CONVENTION GROUNDING: LANGUAGE AS A COORDINATION PRIOR FOR OPEN COORDINATION POINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Connected autonomous vehicles from different manufacturers must coordinate at *open coordination points*—a four-way intersection at simultaneous arrival, or a merge where each stack assumes a different rule—where local information fails to resolve a single order; rules frozen in each proprietary stack leave humans no way to alter which convention governs a mixed fleet. We introduce *convention grounding*: an operator broadcasts a convention as a natural-language sentence, and each receiving vehicle combines that text with its own observation to produce its role-differentiated action—under a reward that never observes the convention. In SUMO, a continuous traffic simulator, we provide, to our knowledge, the first controlled causal demonstration of this mechanism: freezing the scene and flipping only the convention text flips the agent's committed action (flip rate 1.00, zero collisions), while swapping the projected vectors of the two conventions maintains the flip rate but inverts correctness. At a four-way intersection, four independently trained policies—sharing no gradients, policy, training episodes, or projection parameters—coordinate through a common convention signal and their own observations. Generalization to the tested held-out phrasings fails: the role distinction remains unrecovered from the projected representation, and runtime selection is shown only within the grounded convention set. No large generative model is invoked during driving: encoding each broadcast costs one forward pass through a frozen encoder.

1 INTRODUCTION

Coordination among connected autonomous vehicles (CAVs) from different manufacturers is critical to the safety of both CAVs and human-driven vehicles. It must occur during merging, yielding, and intersection negotiation between vehicles developed separately, with no shared training history. This is not a theoretical challenge. On May 31, 2026, a Waymo and a Zoox vehicle entered a motionless standoff at a San Francisco intersection (Torque News, 2026): both were functioning properly and had detected each other, yet neither could commit to a maneuver, and the deadlock ended only when human drivers routed around them. The standoff is not a malfunction but a coordination problem with no built-in answer, and it is one instance of a broader class. An *open coordination point* is one where the rules available to the vehicles do not, taken together, determine a single order of precedence, and it fails in one of two ways. At a *symmetric* coordination point the available rules determine no order: four vehicles reaching an unmarked intersection together cannot be separated by a first-arrival rule, since none arrived first, and "yield to the vehicle on your right" only cycles, since each must yield to the next and none is freed to move. At a *divergent* coordination point the available rules determine more than one order: each vehicle applies the rule its own circumstances invoke, those rules determine different orders, and nothing adjudicates between them, so no single order governs the situation as a whole. These failures have been studied in isolation (Aoki & Rajkumar, 2022; Li et al., 2023; Le Maitre & Prorok, 2020). We introduce a taxonomy on a single structural criterion, the number of orders the available rules determine: none at a symmetric point, more than one at a divergent point.

A shared convention (Shoham & Tennenholtz, 1997) is required to resolve an open coordination point: the convention signal selects among several equally valid solutions to settle the coordination

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.

problem. When vehicles share no such convention, the result is a *negotiation failure*. As mixed-fleet deployment grows we expect negotiation failures to become more prominent, and they are what this paper targets.

The mechanism we consider is this: a roadside unit (RSU) or human operator broadcasts a convention over vehicle-to-everything (V2X) communication at open coordination points. A protocol of this kind must jointly meet three requirements.

(R1) Training independence. An agent must ground the convention alone—no joint optimization, no shared projection or policy parameters, no access to another agent's internal state, and no shared learned encoder or internal embedding space.
(R2) Operator-interpretable convention signal. The broadcast signal must convey the intended convention in terms interpretable to manufacturers and operators rather than as an identifier that requires a predefined mapping.
(R3) Runtime convention selection among grounded conventions. Switching among the grounded conventions must require only a new broadcast—no retraining, no firmware revision, and no coordination with the manufacturers whose vehicles are already deployed.

Standardized numerical codebooks (SAE International, 2024) and one-hot identifiers meet R1 and R3, but fail R2. Neither role decomposition nor learned communication protocols meet R1: both are produced by joint training. A mechanism built on a natural-language broadcast fulfills all three requirements over the grounded convention set: each manufacturer can train its own policy independently using a frozen pretrained sentence encoder, which maps a broadcast string into an embedding space, without co-training with the vehicles the policy will encounter (R1); the broadcast expresses the convention specification in human-interpretable form, not as an identifier that needs a predefined mapping (R2); and because each deployed policy has been trained on the available convention set, an operator can choose among them at runtime (R3).

We introduce *convention grounding*, a new mechanism: an operator or RSU broadcasts a convention as a natural-language string. Each receiving CAV encodes it with a frozen sentence encoder and combines the encoding with its ego-centric observation. The convention is a symmetry-breaking prior, not a per-vehicle command: the same text reaches every vehicle, but the agents produce role-differentiated behavior based on their positions. Language-conditioned single-agent control (Ahn et al., 2022; Brohan et al., 2023) does the opposite by mapping an instruction onto one agent's action sequence. In this work, we define grounding as learning to map a public convention to the vehicle's specific role and action; we do not claim that the learned policy can interpret unseen phrasings of that convention. We evaluate generalization to unseen phrasings separately in Section 4.7. No large language model is invoked in the control loop. A broadcast is encoded once, by a frozen sentence encoder—22M parameters in our main configuration.

Our experiments focus on symmetric coordination points where the observation is insufficient to uniquely determine the crossing order, in controlled SUMO scenes with constant speeds and queue-free conditions. The main contributions of this work are:

- **We introduce a taxonomy of open coordination points** that separates symmetric from divergent failures by a single criterion, the number of orders the available rules determine.
- **We provide the first controlled causal demonstration of convention grounding**, to our knowledge: on a frozen scene with every other input held fixed, changing only the convention text flips the committed action. The reward contains no convention or role label; the convention-following partners provide an indirect learning signal.
- **We demonstrate that the committed action follows the convention input.** Swapping the two conventions' projected vectors maintains switching but inverts correctness, and in our configuration, agents that co-learn instead of grounding a broadcast converge on a wordless habit, acting identically under opposite conventions.
- **We show that grounding holds without a scripted partner**, and is not explained by shared representation: at a two-vehicle merge, two policies trained in separate runs—sharing no encoder, embedding space, gradients, or training episodes—coordinate through the broadcast and cannot differentiate the roles when it is removed.

- **We demonstrate grounding holds beyond the two-vehicle case.** With a scripted partner, it holds at crossing orders over three and four vehicles and at an unregulated four-way intersection under simultaneous arrival.
- **We identify and diagnose a representational limitation**: generalization fails on the tested held-out phrasings, and no tested probe recovers the role assignment difference from the projected representation.

2 Related Work

Natural language as a coordination medium. The case for coordinating vehicles through language instead of a fixed codebook has been argued by Gao et al. (2025), for reasons related to R1–R3: language offers interpretability, can be revised after deployment, and does not require shared training history. That argument is advanced as a position without empirical evaluation. Our contribution is the empirical test left open—whether a learner receiving the convention as a public broadcast grounds it into its own role, and whether the convention, rather than the physical scene, drives action selection (Section 4.2). Closest in methodology is Cui et al. (2026), which jointly learns to generate messages with high-level decisions via self-play among multiple agents. The communication mechanism runs in the opposite direction: linguistic messages emerge bottom-up through training and a generative model is invoked at inference time, whereas convention text is authored offline and no generative model runs during vehicle operation. Language-grounded multi-agent reinforcement learning (MARL) using human-readable messages (Li et al., 2024) aligns with the goal of interpretability and reports zero-shot generalization to unfamiliar teammates and novel task states, but jointly learns the communication protocol, so the protocol is fixed through joint training rather than provided at runtime.

Learned and emergent communication. A broad literature learns inter-agent communication messages (Sukhbaatar et al., 2016; Foerster et al., 2016; Das et al., 2019; Singh et al., 2019), with later approaches introducing selective communication and bandwidth management. The protocols work well within the population that trained on them, which creates the central limitation: they are a product of joint optimization. In the cited training setups, independent protocol learning by each manufacturer is not supported, and the resulting protocol need not be human-interpretable, which may prevent an operator from composing valid messages. Section 4.3 provides the complementary failure mode: under joint adaptation of two agents using our reward formulation, they cease reading the broadcast entirely and coordinate through a fixed learned habit.

