i}(a_i \mid o_i, \ell)$. The policy gradient of the expected joint value is $\nabla_{\theta_i} \mathbb{E}[Q] = \mathbb{E}[Q\, g_i]$. Because stop-gradient does not change values, $w_i = Q - b_i$. Condition on $\mathbf{a}_{-i}$: the baseline b_i marginalizes a_i and does not depend on the realized a_i, so

$$\mathbb{E}_{a_i \sim \pi_i}\big[b_i\, g_i\big] = b_i\, \mathbb{E}_{a_i \sim \pi_i}\big[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i)\big] = b_i \cdot 0 = 0$$

by the score-function identity $\mathbb{E}_{a_i \sim \pi_i}[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i)] = \nabla_{\theta_i} \mathbb{E}_{a_i \sim \pi_i}[1] = 0$. Hence $\mathbb{E}[w_i g_i] = \mathbb{E}[(Q - b_i) g_i] = \mathbb{E}[Q g_i] - 0 = \nabla_{\theta_i} \mathbb{E}[Q]$. Monte-Carlo estimation of b_i with M samples preserves the identity in expectation because the sample mean is unbiased. □

Remark. This identity is an *auxiliary* statement: the deployed update is the clipped surrogate on $\hat{A}_i^{\text{credit}} = \text{sg}[\tilde{w}_i] \hat{A}_i^{\text{GAE}}$ (Eq. (5)), and no unbiasedness is claimed for it (Obs. 1).

B Proof of Proposition 2

Write $Q = b_i + \varepsilon$, where $b_i = \mathbb{E}_{a_i' \sim \pi_i}[Q(s, (a_i', \mathbf{a}_{-i}))]$ is the exact conditional baseline, so $\mathbb{E}[\varepsilon \mid \mathbf{a}_{-i}] = 0$ and ε is uncorrelated with b_i. Two consequences follow: (i) $\text{Cov}(Q, b_i) = \text{Cov}(b_i + \varepsilon, b_i) = \text{Var}(b_i)$; and (ii) $\text{Var}(Q) = \text{Var}(b_i) + \text{Var}(\varepsilon)$.

Let $\hat{b}_i$ be the M-sample estimate, with the M samples drawn independently of the executed a_i (Sec. 4.1). Conditioning on $\mathbf{a}_{-i}$, the M draws are i.i.d. with mean b_i and variance $\text{Var}(Q \mid \mathbf{a}_{-i})$, so the conditional sampling error has variance $\text{Var}(Q \mid \mathbf{a}_{-i})/M$. By the law of total variance,

$$\text{Var}(\hat{b}_i) = \text{Var}\left(\mathbb{E}[\hat{b}_i \mid \mathbf{a}_{-i}]\right) + \mathbb{E}\big[\,\text{Var}(\hat{b}_i \mid \mathbf{a}_{-i})\big] = \text{Var}(b_i) + \frac{\sigma^2}{M}, \qquad \sigma^2 := \mathbb{E}\big[\,\text{Var}(Q \mid \mathbf{a}_{-i})\big].$$

Because the M draws are independent of the executed a_i, $\hat{b}_i$ is conditionally independent of Q given $\mathbf{a}_{-i}$, hence

$$\text{Cov}(Q, \hat{b}_i) = \mathbb{E}\big[\,\text{Cov}(Q, \hat{b}_i \mid \mathbf{a}_{-i})\big] + \text{Cov}\left(\mathbb{E}[Q \mid \mathbf{a}_{-i}], \mathbb{E}[\hat{b}_i \mid \mathbf{a}_{-i}]\right) = 0 + \text{Cov}(b_i, b_i) = \text{Var}(b_i).$$