Zero-shot coordination. Other-Play (Hu et al., 2020) and Any-Play (Lucas & Allen, 2022) study compatibility with independently trained or unfamiliar partners without an externally authored convention. In our setting, a public convention is supplied at runtime, and the question is whether separately trained policies can coordinate under the shared signal.

Role decomposition. Role-based MARL methods group behaviors into role classes and condition each policy on a role index (Wang et al., 2021). This provides the closest structural analogue to our convention grounding—behavior distinguished by a signal external to the observation—and the distinction is in the information carried by the signal. A role index has no intrinsic similarity structure: two indices either differ or match, leaving nothing to relate an unseen role to a learned one. Sentence embeddings encode such similarity structure, but Section 4.7 shows that this structure does not generalize to the tested held-out phrasings. Our claim over a role index lies in R2 rather than in better performance on trained conventions. The procedural difference matters more than the representational contrast: role decomposition occurs during joint training, so role-based coordination inherits the same R1 failure.

Conventions and equilibrium selection. That coordination problems with multiple equivalent solutions are resolved through convention, not deliberation, is well established (Schelling, 1960; Lewis, 1969; Shoham & Tennenholtz, 1997). A public signal selecting among equilibria functions as a correlation device (Aumann, 1974); when the recommendation chooses among specified Nash profiles, the resulting outcome distribution is a mixture of those profiles, so the broadcast performs equilibrium selection, not equilibrium creation. Game-theoretic methods for vehicle merging and

unsignalized intersections (Shu et al., 2023; Rahmati et al., 2022) determine a crossing order under a predefined game. Our question starts one step later: given that an order is selected and communicated in natural language, we ask whether a learned policy trained separately from its counterparts can recover its assigned role (Section 4.6).

3 METHOD

3.1 PROBLEM SETTING

We formalize the environment as a single-step contextual bandit. Let $\mathcal{V}$ denote a set of N vehicle identities and $\mathcal{C}$ represent a finite collection of natural-language convention sentences such as "*Vehicle A yields to Vehicle B*". At the beginning of an episode, a convention $c \sim \mathrm{Uniform}(\mathcal{C})$ is sampled and broadcast to all vehicles; each learner subsequently observes its context and executes a single action, after which the episode concludes. Let $\mathcal{O}(s)$ denote the set of orders implied by the available rules. A coordination point is *open* when the rules do not determine exactly one order, $|\mathcal{O}(s)| \neq 1$; all scenarios evaluated here are symmetric, satisfying $|\mathcal{O}(s)| = 0$.

A convention establishes a role associated with each identity via the deterministic mapping

$$\rho : \mathcal{C} \to (\mathcal{V} \to \mathcal{A}), \tag{1}$$

implemented by parsing the sentence with a fixed grammar: rank is determined by position in the stated ordering, and yield or proceed are assigned according to grammatical subject and object. The parser rejects sentences that are out-of-grammar rather than assigning a default, and permuting vehicle identities induces an identical permutation of the derived roles without modifying any table. The learner has no access to the parser, making role recovery an empirical result; the experiment tests grounding of a given convention, rather than convention discovery. We define $a^\star(c)$ as $\rho(c)[v_{\text{ego}}]$, which represents the convention-correct action associated with the learner.

Two channels. At training time, exactly one vehicle acts as a learner following policy π_θ; the other $N-1$ vehicles are *scripted partners*. The same c reaches the learner and partner pathways differently: partner v observes $\rho(c)[v]$ and acts accordingly, whereas the learner observes only $\phi(c)$ and must determine $a^\star(c)$ based on the embedding and its own state.

Commitment. Each vehicle reaches a predefined *decision line* located at an upstream distance D from the conflict point, at which it observes $s^{(0)} = (\mathbf{x}_{\text{ego}}, \mathbf{x}_{\text{other}}, \mathbf{u}_{\text{ego}}, \mathbf{u}_{\text{other}})$ and selects a single binding action. Each $\mathbf{x}$ comprises four components: normalized vehicle speed, distance-to-conflict and planar coordinates; each $\mathbf{u}$ is a two-component one-hot encoding of vehicle identity, yielding $\dim(s^{(0)}) = 12$ in the two-vehicle setting. The observation contains no partner intent information, no explicit priority signal, and no convention information apart from $\tilde{\mathbf{z}}_c$. The action space is $\mathcal{A} = \{\text{YIELD}, \text{GO}\}$ for the two-vehicle case and $\mathcal{A} = \{1, \ldots, N\}$ under the ordering formulation; a vehicle with rank k waits for all $j < k$ to clear the conflict region, then advances at speed v_{commit}. Precedence is determined by the integer rank comparison only, independent of arrival time and distance.

Condition 1 (Commitment independence)**.** *At the moment of commitment, the learner's non-convention observation is identical under both conventions, though the convention-correct actions differ:*

$$s^{(0)}(c_1) = s^{(0)}(c_2) \quad \textit{and} \quad a^\star(c_1) \neq a^\star(c_2). \tag{2}$$

We establish Condition 1 through invariance instead of randomizing the identity tags. Constraining all vehicle speeds to v_{commit} at the same upstream distance produces byte-identical $s^{(0)}$ observations—including identity tags—for every convention on a frozen scene, while the parser assigns different convention-correct actions, hence (2) holds exactly. This provides convention swap invariance, rather than testing identity reassignment across positions. Vehicles approach with constant velocity instead of decelerating toward the decision line, as a decelerating approach would create timing-dependent positional variation containing information about $a^\star$. The four-way intersection scenario establishes the same condition via four-fold symmetry of the intersection arms.

3.2 Convention Encoding and Gradient Isolation

The encoder ϕ, held frozen, maps $c \mapsto \mathbf{z}_c = \phi(c)$, computed once for each convention. The primary configuration uses MiniLM-L6-v2 (22M parameters, 384-d); Section 4.6 includes configurations where different vehicles operate with different encoders, including all-mpnet-base-v2 (110M, 768-d). A trainable projection g_ψ maps $\mathbf{z}_c$ to the 64-dimensional representation passed to the policy. The experiments use two projection forms, and the choice depends on the scenario, not on the underlying mechanism. For the two-vehicle merge setting, we use one linear projection with no bias term, initialized orthogonally; for the ordering and paraphrase experiments, we use a two-layer perceptron $d \to 128 \to 64$ using ReLU activations, where d is the encoder width. Appendix A specifies the projection for every experiment. The projection is trained only through its own dedicated objective, thus updates to ψ never result from the policy gradient.

Both the actor and critic receive a gradient-detached copy of the same projected vector, and the detachment blocks the gradient path across policy, value, and entropy terms of the PPO objective.

$$\tilde{\mathbf{z}}_c = \mathrm{sg}\big(g_\psi(\mathbf{z}_c)\big), \qquad \frac{\partial \mathcal{L}_{\mathrm{PPO}}}{\partial \psi} = 0. \tag{3}$$

The projection is trained separately using a cosine-margin objective $\mathcal{L}_{\mathrm{proj}}$ (Appendix B) that uses convention-level pair supervision to distinguish conflicting conventions without any explicit role, action, rank, identity, state, or reward labels. The paraphrase experiment (Section 4.7) has a different procedure: there, the projection is pretrained through a supervised classification over convention families, prior to the policy training. The label identifies convention family, not the role or action, while the reward remains convention-agnostic. Detachment is an isolation constraint instead of a performance requirement: Appendix B reports an ablation in which the policy gradient reaches the projection, and neither separation nor flip rate is materially affected. Encoding requires a single forward pass through ϕ for each broadcast, rather than at every control step, and the per-step cost is one actor forward pass within 1% of the 100 ms control budget (Appendix C).

3.3 Scripted Partner and Convention-Agnostic Reward

Partners execute $\rho(c)[v]$ as defined in Equation 1—coordinated behavior arises under a convention only when other vehicles follow it. The reward contains no convention or role label, but the convention influences the learner's outcome via the partner's convention-following behavior. This eliminates the shortcut that two co-learners can exploit. With a convention-agnostic reward every collision-free resolution receives the same return, so a fixed convention-independent behavior constitutes a global optimum, and in our co-learning setting, the agents converge to this behavior (Section 4.3). Since partner behavior and the evaluation criterion are computed from the same convention sentence by the same ρ, the resulting alignment is derived rather than claimed. An unseen paraphrase is authored as part of a convention family and inherits that family's parsed role, because correctness for a sentence outside the parser's grammar must be defined by the family it paraphrases.

The reward is computed only from the joint outcome:

$$r = -w_{\text{coll}}\,\mathbf{1}[\text{collision}] - w_{\text{dead}}\,\mathbf{1}[\text{deadlock}] + w_{\text{comp}}\,\mathbf{1}[\text{completed}], \tag{4}$$

with $w_{\text{coll}} = 15$, $w_{\text{dead}} = 8$, $w_{\text{comp}} = 10$. No reward term depends on c, the role map ρ, or vehicle-identity tags: a reward encoding $\rho(c)$ would collapse convention grounding into reward engineering. The deadlock penalty activates only under *mutual* freezing and therefore does not penalize valid yielding behavior.

3.4 Policy and Training

The learner's policy $\pi_\theta(a \mid s^{(0)}, \tilde{\mathbf{z}}_c)$ is optimized via *single-agent* PPO. Actor and critic are implemented as separate networks sharing the same input, and the critic receives the same detached projection the actor does; it observes neither partner actions nor the joint state. As c fixes partner behavior, this formulation is exact, not an approximation—at training time, no second policy is learned whose value could be estimated by a centralized critic.

Algorithm 1 Convention-grounded training (single-learner PPO, scripted partners)

Require: convention set $\mathcal{C}$, parser ρ, frozen encoder ϕ, actor π_θ, critic V_w, projection g_ψ

precompute $\{\phi(c) : c \in \mathcal{C}\}$ {frozen encoder; computed once}
for episode $e = 1, \ldots, E$ **do**
  $c \sim \mathrm{Uniform}(\mathcal{C})$; $\mathbf{z}_c \leftarrow \phi(c)$
  $\tilde{\mathbf{z}}_c \leftarrow \mathrm{sg}(g_\psi(\mathbf{z}_c))$ (Eq. 3)
  initialize one learner and $N-1$ partners; partner v receives role $\rho(c)[v]$ (Eq. 1)
  fix all vehicles to v_{commit} at an equal upstream distance from the conflict point (Cond. 1)
  advance vehicles to the decision line; observe $s^{(0)}$
  $a \sim \pi_\theta(\cdot \mid s^{(0)}, \tilde{\mathbf{z}}_c)$
  roll out in SUMO; observe r (Eq. 4)
  append $(s^{(0)}, a, r, \tilde{\mathbf{z}}_c)$ to the buffer
  if buffer full **then**
    update (θ, w) with Adam w.r.t. $\mathcal{L}_{\text{PPO}} + c_v\mathcal{L}_V - c_e\mathcal{H}[\pi_\theta]$ (Eq. 5)
    Adam step on ψ w.r.t. $\mathcal{L}_{\text{proj}}$, over the disjoint parameter set
  end if
end for

Given a single commitment in each episode, there is no temporal bootstrapping and the advantage simplifies to

$$A = r - V_w(s^{(0)}, \tilde{\mathbf{z}}_c), \tag{5}$$

without discounting or generalized advantage estimation, standardized within each batch. A first Adam optimizer updates (θ, w); a second, over the disjoint parameter set ψ, optimizes the projection. Resampling c each episode makes the convention use necessary for optimal return in the scripted-partner setting: a policy that ignores $\tilde{\mathbf{z}}_c$ succeeds under only one of the two opposite conventions.

3.5 Evaluation: The Swap Test and Independent Policies

The core evaluation intervenes on c with all remaining inputs fixed. Let $\mathcal{D}$ denote held-out scenes frozen at the commitment instant and c_1, c_2 a pair of opposite conventions. Evaluation is fully deterministic, and the method does not require a scripted partner at deployment. Let $\pi_1, \ldots, \pi_M$ denote policies trained in separate runs of Algorithm 1, one policy for each vehicle, with no shared training episodes. Each has a separate projection $g_\psi^{(k)}$ and encodes through its own encoder ϕ_k, committing $\pi_k(s, c) \triangleq \arg\max_a \pi_k\big(a \mid s^{(0)}, \mathrm{sg}(g_\psi^{(k)}(\phi_k(c)))\big)$.

Section 4.6 includes configurations where ϕ_k varies across vehicles in architecture and embedding dimension; consequently, the arms have no shared encoder, embedding space, gradients, or training episodes. Since $\mathcal{L}_{\text{proj}}$ is seed-deterministic, two arms using the same seed and encoder would yield identical projections; all reported results use cross-seed assignments.

We report the *joint flip rate* across M vehicles,

$$F_M = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}} \mathbf{1}\big[\,\pi_k(s, c_m) = \rho(c_m)[v_k]\;\forall\, k \in \{1, \ldots, M\},\;\forall\, m \in \{1, 2\} \\ \wedge\;\; \pi_k(s, c_1) \neq \pi_k(s, c_2)\;\forall\, k \in \mathcal{K}\big], \tag{6}$$

where $\mathcal{K} = \{k : \rho(c_1)[v_k] \neq \rho(c_2)[v_k]\}$ denotes vehicles whose convention-correct action differs. As a policy that ignores the convention can be correct under at most one of two opposite conventions, $F_M = 1$ is impossible unless every vehicle in $\mathcal{K}$ uses the text; under Condition 1 the observation is unchanged under both conventions, so $F_M \approx 1$ indicates that the change in commitment between conventions is explained by the change in the convention input.

Table 1: Two-vehicle merge setting: the convention-swap evaluation alongside two ablations and the co-learning baseline. Zeroing or shuffling the convention embedding holds all other inputs fixed.

Convention input	Switch rate	Collision	Completion
Intact	1.00	0.00	1.00
Zeroed	0.00	0.17	0.50
Shuffled	0.55	0.26	0.50
Co-learned	0.00	0.00	1.00

4 EXPERIMENTS

4.1 SETUP

We simulate in SUMO 1.26.0 at 0.1 s timestep. Scenes are generated in advance and frozen at the commitment instant. All reported values are obtained from a fresh simulation rollout, or from replaying a precomputed result. Training and held-out scenes from each dataset are split disjointly, and training is repeated with three seeds for each policy, and with four at the four-way intersection. Before training any policy, each dataset passes a conflict gate: we enumerate rank profiles for each scenario and verify that distinct rank assignments yield a collision-free crossing on each scene whereas every duplicated rank results in collision. The gates succeed on $200/200$ scenes for $N = 3$, $N = 4$, and the intersection setting; the $N = 3$ schema logs no emergency-braking field, limiting its gate to collision and completion. Appendix D details scene counts and splits. The perfect rates here denote no failures were observed on the held-out scenes rather than zero underlying risk: the resulting one-sided 95% upper bound for the per-policy failure probability is 0.030 when $n = 100$ scenes and 0.037 when $n = 80$.

We measure the joint flip rate F_M from Equation 6, along with completion and collision rates. The convention-swap experiment tests each policy on the same frozen scene under both c_1 and c_2; the *control* reruns it after zeroing the convention embedding, while no other input is modified.

4.2 CAUSAL DEMONSTRATION AND NECESSITY

In the two-vehicle merge, a single learner coordinates with a scripted partner using $\rho(c)$. Table 1 summarizes the swap test plus two ablations across 300 evaluations (3 seeds $\times$ 100 held-out scenes).