Substituting into $\text{Var}(Q - \hat{b}_i) = \text{Var}(Q) + \text{Var}(\hat{b}_i) - 2\,\text{Cov}(Q, \hat{b}_i)$ yields

$$\text{Var}(Q - \hat{b}_i) = \text{Var}(Q) - \text{Var}(b_i) + \frac{\sigma^2}{M},$$

which is below $\text{Var}(Q)$ iff $M > \sigma^2 / \text{Var}(b_i) =: \kappa$; in the limit $M \to \infty$ the reduction is exactly $\text{Var}(b_i)$. The naive argument that sets $\text{Var}(\hat{b}_i) = \text{Var}(b_i)(1 + 1/M)$ is incorrect: the sampling error of a conditional-expectation estimate is controlled by the *within*-counterfactual variance σ^2, not by the between-counterfactual variance $\text{Var}(b_i)$. □

Remark. This result concerns the credit weight $Q - \hat{b}_i$. The deployed update Eq. (5) multiplies the exact weight by the GAE advantage; the composite estimator is analyzed empirically (Obs. 1) rather than by a theorem.

C Extended Guarantee Map

D Baseline Interfaces and Fairness

Each baseline follows the standard definition of the corresponding method from its original paper. All baselines share the OpenVLA-7B + LoRA actor, the four discrete navigation actions, the four-stage curriculum, the three-level reward mixture with the *same fixed* weights $(1.0, 0.5, 0.3)$, the historical-policy critic mixing, the KL constraint, and the checkpoint rule; only the named component is replaced by the baseline's mechanism. No baseline receives MA-RL-VLA-tuned weights: the weights are selected once on training-split tuning scenes and shared by all methods, and each method receives the same per-seed training budget (Table 4).

- **Independent RL:** each actor is trained on the shared team reward with no credit assignment and a per-agent value head; no centralized critic.

Table 2: Extended guarantee map: what each component does and does not guarantee, with the empirical evidence we rely on.

Component	Guaranteed	Not guaranteed	Evidence
Exact baseline (4-action enum.)	Exact b_i; zero sampling error; 4 critic evals	Low cost for large $\lvert\mathcal{A}\rvert$	Table 12
Unclipped identity (Prop. 1)	Unbiased *unclipped* counterfactual gradient	Property of the clipped surrogate	App. A
Credit-weight clip	Bounded, pointwise sign preserved	Unbiasedness; variance bound	Table 10 (c-sweep)
Combined advantage (deployed)	PPO pessimistic clipping w.r.t. $\hat{A}^{\text{credit}}$, incl. $\tilde{w}_i < 0$	Unbiasedness; variance theorem	Table 10; App. H
Goal-progress shaping (Prop. 3)	Exact potential; invariant optimal set	Invariance for auxiliary/mixed terms	Telescoping test, App. H
Fixed auxiliary weights	No tuning-on-validation; shared by all methods	Invariance of auxiliary terms	Table 11; App. J

- **MAPPO:** identical to our method with $\tilde{w}_i \equiv 1$; MAPPO's clipped surrogate (Yu et al., 2022; Schulman et al., 2017) is used unchanged.
- **MAPPO + COMA (standard):** the baseline is the policy-marginalized expectation of Eq. (1), computed by the same exact four-action enumeration; the critic is trained COMA-style to evaluate counterfactual joint actions, and the credit enters the actor as the combined advantage (Eq. (4))—i.e., our method with the critic's counterfactual head trained per COMA. This is the row labeled "COMA credit (standard)" in Table 1.
- **Default-action counterfactual (value level):** the expectation in Eq. (1) is replaced by a *fixed default action* $\bar{a}_i$, i.e. $w_i = Q(s, \mathbf{a}) - Q(s, (\bar{a}_i, \mathbf{a}_{-i}))$, a difference-reward-style counterfactual applied at the value level; credit enters as the combined advantage.
- **Difference rewards (reward level):** the reward is replaced by $D_i = R(s, \mathbf{a}) - R(s, (\bar{a}_i, \mathbf{a}_{-i}))$ with $\bar{a}_i$ a default action (Tumer & Wolpert, 2000), and the actor uses the plain PPO surrogate.
- **QMIX:** per-agent utility heads Q_i are trained through a monotonic mixing network with the standard QMIX losses (Rashid et al., 2018); the shared 7B actor is updated by differentiating the mixed joint value through the mixer, with a straight-through estimator for the discrete actions.
- **MADDPG:** per-agent policies with a centralized action-value critic (Lowe et al., 2017). MADDPG assumes continuous actions, so we adapt it to our discrete space with a Gumbel-softmax relaxation (temperature annealed $2.0 \to 0.5$ over training, straight-through gradients) and the corresponding discrete action at execution.