The intact condition gives the causal result: on a frozen scene with no other input modified, the committed action follows the text under each convention and differs between them, without collision. Zeroing the convention signal eliminates flipping entirely—the policy selects the same action irrespective of the broadcast convention—and completion drops to 0.50, because a convention-independent action is correct under exactly one of the two conventions. Shuffling sits at chance for switching and F_M (Appendix E). Swapping the two convention representations, the only corruption that preserves training-seen vectors, keeps switching at 1.00 while driving per-convention completion to 0.00 and correctness to 0.00 (Appendix E).

4.3 CO-LEARNING AGENTS IGNORE THE BROADCAST IN OUR SETTING

The scripted partner is replaced with a second learner, and we train both under MAPPO with two actor networks, a centralized critic, a team reward, and a single joint optimizer. The complete two-vehicle commitment observation is given to both actors. Coordination succeeds while the flip rate is 0.00: their actions are unchanged under opposite conventions. We observe byte-identical trajectories under c_1 and c_2 for all held-out scenes, across all three seeds.

In this setting, the scripted partner design is central to training rather than a simplification. With a convention-agnostic reward, a fixed role allocation constitutes a global optimum, so nothing in the gradient forces either agent to use the broadcast; they achieve coordination through an implicit agreement, and the text is ignored. In the co-learning setting considered in this study, the comparison

Table 2: Coordination among independently trained policies via the broadcast. A scripted partner is not used in any configuration. Control replaces the convention embedding with the zero vector at evaluation. Intact figures are constant across assignments: the minimum, maximum, and mean coincide in every row.

Scenario	Encoders	Assignments	F_M intact	F_M control
Merge, M=2	shared	6	1.00	0.00
Merge, M=2	different	6	1.00	0.00
Intersection, M=2	shared	6	1.00	0.00
Intersection, M=4	shared	24	1.00	0.00
Intersection, M=4	mixed	24	1.00	0.00

indicates what convention grounding needs: not merely the presence of a convention, but action on it by at least one vehicle.

4.4 Crossing Orders over Three and Four Vehicles

This mechanism extends the action space to a rank, assigning each of N vehicles to a position in the crossing sequence. In both scenarios, we use a scripted partner set. Intact, the learner selects the convention-correct rank under each convention and a different rank between them, on all held-out scenes at $N = 3$ and $N = 4$, without collisions; setting the embedding to zero reduces flip rate to 0.00 and completion to 0.50 in both settings.

Two validation checks support the result at $N = 4$. Grouping by arrival jitter sign yields $120/120$ correctness in both directions, so the policy does not read residual arrival order. The commit band—the gap between the farthest and nearest vehicle at the commitment instant—is $[0.42, 2.42]$ m with median 0.95 at $N = 4$, compared with $[0.09, 5.06]$ with median 2.06 at $N = 3$: the typical spread and upper extremes decrease as N increases, while the minimum grows.

Thus, increasing the vehicle count from three to four modifies the geometry over which Condition 1 must hold, not the result.

4.5 Four-Way Intersection

In an unregulated four-way intersection configuration, four vehicles approach from the same upstream distance at the same time, and a first-arrival priority cannot apply, since none arrives first. Trained on this geometry independently of the merge experiments, the learner selects its convention-correct rank under each convention and a different rank between them, on all held-out scenes, without collisions; the control falls to 0.50 completion and 0.00 flip, every vehicle defaulting to rank 1. Grouping by arrival phase yields $102/102$, $51/51$, $57/57$, and $30/30$ in the four respective bins, totaling all 240 evaluations—uniformly correct, so arrival phase does not explain the policy's convention correctness.

That the mechanism holds in a second geometry, under a four-way action space rather than binary, rules out the merge result being attributable to the two-vehicle setting or to a yield/proceed choice. The intersection is also a challenging case for Condition 1: symmetry must be maintained over four arms instead of two, and the phase-grouped audit is consistent with its holding.

4.6 Independently Trained Policies

The preceding experiments do not remove the scripted partner. Table 2 does, over five configurations varying what the coordinating policies share.

We trained the policies in separate runs and brought them together only at evaluation. The cross-encoder rows pair all-mpnet-base-v2 at 768-d versus MiniLM-L6-v2 at 384-d: distinct architectures. In these rows, the arms do not share a common encoder, gradients, embedding space, or training episodes; elsewhere they differ in seed and training run only. The control makes the intact condition informative. Without the broadcast, each policy defaults to a fixed rank, and the pairing succeeds

only if the fallback ranks are complementary—consequently, control completion ranges from 0 to 1 across pairings, against a flip rate of 0.00 throughout.

Only the flip rate reflects convention dependence; completion measures luck. As seed-3 arms revert to a mirrored fallback profile, collision occurs in 20 of 24 assignments in the shared-encoder four-learner run, and every collision is detected using live conflict-box occupancy during simulation, not from rank arithmetic. A single arm-seed cell among sixteen is multi-valued, and it is the only source of the observed partial scene-level splits.

4.7 The Representational Ceiling

The results above evaluate on conventions observed during training. We now test held-out phrasings. Every convention family includes 26 phrasings, 22 used for training and 4 reserved for evaluation, with non-overlapping content vocabulary across families. On held-out phrasings, correctness drops to 0.500 and the paired flip rate to 0.000: each mirror pair receives one role instead of two opposite ones, which forces exactly half correct.

The limitation is not explained by probe capacity: every tested probe fits the training set perfectly yet achieves only 0.500 on held-out projected pairs; the frozen encoder output carries only a limited, seed-unstable trace (Appendix F). Raw cosine similarities within mirror pairs range from 0.975 to 0.984: the two wordings "*A yields to B*" and "*B yields to A*" are placed at nearly the same point before the policy sees them. This experiment employs the higher-capacity projection defined in Section 3.2, so the lower-capacity projection does not explain the failure either.

5 Discussion

Scalability. The obstacles to density are independent of the mechanism. Vehicles arriving at different times provide a convention-independent ordering cue, as arrival time predicts crossing order. Stacked vehicles in the same lane can make global rank ordering unexecutable. One vehicle per stream is a simplifying condition for the current execution rule. Convention count scales differently: the active convention can be changed within the grounded set, and unseen phrasings did not generalize zero-shot in our evaluation (Section 4.7).

Limitations. Nothing forces the convention's English semantics onto the behavior; because the reward is convention-agnostic, only the partner does. Identity tags are determined by road position in the scenarios we evaluate, so we cannot say whether the policy resolves the identity token or keys on the position. Paired policies are independent in training runs, seeds, and (in mixed-encoder arms) encoders. They share the training loop, reward, policy topology, and environment, so different *policy architectures* are outside our evidence. Divergent coordination points and training independence at the three- and four-vehicle merge orderings are not evaluated. Broadcast failure and safe fallbacks remain untested; without the convention signal, 20 of 24 shared-encoder four-policy assignments produced collisions (Section 4.6).

6 Conclusion

This paper introduced *convention grounding*, in which a coordination convention is broadcast as a natural-language sentence; each vehicle combines it with its own observation to produce its role-differentiated action; and the reward contains no convention or role label. To our knowledge, this is the first controlled causal demonstration of the mechanism—on a frozen scene, changing the convention sentence flips the committed action. Swapping the two conventions' projected vectors maintains switching but reverses correctness.

Independently trained policies—at a two-vehicle merge and at a four-way intersection—coordinate through the broadcast without a scripted partner when they ground the same sentence through different frozen encoders. A representational limitation bounds generalization: on the held-out phrasings we evaluated, no probe recovers the role difference from the projected representation. The next step is to test whether an order-sensitive encoder or role-structured objective recovers the role distinction without providing the role to the policy.

Statement on the Use of AI Assistants

Every sentence in the submitted paper was written by the authors, and the authors are responsible for all claims. Large language model assistants were used throughout to resolve grammatical and spelling errors, to draft and revise prose in every section, to execute and audit code in the experimental repository.

We did not use model assistants to generate experimental results. All results come from code in the supplementary material, executed by the authors. Where an assistant proposed an experiment, an ablation, or a correction, the underlying computation was run and recorded in the repository before the claim entered the paper.

We maintained a verification ledger recording, for each numeric claim in the paper, the command that reproduces it and the date it was last checked. Every numeric claim in the submitted paper is backed by a recorded command, and Appendix G is the reader-facing form of that ledger.

Ethics Statement

This work is conducted entirely in simulation. It involves no human subjects, no personal data, and no deployment on physical vehicles.

The application domain is safety-critical, and we want to be explicit about what this paper does and does not establish. We demonstrate that a broadcast natural language convention causally determines role-differentiated behaviour in independently trained policies under a controlled intervention. We do not claim deployment readiness. The scenarios are constructed, the observation is an oracle state vector rather than perception, and the evaluation holds the physical scene fixed by design.

A broadcast coordination channel is also an attack surface. Message loss, delay, conflicting or malformed broadcasts, adversarially crafted text, and authentication of the broadcasting authority are all outside the scope of this paper. A deployed system would require a specified fallback behaviour under message loss and an authentication mechanism for the broadcast; we evaluate neither. We note the gap rather than leaving a reader to infer that the mechanism is complete.

We also disclose a confound in the experimental design: identity tags are tied to road position in every scenario we evaluate, so our experiments do not separate grounding in the identity token from grounding in the position it always co-occurs with. This is discussed in the Limitations.

Reproducibility Statement

The full experimental repository is included as supplementary material: source code, trained checkpoints, generated scene datasets, and the raw logs from every run reported in the paper. It is self-contained apart from the two public pretrained sentence encoders, which are downloaded on first use.

Appendix G maps every number reported in the paper to the command that produces it. Appendix A gives the training configuration per experiment, Appendix B the full projection objective including the one experiment that uses a different one, and Appendix D the scenario geometry, dataset construction, and train/test splits.

Four limits on reproducibility are worth stating directly.

First, the simulator version is not pinned in the code. Every environment resolves the SUMO binary by name, and no version string is recorded in the repository or in checkpoint provenance. All results reported here were produced with Eclipse SUMO 1.26.0; this is a property of the machine we ran on rather than of the artifact, and a reader running a different version may see different low-level dynamics.

Second, one reported quantity is derived rather than independently cross-checked. The commit-band statistics for the four-vehicle merge cross-check against a field written by the scene builder at dataset construction time and agree on all 200 scenes. The three-vehicle merge has no such stored field, so the reported band is computed from the committed commitment observations by the same definition.

Third, the control decomposition in Section 4.6 is verified by recomputing the predicted fallback behaviour live from the committed checkpoints and comparing it against the control outcomes recorded in the run logs. This establishes that the fallback account reproduces the logged controls exactly; it is not an independent re-measurement of those controls, which requires re-running the evaluation scripts.

Fourth, checkpoint provenance recording was added partway through the project. Twenty-seven of fifty-nine checkpoints record the configuration they were trained under; for the remainder, Appendix A reports the configured defaults in the code. We know these can diverge, because one checkpoint that does record its configuration shows a training split its own default would not have produced. Splits for every reported experiment are taken from the recorded commands rather than from defaults.

The repository also contains a superseded earlier version of the three-vehicle crossing-order trainer, retained rather than deleted. It has an internal inconsistency between its function default and its command-line default for the iteration count. The reported three-vehicle results come from the current trainer, which does not, but a reader browsing the repository will encounter both.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2022.

Shunsuke Aoki and Ragunathan (Raj) Rajkumar. A-DRIVE: Autonomous deadlock detection and recovery at road intersections for connected and automated vehicles. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 29–36. IEEE, 2022. doi: 10.1109/IV51971.2022.9827436.

Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, March 1974. doi: 10.1016/0304-4068(74)90037-8.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, Proceedings of Machine Learning Research. PMLR, 2023.

Jiaxun Cui, Chen Tang, Jarrett Holtz, Janice Nguyen, Alessandro G. Allievi, Hang Qiu, and Peter Stone. CoopReflect: Towards natural language communication for cooperative autonomous driving via multi-agent learning. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 744–752, Paphos, Cyprus, 2026. International Foundation for Autonomous Agents and Multiagent Systems. doi: 10.65109/MOAV6406.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1538–1546. PMLR, 2019. URL https://proceedings.mlr.press/v97/das19a.html.

Jakob N. Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2137–2145, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/c7635bfd99248a2cdef8249ef7bfbef4-Abstract.html.

Xiangbo Gao, Keshu Wu, Hao Zhang, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Automated vehicles should be connected with natural language, 2025. URL https://arxiv.org/abs/2507.01059.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-Play" for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4399–4410. PMLR, 2020.

Matthew Le Maitre and Amanda Prorok. Effects of controller heterogeneity on autonomous vehicle traffic. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6. IEEE, 2020. doi: 10.1109/ITSC45102.2020.9294509.

David K. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.

Changwen Li, Joseph Sifakis, Qiang Wang, Rongjie Yan, and Jian Zhang. Simulation-based validation for autonomous driving systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*, pp. 842–853. ACM, 2023. doi: 10.1145/3597926.3598100.

Huao Li, Hossein Nourkhiz Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi-Pari, Michael Lewis, and Katia Sycara. Language grounded multi-agent reinforcement learning with human-interpretable communication. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87908–87933, 2024. doi: 10.52202/079017-2790.

Keane Lucas and Ross E. Allen. Any-Play: An intrinsic augmentation for zero-shot coordination. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 853–861. International Foundation for Autonomous Agents and Multiagent Systems, 2022. doi: 10.5555/3535850.3535946.

Yalda Rahmati, Mohammadreza Khajeh Hosseini, and Alireza Talebpour. Helping automated vehicles with left-turn maneuvers: A game theory-based decision framework for conflicting maneuvers at intersections. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11877–11890, 2022.

SAE International. V2X communications message set dictionary. SAE Standard J2735_202409, SAE International, September 2024. URL https://saemobilus.sae.org/standards/j2735_202409-v2x-communications-message-set-dictionary.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

Yoav Shoham and Moshe Tennenholtz. On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence*, 94(1–2):139–166, 1997.

Keqi Shu, Reza Valiollahi Mehrizi, Shen Li, Mohammad Pirani, and Amir Khajepour. Human inspired autonomous intersection handling using game theory. *IEEE Transactions on Intelligent Transportation Systems*, 24(10):11360–11371, 2023.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rye7knCqK7.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2244–2252, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/55b1927fdafef39c48e5b73b5d61ea60-Abstract.html.

Torque News. Amazon's Zoox and Google's Waymo freeze in San Francisco intersection. https://www.torquenews.com/1/amazon-s-zoox-and-google-s-waymo-freeze-san-francisco-intersection, May 2026. Reported May 31, 2026.

Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning roles to decompose multi-agent tasks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TTUVg6vkNjK.

A HYPERPARAMETERS

Two optimizers run per experiment and they are disjoint. One Adam updates the actor–critic network; a second Adam updates the convention projection and is the only gradient that reaches it. The actor and critic are separate networks that share the first optimizer. There is no discounting and no generalized advantage estimation anywhere in the codebase: every episode is a single decision, so the advantage is $A = r - V(s)$, batch-normalized within each update.

Table 3: Convention projection form by experiment. Four forms occur. The linear map and the MLP are not interchangeable across experiments and the distinction matters for the interpretation of Section 4.7.

Experiment	Projection class	Shape	η_{proj}
E1 (one-hot control)	none	—	—
E1 (MiniLM)	linear, no bias	$384 \to 64$	1×10^{-2}
E2	linear, no bias	$384 \to 64$	inference only
Paraphrase (Sec. 4.7)	MLP	$384 \to 128 \to 64$	2×10^{-3}
Crossing order, N=3	MLP	$384 \to 128 \to 64$	2×10^{-3}
Crossing order, N=4	MLP	$384 \to 128 \to 64$	2×10^{-3}
Intersection, all four arms	MLP	$384 \to 128 \to 64$	2×10^{-3}
Intersection, MPNet arms	MLP	$768 \to 128 \to 64$	2×10^{-3}
On-ramp	MLP	$384 \to 128 \to 64$	5×10^{-3}
Merge, mainline arm	linear, no bias	$384 \to 64$	1×10^{-2}
Merge, MPNet arm	linear, no bias	$768 \to 64$	1×10^{-2}
MAPPO baseline	none	—	—

Table 4: Per-experiment PPO settings. The learning rate, iteration count, and entropy coefficient vary; everything in Table 5 does not.