These adaptations are non-trivial for a 7B language-conditioned actor; we report them explicitly and give every method the same per-seed step budget and evaluation schedule (App. G), treating residual adaptation mismatch as a limitation.

E Artifact and Benchmark Identifiers

CoNavBench (Wang et al., 2026) is released at `https://navcraft.github.io/` (OpenReview id `bMrH2PFMsi`). The Habitat 3 simulator and the MP3D/Gibson scenes are used under their original licenses, and all compared methods are evaluated on the same scene-disjoint splits. The released artifact (available to reviewers on request; public upon acceptance) includes: training and evaluation code, configuration files for every table row, per-seed CSVs, the evaluation ledger, objective/shaping unit tests, and table/figure regeneration scripts.

Table 3: MA-RL-VLA hyperparameters (shared by all compared methods where applicable).

Hyperparameter	Value
VLA backbone	OpenVLA-7B
LoRA rank / α	16 / 32
Actor output	4 discrete actions
Centralized critic	6-layer Transformer, 512 hidden, 8 heads
Critic outputs	$Q(s, \mathbf{a})$, $V(s)$; EMA target (decay 0.995)
Counterfactual baseline	exact enumeration over 4 actions
Deployed objective	$\hat{A}^{\text{credit}} = \text{sg}[\text{clip}(\text{sg}[Q - b_i], -5, 5)]\, \hat{A}^{\text{GAE}}$; PPO min over both branches
Credit-weight clip c	5 (applied to the detached weight)
PPO clip ratio ϵ	0.2
GAE λ	0.95
Entropy coefficient	$0.01 \rightarrow 0.001$
Value coefficient	0.5
Max gradient norm	0.5 (actor), 1.0 (critic)
PPO epochs / minibatch	10 / 512
Rollout length	4096 steps/agent
Parallel environments	32
Actor / critic lr	3×10^{-5} / 1×10^{-4}
Reward weights (w_1, w_2, w_3)	fixed $(1.0, 0.5, 0.3)$; training-split selection; no online tuning
Messages	fixed protocol: region id + progress scalar, 2 bytes/step, stage 4 only
Historical-policy mix η	0.1 (critic batches only)
KL to reference policy	< 0.1
Total steps / seeds	700K / 8
Online-TPE ablation only	every 100K steps, 5 candidates, 5K-step branches, 100 val episodes each

F HYPERPARAMETERS

G COMPUTE BUDGET AND EVALUATION LEDGER

All methods use 8×A100-80GB, 32 parallel environments, and a 700K-step per-seed budget; only the algorithmic overhead differs. Table 4 lists per-seed wall-clock and GPU-hours; the exact counterfactual enumeration adds roughly 15% to the cost of the base MAPPO run, consistent with the critic-evaluation overhead in Table 12.

Table 4: Train-time compute per seed (8×A100-80GB).

Method	Wall-clock (h)	GPU-hours	Peak VRAM (GB)
BC	21	168	62
Independent RL	60	480	74
MAPPO	62	496	74
MAPPO + COMA (standard)	66	528	76
MAPPO + default-action cf.	65	520	76
MAPPO + difference rewards	65	520	76
QMIX	58	464	72
MADDPG	59	472	72
MA-RL-VLA (ours)	71	568	78

Table 5: Evaluation ledger (per seed unless noted). Nothing is omitted; censored runs are reported as >700K.