Experiment	η_{policy}	Iterations	Entropy coef.
E1, both arms	3×10^{-3}	80	0.01
E2	inference on E1 checkpoints		
Paraphrase	3×10^{-3}	150	0.01
Co-learning baseline	3×10^{-3}	120	0.01
Merge, mainline arm	3×10^{-3}	80	0.01
Merge, MPNet arm	3×10^{-3}	80	0.01
On-ramp	1.5×10^{-3}	120	0.01
Crossing order N=3	1.2×10^{-3}	150	0.005
Crossing order N=4	1.2×10^{-3}	150	0.005
Intersection, four arms	6×10^{-4}	150	0.005
Intersection, MPNet arms	6×10^{-4}	150	0.005

There is no gradient clipping anywhere in the codebase; the clipping parameter above is the PPO surrogate ratio clip, and no trainer imports the gradient-norm utilities. The actor and critic are separate networks but submodules of one module, so a single optimizer covers both at one learning rate and a single summed loss trains them, with the value term at coefficient 0.5. The co-learning baseline follows the same pattern with three networks — two actors and a centralized critic — under one optimizer.

The reward function has exactly three weighted terms and no shaping, step cost, or emergency-braking term. A collision returns the flat penalty with no dilution by other outcomes; mutual deadlock is checked next; otherwise the agent is rewarded for completing its own manoeuvre. Reward weights are written into checkpoint provenance on every save.

The paraphrase terms of L_{proj} are present in the loss function but inert in every reported experiment: the paraphrase pair set is empty at all ten call sites, so that branch of the loop never executes and

Table 5: Settings shared by every trainer, and the reward and auxiliary-loss weights.

Quantity	Value
Batch size	256
Epochs per update	4
PPO clip ϵ	0.2
Value-loss coefficient	0.5
Optimizer	Adam, default β and ϵ
Advantage	$A = r - V(s)$, batch-normalized
Discount γ	not used
GAE λ	not used
Gradient clipping	none
Reward, collision	-15.0 (flat, overrides all)
Reward, mutual deadlock	-8.0 (both frozen, neither crossed)
Reward, completion	$+10.0$ (must remain below collision)
L_{proj} conflict weight	2.0
L_{proj} paraphrase weight	0.5 (inert, see below)
L_{proj} paraphrase margin	0.1 (inert, see below)

neither weight has ever affected a gradient. We report the values for completeness and note that they carry no experimental meaning here.

Checkpoint provenance recording was added partway through the project, so 27 of 59 checkpoints record their own training configuration and the remainder do not. The values above are the configured defaults in the training code. Where a checkpoint does record its configuration, the recorded value governs: the merge mainline arm, for instance, records a training split that its own default would not have produced, because the command that ran it passed an override. Train and test splits for every reported experiment come from the recorded commands rather than from defaults and are given in Appendix D.

A.1 Policy architecture and the multi-vehicle observation

The actor and critic are separate two-hidden-layer perceptrons of width 128 with ReLU activations, taking the same input: the commitment observation concatenated with the detached projected convention vector. The actor emits logits over the scenario's action space; the critic emits a scalar. Weights use the framework defaults.

The commitment observation extends to N vehicles by concatenating the same four kinematic components — normalised speed, distance to the conflict point, and planar coordinates — for the ego vehicle and then for each other vehicle, followed by one-hot identity tags for all N. Vehicle order is fixed by the scenario rather than sorted at run time: for the crossing-order scenarios it follows the stream index, and at the intersection it follows the arm order north, east, south, west. The observation dimension therefore grows with N and no padding is used, so a policy trained at one vehicle count is not applied at another.

B The Convention Projection Objective

B.1 Form and pair construction

The projection g_ψ is trained by an auxiliary objective that operates on the projected convention vectors and never on the policy's actions, rewards, or states. For a convention set $\mathcal{C}$ with projected vectors $\{\tilde{z}_c\}_{c\in\mathcal{C}}$, conflict pair set P_{conf}, and paraphrase pair set P_{para}:

$$L_{\text{proj}} \;=\; w_c \sum_{(a,b)\in P_{\text{conf}}} \cos(\tilde{z}_a, \tilde{z}_b)^2 \;+\; w_p \sum_{(a,b)\in P_{\text{para}}} \max(0,\; (1-m) - \cos(\tilde{z}_a, \tilde{z}_b)) \tag{7}$$

with $w_c = 2.0$, $w_p = 0.5$, $m = 0.1$. The first term drives conflicting conventions toward orthogonality; the second would pull paraphrases together but is inert in every experiment reported here, as P_{para} is empty throughout.

We state plainly what supervision this objective carries. It provides *convention-level* separation: the two convention instances are named as distinct and their projected representations are driven apart. It supplies no role, action, rank, identity, state, or reward label. The pair arguments are consumed only as dictionary keys; the loss function contains no reference to rank, role, ordering, action, or any label, and the rank tables that define correct behaviour live in a separate module consumed only by the scripted partners and the evaluation oracle, never imported on the projection path.

Because every scenario here has exactly two conventions, P_{conf} is the complete set of unordered pairs over a two-element set, so the objective does not choose *which* pair among several to separate. That is a statement about pair selection, not an argument that the objective is supervision-free: telling the representation that two conventions are distinct is itself weak contrastive supervision at the convention level.

The consequence for the claim is worth being exact about. This experiment does not show that frozen sentence semantics spontaneously produce the behaviour. It shows that each policy can privately ground the same public sentence under its own incompatible representation, with no direct role or action label anywhere in the training signal and a convention-agnostic environmental reward.

B.2 Gradient isolation

The projected vector is detached before entering the policy, so the PPO objective cannot reshape the projection. We verified this directly: across three seeds and eighty iterations each, the gradient norm arriving at the projection from the PPO objective is exactly zero at every iteration.

We originally justified detachment as a necessity, on the argument that a convention-agnostic reward would make convention-invariant behaviour optimal and collapse the two projected conventions together. An ablation shows this argument is wrong. Removing the detachment and adding the projection to the PPO optimizer so the gradient is actually applied, with everything else held identical, changes nothing: all arms converge from an initial cosine near $+0.96$ to an absolute cosine below 0.016 and reach a flip rate of 1.00 on every seed. The auxiliary objective determines the outcome and absorbs the policy gradient as a perturbation.

We therefore state the design honestly. Detachment is a deliberate isolation choice, not a requirement forced by the reward structure. Its value is that the projection is trained by the auxiliary objective alone by construction, provably at every iteration. We make no claim about how the two objectives would interact at larger scale; the ablation found no harmful interaction at the scale tested.

Three limits on the ablation. It was run with the auxiliary objective active, which is the deployed configuration; we did not test the policy gradient in isolation. The two-convention, two-action setting is the most favourable case for the auxiliary objective to dominate. And the ablation uses a slightly larger training split than the reported causal experiment, so its arms are comparable to each other rather than directly to Table 1.

B.3 The paraphrase experiment uses a different objective

One experiment does not use Equation 7. The paraphrase experiment of Section 4.7 replaces it with a supervised binary cross-entropy objective: a small auxiliary linear classifier reads the projected vector and predicts which of the two convention families a wording belongs to, and that gradient shapes the projection. The classifier has its own parameters, is pretrained for 2,500 steps before the reinforcement learning loop begins, and is never seen by the policy.

The label is family identity — which of the two parent sentences a given wording is a rewording of. It plays the same role as the conflict pair set in Equation 7, extended over a cluster of wordings rather than a single sentence. We do not claim that calling the target a family rather than a role removes task information: family identity is convention-level supervision, and in a two-family setting it may suffice, together with ego identity, to determine the required action. What remains true, and is the distinction the paper rests on, is that no direct role or action label is supplied and the environ-

mental reward stays convention-agnostic. We state the difference explicitly because describing this experiment's objective as the cosine loss above would be inaccurate.