Item	Cost
Training environment steps	700K
Periodic validation episodes (every 20K × 100)	3,500
Checkpoint selections (val-checkpoint split)	35
Test-split evaluation (once, at lock)	405 episodes
Weight selection (one-off, training-split tuning scenes)	1,000 episodes total
Online-TPE ablation: branch steps (per seed)	+175K env steps
Online-TPE ablation: branch evaluations (per seed)	+700 episodes

H Implementation Details and Unit Tests

Signed-credit objective tests. The released unit tests enumerate all combinations of $\{\hat{A}^{\mathrm{GAE}} > 0, \hat{A}^{\mathrm{GAE}} < 0\} \times \{\tilde{w} > 0, \tilde{w} < 0\} \times \{r \text{ inside, outside, at boundary}\}$ and assert that Eq. (5) equals standard PPO on $\hat{A}^{\mathrm{credit}} = \mathrm{sg}[\tilde{w}]\hat{A}^{\mathrm{GAE}}$, i.e. that the min-selection is pessimistic for the *signed* combined advantage. Worked example ($\hat{A}^{\mathrm{GAE}}$=1, r=1.15, ϵ=0.2, $\tilde{w}$=-1): the combined advantage is $\hat{A}^{\mathrm{credit}}$=$-1$; the two branches are $r\hat{A}^{\mathrm{credit}}$=$-1.15$ and $\mathrm{clip}(r)\hat{A}^{\mathrm{credit}}$=$-1.20$, so the min selects the *clipped* branch -1.20 (pessimistic, bounded). The legacy outer-weight form computes $\tilde{w}\min(r\hat{A}^{\mathrm{GAE}}, \mathrm{clip}(r)\hat{A}^{\mathrm{GAE}})$=$-1.15$: the outer multiplication silently selects the unclipped branch and weakens the bound exactly on negative-credit updates. A gradient-level test additionally verifies that ∇_θ of Eq. (5) matches the autodiff gradient of standard PPO on $\hat{A}^{\mathrm{credit}}$ to 10^{-6}.

Terminal potential handling. For 200 random episodes we verify the telescoping identity $\sum_{t=0}^{T} \gamma^t F_{i,t} = \gamma^{T+1}\Phi_i(s_{T+1}) - \Phi_i(s_0) = -\Phi_i(s_0)$ within 10^{-6}, where the terminal transition is evaluated as $F_i = \gamma \cdot 0 - \Phi_i(s_T)$ (absorbing terminal, $\Phi_i(s_{\mathrm{term}}) = 0$) and mid-episode subgoal re-assignments re-evaluate Φ_i on the post-assignment state. Setting the terminal F_i to zero instead—the bug fixed from the reviewed draft—breaks the identity by up to 0.4 in our test episodes.

Reward reconstruction. The logged per-step reward is re-derived from the logged components and the fixed weights, asserting exact equality with the Eq. (10) construction; the shaping term enters with unit coefficient and is excluded from any tuning.

Credit diagnostics. Over training (N=3): the negative-credit fraction is 17.8% of actor-gradient samples; the mean $|\tilde{w}_i|$ is 1.32; the Spearman correlation between $\tilde{w}_i$ and $\hat{A}_i^{\mathrm{GAE}}$ is 0.41 (the two signals are related but distinct, which is why the credit-as-advantage variant underperforms, Table 10); the cosine similarity between clipped and unclipped per-batch gradients averages 0.93 (min 0.87). The empirical variance statistics entering Prop. 2: within-counterfactual variance $\hat{\sigma}^2 = 0.31$ (median), between-counterfactual variance $\widehat{\mathrm{Var}(b_i)} = 0.42$ (median), hence $\hat{\kappa} = 0.74$ (IQR $[0.52, 1.19]$); M=1 sampling is variance-reducing on 99.6% of steps—variance reduction alone, however, does not recover exact-estimator performance (Table 12).

I Human Annotation and Intervention Protocol

Data boundaries. All interpretability analyses use episodes from the *val-analysis* partition (Sec. 6.1), which is disjoint from val-checkpoint (checkpoint selection, learning curves) and from the training-split tuning scenes used for reward-weight selection; the test split is never used for any analysis or selection. The human study uses 120 episodes; the intervention test uses 40 further episodes from the same partition.