C RUNTIME

Encoding latency is measured on a single sentence of eight tokens, batch size one, with the tokenizer inside the timed region, under `no_grad`, with one Torch thread, discarding ten warmup iterations and reporting the median of one hundred, on an Intel Core Ultra 9 185H.

Table 6: Latency. The encoder runs once per broadcast; the actor runs once per control step.

Component	Median latency	Share of a 100 ms control step
MiniLM-L6-v2 encode (22M)	7.0 ms	7.0%
all-mpnet-base-v2 encode (110M)	38.9 ms	38.9%
Actor `act()`, 16 threads	296 µs	0.3%
Actor forward pass, 16 threads	126 µs	0.1%

The thread count matters and we report it rather than leaving it implicit. A sixteen-thread protocol is unstable for this workload on this processor: nominally identical repetitions returned medians of 7.35, 8.66, 14.39, and 18.47 ms, two of them within the same run at idle load. Single-thread measurements reproduce to within 0.13 ms. Thread scaling also runs in opposite directions for the two encoders — the smaller model is slower at sixteen threads than at one, since dispatch overhead dominates an eight-token workload, while the larger model benefits from the parallelism.

The larger encoder consumes roughly two fifths of a control step. This is acceptable because it is paid once per broadcast rather than once per step, and we state it rather than leaving a reader to compute it. The separation between encoding and per-step inference is between one and two orders of magnitude depending on the encoder, not more.

D SCENARIOS, DATASETS, AND SPLITS

Table 7: Scene datasets and held-out splits.

Experiment	Dataset	Total	Train	Held out
E1, E2, paraphrase, MAPPO	`scenes.json`	217	117	100
Crossing order $N{=}3$	`scenes3r.json`	200	100	100
Crossing order $N{=}4$	`scenes_merge4.json`	200	120	80
Intersection	`scenes_intersection.json`	200	120	80

Reported rates are not always over the number of held-out scenes, and we state the denominators. The paraphrase correctness rate is over 2,400 evaluations (eight held-out wordings $\times$ 100 scenes $\times$ 3 seeds) and its paired flip rate over 400 per seed (four mirror pairs $\times$ 100 scenes). The four-learner intersection result covers all $4! = 24$ all-distinct seed assignments over 80 held-out scenes, totalling 7,680 live episodes.

Table 8: Scenario geometry. The conflict point and decision distance are not uniform across scenarios.

Scenario	Conflict point	D (m)	Radius (m)
Two-vehicle merge	$(498.36,\ 95.20)$	55	6.0
Crossing order $N{=}3$	$(504.76,\ 98.40)$	55	6.0
Crossing order $N{=}4$	$(504.76,\ 98.40)$	55	6.0
On-ramp	$(339.87,\ 37.20)$	50	6.0
Four-way intersection	box $(150.0,\ 150.0)$	50	7.0

The commitment speed is 10 m/s and the simulation step length 0.1 s in every scenario without exception. The intersection conflict box is given in post-`netconvert` network coordinates, which differ from the node declaration in the source network file.

The lane topology differs between the two crossing-order scenarios. Both merge a single on-ramp into a continuing lane, but the mainline carries two lanes in the three-vehicle case and three in the four-vehicle case, and in both the mainline and the ramp reduce to a single lane at the conflict point. The two-vehicle merge is not a lane drop: both mainline edges carry two lanes and the ramp joins the rightmost.

A scenario run but not reported. The supplementary material contains a third scenario type, a highway on-ramp merge with its own network and dataset, which we ran and did not include in the main results. It passes its conflict gate and reproduces at a switch rate of 1.00 with clean crossings under both conventions across three seeds and 80 held-out scenes, with no emergency braking, and its depart-jitter leak check is uniform at $120/120$ in both directions. We note it here rather than leaving a working scenario visible in the artifact with no mention in the paper.

The scene generators enforce a conflict gate: every scene admitted to a dataset is one in which distinct role assignments complete cleanly and shared assignments collide. This gate is not identical across scenarios. The four-vehicle merge and the intersection additionally screen for emergency braking on the clean side; the three-vehicle merge has no emergency-braking field in its outcome structure, so its gate tests collision and completion only. We note this because it makes the three-vehicle gate strictly weaker than the other two, not because it affects any reported result.

D.1 Which vehicles carry the flip

The two conventions in each ordering scenario are reversals of one another. A reversal is a derangement only for even N: at three vehicles the middle one holds the same rank under both conventions, and the same holds for the middle vehicle at any odd count. The joint flip criterion must therefore quantify over the vehicles whose convention-correct roles actually differ, not over all N; otherwise the action-changes clause is unsatisfiable for the fixed vehicle and the criterion is identically zero at odd N.

Table 9: Vehicles whose convention-correct rank differs between the two conventions. A reversal moves $2\lfloor N/2 \rfloor$ of them.

Scenario	N	Vehicles carrying the flip
Crossing order	3	2
Crossing order	4	4
Four-way intersection	4	4

In the three-vehicle case the convention pair is a transposition of the outer two vehicles with the middle held fixed by construction, so the flip claim there covers two of the three. At the four-way intersection, which carries the strongest result, the moved set is all four vehicles and the criterion quantifies over everything.

E Additional Controls

The controls reported in the body divide into two kinds, and the distinction governs which columns carry evidence.

A *convention-independent* control replaces the projected convention vector with one vector used under both conventions. The policy's input is then byte-identical in the two conditions, so its greedy action is identical, so both the switch rate and F_M are zero by construction rather than by measurement. This is true of the zeroed control and of the two replacement arms below.

A *convention-dependent* control corrupts or permutes the vector while leaving it distinct across conventions. Here the two columns come apart, and the separation is informative. The shuffled

control is of this kind, and so is the exchange control introduced below, which is the strongest of the set.

Table 10: Control arms, three seeds × 100 held-out scenes. *Switch* is the raw rate at which the committed action differs between the two conventions; F_M additionally requires correctness under each. Completion separates the arms into three levels.

Arm	Kind	Switch	F_M	Completion	Corr. c_1	Corr. c_2
Intact	—	1.00	1.00	1.00	1.00	1.00
Zeroed	independent	0.00	0.00	0.50	0.67	0.33
Norm-matched random	independent	0.00	0.00	0.50	0.57	0.43
Midpoint	independent	0.00	0.00	0.50	0.67	0.33
Shuffled	dependent	0.55	0.27	0.50	0.48	0.52
Exchanged	dependent	1.00	0.00	0.00	0.00	0.00

E.1 Replacement controls

The objection these answer is that a zero vector may never occur during training, so the collapse under it could be ordinary out-of-distribution behaviour rather than removal of the convention signal. We therefore added two replacements closer to the observed vectors: one matching the mean norm of the two projected conventions, drawn twenty times, and one taking their elementwise mean. Neither is on the learned manifold in any strong sense, and we do not claim they are; they match different properties of it.

Neither recovers any convention-dependence. Both collapse as the zero vector does, and correctness under the two conventions sums to 1.00 in every corrupted arm against 2.00 intact — the signature of a policy committing to one fixed action and therefore being right under exactly one of two opposing conventions. Across the twenty norm-matched draws the flip rate has standard deviation zero and that sum is 1.000 on every draw.

The shuffled arm sits at its analytic chance level in both columns. It redraws a convention identifier per call and the policy acts correctly on whatever vector it is handed, so the drawn pair decides the outcome: the action switches when the two draws differ, with probability $1/2$, and F_M fires only when both draws happen to match the active convention, with probability $1/4$. Observed 0.55 and 0.27 over 300 evaluations. The corrupted embedding therefore carries no recoverable convention information rather than a degraded amount of it.

Note also that the two ways of reaching $F_M = 0$ are different, and the switch column is what separates them. The convention-independent arms fail the action-change clause: one vector under both conventions cannot produce two actions. The exchange arm fails the correctness clause instead, and its actions switch on every scene.