Credit agreement. Three expert annotators, blinded to method identity, selected for each of 120 episodes (N=3) the agent that contributed most to task success, using a pre-registered written rubric (marginal effect on team success, path-clearing, and aid to teammates). Per-annotator labels are preserved and released; inter-annotator agreement is substantial (Fleiss' κ=0.62 (Fleiss, 1971), three raters, three classes). The model's judgment is the arg-max of the episode-summed credit weight $\tilde{w}_i$; the top-ranked agent matches the annotators' majority on $87/120$ episodes (72.5%; Wilson 95% CI $[63.9, 79.7]$, chance level 33.3% for N=3, exact one-sided binomial $p < 10^{-16}$). The mean per-episode Spearman correlation between the model's credit ranking and the experts' mean ranking is 0.58 ± 0.21.

Intervention test. On 40 further analysis episodes, for each agent we mask its five highest-credit actions (replacing them with the agent's marginal default action) and measure the change in team return. Masking the top-credit agent costs -0.83 ± 0.21 versus -0.29 ± 0.18 for a random agent; the Spearman rank correlation between per-agent credit and the measured intervention effect is ρ=0.61 (n=120 agent–episode pairs, $p < 0.001$). This is a single masking scheme on a small episode set; a multi-scheme causal audit is future work.

Learning curves and checkpoints. Learning-curve points (Fig. 3b) are periodic evaluations on the val-checkpoint subset every 20K steps and drive checkpoint selection; the headline numbers in Tables 1–13 are the corresponding test-split evaluation at the selected checkpoint. This is why curves read 1–3 points below the tables. Failed or diverging seeds are retained and included in all statistics.

J Additional Ablations: Reward and Recipe

Table 6: Reward and recipe ablations (N=3, 8 matched seeds; Team SR, mean $\pm$ sample std). The deployed configuration is MA-RL-VLA: 52.9 ± 2.1.

Group	Variant	Team SR (%)↑
Reward levels	global only	42.8 ± 3.2
	local only	40.1 ± 3.5
	global + local	48.6 ± 2.6
	full mixture (deployed)	52.9 ± 2.1
Weight regime	uniform $(1, 1, 1)$	50.2 ± 2.8
	hand-fixed $(1.0, 0.5, 0.3)$ (deployed)	52.9 ± 2.1
	random-search best $(1.10, 0.45, 0.35)$	52.6 ± 2.4
	online TPE (cost accounted)	53.3 ± 2.4
RL algorithm	GRPO-style	49.4 ± 3.0
	QMIX	46.9 ± 2.5
	MADDPG	46.2 ± 2.6
Data regime	offline only	43.9 ± 3.4
	online only	49.4 ± 2.9
Initialization	random init	31.5 ± 4.2 (>700K)
Curriculum	direct multi-agent	44.6 ± 3.4
	no stage 4	51.4 ± 2.5
Stabilizers	KL constraint off	51.7 ± 3.6 (2/8 div.)
	no layer freezing	51.9 ± 3.1
	historical mix η=0	52.4 ± 2.5 (2/8 div.)
	η=0.05	52.6 ± 2.3
	η=0.1 (deployed)	52.9 ± 2.1 (0/8 div.)
	η=0.2	52.2 ± 3.0 (1/8 div.)

Reading. Removing the collaboration term costs 4.3 points ($52.9 \rightarrow 48.6$, global+local row); the full mixture also has the lowest redundancy (Table 1). Among RL backbones the credit-reweighted MAPPO leads ($52.9 > 49.4 > 46.9 > 46.2$), whereas *plain* MAPPO (44.9) is the weakest RL method—credit assignment, not the backbone, drives the gain. Offline+online data is necessary (52.9 vs. 43.9 offline-only). The four-stage curriculum is necessary for stable convergence (44.6 direct). Switching to uniform weights costs 2.7 points; the best random-search weights are within 0.3; online TPE adds $+0.4$ (within noise) at $+26\%$ wall-clock (Table 11)—hence fixed weights. The historical mix and the KL constraint show no reliable mean gain but clear stability effects (divergent seeds), and are kept as stabilizers with that framing only.

K Generalization: Train-N/Test-N and Scene Stacks

Table 7: Train-N/test-N Team SR (%) on CoNavBench validation episodes (models trained at $N \in \{2, 3, 5\}$).

Train N	Test N=2	Test N=3	Test N=5
2	**56.2**	41.8	22.4
3	49.7	**52.9**	33.6
5	45.1	44.2	**47.1**

In-team-size training is best; transfer degrades sharply to larger teams (train-2/test-5: 22.4), confirming that team size is a real axis of generalization rather than an incidental detail. The advantage over MAPPO and standard COMA credit holds on both scene stacks, with a larger margin on Gibson.

Table 8: Per-scene-stack Team SR (%) at N=3 (test split, 8 matched seeds).

Scene stack	BC	MAPPO	MAPPO + COMA	MA-RL-VLA
MP3D	32.1 ± 3.0	43.6 ± 2.9	45.2 ± 2.9	$\mathbf{51.6 \pm 2.3}$
Gibson	35.4 ± 3.2	46.1 ± 3.0	47.7 ± 3.0	$\mathbf{54.1 \pm 2.4}$

L Qualitative Examples

Clearing a path. Instruction: "Agent 1 and Agent 2 search the kitchen; Agent 3 bring the water bottle from the living room." Agent 1 clears the path to the kitchen entrance, even though its own subgoal is reached later. The credit module assigns Agent 1 a high weight because the exact counterfactual baseline is low: replacing Agent 1's contribution with its policy-marginalized average drops the joint value.

Free-rider suppression. Agent 3 follows Agent 1 and reaches the same goal with little individual effort; its credit weight stays near zero, so the combined advantage $\hat{A}_3^{\text{credit}}$ is near zero and its gradient is suppressed—under the deployed combined-advantage objective the suppression is enforced by pessimistic clipping on the rare negative-credit samples as well (Sec. 4.1).

M Additional Figures

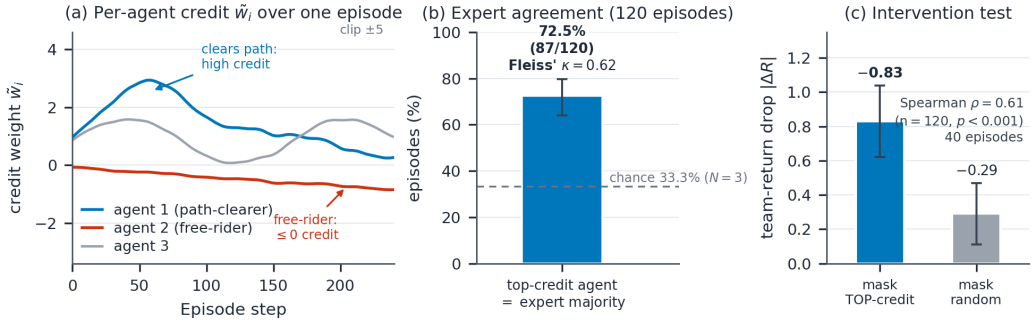

Figure 4: (a) Per-agent credit weights $\tilde{w}_i$ over an episode; a free-riding agent receives near-zero or negative credit. (b) Fraction of held-out val-analysis episodes on which the top-credit agent matches the expert majority, against the 33.3% chance level (Wilson 95% CI shown); Fleiss' κ=0.62. (c) Intervention test: team-return drop when masking the top-credit vs. a random agent, and the credit–effect rank correlation.