Two qualifications. The arms bracket rather than cover: because the auxiliary objective drives the two conventions toward orthogonality, their midpoint has norm approximately $\|z\|/\sqrt{2}$, so one arm matches norm and not region while the other does the reverse. And the collision rate under the second convention is zero in every arm including the intact one, because there the partner yields and a convention-blind agent deadlocks rather than colliding. That asymmetry belongs to the scenario, not to the corruption.

E.2 The exchange control

The strongest control exchanges the two conventions' projected vectors rather than corrupting either: the policy receives the vector for c_2 while c_1 is the live broadcast, and conversely. The scripted partner and the correctness oracle continue to follow the true active convention.

Both vectors are ones the policy encountered in training, at the trained norm and direction, so the corruption is in-distribution at the level of the embedding. The pairing of vector to world state is what is inverted, and that joint configuration is of course not one training produced. It is also the only control that makes a positive prediction rather than predicting degradation: if the committed action

is a function of the convention channel, exchanging the channel should leave the action switching intact while exactly reversing correctness.

It does, on every seed and every scene. The raw switch rate stays at 1.00, so the policy remains fully sensitive to which vector it receives. Correctness under each convention falls from 1.00 to 0.00, and cross-correctness — whether the action taken under one convention matches what the other requires — rises from 0.00 to 1.00. F_M is therefore 0.00, since it requires correctness as well as a switch. The policy is not confused by the exchange; it acts correctly on the convention it was handed. Separating the two columns is what makes this readable: sensitivity to the convention representation and correct grounding of that representation are distinct properties, and the exchange arm has the first without the second.

Completion separates the three levels cleanly. The intact arm completes. Every convention-independent arm completes half the time, because a fixed action is correct under one of two opposing conventions by luck. The exchange arm completes *never*, because it is wrong under both. No other control reaches zero.

Its failure mode is asymmetric and the asymmetry is diagnostic. Under the first convention the learner proceeds while the convention-obeying partner also proceeds, and they collide. Under the second both yield, and they deadlock. Collision and deadlock must be read as a pair for this row; either alone understates it.

F Probe Capacity Sweep

The ceiling reported in the body rests on a linear readout. A reader may reasonably ask whether a single linear probe failing establishes that the distinction is absent rather than merely hard to read. We therefore swept probe capacity, and we did so in two representation spaces: the projected vector the policy actually consumes, and the frozen encoder output upstream of the trained projection.

Table 11: Held-out accuracy on four mirror pairs, mean over three seeds. Every cell reaches training accuracy 1.000. Chance is 0.500.

Space	Probe	Params	Held-out	Pairs collapsed
Projected	linear	65	0.500	3/3
Projected	MLP-64	4,225	0.500	3/3
Projected	MLP-128	8,449	0.500	3/3
Projected	MLP-256	16,897	0.500	3/3
Projected	MLP-128-128	24,961	0.500	3/3
Raw	linear	385	0.500	3/3
Raw	MLP-64	24,705	0.500	2/3
Raw	MLP-128	49,409	0.542	2/3
Raw	MLP-256	98,817	0.542	2/3
Raw	MLP-128-128	65,921	0.583	1/3

In the representation the policy consumes the result is flat and total. Five capacities spanning 65 to 24,961 parameters, three seeds each, and every cell gives training accuracy 1.000 with held-out accuracy at chance and both members of every mirror pair assigned the same label. Capacity is not the limiting factor there.

Upstream in the frozen encoder the picture is not flat, and we state it narrowly because the evidence is thin. Three cells reach 0.625, which on eight held-out items means one of the four mirror pairs separated correctly. Nothing exceeds that, no configuration is consistent across seeds, and the largest raw probe collapses the pairs on only one seed of three rather than all three. The defensible statement is therefore that the distinction is not recovered at any probe capacity we tried in the space the policy consumes, and that the frozen encoder output upstream carries only a limited, seed-unstable trace — never above five of eight items, and not consistent across seeds.

Each probe was given its best chance rather than a single fixed schedule. A first pass at the learning rate used by the committed linear readout left several higher-capacity probes in the raw space un-

able to fit their own training set, and a probe that cannot fit forty-four training points says nothing about generalization. Each cell is therefore fit over a small learning-rate grid with restarts, keeping the trial with the best training accuracy, which makes a ceiling result the strongest rather than the weakest version of itself. The schedule of the committed linear readout is attempted first, so that cell reproduces the body result exactly.

The held-out set is eight wordings forming four mirror pairs, so a difference of 0.125 in any single cell is one test item. The sweep is strong evidence about the projected space precisely because it is flat across every capacity, seed, and pair; it is weak evidence about the raw space in either direction.

G Commands for Recomputing Reported Results

Each command runs from the repository root. These commands recompute the reported experimental outcomes from the supplied checkpoints and datasets; they do not retrain the checkpoints. Quantities not listed here — model dimensions, hyperparameters, reward weights, dataset sizes — are given in the appendix tables above rather than by a command.

A note on naming. The paraphrase experiment is called `e2b` throughout the repository; scripts, checkpoints, and logs use that name rather than the section numbering used in the paper.

Two entries carry a qualification. The conflict-gate script for the three-vehicle merge defaults to an eight-seed preview; the reported result is the two-hundred-seed form in Table 12. The fallback decomposition recomputes the predicted behaviour live from the committed checkpoints and compares it against the control outcomes recorded in the run logs, so it confirms that the account reproduces those logs exactly rather than re-measuring the controls.

Table 12: Reported quantity to reproducing command.

Reported quantity	Command
Causal demonstration, all rows	`python3 -m eval.negative_control --seeds 0 1 2 --n_train 117`
Additional controls (Table 10)	`python3 -m eval.control_ablations --seeds 0 1 2 --n_train 117 --draws 20`
Probe capacity sweep (Table 11)	`python3 -m checks.representation_ceiling --seeds 0 1 2 --sweep`
Co-learning baseline	`python3 -m eval.mappo_eval --seeds 0 1 2 --n_train 117`
Co-learning trajectory invariance	`python3 -m checks.mappo_invariance_md5 --n_train 117`
Paraphrase flip rate and correctness	`python3 -m eval.e2b_eval --seeds 0 1 2 --n_train 117`
Representational ceiling	`python3 -m checks.representation_ceiling`
Mirror-pair raw cosines	`python3 -m checks.mirror_cosines`
Paraphrase vocabulary overlap	`python3 -m checks.paraphrase_vocab_overlap`
Crossing order $N{=}3$	`python3 -m eval.exp5r_eval --seeds 0 1 2 --n_train 100`
Crossing order $N{=}4$	`python3 -m eval.merge4_swap --seeds 0 1 2 --n_train 120`
Intersection swap test	`python3 -m eval.intersection_swap --seeds 0 1 2 --n_train 120`
Conflict gate, $N{=}3$	`python3 -m checks.merge3r_conflict 200`
Conflict gate, $N{=}4$ and intersection	`python3 checks/merge4_conflict.py 200; python3 checks/intersection_conflict.py 200`
Commit-band statistics	`python3 -m checks.commit_band`
Independent policies, merge	`python3 -m eval.r1_cross_pairs --n_train 117`
Independent policies, merge, mixed encoders	`python3 -m eval.r1_cross_pairs_mpnet --n_train 117`
Independent policies, intersection, $M{=}2$	`python3 -m eval.r1_intersection`
Independent policies, intersection, $M{=}4$	`python3 -m eval.r1_intersection4`
Independent policies, intersection, $M{=}4$, mixed encoders	`python3 -m eval.r1_intersection4_mixed`
Fallback control decomposition	`python3 -m checks.fallback_algebra`
Grounding divergence across encoders	`python3 -m checks.grounding_divergence_mpnet`
Identity-position association	`python3 -m checks.tag_position_independence`
Parser ground truth	`python3 -m checks.parse_ground_truth`
Encode latency	`python3 -m checks.encode_runtime`
Actor latency	`python3 checks/actor_forward_runtime.py`
Gradient isolation ablation	`python3 -m checks.detach_ablation --seeds 0,1,2 --arms detached,nodetach,strict`