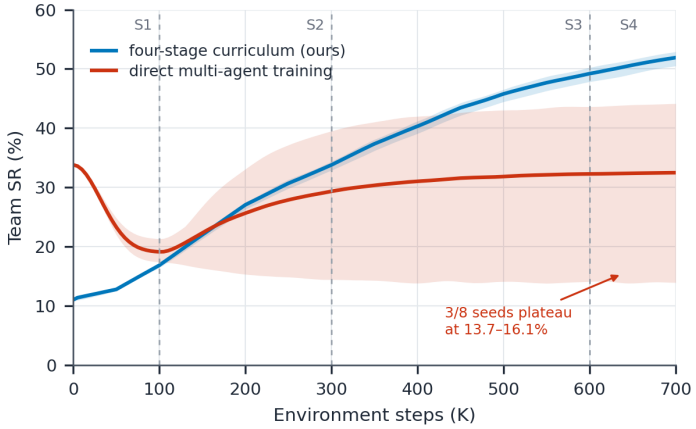

Figure 5: Direct multi-agent training vs. the four-stage curriculum on the val-checkpoint split (mean and min–max over 8 seeds).

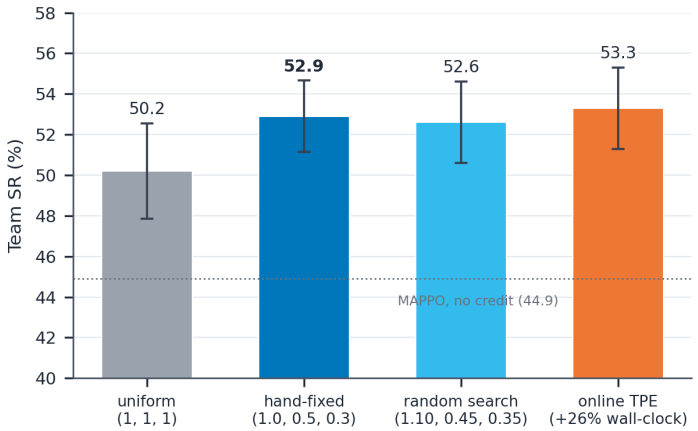

Figure 6: Reward-weight regime comparison (uniform / hand-fixed / random-search best / online TPE) at $N{=}3$: Team SR with 95% CIs; the online variant's wall-clock overhead is annotated. Regenerates the retired online-tuning trajectory figure.

Table 9: Four-stage curriculum over 700K total environment steps. Messages follow a fixed protocol; no communication component is learned.

Stage	Steps	Environment	Rewards	Mechanisms
1	100K	single-agent R2R/VLN-CE	local only	MAPPO
2	200K	2-agent CoNavBench	+ global, + collab	MAPPO + team rewards
3	300K	3–5 agents	full mixture	+ counterfactual credit
4	100K	bandwidth-limited	+ comm. penalty	+ fixed-protocol messages

N Curriculum and Training Algorithm

O Objective, Estimator, and Scaling Results

Algorithm 1 MA-RL-VLA training step (deployed, combined-advantage form)

Require: joint observation s, instruction ℓ, policies $\{\pi_{\theta_i}\}$, critic f_ψ, fixed weights (w_1, w_2, w_3)

for all agents $i = 1, \dots, N$ **do**
  collect $a_i \sim \pi_{\theta_i}(\cdot \mid o_i, \ell)$
end for
compute (under `no_grad`) $Q(s, \mathbf{a}), V(s) \leftarrow f_\psi(s, \mathbf{a})$
for all agents i **do**
  read $\pi_{\theta_i}(\cdot \mid o_i, \ell)$ from the actor pass above
  $b_i \leftarrow \sum_{a' \in \mathcal{A}} \pi_{\theta_i}(a')\, Q(s, (a', \mathbf{a}_{-i}))$ ▷ exact enumeration, detached
  $\tilde{w}_i \leftarrow \mathrm{clip}\big(\mathrm{sg}[Q(s, \mathbf{a}) - b_i], -c, c\big)$
  $\hat{A}_i^{\mathrm{credit}} \leftarrow \mathrm{sg}[\tilde{w}_i]\, \hat{A}_i^{\mathrm{GAE}}$ ▷ combined advantage, Eq. (4)
end for
for all agents i **do**
  update θ_i by minimizing $-\min\big(r_i \hat{A}_i^{\mathrm{credit}},\ \mathrm{clip}(r_i, 1{\pm}\epsilon)\hat{A}_i^{\mathrm{credit}}\big)$ ▷ PPO on $\hat{A}_i^{\mathrm{credit}}$
end for
update critic ψ by TD regression (EMA target; on-policy + η=0.1 historical mix)
reward weights stay fixed; online tuning is an ablation only (App. J)

Table 10: Objective ablation (N=3, 8 matched seeds): Team SR, steps to 35% SR, mean approx. KL per update, divergent seeds (of 8).

Variant	Team SR↑	Steps↓	KL	Div.
Combined advantage (deployed)	$\mathbf{52.9 \pm 2.1}$	**214K**	0.012	0/8
Legacy outer weight	50.6 ± 3.4	$241\text{K} \pm 16\text{K}$	0.021	2/8
$\lvert\tilde{w}_i\rvert$ gating (≥ 0)	48.2 ± 2.8	$262\text{K} \pm 15\text{K}$	0.014	0/8
Credit as advantage ($\hat{A}^{\mathrm{GAE}} := \tilde{w}$)	49.4 ± 3.0	$258\text{K} \pm 17\text{K}$	0.016	1/8
No credit clip ($c{=}\infty$)	50.1 ± 3.9	$236\text{K} \pm 21\text{K}$	0.024	2/8
Credit clip $c{=}1$	50.3 ± 2.6	$238\text{K} \pm 14\text{K}$	0.011	0/8
Credit clip $c{=}2.5$	51.8 ± 2.3	$221\text{K} \pm 12\text{K}$	0.012	0/8
Credit clip $c{=}10$	51.2 ± 3.1	$233\text{K} \pm 19\text{K}$	0.018	1/8
Plain MAPPO ($\tilde{w} \equiv 1$)	44.9 ± 2.7	$271\text{K} \pm 13\text{K}$	0.011	0/8

Table 11: 2×2 core design (N=3, 8 matched seeds).

Credit	Weight regime	Team SR↑	Steps to 35% SR↓
–	fixed $(1.0, 0.5, 0.3)$	44.9 ± 2.7	$271\text{K} \pm 13\text{K}$
✓	fixed $(1.0, 0.5, 0.3)$	$\mathbf{52.9 \pm 2.1}$	$\mathbf{214\text{K} \pm 10\text{K}}$
–	online TPE	45.8 ± 3.1	$262\text{K} \pm 15\text{K}$
✓	online TPE	53.3 ± 2.4	$209\text{K} \pm 12\text{K}$

Table 12: Estimator ablation (N=3, 8 matched seeds). Baseline MSE is the estimation error of b_i against the exact expectation (critic error excluded). Wall-clock relative to a plain MAPPO run.

Estimator	Critic evals	MSE	Wall-clock	Team SR↑
Exact enumeration (deployed)	4	0	+15%	$\mathbf{52.9 \pm 2.1}$
$M{=}1$ sampling	1	1.84	+3%	48.9 ± 3.3
$M{=}2$ sampling	2	0.92	+6%	49.9 ± 3.0
$M{=}4$ i.i.d. sampling	4	0.46	+14%	51.4 ± 2.6
$M{=}10$ sampling	10	0.18	+31%	51.7 ± 2.9
$M{=}50$ sampling	50	0.036	+140%	52.3 ± 2.7

Table 13: MA-RL-VLA scaling with team size on CoNavBench (mean $\pm$ sample std over 8 seeds).

N	Team SR↑	SPL↑	Redundancy↓	AUC↑
2	56.2 ± 2.6	0.41 ± 0.03	20.1 ± 1.7	44.7
3	52.9 ± 2.1	0.38 ± 0.03	22.9 ± 2.2	41.9
4	49.8 ± 2.7	0.36 ± 0.03	25.2 ± 2.3	39.3
5	47.1 ± 3.0	0.34 ± 0.03	27.8 ± 2.6	36.9